Sensitivity Analyses over the Service Area for Mobility Allowance Shuttle Transit (MAST) Services

Luca Quadrifoglio and Maged M. Dessouky

Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, California 90089-0193
maged@usc.edu

Summary. A Mobility Allowance Shuttle Transit (MAST) system is an innovative concept that merges the flexibility of Demand Responsive Transit (DRT) systems with the low cost operability of fixed-route bus systems. It allows vehicles to deviate from the fixed path so that customers within the service area may be picked up or dropped off at their desired locations. In this paper, we summarize the insertion heuristic presented by Quadrifoglio *et al.* (2007) for routing and scheduling MAST services, and we carry out a set of simulations to show a sensitivity analysis of the performance of the algorithm and the capacity of the system over different shapes of the service area. The results show that a *slim* service area performs better in general, but also that the positive effects of a proper setting of the control parameters of the heuristic is much more evident for wider service areas. In addition, a performance comparison shows that MAST systems can provide a better service to customers than fixed-route ones even for a slim service area.

1 Introduction

The Mobility Allowance Shuttle Transit system is an innovative concept in transportation that merges the flexibility of Demand Responsive Transit systems with the low cost operability of fixed-route bus systems, in order to satisfy the current needs of transit agencies, which are seeking ways to improve their service flexibility in a cost efficient manner. A MAST system is characterized by one or more vehicles driving along a base fixed-route covering a specific geographic zone, with one or more mandatory checkpoints conveniently located at major transfer points or high demand density zones. Given a proper amount of slack time, vehicles are allowed to deviate from the fixed path to serve (pick-up and/or drop-off) customers at their desired locations, as long as they are within a service area. Customers can make reservations before or during the service, thus the MAST system works under a dynamic environment.

Line 646 of the Metropolitan Transit Authority (MTA) of Los Angeles County offers a MAST *nightline* service. The vehicle drives nine times back and forth between two terminal checkpoints, passing by a third intermediate checkpoint in each trip. The vehicle is allowed to deviate from the fixed-route to serve customers as long as their service stops are within half a mile from either side of the main route. The demand of Line 646 is currently low enough to allow the bus operator to make all the decisions concerning accepting/rejecting requests and routing the vehicle. Quadrifoglio *et al.* (2007) developed a customized insertion heuristic algorithm to handle heavier demand in a potential daytime MAST operation and several requests for deviations. The vehicle's route and schedule are updated shortly after each request and customers are notified whether their request has been accepted and are provided with a time window for their pick-up and/or drop-off stops. The main characteristic of their algorithm is the development of efficient control parameters as a function of the future expected demand that, if properly set, significantly enhances the performance of the algorithm.

The purpose of this paper is to evaluate the sensitivity to the shape of the service area of the performance of MAST systems and of the effectiveness of the control parameters of the above mentioned algorithm. In particular we will show how a proper setting of those parameters is able to raise the saturation demand level in each configuration, allowing the system to serve more customers with a comparable service level. In addition, we perform a simulation comparison to test the competitiveness of hybrid systems like MAST versus conventional fixed-route types of services in a *slim* service area, apparently more suitable for the latter services.

Hybrid types of transportation systems have just lately been approached by researchers. Daganzo (1984) describes a checkpoint DRT system that combines the characteristics of both fixed route and door-to-door service. Malucelli *et al.* (1999) provide a general overview of flexible transportation systems. Crainic *et al.* (2001) incorporate the hybrid fixed and flexible concept in a more general network setting. Zhao and Dessouky (2004) study the optimal service capacity of a MAST system through a stochastic approach. Quadrifoglio *et al.* (2006) look at MAST systems from a design point of view, evaluating the relationship between the longitudinal velocity of the vehicle and the demand density, in order to allocate slack time and set other system parameters.

Some work approached hybrid systems in which different vehicles perform the fixed and variable portions. Aldaihani *et al.* (2004) develop a continuous approximation model for designing such a service. Scheduling heuristics based on a hybrid system include the decision support system of Liaw *et al.* (1996), the insertion heuristic of Hickman and Blume (2001) and the tabu heuristic of Aldaihani and Dessouky (2003). Another work studying a combination of fixed and flexible service can be found in Cortés and Jayakrishnan (2002).

Savelsbergh and Sol (1995), Desaulniers *et al.* (2000) and Cordeau and Laporte (2003) provide reviews on the Pickup and Delivery problem and Dial-a-Ride systems. Wilson *et al.* (1971) formulate the problem as a dynamic search procedure. Continuing work is presented by Wilson and Hendrickson (1980). Stein (1977), Stein (1978b), Stein (1978a) develops a probabilistic analysis of the problem and Daganzo (1978) presents a model to evaluate the performance of a Dial-a-Ride system. Theoretical studies of the problem case include the work by Psaraftis (1980), Psaraftis (1983), Sexton and Bodin (1985a), Sexton and Bodin (1985b), Sexton and Choi (1986), Desrosiers *et al.* (1986) and Lu and Dessouky (2004).

Heuristics to solve multi-vehicle problems have been proposed by Psaraftis (1986), Jaw *et al.* (1986), Bodin and Sexton (1986), Desrosiers *et al.* (1988) and Madsen *et al.* (1995). Parallel insertion heuristics are proposed by Toth and Vigo (1997), Diana and Dessouky (2004) and Lu and Dessouky (2006). Diana (2006) assesses by simulation the effectiveness of the latter algorithm. Horn (2002a) develops an algorithm for the scheduling and routing of a fleet of vehicles that is embedded in a modeling framework for the assessment of the performance of a general public transport system with the latter being presented in Horn (2002b).

This paper is organized as follows. In Section 2 we describe the model for a MAST system. In Section 3 we briefly summarize the insertion heuristic algorithm described by Quadrifoglio *et al.* (2007), that we utilize to perform the simulation analysis described in Section 4, where a sensitivity over the shape of the service area is presented. Section 5 provides a MAST/fixed-route comparison and Section 6 the conclusions.

2 MAST System Model

The MAST system model is described by a service area shaped as a rectangular region L×W. C checkpoints are distributed along the x axis in the middle of the rectangle with a y coordinate W/2. Checkpoints 1 and C are at the extremities of the rectangle and the remaining C-2 checkpoints are within it (see Fig. 1). A single vehicle is assigned to this service area. A trip r begins at checkpoint 1 (or C) and ends at checkpoint C (or 1), after visiting in a predefined order all the intermediate checkpoints, which have fixed departure times. If R is the total number of trips, the total number of stops at the checkpoints is TC = (C-1)R+1. Hence, the initial vehicle's schedule is represented by an ordered sequence of stops from 1 to TC. We assume that the vehicle follows a rectilinear metric and has infinite capacity.

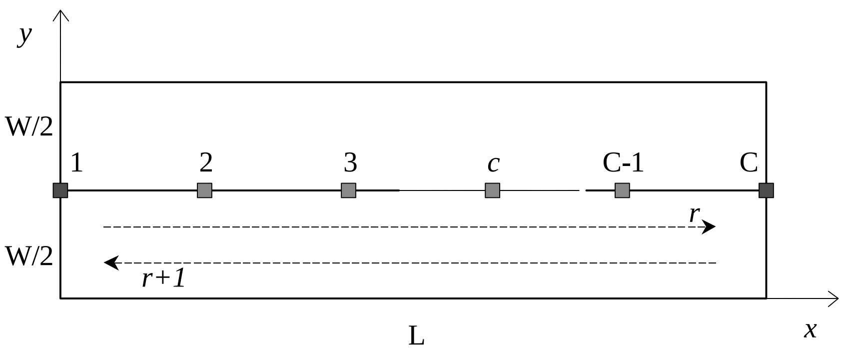

Fig. 1. MAST System Model

The demand is defined by a set of requests, which can be of three types: "hybrid" (having one service point at a non-checkpoint location in the service area and the other one at the checkpoints), "regular" (both service points at the checkpoints) or "random" (both service points located at non-checkpoint stops). We assume that the total demand rate θ is constant over time and that the non-checkpoint stops are uniformly distributed in the service area. At any moment a customer may call in (or show up at the checkpoints), specifying the locations of pick-up and/or drop-off points. "Regular" customers do not need a booking process to use the service.

In order to allow deviations from the main route to serve non-checkpoint requests, there needs to be a certain amount of *slack time* in the schedule. The initial slack time between any pair of consecutive checkpoints in the schedule is given by the difference between their scheduled departure times minus the time needed by the vehicle to travel from one to the other. The slack time is dynamically consumed by the insertion procedure when the demand arises. The amount of it to be allocated depends on the amount and type of demand and it may be adjusted properly to fit particular situations; see Quadrifoglio *et al.* (2006) and Zhao and Dessouky (2004) for more detailed analyses on the matter. In this paper we assume a slack time larger than the actual one in the MTA Line 646, where the demand is very low.

3 Algorithm Description

In this section we summarize the main features of the insertion heuristic algorithm described in Quadrifoglio *et al.* (2007) that will be utilized to perform the sensitivity analyses described in the following Section 4.

A *bucket* of a checkpoint c is the portion of the schedule beginning at one occurrence of c in the schedule and the following one. Since "hybrid" customers rely on a checkpoint c for either their pick-up or drop-off stop, the algorithm checks the schedule for possible insertion of their non-checkpoint stop "bucket by bucket" of c, until feasibility is found (for "random" requests, buckets are represented by trips). The following flowchart in Fig. 2 summarizes the insertion procedure.

All customers, once their request is placed in the schedule, are provided with time-windows for both their pick-up and drop-off stops. These time-windows depend on the current schedule at the time of the request and are naturally bounded by the hard time constraints of the checkpoints.

The cost function needed to select the best insertion among the feasible ones is given by

$$\text{COST } = w_1 \times \Delta t + w_2 \times \Delta RT + w_3 \times \Delta WT \tag{1}$$

where Δt is the slack time consumed by the insertion. ΔRT is the sum over all passengers of the additional ride time, including the whole ride time of the requesting customer, caused by the insertion. In fact, a new inserted request would cause the passengers onboard to be delayed if the insertion takes place before and within the same pair of consecutive checkpoints of their drop-off. Also "regular" onboard passengers may be affected by this caused delay, because the *arrival* time at their

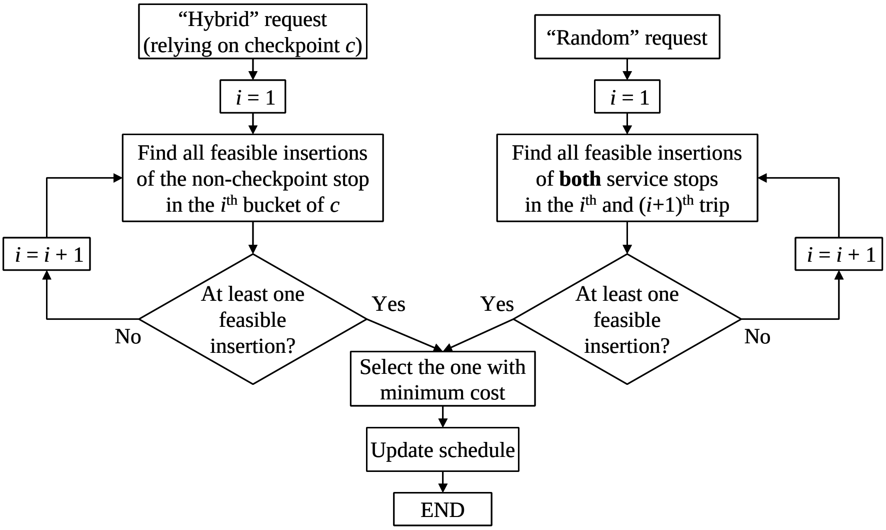

Fig. 2. Insertion Algorithm for MAST Systems, Quadrifoglio *et al.* (2007)

checkpoints is not fixed (the *departure* time at the checkpoints is) and depends on how much slack time is consumed in that portion of the schedule. ΔWT is the sum over all passengers of the additional waiting time caused by the insertion. In fact, customers that are already scheduled and are waiting for their pick-up at the time initially agreed might have to wait longer if the new insertion is placed before them and in between the same pair of consecutive checkpoints. w_1, w_2 and w_3 are the weights, which can be modified as needed to emphasize one factor over the others.

Insertion feasibility and control parameters The "myopic" consumption of the slack time could prevent future requests to be properly satisfied, worsening the overall performance of the system. In order to prevent and solve this problem the heuristic makes use of two control parameters that are a function of the expected future demand and the relative position of the new request with respect to the already scheduled stops. The control parameter $\pi^{(0)} \leq 1$ is multiplied by the initial slack time and sets a cap on how much slack time each insertion may require. The BACK parameter (in miles) defines the maximum allowable backtracking distance available for each insertion. A proper setting of these two parameters (to be determined by simulation analysis) allows the system to control the consumption of slack time and improves the overall performance significantly, especially reducing the total mileage driven and allowing the system to serve more demand, raising the saturation level.

Thus, a candidate insertion is feasible if the customer precedence constraints are met, the slack time consumed is less than the current available and less than the maximum allowed (controlled by $\pi^{(0)}$), and the potential backtracking distance is less than the maximum allowed (controlled by BACK).

4 Sensitivity Over Service Area

In this section we perform a simulation analysis to observe the behavior of the system when modifying the shape of the service area, maintaining constant the total square mileage. In particular we want to observe the effect of the control parameters in each configuration over their saturation level.

The service area considered is described by Fig. 1. The time interval between the scheduled departure times of the two terminal checkpoints is assumed to be 50 minutes. We consider two different cases: C = 3, as for the MTA Line 646, therefore with only one intermediate checkpoint placed in the middle of the area (25 minutes between each pair of consecutive checkpoints) and C = 5, with three intermediate checkpoints (12.5 minutes between each pair of consecutive checkpoints). The initial slack time available between any pair of consecutive checkpoints will vary depending on the assumed proportion between W and L. With smaller L, the amount of slack time is larger because the checkpoints are closer.

The vehicle is riding back and forth between the two terminal checkpoints for a total simulated time of 50 hours, without interruption and therefore the total number of trips is R = 60. The simulation time has been chosen to ensure that the system reaches a steady state. The speed of the vehicle is assumed constant and equal to 25 miles/hour.

Demand is arising dynamically during the trip; we assume that the demand rate θ (customers/hour) is constant over time and that the customer types are distributed as shown in Table 1, as it is for MTA Line 646. In addition, we assume that checkpoint requests (P for pick-up and D for drop-off) are uniformly distributed among the checkpoints and that non-checkpoint requests (NP and ND) are uniformly distributed over the service area.

Table 1. Customer Type Distribution

Type	PD	PND	NPD	NPND
%	10%	40%	40%	10%

The weights in the COST function are w_1 = w_2 = 0.25, w_3 = 0.5, reasonably assuming that customers would rather stay onboard (w_2) than waiting (w_3) at the bus stop and assigning the same value to w_1 (slack time consumed) and w_2. These values can be modified accordingly depending on the objective function of the transit agency.

The main purpose of the analysis is to determine the demand saturation level of the system for each configuration, by running several simulation experiments: first, with no control (BACK = L and $\pi^{(0)}$ = 1, which allow for any backtracking and any consumption of slack time, if available; therefore, giving the maximum freedom to the algorithm when checking for insertion feasibility); then, with the best setting of the control parameters that we could find, in order to maximize the saturation demand level. In addition, we compute the following performance parameters (directly related to the corresponding terms in the COST function) to compare the efficiency of the algorithm and the service level among the cases:

- M: total miles driven by the vehicle
- RT: average ride time per customer
- WT: average extra waiting time per customer

Configuration A: W = 1; L = 12 The first analysis is done over a *slim* service area with L = 12 and W = 1, both in miles. The distance between checkpoints is 6 miles and the slack time available between any consecutive pair is therefore about 10.5 minutes for C = 3 and 5 minutes for C = 5. The saturation levels of this system configuration with BACK = L and $\pi^{(0)} = 1$ (no control) and with the best setting of the parameters to maximize demand are shown in Table 2.

Table 2. Configuration A – Saturation Demand Levels: No Control / Best Control

C	3		5	
Control	None	Best	None	Best
BACK (miles)	L	0.2	L	0.2
$\pi^{(0)}$	1	0.3	1	0.6
θ (customers/hour)	18	21	15	18
WT (min)	0.99	1.43	0.34	0.46
RT (min)	25.33	25.42	27.04	25.97
M (miles)	1049.8	1018.2	1020.5	981.9

The system becomes unstable with θ greater than the values shown, that are approximately the saturation levels of these configurations.

For C = 3, the system is able to handle up to about 21 customers/hour, with a proper setting of the control parameters, namely BACK = 0.2 and $\pi^{(0)} = 0.3$. For C = 5 instead, the system capacity is about 18 customers/hour, with BACK = 0.2 and $\pi^{(0)} = 0.6$. The improvement on the capacity of the system is only 3 customers/hour for both cases (about 15-20% increase), but the improved efficiency of the algorithm is evident on the total mileage M as well, that has decreased by approximately 30-40 miles despite the increased demand. Note that the cases with C = 5 have lower capacities than the ones with C = 3, because of the additional constraints of the two extra checkpoints. From "None" to "Best" control cases, the ride time (RT) remains about the same, while the extra waiting time at stops (WT) slightly increases, due to the heavier demand that leads to an increased number of insertions and postponement of NP pick-ups. Also, the WT is lower for the cases with C = 5, because the number of possible insertions between consecutive checkpoints is smaller due to the checkpoints that are closer to each other and less slack time is allocated between each pair.

Configuration B: W = 2; L = 6 A similar analysis is performed over a service area with W = 2 and L = 6, always referring to the model in Fig. 1. The total square mileage is still 12 and all the other parameters of the system are kept the same. However, given the different shape of the area, checkpoints are closer to each other and therefore the initial slack time available between any pair of consecutive checkpoints is larger, namely equal to about 18 minutes for C = 3 and about 9 minutes for C = 5. Table 3 shows the figures for the saturation levels of this configuration.

Table 3. Configuration B – Saturation Demand Levels: No Control / Best Control

C	3		5	
Control	None	Best	None	Best
BACK (miles)	L	0.3	L	0.2
$\pi^{(0)}$	1	0.3	1	0.6
θ (customers/hour)	12	20	10	18
WT (min)	1.36	1.94	0.20	0.54
RT (min)	20.59	22.81	25.04	29.57
M (miles)	1054.5	933.5	909.8	917.8

In this case the improvement due to control parameter adjustment is more significant: the saturation level jumps from 12 to 20 customers/hour for C = 3 and from 10 to 18 for C = 5 (65-80% increase). The mileage (M) is reduced by about 120 miles for C = 3 and slightly increases for C = 5, even with the increased demand. The values of RT increase slightly more than in Configuration A.

Configuration C: W = 3; L = 4 We consider now a service area with W = 3 and L = 4. The total square mileage is again still 12 and all the other parameters of the system are kept the same, but checkpoints are even closer to each other and the initial slack time available between any pair of consecutive checkpoints is now about 20 minutes for C = 3 and about 10 minutes for C = 5. Results are in Table 4.

Table 4. Configuration C – Saturation Demand Levels: No Control / Best Control

C	3		5	
Control	None	Best	None	Best
BACK (miles)	L	0.5	L	0.2
$\pi^{(0)}$	1	0.5	1	1
θ (customers/hour)	12	18	10	15
WT (min)	1.73	1.68	0.38	0.51
RT (min)	17.37	22.17	21.62	24.86
M (miles)	1047.3	964.0	955.4	896.8

The increase in the saturation level due to control parameter adjustments is significant, from 12 to 18 customers/hour for C = 3 and from 10 to 15 for C = 5 (50% increase) and the mileage (M) also is reduced by about 80 and 60 miles, respectively. As for Configuration B, a more significant increase of the RT value is observed. Fig. 3 summarizes the findings shown in the previous tables.

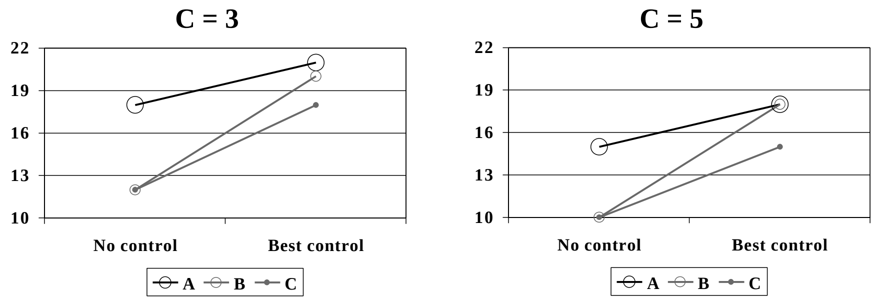

Fig. 3. Saturation Levels (Customers/Hour)

The analysis shows that a proper setting of the control parameters could significantly improve the performance of the system for every configuration. The results also show that the *slim* Configuration A performs better with or without the involvement of the control parameters, even though with different emphasis in the two cases.

With no control (BACK = L and $\pi^{(0)} = 1$) Configuration A outperforms Configurations B and C in terms of system capacity (18 vs. 12 customers/hour for C = 3 and 15 vs. 10 for C = 5), meaning that the insertion procedure is able to perform better in case of a slimmer service area and consequently a lesser amount of slack time. This is due to the fact that a "wild" consumption of the slack time is less likely to happen when there is a smaller amount of it available to begin with and the system is able to control itself better.

When properly setting the control parameters, every configuration benefits from it, but the improvements shown in Configurations B and C are much more evident than those in Configuration A, and while the *slim* case still performs better, the three "controlled" systems are comparable in terms of capacity and performance.

In addition, we note that the longitudinal velocity (along the x axis in Fig. 1) of the vehicle decreases with the widening of the service area (Configurations B and C), because of the increased amount of time needed by the vehicle to serve points along the larger width. Customers traveling to/from checkpoints could perceive this slowness unfavorably because on average they would experience ride times increasingly larger than the direct time needed to travel between their pick-up and drop-off. Therefore, only slimmer service areas, such as Configuration A, would be suitable for public transportation purposes where the longitudinal velocity of the vehicle is not much slower than a fixed route line traveling between checkpoints. However, configurations with wider service area could very well be appropriate for transportation of goods instead of people.

5 MAST/Fixed-Route Comparison

It could be noted that slimmer service areas, such as Configuration A, would be more suitable for a regular fixed-line service. For this purpose we perform a comparison between the MAST service (Configuration A, with C = 3) and a fixed-route bus service serving the same service area. Both systems serve the same demand of 21 customers/hour; with the distribution of Table 1. We assume the same vehicle speed v = 25 miles/hour and a service time of 18 seconds at each stop for both systems. The fixed-route line has C = 25 fixed stops evenly distributed along the x axis (one stop every 0.5 miles), therefore the headway is 72 minutes and the scheduled/actual travel time between two consecutive stops is 1.5 minutes. We assume that there is no variability in the travel time between two consecutive stops for the fixed line. The only variability for the MAST system is due to the random locations of the non-checkpoint demand. Fig. 4 illustrates the geometry and the features of the systems.

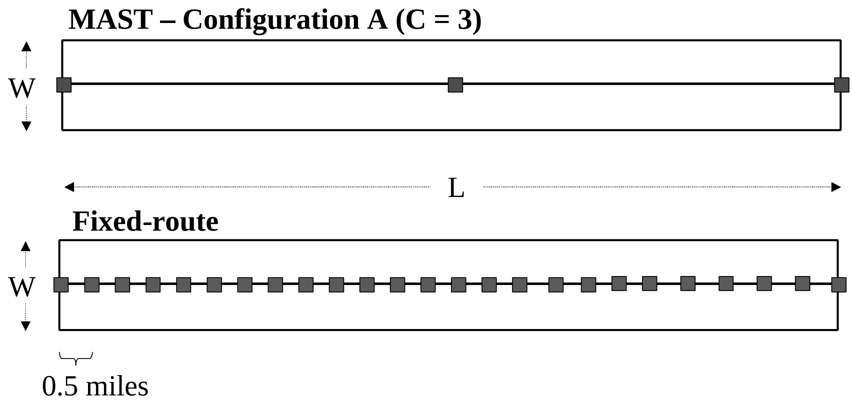

Fig. 4. MAST/Fixed-route Comparison

In order to perform the comparison we define WKT, being the average walking time per passenger (assumed walking speed = 3 miles/hour). While the MAST system serves its customers point to point and no walking occurs, a fixed-route system forces non-checkpoint requests to walk to/from the nearest fixed stop in order to use the service. Note that checkpoint requests could have a certain amount of walking time associated with it, but considering the same demand it would be equivalent for both systems and consequently we assume it to be zero.

We observe that for headways larger than 12-13 minutes the majority of the customers are aware of the schedule (Okrent (1974)) and this is true for all requests showing up at bus stops (for both systems). Therefore, we do not consider the waiting time until the pick-up as a valid parameter for this comparison. WT measures instead the extra waiting time that MAST customers have to wait at their stops, because of other insertions occurring after their requests.

Thus, the overall performance Z (in time units) is defined as follows:

$$Z = w_1 \times M/v + w_2 \times RT \times NC + w_3 \times WT \times NC + w_4 \times WKT \times NC \quad (2)$$

where NC is the total number of customers served by the system and the last term represents the contribution to Z of the amount of walking time. We assume that the weight for walking time (w_4) is conservatively equal to w_3 (even though customers would probably perceive walking time with more discomfort than waiting time at a bus stop, especially during nighttime for safety reasons). Hence the weights in Z are set as follows: $w_1 = w_2 = 0.25$ and $w_3 = w_4 = 0.5$.

We ran the simulations (using Common Random Numbers for the two systems) for 45 hours, so that for the fixed-route service R = 75 and for the MAST system R = 54 (since the headway is 100 minutes). The results are shown in Table 5.

Table 5. MAST/Fixed-route Comparison

θ (customers/hour)	21	
System	MAST Conf. A (C = 3)	Fixed
WT (min)	1.56	0
RT (min)	25.53	16.6
WKT (min)	0	7.5
M (miles)	926.3	900
Z	6.804	7.831

The figures show that the MAST system compared to the fixed-route results has a small WT (< 2 minutes) and a RT bigger by approximately 10 minutes, but M is lower and there is no walking for the customers as opposed to the fixed-route system where on average customers walk 7.5 minutes.

6 Conclusions

In this paper we summarize the insertion heuristic algorithm developed for the Mobility Allowance Shuttle Transit services presented by Quadrifoglio *et al.* (2007) and we utilize it to carry out a sensitivity analysis of its performance over the shape of the service area. The algorithm makes use of proper control parameters, aiming to cherish the consumption of the slack time. A proper setting of them allows the system to increase its capacity, maintaining an analogous service level for the customers. In particular, we show that this positive control effect is more evident in a wider service area with more slack time. The results also show that slimmer configurations perform better in terms of capacity and are more suitable for public transportation purposes. In addition, the findings show that MAST services are competitive with fixed-route ones and perform better under certain demand distributions, even for slim service areas.

Acknowledgement: The research reported in this paper was partially supported by the National Science Foundation under grant NSF/USDOT-0231665.

References

Aldaihani, M. M. and Dessouky, M. (2003). Hybrid scheduling methods for paratransit operations. *Computers & Industrial Engineering*, **45**, 75–96.

Aldaihani, M. M., Quadrifoglio, L., Dessouky, M., and Hall, R. W. (2004). Network design for a grid hybrid transit service. *Transportation Research*, **38A**, 511–530.

Bodin, L. and Sexton, T. (1986). The multi-vehicle subscriber dial-a-ride problem. *TIMS Studies in the Management Sciences*, **22**, 73–86.

Cordeau, J. F. and Laporte, G. (2003). The dial-a-ride problem (DARP): variants, modeling issues and algorithms. *4OR*, **1**(2), 89–101.

Cortés, C. E. and Jayakrishnan, R. (2002). Design and operational concepts of a high coverage point-to-point transit system. *Transportation Research Record 1783*, pages 178–187.

Crainic, T. G., Malucelli, F., and Nonato, M. (2001). Flexible many-to-few + few-to-many = an almost personalized transit system. In *TRISTAN IV, São Miguel Azores Islands*, pages 435–440.

Daganzo, C. F. (1978). An approximate analytic model of many-to-many demand responsive transportation systems. *Transportation Research*, **12**, 325–333.

Daganzo, C. F. (1984). Checkpoint dial-a-ride systems. *Transportation Research*, **18B**, 315–327.

Desaulniers, G., Erdmann, A., Solomon, M. M., and Soumis, F. (2000). The VRP with pickup and delivery. Technical report, Cahiers du GERARD G-2000-25, Ecole des Hautes Etudes Commerciales, Montréal.

Desrosiers, J., Dumas, Y., and Soumis, F. (1986). A dynamic programming solution of the large-scale single-vehicle dial-a-ride problem with time windows. *American Journal of Mathematical and Management Sciences*, **6**, 301–325.

Desrosiers, J., Dumas, Y., and Soumis, F. (1988). The multiple dial-a-ride problem. In *Computer Aided Transit Scheduling*, volume 308 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin.

Diana, M. (2006). The importance of information flows temporal attributes for the efficient scheduling of dynamic demand responsive transport services. *Journal of Advanced Transportation*, **40**(1), 23–46.

Diana, M. and Dessouky, M. (2004). A new regret insertion heuristic for solving large-scale dial-a-ride problems with time windows. *Transportation Research*, **38B**, 539–557.

Hickman, M. and Blume, K. (2001). A method for scheduling integrated transit service. In S. Voss and J. Daduna, editors, *Computer Aided Scheduling of Public Transport, Lecture Notes in Economics and Mathematical Systems 505*, pages 233–251. Springer, Berlin.

Horn, M. E. T. (2002a). Fleet scheduling and dispatching for demand-responsive passenger services. *Transportation Research*, **10C**, 35–63.

Horn, M. E. T. (2002b). Multi-modal and demand-responsive passenger transport systems: a modeling framework with embedded control systems. *Transportation Research*, **36A**, 167–188.

Jaw, J. J., Odoni, A. R., Psaraftis, H. N., and Wilson, N. H. M. (1986). A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows. *Transportation Research*, **20B**(3), 243–257.

Liaw, C. F., White, C. C., and Bander, J. L. (1996). A decision support system for the bimodal dial-a-ride problem. *IEEE Transactions on Systems, Man, and Cybernetics*, **26**(5), 552–565.

Lu, Q. and Dessouky, M. (2004). An exact algorithm for the multiple vehicle pickup and delivery problem. *Transportation Science*, **38**, 503–514.

Lu, Q. and Dessouky, M. (2006). New insertion-based construction heuristic for solving the pickup and delivery problem with hard time windows. *European Journal of Operational Research*, **175**, 672–687.

Madsen, O. B. G., Raven, H. F., and Rygaard, J. M. (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.

Malucelli, F., Nonato, M., and Pallottino, S. (1999). Demand adaptive systems: some proposals on flexible transit. In T. Ciriania, E. Johnson, and R. Tadei, editors, *Operations Research in Industry*, pages 157–182. McMillan, London.

Okrent, M. M. (1974). Effect of transit service characteristics on passenger waiting time, MS thesis. Department of Civil Engineering, Northwestern University, Evanston.

Psaraftis, H. N. (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. N. (1983). An exact algorithm for the single vehicle many-to-many dial-a-ride problem with time windows. *Transportation Science*, **17**, 351–357.

Psaraftis, H. N. (1986). Scheduling large-scale advance-request dial-a-ride systems. *American Journal of Mathematical and Management Sciences*, **6**, 327–367.

Quadrifoglio, L., Hall, R. W., and Dessouky, M. M. (2006). Performance and design of mobility allowance shuttle transit services: Bounds on the maximum longitudinal velocity. *Transportation Science*, **40**, 351–363.

Quadrifoglio, L., Dessouky, M. M., and Palmer, K. (2007). An insertion heuristic for scheduling mobility allowance shuttle transit (MAST) services. *Journal of Scheduling*, **10**, 25–40.

Savelsbergh, M. W. P. and Sol, M. (1995). The general pickup and delivery problem. *Transportation Science*, **29**, 17–29.

Sexton, T. R. and Bodin, L. D. (1985a). Optimizing single vehicle many-to-many operations with desired delivery times: 1. Scheduling. *Transportation Science*, **19**, 378–410.

Sexton, T. R. and Bodin, L. D. (1985b). Optimizing single vehicle many-to-many operations with desired delivery times: 2. Routing. *Transportation Science*, **19**, 411–435.

Sexton, T. R. and Choi, Y. (1986). Pickup and delivery of partial loads with soft time windows. *American Journal of Mathematical and Management Sciences*, **6**, 369–398.

Stein, D. M. (1977). Scheduling dial-a-ride transportation systems: an asymptotic approach. Technical report, No. 670, Harvard University, Division of Applied Science.

Stein, D. M. (1978a). An asymptotic probabilistic analysis of a routing problem. *Mathematics of Operations Research*, **3**, 89–101.

Stein, D. M. (1978b). Scheduling dial-a-ride transportation problems. *Transportation Science*, **12**, 232–249.

Toth, P. and Vigo, D. (1997). Heuristic algorithm for the handicapped persons transportation problem. *Transportation Science*, **31**, 60–71.

Wilson, N. H. M. and Hendrickson, C. (1980). Performance models of flexibly routed transportation services. *Transportation Research*, **14B**, 67–78.

Wilson, N. H. M., Sussman, J. M., Wong, H. K., and Higgonet, B. T. (1971). Scheduling algorithms for a dial-a-ride system. Technical report, USL TR-70-13, M.I.T, Urban Systems Laboratory.

Zhao, J. and Dessouky, M. (2004). Optimal service capacity for a single bus mobility allowance shuttle transit (MAST) system. Submitted for publication.