

## ABSTRACT

Title of dissertation: HEURISTICS AND PERFORMANCE METAMODELS  
FOR THE DYNAMIC DIAL-A-RIDE PROBLEM

Ying Luo, Doctor of Philosophy, 2006

Directed by: Professor Paul M. Schonfeld  
Department of Civil and Environmental Engineering

Explicit performance models of a transit system are often very useful in facilitating system design, optimization, alternative comparison, and gaining insights into the system relations. In this dissertation, three performance metamodels have been developed using the response surface metamodeling approach for the dynamic many-to-many dial-a-ride problem. The models predict, respectively, the minimum vehicle fleet size requirement, the average passenger time deviation from desired time, and the average passenger ride time ratio. The metamodeling approach incorporates in its simulation experiments a detailed vehicle routing algorithm and passenger time constraints, which are oversimplified or omitted by analytical approaches.

A new rejected-reinsertion heuristic has been developed for the static dial-a-ride problem. The heuristic achieves vehicle reductions of up to 17% over the parallel insertion heuristic and of up to 12% over the regret insertion heuristic. The static heuristic has been extended to two online heuristics for the dynamic large-scale dial-a-ride problem, the

immediate-insertion online heuristic and the rolling horizon online heuristic. The rolling horizon heuristic outperforms the immediate insertion heuristic by up to 10% vehicle reduction for demand scenario in which different demand lead times exist. Their computational efficiency makes them usable in real dynamic applications. The rolling horizon heuristic with an improvement procedure is employed in the simulation experiments upon which the metamodels are based. It is simple in concept, and it does not involve complex algorithm parameter calibration.

The response surface methodology models the functional relation between an output of a process and its input factors through well designed experiments and statistical analysis. A face-centered central composite design is used in this study to determine the design points. Models are based on data collected from the simulation experiments and fitted using SPSS's linear regression function. The metamodels are validated using an additional set of randomly generated data. The resulting models are relatively simple in structure, inexpensive to use and fairly robust. The applications of the performance models are illustrated through the parametric analysis and optimization of a dial-a-ride service considering the tradeoff between operator cost and user cost.

HEURISTICS AND PERFORMANCE METAMODELS FOR THE DYNAMIC  
DIAL-A-RIDE PROBLEM

By

Ying Luo

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2006

Advisory Committee:

Professor Paul M. Schonfeld, Chair  
Professor Michael O. Ball  
Professor Kelly Clifton  
Professor Ali Haghani  
Professor Hani S. Mahmassani

## DEDICATION

To my dear parents

## ACKNOWLEDGEMENTS

I would like to express my profound appreciation to my advisor, Professor Paul Schonfeld, for initiating this research, providing me freedom to follow my own interests, and giving me guidance and encouragement throughout the entire research. I am very grateful for his availability whenever I needed to talk, and the discussions were always helpful and inspiring. It has been a real pleasure and privilege working with him.

I would like to thank my advisory committee members, Professor Michael Ball, Professor Kelly Clifton, Professor Ali Haghani, and Professor Hani Mahmassani, for their valuable comments and suggestions on this research. Special thanks go to Professor Hani Mahmassani and Professor Michael Ball for having interesting and fruitful discussions with me during my study.

My gratitude also goes to Dr. Marco Diana and Professor Maged M. Dessouky at the University of Southern California for providing their datasets and helping clarify some data issues.

Lastly but not least, I wish to express my deep gratitude to my family for their support. I would like to thank my husband, Xiaohu, for always being there when I needed him the most. I would like to thank my parents, my brother, and my sister-in-law for encouraging me to pursue my goal overseas.

## TABLE OF CONTENTS

<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Abbreviations</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem and Motivations	1
1.2 Background of Dial-a-Ride Services	4
1.3 Objectives and General Methodology	7
1.4 Organization of the Dissertation	9
<b>2 Literature Review</b>	<b>10</b>
2.1 Existing Performance Models for Dial-a-Ride Services	10
2.1.1 Theoretical analysis	11
2.1.2 Statistical methods based on real or simulation data	14
2.1.3 Findings	17
2.2 Description of the Static and Dynamic Dial-a-Ride Problems	18
2.3 Dial-a-Ride Problem Algorithms	20
2.3.1 Insertion-based	22
2.3.2 Cluster-first route-second or cluster-based	26
2.3.3 Metaheuristics	28
2.3.4 Local improvement procedures	29
2.3.5 Dynamic dial-a-ride problem algorithms	31
2.3.6 Findings	35
<b>3 A Rejected-Reinsertion Heuristic for the Static Dial-a-Ride Problem</b>	<b>37</b>
3.1 Operating Scenario of the Service	38
3.2 Proposed Insertion-Based Rejected-Reinsertion Heuristic	40
3.2.1 Rejected-reinsertion operator	41

3.2.2	Improvement procedure	45
3.2.3	Variable vs fixed fleet size	46
3.2.4	Complete heuristic procedure	47
3.2.5	Feasibility check for inserting a request	47
3.2.6	Feasibility check for removing a request	51
3.2.7	Insertion criterion	52
3.2.8	Vehicle scheduling	53
3.3	Computational Study	54
3.3.1	Randomly generated problems	55
3.3.2	Diana and Dessouky's test problems	66
3.4	Conclusions	73
<b>4</b>	<b>Online Heuristics for the Dynamic Dial-a-Ride Problem</b>	<b>75</b>
4.1	Operating Scenario of the Dynamic Problem	75
4.2	Online Insertion Heuristics	77
4.2.1	Immediate online insertion heuristic	77
4.2.2	Rolling horizon online insertion heuristic	78
4.3	Computational Study	81
4.3.1	Comparison of rolling horizon and immediate online insertion heuristics	81
4.3.2	Performance of the periodical improvement procedure	87
4.3.3	Test of the effectiveness of the rejected-reinsertion operation	90
4.3.4	Effect of advance information	91
4.3.5	Sensitivity analysis of parameter settings of the rolling horizon online heuristic	96
4.3.6	Comparison of two vehicle scheduling policies	99
4.4	Conclusions	101
<b>5</b>	<b>Development of Performance Metamodels</b>	<b>102</b>
5.1	Introduction of Response Surface Methodology	103
5.2	Experimental Design	105

5.2.1	Input factors	106
5.2.2	Region of interest	109
5.2.3	Factorial design and face-centered composite design	109
5.2.4	Generation of demand scenarios	112
5.3	Regression Analysis	113
5.3.1	Vehicle resource requirement model	114
5.3.2	Time deviation model	123
5.3.3	Ride time ratio model	131
5.4	Metamodel Validation	141
5.4.1	Vehicle resource requirement model	142
5.4.2	Time deviation model	143
5.4.3	Ride time ratio model	144
5.5	Model Summary	150
<b>6</b>	<b>Sensitivity Analysis and Model Applications</b>	<b>152</b>
6.1	Sensitivity Analysis	153
6.1.1	Shape of the service area	153
6.1.2	Demand distribution in space	158
6.1.3	Percentage of passengers specifying desired pickup time	161
6.2	Model Applications	164
6.2.1	Parametric analysis	164
6.2.2	Optimization of the dial-a-ride service	171
<b>7</b>	<b>Conclusions and Future Work</b>	<b>178</b>
7.1	Conclusions	178
7.2	Future Work	182
	<b>Notation</b>	<b>184</b>
	<b>References</b>	<b>186</b>



## List of Tables

Table 2-1. Comparison of multi-vehicle dial-a-ride algorithms	33
Table 3-1. Constraint settings for four service quality scenarios	56
Table 3-2. Results of six algorithm variations for scenario L	58
Table 3-3. Results of six algorithm variations for scenario M	58
Table 3-4. Results of six algorithm variations for scenario H	59
Table 3-5. Results of six algorithm variations for scenario VH	59
Table 3-6. Comparison of the rejected-insertion heuristics with parallel insertion heuristic	61
Table 3-7. Variability of number of vehicles over five replications	63
Table 3-8. Comparison of algorithms with and without rejected-reinsertion operator	64
Table 3-9. Effects of initial fleet size on final fleet size with Algorithms 4 and 4w	65
Table 3-10. Constraint settings for three service quality scenarios	69
Table 3-11. Computational results for scenario L of the 1000-request problem	70
Table 3-12. Computational results for scenario M of the 1000-request problem	70
Table 3-13. Computational results for scenario H of the 1000-request problem	71
Table 4-1. Comparison of rolling horizon vs immediate online insertion heuristics for scenario L	83
Table 4-2. Comparison of rolling horizon vs immediate online insertion heuristics for scenario M	84
Table 4-3. Comparison of rolling horizon vs immediate online insertion heuristics for scenario H	85

Table 4-4. Variability of number of vehicles over five replications	87
Table 4-5. Comparison of online heuristics without and with periodical improvement for scenario L	88
Table 4-6. Comparison of online heuristics without and with periodical improvement for scenario M	89
Table 4-7. Comparison of online heuristics without and with periodical improvement for scenario H	89
Table 4-8. Comparison of heuristics without and with rejected-reinsertion operation	90
Table 4-9. Sensitivity to the horizon	97
Table 4-10. Sensitivity to the rolling interval $\alpha$	98
Table 4-11. Sensitivity to the improvement interval	98
Table 4-12. Effects of two scheduling policies on fleet size	100
Table 4-13. Effects of two scheduling policies on average passenger time deviation	100
Table 5-1. Lower and upper values of the factors	109
Table 5-2. Factor combinations for a face-centered CCD for $k = 3$	112
Table 5-3. Observed vs estimated values from the performance models for the 30 validation experiments	148
Table 6-1. Expected Euclidean distance $D$ with different aspect ratios	154
Table 7-1. Vehicle reductions due to rejected-reinsertion heuristics for static problem	179
Table 7-2. Vehicle reductions due to rejected-reinsertion rolling horizon heuristics for dynamic problem	179
Table 7-3. Vehicle reductions due to rolling horizon heuristics for dynamic problem	180

## List of Figures

Figure 3-1. Transformation of the time deviation and maximum ride time constraints into time windows	40
Figure 3-2. Illustration of the rejected-reinsertion method	43
Figure 3-3. Removing a request from one schedule block	52
Figure 3-4. Demand distribution over time	67
Figure 3-5. Direct travel time distribution	68
Figure 4-1. Definition of the lead time	76
Figure 4-2. Schematic representation of the rolling horizon online insertion principle	80
Figure 4-3. Vehicle fleet size requirement vs average lead time for scenario H	92
Figure 4-4. Vehicle fleet size requirement vs average lead time for scenario M	92
Figure 4-5. Vehicle fleet size requirement vs average lead time for scenario L	93
Figure 4-6. Effect of minimum lead time for scenario H	94
Figure 4-7. Effect of minimum lead time for scenario M	95
Figure 4-8. Effect of minimum lead time for scenario L	95
Figure 5-1. Central composite design (CCD) for $k = 3$	111
Figure 5-2. Normal probability plot of $\log_{10} F$ of the multiplicative Model F1	117
Figure 5-3. Residual vs the predicted $\log_{10} F$ of the multiplicative Model F1	117
Figure 5-4. Estimated vs observed $\log_{10} F$ of the multiplicative Model F1	118
Figure 5-5. Estimated vs observed $F$ of the multiplicative Model F1	118
Figure 5-6. Normal probability plot of $\log_{10} F$ of the first-order Model F2	119

Figure 5-7. Residual vs the predicted $\log_{10} F$ of the first-order Model F2	119
Figure 5-8. Estimated vs observed $\log_{10} F$ of the first-order Model F2	120
Figure 5-9. Estimated vs observed $F$ of the first-order Model F2	120
Figure 5-10. Normal probability plot of $\log_{10} F$ of the second-order Model F3	121
Figure 5-11. Residual vs the predicted $\log_{10} F$ of the second-order Model F3	121
Figure 5-12. Estimated vs observed $\log_{10} F$ of the second-order Model F3	122
Figure 5-13. Estimated vs observed $F$ of the second-order Model F3	122
Figure 5-14. Normal probability plot of time deviation of Model D1	125
Figure 5-15. Residual vs the predicted time deviation of Model D1	125
Figure 5-16. Estimated vs observed time deviation of Model D1	126
Figure 5-17. Normal probability plot of time deviation of Model D2	127
Figure 5-18. Residual vs the predicted time deviation of Model D2	127
Figure 5-19. Estimated vs observed time deviation of Model D2	128
Figure 5-20. Normal probability plot of time deviation of Model D3	129
Figure 5-21. Residual vs the predicted time deviation of Model D3	129
Figure 5-22. Estimated vs observed time deviation of Model D3	130
Figure 5-23. Normal probability plot of ride time ratio of the first-order Model R1	133
Figure 5-24. Residual vs the predicted ride time ratio of the first-order Model R1	133
Figure 5-25. Estimated vs observed ride time ratio of the first-order Model R1	134
Figure 5-26. Normal probability plot of ride time ratio of the first-order Model R2	135
Figure 5-27. Residual vs the predicted ride time ratio of the first-order Model R2	135
Figure 5-28. Estimated vs observed ride time ratio of the first-order Model R2	136

Figure 5-29. Normal probability plot of ride time ratio of the second-order Model R3	137
Figure 5-30. Residual vs the predicted ride time ratio of the second-order Model R3	137
Figure 5-31. Estimated vs observed ride time ratio of the second-order Model R3	138
Figure 5-32. Normal probability plot of ride time ratio of the multiplicative Model R4	139
Figure 5-33. Residual vs the predicted ride time ratio of the multiplicative Model R4	139
Figure 5-34. Estimated vs observed ride time ratio of the multiplicative Model R4	140
Figure 5-35. Model validation: Estimated vs observed vehicles of the multiplicative Model F1	142
Figure 5-36. Model validation: Estimated vs observed time deviation of the Model D3	143
Figure 5-37. Model validation: Estimated vs observed ride time ratio of the first-order Model R1	145
Figure 5-38. Model validation: Estimated vs observed ride time ratio of the first-order Model R2	145
Figure 5-39. Model validation: Estimated vs observed ride time ratio of the second-order Model R3	146
Figure 5-40. Model validation: Estimated vs observed ride time ratio of the multiplicative Model R4	146
Figure 6-1. Average direct travel time vs aspect ratio	155

Figure 6-2. Effect of area shape on vehicles needed for service scenario H	156
Figure 6-3. Effect of area shape on vehicles needed for service scenario M	157
Figure 6-4. Effect of area shape on vehicles needed for service scenario L	157
Figure 6-5. Probability density function for linear distribution of demand density along one side of the service area	159
Figure 6-6. Effect of non-uniform demand distribution on vehicles needed for service scenario H	159
Figure 6-7. Effect of non-uniform demand distribution on vehicles needed for service scenario M	160
Figure 6-8. Effect of non-uniform demand distribution on vehicles needed for service scenario L	160
Figure 6-9. Effect of percentage of passengers with desired pickup time on vehicles required	162
Figure 6-10. Effect of percentage of passengers with desired pickup time on average time deviation	162
Figure 6-11. Effect of service demand density on vehicles required	165
Figure 6-12. Effect of service demand density on vehicle productivity	166
Figure 6-13. Effect of service area size on vehicles required	167
Figure 6-14. Effect of maximum time deviation on vehicles required	168
Figure 6-15. Effect of maximum time deviation on vehicle productivity	169
Figure 6-16. Effect of maximum ride time ratio on vehicles required	170
Figure 6-17. Effect of maximum ride time ratio on vehicle productivity	170
Figure 6-18. Effect of both time constraints on vehicle productivity	171

Figure 6-19. Cost components of the dial-a-ride service in the case study	175
Figure 6-20. Total cost of the dial-a-ride service in the case study	175
Figure 6-21. Costs vs demand density of the dial-a-ride service in the case study	176

## List of Abbreviations

DAR	Dial-A-Ride
DARP	Dial-A-Ride Problem
PDP	Pickup and Delivery Problem
SB	Schedule Block
VRP	Vehicle Routing Problem



# **Chapter 1 Introduction**

## **1.1 Problem and Motivations**

At the planning or design stage of transportation systems, explicit performance models of a proposed system are often very useful in facilitating optimization of the system in terms of its controllable variables, comparing and/or selecting of alternatives, or gaining insights into the system relations. This dissertation develops performance models for dynamic many-to-many dial-a-ride (DAR) paratransit service. The main performance model is a vehicle resource requirement model, which predicts the minimum vehicle fleet size required to provide a given level of service to a given demand level. The other two models estimate level of service attributes, which predict respectively the average passenger time deviation from their desired pickup or delivery time and the average ratio of the passenger actual ride time to the direct ride time.

There are always tradeoffs between operating cost and service quality for the transit service. As more active vehicles operate in the system, the operating cost and the service quality to the users also increase. For conventional bus services with fixed routes and schedules, the relation between operating cost and service quality are relatively

straightforward and easy to quantify. The average route spacing and headway for each route can determine the number of vehicles required. Average passenger access time, waiting time at bus stops and ride time can all be estimated, given the fixed routes and headways. For DAR services, the routes and schedules are not fixed; they are determined to accommodate the transportation requests with different origins and destinations and desired service times. In a dynamic system, the routes and schedules are determined in real-time to accommodate demand occurring during the service time period. The relations between operating cost and service quality for DAR services are not easy to quantify due to the complex nature of the DAR operations. Furthermore, the relations among the system parameters (e.g. vehicle operating speed, service area size, etc.) are not fully understood. With developed performance models, tradeoffs between service quality and operating cost can thus be evaluated quantitatively.

The explicit performance models are intended to be used at the high-level system planning stage. Therefore, they should be inexpensive to use and not require excessive data which might not yet be available at the planning stage. Model prediction should be reasonably accurate and sensitive to relevant policy alternatives (i.e. maximum waiting time, vehicle operating speed and etc.). The form of the models should be relatively simple to use and facilitate the understanding the causal relations.

The models can be used as part of the formulation of an optimization models for DAR systems or integrated systems (conventional bus and DAR) to determine the system parameters considering both operator cost and user cost. For example, the analyst could

determine the optimal upper limits for the passenger time deviation and ride time ratio and determine the resulting vehicle fleet requirement that would minimize the combined operator cost and user cost. In a potential integrated system (e.g. use the entire fleet for conventional bus service during peak hours and use the excess fleet during off-peak to provide DAR service with higher service quality to low-density surrounding areas), the performance models can be input to a vehicle resource allocation model to determine how many vehicles should be allocated at various times to the DAR and conventional bus services.

The models can be used to determine the threshold demand level separating the domains in which DAR service and conventional bus service will operate more cost-effectively. It is generally thought that DAR services are suitable in areas or time periods with low demand densities. A threshold analysis can determine the approximate numerical value for the demand level.

The performance models assist in the demand forecasting for DAR systems under a demand equilibrium environment. Since the demand for passenger transportation services is quite sensitive to the level of service provided (Wilson and Hendrickson, 1980), the level of service predicted by the performance models can be used to forecast the elastic demand.

An explicit performance model is also useful in quantifying the effects of system parameters (such as vehicle speed and maximum ride time ratio) on the performance. For

example, the operator may increase the vehicle operating speed to lower the vehicle fleet requirement. With a performance model, the operator may estimate how many vehicles would be saved. Detailed knowledge or analysis of what interactions exist among the parameters and how these interactions affect the performance of a system would be available through sensitivity analysis.

Few performance models for DAR systems have been developed. The literature on the existing performance models is reviewed in Section 2.1. The development of the performance models for DAR services is desirable and useful, especially for the service planning stage.

## **1.2 Background of Dial-a-Ride Services**

DAR paratransit is one of the public transit services which can provide shared-ride door-to-door service with flexible routes and schedules. DAR was initially designed for service to the general public. It generally provides a higher quality of service (e.g. negligible access time, wait at home and no transfers) but increases operating cost due to a lower vehicle productivity (e.g., passenger trips per vehicle hour) than conventional bus services. Due to the required subsidy, DAR service to the general public is usually limited to suburban areas or time periods with low demand densities and service as a feeder to line-haul systems, in those situations where they operate more cost-effectively. In some cities, DAR is limited to use by a special group of persons with mobility

difficulties (handicapped or elderly persons) who are not able to access other public transportation services.

A DAR system is made up of a control center and a fleet of small vehicles (usually < 20 seats) compared with conventional buses. The vehicles, operating with flexible routes and schedules, respond to requests for transportation as they are received by the control center. Each customer will provide information about the locations of his/her origin and destination, the desired time of pickup or delivery, and number of riders. The dispatcher in the control center will combine the customer information with information regarding vehicle positions and their tentative routes to plan the new routes for vehicle using manual or automated dispatching techniques. The passengers are provided the expected pickup time. Unlike taxi service, DAR services allow ridesharing and thus reduce cost per passenger. Early demand-responsive systems used manual dispatching techniques. With technological advances in computer hardware and software, automatic computer dispatching algorithms are available to many current services. Furthermore, Advanced Vehicle Location (AVL), Global Position Systems (GPS), Geographical Information Systems (GIS) and similar systems are making the real-time dispatching more feasible.

In the existing systems or algorithms, two types of service requests are considered: advance requests and immediate requests. The advance requests usually refer to those received at least one day before the service is provided, so that routes and schedules can be planned at the start of the day of service. Service provided to handicapped persons often requires advance requests. If all the requests are advance requests (and assuming all

other factors, such as traffic conditions, are predictable), then the determination of the routes and schedules is a static Dial-A-Ride Problem (DARP). Immediate requests are those asking for service as soon as possible without a designated desired pickup or delivery time, such as in Wilson et al. (1971). In practice, a reasonable service provided to the general public would allow the service requests with specified desired pickup or delivery time throughout the service period (without the requirement of 24-hour reservation), probably at least some time in advance (e.g. 20 minutes) for the efficient route and schedule planning. This kind of service is considered in this study. Except when serving only previous-day advance requests, the routing and scheduling of a DAR service is a dynamic problem, in which decisions for requests coming throughout the operating period are made in real time.

DAR services may be classified as many-to-many, many-to-few and many-to-one, depending on the demand patterns and the service quality to be achieved. Many-to-many means passengers can differ in their origins and destinations. If all the passengers are picked up or delivered at the same location (e.g. a shopping center), the service is many-to-one. One example of many-to-one service is the feeder service in a local area, in which all passengers are collected to feed a metro station. Many-to-few lies between those two extreme conditions, where there are a few common origins and/or destinations.

### **1.3 Objectives and General Methodology**

The main research objective is to develop analytical performance models for dynamic many-to-many DAR services, which predict the fleet size requirement and the attributes related to passenger time. The main purpose of the models is to assist in system planning or alternative evaluation for DAR or integrated systems. Therefore, the models should be easy to acquire and use (e.g. in explicit form other than running the simulation to get the performance results), and comprehensively take into account the effect of various system parameters (e.g. area covered, maximum time deviation, vehicle operating speed, etc.) on the performance measures.

The prediction of the performance measures for DAR services is not as straightforward as for fixed-route bus systems due to the complex nature of the DAR operations: passenger requests come-in in real-time, the DAR routes and schedules are flexible and change day by day, the operation of DAR requires solving the DARP problem considering special passenger precedence and travel time constraints, the solution of the problem is usually near-optimal, and the performance measures are closely related with vehicle operating speed, time constraints, service coverage etc. Manual dispatching is relatively simple for very small system, but is much less efficient than computerized dispatching with sophisticated algorithms, and seems obsolete especially when more powerful and inexpensive computers are available these days. From the literature review in Section 2.1., it is found that some existing performance models were developed through theoretical analysis (e.g. based on geometric probability or queuing theory) and they generally suffer from the limitations by using manual or very simple vehicle routing

algorithm and not taking into account the passenger time constraints. Meanwhile, real DAR data are rare and not easy to acquire, and most of them are those for handicapped DAR services. (The handicapped DAR services differ from the general DAR services in that the former have lower demand, and usually allow longer time deviation and require 24-hour advance reservation.) Thus, regression models developed with extensive real data are not practical, at least yet.

Simulation remains the most effective and accepted approach to represent complex systems. However, simulation models are not directly suitable for high-level decision making. In this dissertation, response surface metamodeling approach has been used to develop the performance models, in which simulation experiments are designed and executed, and simulation data are collected and used in the regression analysis. The metamodels developed are much less expensive to use than running simulations directly each time. On the other hand, more sophisticated computerized routing and scheduling algorithm can be incorporated into the simulation experiment to better represent the complex operation of the DAR service.

Thus, the second main objective of this research is to develop an advanced online heuristic for the large-scale dynamic DARP, which could efficiently assign real-time requests into vehicle routes and determine their schedules. The dynamic algorithm should be advanced in terms of the performance, computationally efficient, and reasonably applicable to real dynamic systems. DARP algorithms have been proposed since the 70's, mostly for the static version of the problem. In this study, two online heuristics, online



immediate insertion heuristic and online rolling horizon heuristics are developed and compared. Real-time requests with different lead times (a measure of how far in advance the request is made) are considered. The effect of lead time of the requests on the operating efficiency is also examined.

Parametric analysis of the model is performed as a single parameter is varied to better understand the interrelations in the system. Especially important is the tradeoff relation between the vehicle fleet size requirement and the level of service provided, which are closely related with the system operating cost and user cost. The tradeoffs can be quantified and evaluated by the performance models. The models are also applied in the optimization of the service considering the combined operator cost and user cost.

#### **1.4 Organization of the Dissertation**

Chapter 2 is devoted to the literature review of the existing performance models and the algorithms for the static and dynamic DARPs. Chapter 3 develops an insertion-based rejected-reinsertion heuristic for the static dial-ride problem, which is the basis for the online heuristics developed for the dynamic version of the problem in Chapter 4. In Chapter 5, performance metamodels are developed using response surface metamodeling approach. Model validation is also addressed in this chapter. Chapter 6 analyzes the sensitivity of the performance to two of the assumptions made in the development of the models and illustrates two model applications. Finally, Chapter 7 provides the conclusions of the dissertation and discusses further research directions.

## **Chapter 2 Literature Review**

In this chapter the literature is reviewed in two areas: existing performance models for DAR services, and algorithms for both static and dynamic DARPs.

### **2.1 Existing Performance Models for Dial-a-Ride Services**

Computer simulation is the first and the most generally accepted approach to predict system performance of demand responsive systems (Wilson and Hendrickson, 1980). Simulation models are capable of generating individual service requests from specified time and space distributions, employing the specified routing and scheduling algorithm and get disaggregate measures, which are summarized statistically to indicate the system performance. In this way, simulation models are able to replicate the complex nature of the DAR operating system and get a reliable estimate of the true performance measures. Simulation models for DAR systems have been developed by Heathington et al. (1968), Wilson et al. (1970, 1976), and Fu (2002a). However, simulation models tend to be difficult to acquire and typically require fairly sophisticated planners with no pressing time constraints to use successfully (Wilson and Hendrickson, 1980). They are also time-

consuming to develop for specific cases. Especially if elastic demand is considered the simulation models must be executed repeatedly. The simulation models are not quite suitable at the planning level since no explicit relation exist between system outputs and inputs. Despite that, simulation still plays an important role in producing simulated data and calibrating analytical performance models, as will be seen in Section 2.1.2.

The following literature review on existing performance models will be confined to analytical models. Research on analytical DAR performance models is limited. There are two general methods to obtain the analytical performance models: theoretical analysis and statistical methods based on real or simulated data.

### 2.1.1 Theoretical analysis

Stein (1978a, 1978b) conducted an analytic investigation into the lengths of optimal bus tours for the DAR transportation systems. He has developed asymptotic equation for a many-to-many DAR system with a single bus:

$$\lim_{n \rightarrow \infty} Y_n^* = \frac{4}{3} \sqrt{2} \cdot b \sqrt{nA} \quad (2-1)$$

$Y_n^*$  is the length of an optimal bus tour through  $n$  random demand pairs.  $b$  is a constant.  $b$  has been roughly estimated at 0.75 for the Euclidean distance by manually constructing tours for a 202- and 400-city instances by Beardwood et al. (1959). The value was later corrected to be 0.7124 by more substantial experimental studies (Johnson et al., 1996). Furthermore, by assuming that transfers are permitted and take no time, Stein is able to develop an asymptotic equation for the multiple-bus case.

$$\lim_{n \rightarrow \infty} Y_n^k = \frac{4}{3} \sqrt{2} \frac{b}{k} \sqrt{nA'} \quad (2-2)$$

$Y_n^k$  is the length of an optimal bus tour for case with  $k$  buses. The approach taken for the multiple-bus case is based upon the decomposition of the area into regions and upon the specialization of buses within these regions. The objective is to minimize time to completion as well as total travel time. Equation (2-2) is based on a quite idealized transfer condition and no user inconvenience is considered in the objective function. However, the equation is asymptotical in the number of demand points and does not take into account any user time constraints.

Daganzo et al. (1977) present analytical models for waiting time, ride time and total service time of a many-to-one DAR system where buses periodically visit a fixed point, which is either the origin or the destination of every trip. The models are for zones instead of the whole service area, in which one vehicle operates in one zone. The approach taken is to develop a steady-state deterministic model of the single vehicle operation using a fluid queuing approximation with service rates derived from geometrical probability for the expected distance between a random point and the nearest of a set of randomly distributed points in a zone. The simple next-nearest routing algorithm is employed. The collected and distributed passengers are treated in two separate phases, which is not efficient. The resulting models are adjusted, to some extent, to reflect the stochastic nature of the demand process and the integer nature of customer service. The same expression for the expected distance is used to derive an approximate analytical model of many-to-many demand responsive service using three variants of the next-nearest strategy (Daganzo, 1978). He further approximately models the request

arrival process as a time-homogeneous Poisson process and the service rates as mutually independent negative exponential variables independent of the arrival process, in order to handle the stochastic nature of the problem. The use of the next-nearest routing algorithm in both studies is fairly restrictive, since it is not able to take into account the time distribution of the passenger requests. The passengers may experience intolerably long waiting time and ride time under the next-nearest routing strategy.

Later, Daganzo (1984a, 1984b) provides an expression for predicting the tour length of a vehicle visiting a set of demand points in a zone served by a single vehicle by using a simple manual routing strategy. The depot influence area is first partitioned into districts containing clusters of stops; one vehicle route is then constructed to serve each cluster. For each vehicle route, first a swath is cut covering the whole zone, and then the tour moves forward along the swath. Again, the manual routing strategy seems too restrictive and no time constraints are considered.

Lerman and Wilson (1974) have modeled the many-to-many service by using an M/M/1 system, in which the mean of the exponentially distributed service time is based on a linear function of trip length and productivity. The linear function is calibrated using simulation results. Wait time is based on the average distance between the vehicle assigned and the passenger's origin, which is assumed to be a linear function of the vehicle density and the demand density. The predictions are considered valid only in relatively uncongested systems, and an assumption of linearity in interstop distance with productivity certainly suggests that the model would at best be useful only within a

narrow range (Wilson and Hendrickson, 1980). The supply model is part of the first attempt to model demand responsive systems in an equilibrium framework.

### **2.1.2 Statistical methods based on real or simulation data**

Wilson et al. (1971) have developed the first empirical model for many-to-many DAR service using an intuitive model form, calibrated with data from simulation experiments with combined many to several and many-to-many demands:

$$N = \frac{A(0.68 + 0.072D)}{(LOS - 1)^{1/2}} \quad (2-3)$$

where  $N$  = number of operating vehicles,  $A$  = area size in square miles,  $D$  = demand density in trips per square mile per hour,  $LOS$  = mean ratio of total service time (waiting + travel) to direct driving time ( $LOS > 1$ ). The results are based on simulation experiments with limited variations in operating conditions (e.g. area sizes of  $3 \times 3$  and  $5 \times 5$  square miles, minimum demand density as 10 demands per square miles per hour). The demands are assumed to be served as soon as possible. Furthermore, variations in vehicle speed, time constraints are not considered in Equation (2-3).

Arillaga and Medville (1974) have developed demand, supply and cost models by fitting a simple linear form to observed operating data, based on results from thirteen existing systems (data of three of sixteen surveyed sites are not included in the models) of various operating types (i.e. many-to-many, many-to-few and many-to-one). Thus, the models do not reflect the differences in operating systems and the thirteen sets of data used for estimation are very limited. Furthermore, the models fail to capture the critically

important non-linear character of the performance relations (Wilson and Hendrickson, 1980).

The following models for many-to-many DAR service have been developed by Flusberg and Wilson (1976) with separate prediction of waiting time and ride time, and calibrated with the MIT simulation model (Wilson et al., 1971):

$$\text{Wait time: } WT = \frac{f_a}{2V_{eff}} \sqrt{\frac{A}{N}} \exp\left(0.22 \cdot \sqrt{\frac{A+4}{N+12}} \cdot \lambda^{0.9}\right) \quad (2-4)$$

$$\text{Ride time: } RT = \frac{f_a L}{V_{eff}} \exp\left(0.084 \cdot \left(\frac{A\lambda}{N}\right)^{0.7}\right) \quad (2-5)$$

where  $f_a$  = ratio of street distance to airline distance,  $V_{eff}$  = effective vehicle speed including passenger loading and unloading times,  $\lambda$  = vehicle productivity,  $L$  = mean direct trip length. The model form is developed through observation of the relationship between service levels and the parameters, in both actual systems and experience with the simulation model. The model is developed as part of a combined supply/demand/equilibrium model of many-to-many DAR and shared ride taxi systems (Lerman et al., 1977; cited from Flusberg and Wilson, 1976). A set of adjustments are developed to model service times under manual dispatching and alternative computer control algorithms, when waiting and ride time are not weighted equally, as well as for the degradation in service to immediate-request passengers due to the priority given to advanced request passengers (Menhard, et al., 1978).

Recently, Fu (2003) has developed an analytical model which predicts the minimum fleet size requirement for many-to-many static DAR service by calibrating the proposed model form with simulation data:

$$FS = \frac{\lambda_T}{E^{0.2}} \left( \tau + \frac{4.62}{V} \left( \frac{A}{\lambda_T \cdot T} \right)^{0.31} \right) \quad (2-6)$$

where  $FS$  = minimum fleet size,  $\lambda_T$  = peak trip rate,  $E$  = maximum allowable ratio of excess ride time (the difference between the actual ride time and the direct ride time) to direct ride time,  $\tau$  = boarding plus alighting time,  $A$  = size of the service area, and  $T$  = trip service (pickup/delivery) time window. A sequential insertion heuristic is used and the objective is to minimize the number of fleet size while satisfying all the demand for given service quality constraints (time window and ride time constraints). The model is based on three demand density settings of 1.29, 2.59 and 3.88 trips/mi<sup>2</sup>/hour (0.5, 1.0 and 1.5 trips/km<sup>2</sup>/hour). All the origins and destinations are uniformly distributed in square areas.

Tour length expression for a vehicle visiting a set of demand points in a zone served by a single vehicle has been developed from simulation experiments by Mason and Mumford (1972). The tour length expressions can assist the design of many-to-one systems, in which the whole service region may be partitioned into service zones each served by one vehicle. The limitation is that the partition is not always efficient if time constraints and dynamic demands are considered.



### 2.1.3 Findings

1. Limited number of performance models for DAR systems are available and most of them were developed in the 70's and 80's.
2. Theoretical analysis is usually based on notions of geometric probability or queuing theory and does not take into account the time constraints of passenger requests. Two of the difficulties in analytically modeling DAR systems are inability to represent the vehicle routing algorithm adequately and to accommodate the time constraints (i.e. simple next-nearest strategy without time constraints).
3. Real data for DAR operations are rare and the operations differ considerably in operating conditions such as area covered, form of DAR implemented and routing algorithm used.
4. Simulation is still a promising method to replicate the complex DAR operation since it can represent the vehicle routing algorithm and take into account other constraints and randomness in the system. It is useful to generate simulation data if real data are rare or unavailable.
5. Available models based on statistical methods and simulation data are limited. They are developed for many-to-many service or combined service only. Most models are based on an MIT simulation model (Wilson et al., 1971), in which the passengers are assumed to be picked up as soon as possible. (In practice, passengers may want to be picked up or delivered close to their desired time.) Variations of some of the system parameters are not sufficiently considered. No relations on the tradeoff between cost and service have been analyzed for the

spectrum of DAR services from many-to-many to many-to-one, where demands are more clustered.

## **2.2 Description of the Static and Dynamic Dial-A-Ride Problems**

In a DAR context, passengers specify transportation requests between given origins and destinations, either with a desired pickup time or delivery time. Transportation is supplied by a fleet of vehicles usually based at a common depot. The aim is to design a set of vehicle routes and schedules capable of accommodating the requests, in order to minimize a certain cost under a set of constraints. General objective functions include those that to minimize the total vehicle travel time/distance to service providers, to minimize passenger inconvenience or dissatisfaction represented by the desired time deviations and/or passenger excess ride times. The most common constraints relate to customer-desired time deviation (the difference between the desired pickup/delivery time and the actual pickup/delivery time should be less than or equal to a pre-specified value), excess ride time (the difference between the actual ride time and direct ride time should be less than or equal to a pre-specified value), and vehicle capacity. One other common constraint considered in the passenger service is that a vehicle is not allowed to wait while carrying passenger(s). Precedence constraints and pairing constraints are implied in the problem. Precedence constraints require that the pickup location of one passenger has to be visited before the delivery location of the same passenger. Pairing constraints require that the passenger should be picked up and delivered by the same vehicle.

The desired time deviation and excess ride time constraints can usually be transformed into time windows on pickups and deliveries as in Jaw et al. (1986). Since DAR is a highly restricted problem, it is possible that not all requests can be served without either violating the time constraints or increasing the given fleet size. Consequently, at least one of the following has to be allowed in the algorithm development:

- increase of the fleet size
- rejection of part requests
- use of soft time windows.

Researchers classify problems as static and dynamic based on whether the problem is fully known with all its input information (e.g. demand, travel times) for the time period considered. In a static version of DARP, all the requests are known in advance (e.g. all the passengers call at least one day before their desired trips in a DAR service) and travel times are predictable. The algorithm can be executed once at the beginning of service. In a dynamic version, passengers call for trip requests throughout the day. Thus, the vehicle routes and schedules are adjusted in real-time.

The Dial-A-Ride Problem (DARP) is a generalization of the Pickup and Delivery Problem (PDP) and the Vehicle Routing Problem (VRP), which are NP-hard. The DARP is a PDP in which the loads to be transported represent people. In DARP, maximum excess ride time constraints are usually considered. The DARP is different from and somewhat more difficult than most other routing problems due to the above mentioned precedence and travel time constraints, and also because operator cost and user

inconvenience must be weighted against each other when designing a solution instead of considering the operator cost alone. For overviews, see Bodin et al. (1983) for general routing and scheduling of vehicles and crews, Solomon (1987) and Desrosiers et al. (1995) for vehicle routing and scheduling problems with time window constraints, Savelsbergh et al. (1995), Mitrovic-Minic (1998, 2001) and Desaulniers et al. (2002) for general pickup and delivery problem, and Cordeau and Laporte (2003) for DARP. The following review will focus on the scientific literature specific to the DARP.

### **2.3 Dial-A-Ride Problem Algorithms**

Algorithms for the DARP can be categorized based on whether they are designed for the static or dynamic version of the problem, for single- or multiple-vehicle system, with or without time windows, and exact or heuristic. Below, algorithms for the single-vehicle problem will be reviewed first. Then the algorithms for the multiple-vehicle problem with time windows are categorized based on the general methods used: insertion-based, cluster-first route-second, metaheuristics and post-improvement.

A single-vehicle problem is rarely applicable in practice. However, it is considered as a sub-problem of some multi-vehicle DARP (especially in cluster-first route-second algorithms). Psaraftis (1980) has developed an exact dynamic programming algorithm for the single-vehicle many-to-many static version of the problem without time window. User inconvenience is controlled through a “maximum position shift” constraint limiting the difference between the user’s position in the list of requests and that position in the

vehicle route. The objective function is a weighted combination of the time to service all customers and the total degree of dissatisfaction experienced by customers while waiting for service. The dissatisfaction is assumed to be a linear function of each customer's waiting and ride times. The solutions for the dynamic version are based on reoptimization every time a new request was received. Psaraftis (1983) later modified the exact dynamic programming algorithm to be applicable to a similar problem with time windows on each pick-up and drop-off. The computational effort of both algorithms varies exponentially with the size of the problem, and therefore only very small problems can be handled. Less than 10 customers are considered in Psaraftis' example.

Sexton and Bodin (1985a, 1985b) propose a heuristic for a static single vehicle problem. They apply Benders' decomposition procedure to a mixed binary nonlinear formulation of the problem, which separates the routing and scheduling components allowing each to be attacked individually. User inconvenience is measured as a weighted sum of two terms, the excess ride time and the deviation of the desired delivery time. One of the limitations is that all desired delivery times or all desired pickup times must be specified instead of mixing them. Results are reported for up to 7 vehicles and 20 users.

Desrosiers et al. (1986) solve the single-vehicle problem by formulating it as an integer problem and solving it exactly through dynamic programming. It is applied to the solution of instances with up to 40 users.

The algorithms to solve the multi-vehicle DARP with time windows can be categorized into four groups: insertion-based, cluster-first route-second, meta-heuristics and improvement. Those methods might be combined and used in one algorithm (e.g. insertion + improvement). Note that all the following algorithms are exclusively heuristic due to the NP-hard nature of the problem.

### **2.3.1 Insertion-based**

Insertion heuristics have proven to be popular methods for solving a variety of vehicle routing and scheduling problems. They are popular because they are fast, produce decent solutions, are easy to implement, and can easily be extended to handle complicating constraints (Campbell and Savelsbergh, 1998).

In general, an insertion-based algorithm is a method that inserts one passenger request into the vehicle routes at a time, at a position that is feasible to the new passenger and all the passengers already assigned, and results a minimum increase of a pre-specified objective function. A sequential insertion procedure (Kikuchi and Rhee, 1989) constructs one route at a time until all customers are scheduled. A parallel insertion procedure (Jaw et al., 1986; Madsen et al., 1995; Toth and Vigo, 1997; Diana and Dessouky, 2004) is characterized by the simultaneous construction of a number of routes (Solomon, 1987). The disadvantage of the sequential insertion procedure is that workloads of vehicles are uneven: the vehicle whose schedule is built first tends to receive the maximum workload, while the following vehicles receive less workload gradually.

The DARP was first examined by Wilson et al. (1971, 1976 and 1977) in the development of real-time algorithms for the DAR systems of Haddonfield, Jew Jersey and Rochester, New York. The fundamental concept of sequential insertion of customers is developed in those studies. The main requests considered are immediate-requests, which makes the scheduling part of the problem trivial since the requests are satisfied as soon as possible. While these studies sought real-time solutions to the dynamic DARP, it seems that thereafter most work has concentrated on the static version.

Jaw (1984) and Jaw et al. (1986) are among the first few to develop a parallel insertion heuristic for multi-vehicle advance request DARP with time windows. The quality of a solution is measured through a non-linear objective which is a weighted sum of disutility to the system's customers due to excess ride times and desired time deviation and of system operator cost. The problem is solved by sequentially inserting passengers into vehicle routes so as to yield the least possible increase in the objective function value. The core parts of the algorithm are a feasibility check for the attempted insertion and an optimization process to determine the insertion position once the attempted insertion vehicle and insertion sequence have been given. The concept of a "schedule block" is proposed for facilitating the feasibility check. Computation results are included for a real-time dataset with 2617 users and some 20 simultaneously active vehicles covering 16 hours of operation. They also reported that none of the variations of the algorithm they attempted (e.g. considering a group of two or more customers as candidates for the next insertion) have resulted in significant and consistent improvements to the solution obtained through the basic version of the algorithm.

Potvin and Rousseau's (1992) heuristic looks very similar to that of Jaw et al. (1986). The big difference is that instead of inserting the customer into the position with the minimum cost, they maintain the  $W$  (heuristic parameter, called as beam search width) best alternative solutions in parallel at each state and those  $W$  solutions are considered for further expansion. The process is repeated and the best solution out of several final parallel solutions is selected as the result. The solution can be further improved by a post-optimization phase. In this way, they try to alleviate the "myopia" of the insertion heuristic at the expense of greater computation time. The heuristic achieved slightly better solutions for small instances with 90 customers in terms of number of vehicles required, customer ride time and time deviation. The computation time is 2-5 times greater than in the heuristic of Jaw et al. (1986). The performance of the heuristic on the large problems needs further exploration.

Madsen et al. (1995) describe a system for the solution of a static DARP with multiple vehicle capacities and multiple objectives, based on the insertion heuristic proposed by Jaw et al. (1986). The requests are pre-ranked based on some priority parameters. The system does not operate a schedule consisting of blocks in order to reduce the running times for the algorithm. The computation time is relatively low (a few seconds for a problem with 300 customers and 7 vehicles), enabling the algorithm to be implemented in a dynamic environment for on-line scheduling. No detailed description is provided for the online implementation of the algorithm.



Toth and Vigo (1997) describe another parallel insertion heuristic. The heuristic is based on the relaxation of the desired service time constraints by the introduction of a piecewise linear user inconvenience penalty in the objective function. They define a set of parameters which help to initialize a small set of routes each with a single pivot, and then iteratively insert unrouted trips into existing routes, solving at each iteration an assignment problem on a cost matrix obtained by using a modified cheapest insertion criterion based on locally optimal choices.

Diana and Dessouky (2004) have developed a regret insertion heuristic for solving static DARP with time windows. The basic idea is, for all unrouted requests, to calculate a regret matrix, whose rows correspond to unrouted requests and whose columns correspond to routes. Each element of the matrix is defined as the incremental cost by the insertion of the unrouted request to the corresponding route. The request with the largest regret will be inserted into the previously computed position. The regret cost is a measure of the potential price that could be paid if a given request were not immediately inserted. The calculation of the regret matrix here and calculation of the cost matrix in Toth and Vigo (1997) are expensive.

Teodorovic and Radivojevic (2000) combine fuzzy logic reasoning in the insertion procedure to make the decision about which vehicle will accept the new request and to design the new route and schedule for the vehicle chosen to serve the new request. The reasoning process needs the subjective perception of the dispatchers (e.g. extra distance to be traveled by the vehicle by inserting a new request into a vehicle route in terms of

small, medium and big). Though only the fuzzy judgment is needed, it is still beyond the capability of a person for a relatively large problem.

The basic idea of the insertion method was applied in other research with special objectives (Dessouky et al., 2003; Fu, 2002b, 2003). Dessouky et al. (2003) jointly optimize the operator and user cost as well as environmental impact for demand responsive paratransit system. Fu (2002b) schedules the DAR paratransit for time-varying, stochastic condition.

### **2.3.2 Cluster-first route-second or cluster-based**

Cluster-first route-second is a commonly used technique in various VRPs. To be applied in DARP, the cluster phase needs special considerations due to the pairing constraints and time window constraints of DARP.

Bodin and Sexton (1986) develop a cluster first, route and schedule second and swap the third heuristic for the problem, employing a space time heuristic to form a route for customers in a cluster. No detailed procedure is provided for the initial breakdown of customers into vehicle clusters. The heuristic can only handle the condition that every request has a desired pickup time or every request has a desired delivery time. The objective is to minimize total customer inconvenience, which is the weighted sum of the customer delivery time deviation and excess ride time.

Desrosiers et al. (1988) solve the multiple-vehicle DARP by mini-clustering first, routing second. At the first stage, mini-clusters group together nearby customers who can be transported by the same vehicle over a route segment. This grouping into mini-clusters of similar requests deals with local temporal and spatial considerations only. The mini-clusters are obtained by breaking down the routes in an initial solution into segments each time the vehicle becomes empty. At the second stage, routes for all the vehicles are constructed simultaneously by column generation algorithm. This step deals with global considerations by assigning mini-clusters to vehicles. Ioachim et al. (1995) improve the mini-clustering phase by using a mathematical optimization technique to form the mini-clusters and solving the problem by column generation. Borndorfer et al. (1997) use a set partitioning approach for the solution of the problem in both of the clustering step and chaining step. Both set partitioning problems are solved by a branch-and-cut algorithm. Total vehicle travel distance is minimized in both steps. The customer inconvenience is not considered in the objective functions in either Ioachim et al. (1995) or Borndorfer et al. (1997). Incorporating the customer inconvenience is difficult because it is harder to formulate the passenger-related costs than link-related costs in the mathematical programming and also solve the problem efficiently. Baugh et al. (1998) approach the problem by using simulated annealing for clustering and a modified space-time nearest neighbor heuristic for developing the routes within the clusters.

It should be mentioned that the methods of Desrosiers et al. (1988), Ioachim et al. (1995) and Borndorfer et al. (1997) can also be categorized as mathematical programming methods in that either the routing subproblem or both the clustering and routing

subproblems are formulated as an integer nonlinear programming problem or as a set partitioning problem.

### **2.3.3 Metaheuristics**

Metaheuristics such as tabu search and simulated annealing have been tried by researchers in the area of DARP. In metaheuristics, the emphasis is on performing a deep exploration of the most promising regions of the solution space. The methods typically combine sophisticated neighborhood search rules and memory structures. The main disadvantages of such methods are that they are computationally expensive. Gendreau et al. (1992) pointed out that heuristics such as tabu search and simulated annealing are open-ended improvement procedures whose performance is directly related to running time. They are usually context-dependent and need careful calibration of the algorithm parameters to the specific problem in order to produce good results. Generally, those heuristics can produce near-optimal solution if the running time is long enough. Therefore, in absence of the optimal solution for the DARP, solutions obtained from modern heuristics might be used as comparison bases for solution obtained with other heuristics.

Cordeau and Laporte (2003) use a tabu search heuristic for the static DARP. To model the time constraints, they assume that users impose a time window of a pre-specified width on the arrival time of their outbound trip or the departure time window of their inbound trip and that a maximum ride time is associated with each user. The scheduler determines the most suitable pickup and delivery times for the outbound and inbound

trips. The objective function during the search include the total routing cost of the vehicle and the total violation of load, duration, time window and ride time constraints. The algorithm iteratively removes a request and reinserts it into another route. Intermediate infeasible solutions are allowed through the use of the penalized objective function.

Toth and Vigo (1997) have also developed a tabu thresholding procedure, which can improve the solution obtained by their insertion solution. Tabu thresholding is based on the alternation of an improve phase used to reach a local optimum and a mixed phase used to try to escape from it. The neighborhood of the current solution used for the search is subdivided into subsets of moves. At each iteration, one of the subsets is chosen and the best admissible move belonging to the subset, if any, is performed. Trip insertion, trip exchange and trip double insertion are considered as the movements in the local search process. Baugh, et al. (1998) use the simulated annealing in the cluster stage of the cluster-first route-second strategy. Hart (1996) has developed a simulated annealing based solution heuristic for the DARP. The heuristic is computationally expensive (e.g. a 30 or 40 customer problem will require thousands of seconds). The test cases are specially designed without time windows so that the optimal solution is known in order to compare the results.

#### **2.3.4 Local improvement procedures**

Local improvement procedures for the general vehicle routing problem are those that re-sequence stops already assigned within the same route (intra-route) or reassign requests to different routes (inter-route) for a given solution in order to achieve a better solution. If

a given change improves the quality of the solution, it is made and a new solution is obtained. The procedure can be applied until the solution that can no longer be improved. Tour improvement procedures can be applied in the vehicle routing problem after the use of the constructive heuristics.

Van Der Bruggen, et al. (1998) develop a local search method for the single-vehicle pickup and delivery problem with time windows based on a variable-depth search, similar to the Lin-Kernighan algorithm (Lin and Kernighan, 1973) for the traveling salesman problem. They tested the algorithm for problems from 5 to 50 demand pairs with known optimal solutions. For 50-demand problems, the computation times range from 47 to 1035 seconds in increasing order of the time window width, and the maximal relative error compared with optimal value is 3.4%.

Bodin and Sexton (1986) employ a swapper algorithm to reassign customers to form different vehicle clusters in their cluster-first route-second iterations. The swapper algorithm attempts to move customers among the specified vehicle clusters in order to find a final set of vehicle clusters with reduced customer inconvenience. Toth and Vigo (1996) describe local search refining procedures, which can be used to improve the solutions of large-size instances obtained by a parallel insertion heuristic. Intra-route movements are obtained by moving a single stop to a different position of the route, while preserving route feasibility. Inter-route movements include trip insertion, trip exchange and trip double insertion.

### 2.3.5 Dynamic dial-a-ride problem algorithms

As in most combinatorial optimization problems, dynamic aspects of the DARP are not well studied. A straightforward method to deal with the dynamic aspects of the problem is to adapt the static approaches (Wilson et al., 1971; Psaraftis, 1980, 1988, 1995; Mitrovic-Minic et al. 2004; Attanasio et al., 2004). The dynamic problem is solved as a sequence of static problems. Each time an input update occurs, a modified instance of the static problem is solved to update the current solution. One practical problem with this approach is the difficulty of solving the problem in a shorter time interval than the updating interval.

Algorithm variations exist where different updating mechanisms (e.g. eligible requests to be considered in the updated problem, time horizon) and different objective functions are used. In Psaraftis's (1980) study, new passenger requests are automatically eligible for consideration at the time they occur. Psaraftis (1988) describes an algorithm for the dynamic routing of cargo ships. The algorithm is based on a rolling horizon principle. At  $t_k$ , the time at the  $k^{\text{th}}$  iteration, the algorithm considers only those known cargoes whose earliest pickup times are between  $t_k$  and  $t_k + L$ , where  $L$  is the length of the rolling horizon. It then makes a tentative assignment of those cargoes to eligible ships. Only cargoes within the front end of  $L$ ,  $[t_k, t_k + aL]$  for  $a \in (0,1)$ , are considered for permanent assignment. In this way, the algorithm places less emphasis on the less reliable information on future cargo movements. A double-horizon based heuristic for the dynamic pickup and delivery problem with time windows has been developed based on the rolling horizon principle (Mitrovic-Minic, 2001; Mitrovic-Minic et al. 2004). The

heuristic solves a dynamic problem over two time horizons using two goals, with the short-term goal of reducing travel distance and long-term goal of maintaining the routes in a state that will enable them to easily respond to future requests. The heuristic is said to be useful in contexts where problem solutions span a significant part of the service period (e.g. same-day parcel pickup and delivery with wide time windows).

Attanasio et al. (2004) use the parallel computing technique to speed up computation time of the tabu search heuristic of Cordeau and Laporte (2003) in order to be applicable in a dynamic environment. The dynamic algorithm works as follows. A static solution is constructed on the basis of the requests known at the start of the planning horizon. When a new request arrives, the algorithm searches for a feasible solution and then the algorithm performs a post-optimization which is the parallel implementation of the static tabu search of Cordeau and Laporte (2003).

A summary of the main multi-vehicle DARP algorithms is provided in Table 2-1. The computation time of each algorithm (if available) is listed in order to evaluate the computational efficiency of the algorithms and help in choosing the basic category of algorithm to be used for this purposed research.



**Table 2-1. Comparison of multi-vehicle dial-a-ride algorithms**

<b>Authors</b>	<b>Method</b>	<b>Features</b>	<b>Time Window</b>	<b>Objective function</b>	<b>Problem Size</b>	<b>Computer Used</b>	<b>Computation Time</b>
Jaw, et al. (1986)	parallel insertion	static	hard	min weighted sum of disutility to the system's customers and of operator costs	2617 requests * some 20 vehs during 16 hours	VAX 11/750	12 minutes (<0.3 seconds per customer)
Madsen et al. (1995)	parallel insertion	static <sup>[a]</sup>	hard	min weighted goals of driving time, user waiting time, deviation, and capacity utilization	300 requests * 24 vehs	HP-735/9000	< 10 seconds
Toth and Vigo (1997)	parallel insertion	static	soft	min fixed and routing costs (for taxis) and user inconvenience penalties (desired time deviation)	about 300 requests	IBM 486/66	less than 30 seconds
Diana and Dessouky (2003)	regret insertion	static	hard	min weighted sum of travel distance, excess ride time and idle time	500 requests 1000 requests during 24 hours	Pentium III	26 minutes 3.25 hours
Bodin and Sexton (1986)	cluster first, route and schedule second, and swap third	static	hard, one-sided windows	min total customer inconvenience (delivery time deviation, excess ride time)	85 requests * 7 vehs during the afternoon	Univac 1108	roughly 2 or 3 minutes
Desrosiers et al. (1988)	mini-cluster first, route second	static	hard	min number of pieces of work and the travel time	190 requests 880 requests	Cyber173	181 seconds 1305 seconds
Ioachim et al. (1995)	optimization-based mini-clustering	static	hard	min total travel time	250 requests 2545	SUN 4/330-32	5133 seconds N/A
Borndorfer et al. (1997)	cluster first, chain second (set partitioning for solutions)	static	hard	min traveling distances	up to 1771 requests	Sun Ultra Sparc 1 Model 170E	about 7200 seconds

**Table 2-1. Comparison of multi-vehicle dial-a-ride algorithms (Cont')**

<b>Authors</b>	<b>Method</b>	<b>Features</b>	<b>Time Window</b>	<b>Objective function</b>	<b>Problem Size</b>	<b>Computer Used</b>	<b>Computation Time</b>
Baugh et al. (1998)	cluster first (simulated annealing), route second	static	soft	min total travel distance, customer disutility, number of vehicles	over 300 requests a day	N/A	N/A
Hart (1996)	simulated annealing	static	soft	multiple objective functions (e.g. min number of vehicles, min average customer travel time)	40 customers	IBM 486DX2 50 MHz	6252 seconds
Toth and Vigo (1997)	parallel insertion + tabu thresholding	static	soft	min fixed and routing costs (for taxis) and user inconvenience penalties (desired time deviation)	about 300 requests	IBM 486/66	almost one hour
Cordeau and Laporte (2003)	tabu search	static	soft	min total vehicle distance	up to 295 requests * 20 vehs	Pentium 4, 2 GHz	up to 29, 50, 268 minutes <sup>[b]</sup>
Atanasio et al. (2004)	Parallel tabu search	dynamic	soft	min total vehicle distance	N/A	N/A	N/A

a. Static algorithm is presented and it is implemented in a dynamic environment.

b. Times correspond to the two steps, six steps and full procedure algorithms.

### 2.3.6 Findings

1. Exact solutions to the DARP have been limited to relatively small problems and heuristics are widely employed for medium and large problems. Heuristics will be developed in this research for the routing and scheduling of large scale DARP in which the service is provided to the general public.
2. The pure insertion heuristics are generally quite fast, while metaheuristics or optimization-embedded methods (e.g. Optimization-based mini-clustering by Ioachim, et al. 1995) are computationally expensive. The performance of the metaheuristics or post-improvement procedures depends on the available running time.
3. The dynamic aspects of the DARP are not very well studied, as in most combinatorial optimization problems. The basic idea underlying the available dynamic algorithms for the DARP is to solve a static problem each time a new request arrives. The updating mechanism based on each new request might be appropriate for serving only immediate requests, as adopted by Wilson et al., (1971) and Psaraftis (1980); however, it may not be the most efficient method if requests with different lead times (which indicate how advance the passengers make the requests compared with their earliest pickup times) are considered. In this research, the algorithm will be designed to use the advance information available.
4. The insertion-based heuristics are most suitable for adaptation to the dynamic version, in which only waiting requests or some of the waiting requests need to be

inserted into the vehicle tours. In other methods, the whole problem with the updated information must be run again.

5. Therefore, an insertion-based online heuristic for the dynamic DARP will be developed which will take into account the different lead times of the requests.

## **Chapter 3 A Rejected-Reinsertion Heuristic for the Static Dial-a-Ride Problem**

In this chapter an insertion-based rejected-reinsertion heuristic for the multi-vehicle static DARP with service quality constraints is developed. It is the basis for the online heuristics developed for the dynamic DARP described in the next chapter. This study analyzes a static problem, in which all the demands are known at the time when the vehicle routes are planned. Most current DAR services for the elderly and disabled operate in the static mode. The main objective is to minimize the number of vehicles that satisfies all the demand, thus maximizing the vehicle productivity.

This chapter is organized as follows. Section 3.1 describes the basic operating scenario of the DAR service. Section 3.2 presents the proposed rejected-reinsertion heuristic, in which the rejected-reinsertion operator, improvement procedure, variable fleet size, feasibility check of inserting and removing a request, and scheduling are discussed. Two sets of problems are tested and the results are summarized in Section 3.3. The final Section 3.4 contains some concluding remarks.

### 3.1 Operating Scenario of the Service

The operating scenario considered in this study is similar to the one described by Jaw et al. (1986). More specifically,

1. Each passenger  $i$  specifies *either* a desired pick-up time  $DPT_i$  at his/her origin *or* a desired delivery time  $DDT_i$  at his/her destination.
2. Deviation constraint from desired time: A passenger with a desired pick-up time will be picked up during time period  $[DPT_i, DPT_i + TW_i]$  and a passenger with a desired delivery time will be delivered during time period  $[DDT_i - TW_i, DDT_i]$ .  $TW_i$  is the pre-specified maximum deviation from desired time and it is usually the same for all the passengers.
3. Ride time constraint: A passenger's actual ride (in-vehicle) time will not exceed a given maximum ride time  $MRT_i$ , which is usually a function of the passenger's direct ride time  $DRT_i$ .
4. A vehicle is not allowed to wait idly while carrying passengers.
5. Vehicle capacity should not be violated. In the DAR context, due to the low vehicle productivity, the vehicle capacity is usually not a relevant constraint.

The level of service is guaranteed by both the constraint on deviation from desired pickup or delivery time and maximum ride time constraint, which limit the worst case bounds for the service quality. The average service quality will be better than those bounds allow.

The fourth constraint assures that the passengers do not sit in an idle vehicle during their

trips just waiting for other passengers (except to board or exit), which would deteriorate the DAR service quality.

The deviation constraint from desired time and maximum ride time constraint are usually transformed into time windows for pickup and delivery for facilitating the feasibility check for the insertion (Jaw et al. 1986). Define  $EPT_i$  and  $LPT_i$  as the earliest and latest pickup times for request  $i$ , and  $EDT_i$  and  $LDT_i$  as the earliest and latest delivery times for request  $i$ .

For customers specifying desired pickup time ( $DPT$ ):

$$EPT_i = DPT_i \quad (3-1a)$$

$$LPT_i = EPT_i + TW_i \quad (3-1b)$$

$$EDT_i = EPT_i + DRT_i \quad (3-1c)$$

$$LDT_i = LPT_i + MRT_i \quad (3-1d)$$

For customers specifying desired delivery time ( $DDT$ ):

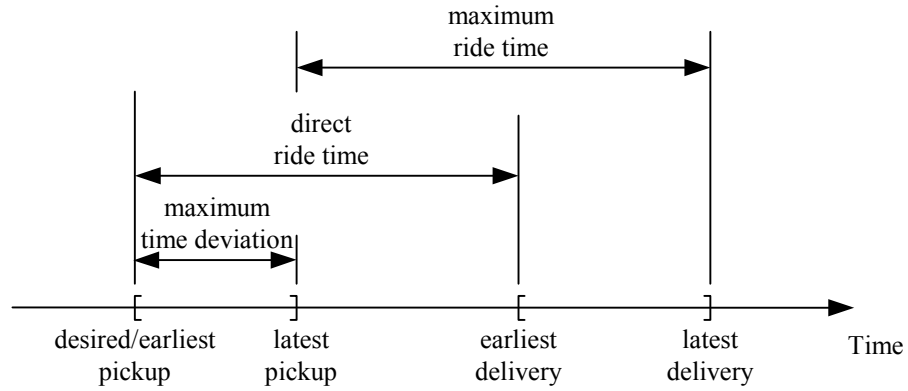
$$LDT_i = DDT_i \quad (3-2a)$$

$$EDT_i = LDT_i - TW_i \quad (3-2b)$$

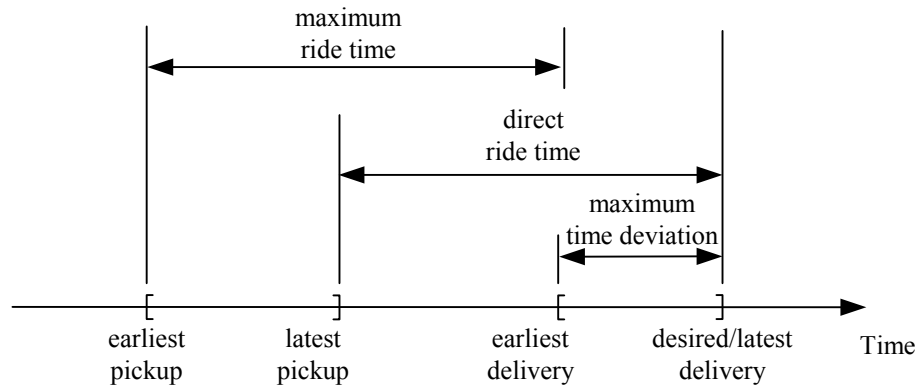
$$LPT_i = LDT_i - DRT_i \quad (3-2c)$$

$$EPT_i = EDT_i - MRT_i \quad (3-2d)$$

Figure 3-1 illustrates how the deviation constraint from desired time and maximum ride time constraint are transformed into time windows for pickup and delivery.



(a) Desired Pickup



(b) Desired Delivery

Figure 3-1. Transformation of the time deviation and maximum ride time constraints into time windows

### 3.2 Proposed Insertion-Based Rejected-Reinsertion Heuristic

Before proceeding further, the basic parallel insertion algorithm (Jaw et al. 1986) is summarized first since it is the basis of the proposed algorithm.



Consider  $N$  passenger requests for service and  $M$  available DAR vehicles. The parallel insertion algorithm (Jaw et al. 1986) first sorts the passengers in sequence (i.e. based on their earliest pickup times). Then each customer is processed in the list in sequence, and assigned to a vehicle until the list of customers is exhausted.

For each customer  $i$  ( $i = 1, 2, \dots, N$ ),

Step 1: For each vehicle  $j$  ( $j = 1, 2, \dots, M$ )

- a) Find all the feasible insertion sequences in which customer  $i$  can be inserted into the work-schedule of vehicle  $j$ . If it is infeasible to assign customer  $i$  to vehicle  $j$ , examine the next vehicle  $j+1$ , and restart Step 1; Otherwise:
- b) Find the insertion of customer  $i$  into the work-schedule of vehicle  $j$  that results in minimum additional cost. Call this additional cost  $C_j$ .

Step 2: If it is infeasible to insert  $i$  into any vehicle  $j$ , then declare a “rejected customer”; otherwise, assign  $i$  to the vehicle  $j^*$  for which  $C_{j^*} \leq C_j$  for all  $j$  ( $j = 1, 2, \dots, M$ ).

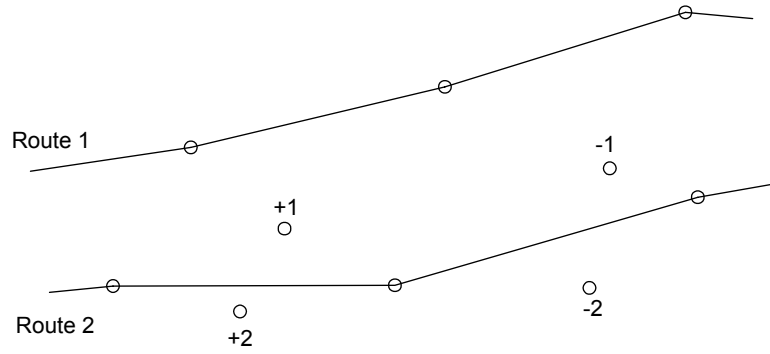
Algorithm variations exist, depending mostly on the sorting scheme, insertion criteria and the determination of the vehicle schedules once an insertion sequence is determined. We sort the passengers by their earliest pickup times. Insertion criteria and vehicle scheduling will be discussed in later sections.

### **3.2.1 Rejected-reinsertion operator**

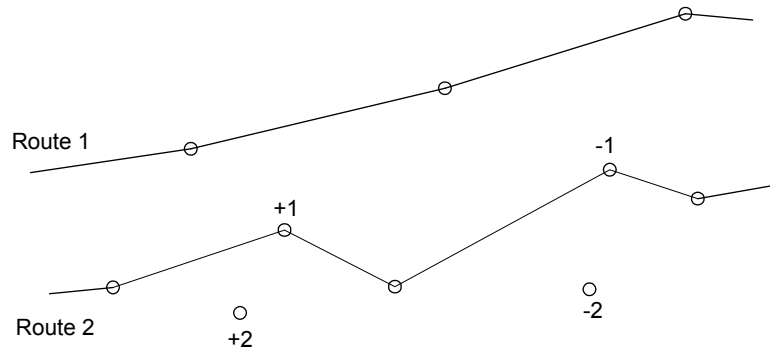
The main disadvantage of the insertion method is that it works in a myopic way in that each request is inserted into its current best position without having an overview of all the

requests. The regret insertion heuristic (Diana and Dessouky, 2004) alleviates the problem by calculating for each unassigned request its regret, which is a measure of the potential cost that could be paid if the given request were not immediately inserted, and inserting the request with the largest request. Local improvement procedures such as swapping the customers into different routes or reinserting the customer could also improve the routing and scheduling in terms of an explicit objective function (i.e. total vehicle travel distance).

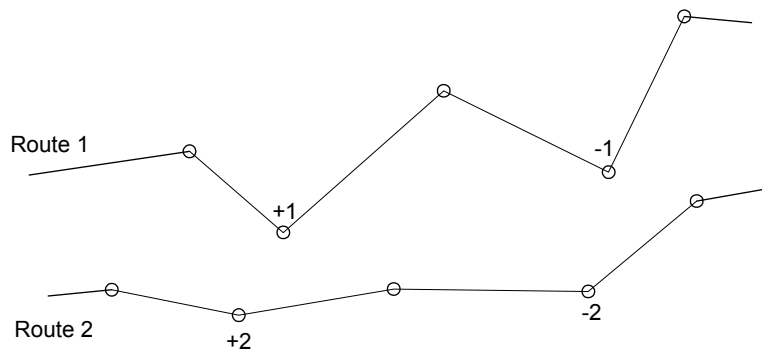
The basic idea of the rejected-reinsertion operation can be illustrated in Figure 3-2 using a simplified scenario. Consider the scenario in Figure 3-2(a) with two vehicle routes and two new requests 1 and 2 to be scheduled. '+' and '-' represent the origin and destination of a request, respectively. Assume request 1 has the earlier pickup time so that it will be scheduled first. Also assume that it is feasible to insert request 1 into either route 1 or route 2. Using the basic insertion method, request 1 is inserted into route 2 which produces a smaller insertion cost, as shown in Figure 3-2(b). When turning to schedule request 2, we might find that it is infeasible to insert request 2 into route 2 because the schedule of route 2 during the time windows of request 2 is filled. It might also not be inserted into route 1 because it is too far from route 1 to make that insertion feasible. Under this condition, request 2 is either rejected or more vehicles are needed.



(a) before the insertion



(b) basic insertion



(c) rejected-reinsertion

Figure 3-2. Illustration of the rejected-reinsertion method

In the case in Figure 3-2, whenever an infeasible insertion occurs (e.g., insertion of request 2), we attempt to vacate some slot for the infeasible request by removing another request that is similar in terms of time frame and geographic location from its current route and reinserting it into some other route. If the new request can be inserted into the available vacancy and the removed request can be reinserted somewhere else, then the insertion algorithm proceeds to schedule the next request. If either of the requests cannot be inserted, the above search is repeated with another previously assigned request. A deeper search, incorporated in all the heuristics tested below, considers all previously assigned requests, instead of stopping after finding the first feasible one. The “least cost” set of moves is selected for implementation. In Figure 3-2(c), using the rejected-reinsertion operation, request 1 is removed from route 2 and reinserted into route 1 and request 2 is inserted into route 2. In this way, some of the myopic behavior of the insertion method is alleviated. The concept of rejected-reinsertion is simple and straightforward but is very effective in reducing the number of vehicles used, as will be shown in the computational study. The detailed procedure for the rejected-reinsertion operation is as follows:

Assume that requests up to  $k - 1$  have been scheduled. For new request  $k$ , if it is infeasible to insert the new request,

1. For each request  $i = 1, \dots, k - 1$
2. If request  $i$  and request  $k$  satisfy time proximity criterion 1 defined as

$$EPT_i \leq LDT_k \quad \text{and} \quad EPT_k \leq LDT_i, \text{ go to step 3; else go to step 1;}$$

3. Remove request  $i$  from its planned route  $R_i$ , and calculate the associated removal cost as  $C_{remove}^i$ . (It is actually a saving and the value should be negative);
4. Insert request  $k$  into route  $R_i$ . If it is feasible, calculate the associated insertion cost as  $C_{insert}^k$ ; else recover request  $i$ , go to step 1;
5. Insert request  $i$  considering all the available vehicles. If it is feasible, calculate the associated insertion cost as  $C_{insert}^i$ , and the total cost  $C_{total} = C_{remove}^i + C_{insert}^k + C_{insert}^i$ ; else recover requests  $i$  and  $k$ ;
6. Go to step 1.
7. Make the move with the minimum total cost  $C_{total}^*$ .

Note that it is still possible that a request may be infeasible to schedule. It will then be rejected or served by additional vehicles.

### 3.2.2 Improvement procedure

One option of the heuristic is to add a local improvement procedure periodically or after a certain number of insertions. Two inter-route reassignment operators (Toth and Vigo 1996) are considered in the local improvement procedure: (1) *Trip reinsertion* operator: remove trip  $i$  from its current route and reinsert it into all the vehicle routes (the final route could be the same as the current one); (2) *Trip exchange* operator: remove trip  $i$  from its route  $r$  and remove trip  $j$  from its route  $s$  ( $r \neq s$ ); insert the two stops of trip  $i$  in the best positions of route  $s$  and insert the two stops of trip  $j$  in the best positions of route  $r$ .

In this study, one iteration of the trip reinsertion is implemented as follows: examine all assigned requests in sequence; when a trip reinsertion results a total cost below zero, apply it and examine the next assigned request. Due to the high computational cost, the trip exchange operation is performed only on the restricted neighborhoods. For one iteration of the trip exchange procedure, we examine assigned requests  $i$  from 1 to  $N - 1$ . Only those assigned requests  $j$  ( $j = i + 1, \dots, N$ ) are considered for exchange that satisfy time proximity criterion 2 defined as:

$$\text{pickup time window overlap } EPT_i \leq LPT_j \quad \text{and} \quad EPT_j \leq LPT_i,$$

$$\text{delivery time window overlap } EDT_i \leq LDT_j \quad \text{and} \quad EDT_j \leq LDT_i,$$

Whenever the trip exchange results in a total cost less than zero, the trip exchange operation is implemented. The implementation of the improvement procedure consists of iterating the trip exchange procedure until no further improvement is possible, followed by the iteration of the trip reinsertion procedure until no further improvement is possible. The whole procedure is repeated until no change occurs or some prescribed number of iterations is reached. Based on our computational experiments, the number of iterations of the whole procedure is usually 2 to 4.

### **3.2.3 Variable vs fixed fleet size**

In order to satisfy all the demand, either a sufficient fleet size should be provided initially if fleet size is fixed throughout the planning process, or fleet size should be increased during the insertion process to serve the infeasible demand. In the former case, the minimum number of vehicles required to serve all the demand is usually obtained by tentatively using different numbers of vehicles and running the algorithm repeatedly in

order to find the minimum number which satisfies all the demand. The algorithms with variable and fixed fleet size will be compared in the computational study. The initial fleet size for the variable fleet size will also be tested.

### **3.2.4 Complete heuristic procedure**

The complete rejected-reinsertion heuristic procedure can be described as follows:

1. Sort the passengers in the order of their earliest pickup times.
2. Set the initial fleet size  $F_0$ .
3. Insert the passengers in sequence.

For each passenger, if insertion into the current fleet is infeasible, perform the rejected-reinsertion operation specified in Section 3.1. If it is still infeasible to insert the passenger into the current fleet, add one new vehicle into the fleet and insert the passenger into it. The new vehicle can serve all subsequent requests.

4. (optional) Perform the improvement procedure periodically.

The setting of the initial number of vehicles  $F_0$  in step 2 is not essential. Sensitivity analysis in the computational study will show that the resulting minimum number of vehicles required to serve all the demand is quite insensitive to  $F_0$ . Basically, the value of  $F_0$  should be less than the required number of vehicles to serve all the demand.

### **3.2.5 Feasibility check for inserting a request**

For each insertion of the origin and destination stops of a request, all the constraints including those on vehicle capacity, time windows and maximum ride time of all

passengers should be satisfied. If  $n$  is the average number of the stops in a vehicle route and  $M$  is the number of operating vehicles, the number of possible insertions is of  $O(M \cdot n^2)$ . Jaw et al. (1986) proposed four statistics to expedite the time window feasibility check (which is the most difficult and time-consuming). A “schedule block” (SB) concept was first proposed by Jaw et al. (1986) in facilitating the feasibility check of each attempted insertion. This concept applies to the version of DARP in which no vehicle can be idle while there are passengers onboard. It is defined as a continuous period of active vehicle time between two successive periods of vehicle slack (idling) time, starting and ending with empty vehicle. For each stop  $\alpha$  within schedule block  $k$  they define four statistics  $BUP_\alpha$ ,  $BDOWN_\alpha$ ,  $AUP_\alpha$  and  $ADOWN_\alpha$  to facilitate the feasibility check of inserting the origin and destination of a request into the same schedule block.  $BUP_\alpha$  ( $BDOWN_\alpha$ ) represents the maximum amount of time by which stop  $\alpha$  and all its preceding stops in the same schedule block can be advanced (delayed) without violating the time window constraints.  $AUP_\alpha$  ( $ADOWN_\alpha$ ) similarly represents the maximum amount of time by which stop  $\alpha$  and all its following stops can be advanced (delayed).

The statistics can only be used when inserting the origin and destination of a request into the *same* schedule block. The maximum shifts are bounded by the available slack times at the ends of the schedule block. However, the insertion of the origin and destination of a request should not be unnecessarily constrained to the same schedule block, especially for such a highly constrained problem. The statistics can be easily generalized for the case considering the whole route instead of one schedule block, as follows:



$$BUP_i = \begin{cases} \min(BUP_{i-1} + Idle_k, AT_i - ET_i) & \text{if stop } i \text{ is the 1st stop of one SB } k \\ \min(BUP_{i-1}, AT_i - ET_i) & \text{otherwise} \end{cases} \quad (3-3a)$$

$$BDOWN_i = \begin{cases} LT_i - AT_i & \text{if stop } i \text{ is the 1st stop of one SB } k \\ \min(BDOWN_{i-1}, LT_i - AT_i) & \text{otherwise} \end{cases} \quad (3-3b)$$

$$AUP_i = \begin{cases} AT_i - ET_i & \text{if stop } i \text{ is the last stop of one SB } k \\ \min(AUP_{i+1}, AT_i - ET_i) & \text{otherwise} \end{cases} \quad (3-3c)$$

$$ADOWN_i = \begin{cases} \min(ADOWN_{i+1} + Idle_{k+1}, LT_i - AT_i) & \text{if stop } i \text{ is the last stop of one SB } k \\ \min(ADOWN_{i+1}, LT_i - AT_i) & \text{otherwise} \end{cases} \quad (3-3d)$$

In the above definition,  $BUP_i(BDOWN_i)$  represents the maximum amount of time by which stop  $i$  and all its preceding stops on the same *vehicle route* can be advanced (delayed) without violating the time window constraints.  $AUP_i(ADOWN_i)$  represents the maximum amount of time by which stop  $i$  and all its following stops can be advanced (delayed).  $AT_i$ ,  $ET_i$  and  $LT_i$  are the actual, earliest and latest times (either pickup or delivery) for stop  $i$ , respectively.  $Idle_k$  is the idling (slack) time before schedule block  $k$ .

If pickup stop  $+i$  of a new request is inserted between stop  $p$  and  $p+1$  and delivery stop  $-i$  is inserted between stops  $p$  and  $p+1$ , then the necessary time window feasibility conditions include:

$$T_p + BDOWN_p + T_{p,+i} \geq EPT_i, \text{ if stops } p \text{ is not the last stop of one SB} \quad (3-4)$$

$$T_p - BUP_p + T_{p,+i} \leq LPT_i \quad (3-5)$$

$$T_{q+1} + ADOWN_{q+1} - T_{-i,q+1} \geq EDT_i \quad (3-6)$$

$$T_{q+1} - AUP_{q+1} - T_{-i,q+1} \leq LDT_i, \text{ if stops } q \text{ is not the last stop of one SB} \quad (3-7)$$

$$T_{detour}^i \leq BUP_p + ADOWN_{q+1} + Idle_{p+1,q+1} \quad (3-8)$$

In Equations (3-4) to (3-7),  $T_i$  denotes the scheduled time for stop  $i$  and  $T_{i,j}$  denotes the direct ride time from stop  $i$  to stop  $j$ . In Equation (3-8),  $Idle_{p+1,q+1}$  is the total idling time between stop  $p$  and  $q + 1$ .  $T_{detour}^i$  is the additional travel time due to inserting both stops  $+i$  and  $-i$ .

$$T_{detour}^i = T_{p,+i} + T_{+i,-i} + T_{-i,p+1} - T_{p,p+1}, \quad \text{if } p = q \quad (3-9a)$$

$$T_{detour}^i = T_{p,+i} + T_{+i,p+1} + T_{q,-i} + T_{-i,q+1} - T_{p,p+1} - T_{q,q+1}, \text{ if } p \neq q \quad (3-9b)$$

Note that the idling time between stops  $p + 1$  and  $q + 1$  should be eliminated if idling is not permitted while passengers are onboard. If insertion of both the origin and destination of a request are feasible in terms of time window constraints, the maximum ride time constraints of assigned passengers (and the capacity constraint if necessary) should also be checked by scanning through the list of customers and comparing the attempted ride times with the maximum ride times.

### 3.2.6 Feasibility check for removing a request

The rejected-reinsertion, trip reinsertion and trip exchange operations all involve the removing of a request from its assigned route. Special caution should be taken when removing a request from its current route because it might cause a time window violation for some other passengers already assigned to the route. This only applies to the operating scenario considered in which a vehicle is not allowed to idle while carrying passengers. Figure 3-3 shows one complete schedule block from which one request is to be removed. For illustration purposes, only removal of one stop (stop  $b$ ) will be discussed and only the time windows of some stops are shown in Figure 3-3. Assume stop  $a$  would not be the last stop of a possible new schedule block after stop  $b$  is removed from the route, then stop  $c$  should be visited directly from stop  $a$  without idling. If there is more time between stop  $a$  and  $c$  than the direct ride time, the stops preceding stop  $a$  could be pushed forward and/or the stops following stop  $c$  could be pushed backward to reduce the time gap between  $a$  and  $c$ . In Figure 3-3, the ‘max delay’ is the maximum amount of time that all stops preceding stop  $b$  could be delayed without violating the time window constraints, and the ‘max pushback’ is the maximum amount of time that all stops following stop  $b$  could be pushed backward. If the direct travel time from stop  $a$  to  $c$  (plus the service time at stop  $a$ , if that is considered) is less than the time interval between stops  $a$  and  $c$  (i.e.,  $T_c - T_a$  in Figure 3-3), then idling time before stop  $c$  is necessary or the time window constraints of some stops within the current schedule block will be violated if no idling time between  $a$  and  $c$  is provided. Thus, the removal of a request may cause a time window violation.

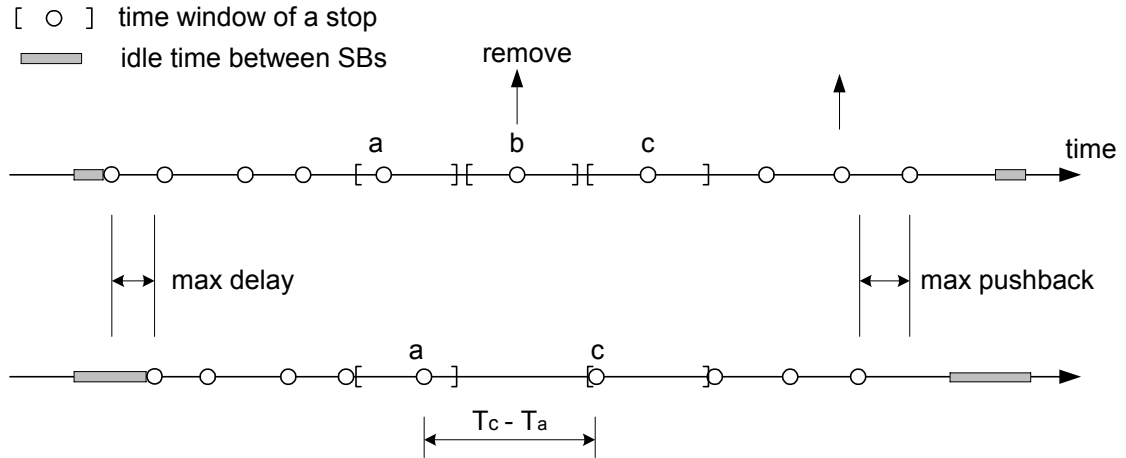


Figure 3-3. Removing a request from one schedule block

When the time window violation occurs in the removal process, the route may become feasible again if at least one stop is inserted into the same schedule block (i.e. after stop *a*) in the reinsertion step (i.e. the insertion of the previously rejected passenger into the removed route in the rejected-reinsertion operation).

### 3.2.7 Insertion criterion

In the insertion heuristic, the insertion decision is made based on the additional increase of the objective function. The insertion with the least incremental cost will be chosen. In the context of service operations in the public sector, there are always tradeoffs between minimizing the operating cost and the passenger inconvenience cost. A general form of the objective function might include active vehicle travel time (moving time)/distance, excess ride time (the difference between the actual ride time and direct ride time) of all current passengers, time deviation (the difference between the actual pickup/delivery time and desired time) of all current passengers and vehicle idling time.

Although selecting the weights of components is up to the system operating managers and the proposed heuristic does not depend on the objective form chosen, the ultimate objective here is to minimize the number of vehicles required in order to maximize the vehicle productivity, which is usually very low for DAR systems due to their high quality of service (i.e. door-to-door service) and dispersed demand. Also, because the passenger inconvenience (i.e. waiting time, excess ride time) is already formulated through hard constraints, it seems unnecessary to include it in the objective function at the cost of more vehicles used. However, the number of vehicles is an input to the algorithm and cannot be expressed in the objective function explicitly. A common alternative way is to minimize the vehicle travel time/distance. Some studies (e.g., Jaw et al. 1986) implicitly suggest including other components, such as vehicle idling time, in the objective function, as that reserves some flexibility for future demand. Thus, the components in the objective function work more like heuristic parameters.

### **3.2.8 Vehicle scheduling**

Scheduling refers to the determination of the actual pickup and delivery times of the new insertion and the corresponding modification of the actual pickup and delivery times of the affected passengers assigned once the insertion sequence is determined. The scheduling will affect the passenger time deviation, but will not affect the passenger ride time and vehicle travel time/distance. The schedules can be formed as soon as possible (Diana and Dessouky 2004), or can be optimized based on the incremental cost (Jaw et al. 1986). For a congested system, the two methods may lead to similar results. Our experimental tests show that the above two scheduling methods achieve very similar

results. In this study, schedules are sought that minimize the time deviation of the passengers. A more detailed discussion of schedule optimization based on general cost functions can be found in Jaw (1984).

### **3.3 Computational Study**

Although static DARPs have been studied by many researchers, there are very few benchmark problems available for comparison. One reason might be that there is far less research on DAR than on general VRPs. Another reason is that different operational scenarios (i.e. whether or not vehicles are allowed to be idle while carrying passengers) or objectives are considered for different studies, which further reduces the available test problems in each category.

Below, we test our heuristics with our own randomly generated problems and with test problems from Diana and Dessouky (2004). The latter problems are the latest found in the literature that consider operational scenarios very similar to ours. The randomly generated problems have smaller service areas and average direct travel distances compared with the second set of problems. Although both problem categories consider the time-dependent demand, in the randomly generated problems, the demand is relatively stable, which might justify the usage of the same fleet size throughout the service period. To deal with the randomness of the demand, five replications are generated for each problem, and the statistics reported are the average over five

replications. The computer program is coded using visual C++ and is run on a personal laptop with a 1.6 GHz Pentium M and 768M of RAM.

### 3.3.1 Randomly generated problems

An 8 mile  $\times$  8 mile service area with the depot located in the center of the area is studied. The Euclidean distance metric is used with a circuitry factor of 1.3 (by which each direct distance is multiplied). Vehicle speed is assumed to be constant at 15 mph. The locations of origins and destinations of all the demand are uniformly and independently distributed in the area. The time intervals between consecutive earliest pickup times follow a negative exponential distribution. We simulate 9 hours of service with the hourly demand as 120, 120, 160, 200, 200, 160, 160, 120, 120 requests per hour. The departure times from and return times to the depot are not restricted to the 9-hour period. Vehicle capacity is assumed to be a large number. The maximum number of passengers onboard simultaneously will be recorded, which indicates the minimum vehicle size should be provided. Four service quality scenarios as constrained by time window and maximum ride time are considered. The following linear maximum ride time equation is used:

$$MRT = a_0 + a_1 \cdot DRT \quad (3-10)$$

Table 3-1 shows the parameter settings for the four scenarios ‘L’, ‘M’, ‘H’ and ‘VH’, which stand for low, medium, high and very high service quality, respectively. The service quality improves from ‘L’ to ‘VH’.

Table 3-1. Constraint settings for four service quality scenarios

Scenario	Maximum time deviation (min)	Constant term $a_0$ in Equation (3-10) (min)	Slope $a_1$ in Equation (3-10)
L	30	5	2.5
M	20	5	2.0
H	10	5	1.5
VH	5	5	1.3

We include active vehicle travel time (when the vehicle is moving) and passenger excess ride time in the objective function. The component of the passenger excess ride time works somewhat like a heuristic parameter. Based on some experimental tests, we found that the minimum number of vehicles used is not very sensitive to the weight assigned to the passenger excess ride component as long as the weight is below 0.5 for the two sets of problems. For the passenger excess ride time, a weight of 0.2~0.3 yields slightly better solutions than a weight of zero. The values of the weights used in this study are 0.7 for the active vehicle travel time and 0.3 for the passenger excess ride time.

For each scenario considered, six algorithm variations are implemented and their results are shown in Tables 3-2 to 3-5. Algorithm 1 is the basic parallel insertion heuristic similar to that of Jaw et al. (1986) except that insertions across multiple schedule blocks are allowed and insertion schedules are determined to minimize the time deviation. The fleet size is fixed throughout the planning process. Algorithm 2 is similar to Algorithm 1. The difference is that one vehicle is added to the fleet each time it is infeasible to insert a new request into the current fleet. Algorithm 3 differs from Algorithm 1 in that rejected-



reinsertion is implemented for those rejected requests. Algorithm 4 combines features of Algorithms 2 and 3, in which the rejected-reinsertion is implemented for rejected requests and fleet size is added after a request is rejected by the rejected-reinsertion operation. In Algorithms 2w and 4w, a periodical improvement procedure at 30-min time intervals is implemented upon Algorithms 2 and 4. The starting fleet sizes for Algorithms 2, 2w, 4 and 4w are 30, 40, 50 and 65 for scenarios L, M, H and VH, respectively. For Algorithms 1 and 3, the number of vehicles required is obtained by running the program repeatedly using different fleet sizes and finding the smallest fleet that satisfies all the demand.

The notation in the following tables is as follows. ‘Vehicle miles’ is the total vehicle travel distance in miles. ‘Vehicle prod.’ is the vehicle productivity defined as the number of served trips divided by the total vehicle service time (including idling time), in trips per vehicle hour. The sixth column reports the total passenger miles. The average passenger time deviation from the desired times and average passenger ride ratio are reported in the next two columns. ‘Max passengers onboard’ indicates the vehicle capacity actually required since a large vehicle capacity is initially assumed. Finally, the last column indicates the average computation time in seconds.

Table 3-2. Results of six algorithm variations for scenario L

Algo.	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)
1	40.2	4,813	3.95	11,328	14.21	1.475	10.2	16
2	43.8	4,857	3.869	11,483	14.14	1.494	9.6	16
2w	37.0	4,250	4.38	10,312	14.93	1.351	9.6	1,381
3	37.6	4,769	4.07	11,507	14.38	1.496	9.6	31
4	38.4	4,791	4.09	11,859	14.06	1.546	9.4	83
4w	34.0	4,223	4.57	10,862	14.69	1.420	9.6	1,686

Table 3-3. Results of six algorithm variations for scenario M

Algo.	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)
1	49.4	5,323	3.40	10,414	8.89	1.362	8.0	13
2	50.2	5,371	3.379	10,486	8.99	1.373	8.2	13
2w	44.0	4,769	3.78	9,973	9.68	1.311	7.6	461
3	45.4	5,319	3.54	10,555	9.04	1.380	7.4	19
4	45.6	5,339	3.55	10,670	9.00	1.396	8.2	40
4w	41.2	4,769	3.86	10,074	9.54	1.323	7.8	497

Table 3-4. Results of six algorithm variations for scenario H

Algo.	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)
1	62.2	6,242	2.73	9,308	4.30	1.231	5.6	11
2	63.4	6,302	2.725	9,314	4.22	1.232	5.6	11
2w	58.8	5,846	2.91	9,180	4.49	1.216	6.2	120
3	58.4	6,245	2.83	9,327	4.34	1.234	5.6	15
4	58.0	6,298	2.85	9,394	4.26	1.243	5.8	23
4w	55.4	5,885	3.00	9,252	4.51	1.224	6.2	126

Table 3-5. Results of six algorithm variations for scenario VH

Algo.	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)
1	78.2	7,101	2.23	8,589	1.85	1.146	4.2	11
2	76.4	7,158	2.26	8,609	1.88	1.149	4.2	11
2w	75.8	6,882	2.29	8,609	1.91	1.148	5.0	47
3	72.8	7,138	2.32	8,603	1.9	1.148	4.0	14
4	70.6	7,135	2.36	8,627	1.92	1.151	4.2	19
4w	70.6	6,916	2.38	8,601	1.95	1.147	5.0	62

Results in Tables 3-2 to 3-5 are rearranged in Tables 3-6 and 3-8 and analyzed in the following sections for different comparison purposes.

*(1) Comparison of the rejected-insertion heuristics with basic insertion heuristic*

Table 3-6 only shows the results by the basic insertion heuristic (Algorithm 1) and the rejected-insertion heuristics without and with the periodical improvement (Algorithms 4 and 4w). The performance differences between Algorithms 1 and 4 and between Algorithms 1 and 4w are also shown. Based on Table 3-6, Algorithm 4 outperforms Algorithm 1 in terms of number of vehicles (up to -9.7%) and vehicle productivity (up to +5.8%) at a cost of slightly increased passenger time deviation and ride time ratio. The vehicle productivity increases as the number of vehicles decreases. The average passenger time deviation is slightly less than half of the maximum deviation from desired time. As constraints become more restrictive, Algorithm 4 provides solutions increasingly superior to those of Algorithm 1. Algorithm 4 is still very efficient computationally, although its computation time is approximately doubled in the VH scenario and quintupled in the L scenario compared to Algorithm 1.

Table 3-6. Comparison of the rejected-insertion heuristics with parallel insertion heuristic

Scenario	Algo.	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Comp. time (sec)
L	1	40.2	4,813	3.95	11,328	14.21	1.475	16
	4	38.4	4,791	4.09	11,859	14.06	1.546	83
	4 vs 1	-4.5%	-0.5%	+3.5%	+4.7%	-1.1%	+4.8%	
	4w	34.0	4,223	4.57	10,862	14.69	1.420	1,686
4w vs 1	-15.4%	-12.3%	+15.7%	-4.1%	+3.4%	-3.7%		
	4w vs 4	-11.5%	-11.9%	+11.7%	-8.4%			
M	1	49.4	5,323	3.40	10,414	8.89	1.362	13
	4	45.6	5,339	3.55	10,670	9.00	1.396	40
	4 vs 1	-7.7%	+0.3%	+4.4%	+2.5%	+1.2%	+2.5%	
	4w	41.2	4,769	3.86	10,074	9.54	1.323	497
4w vs 1	-16.6%	-10.4%	+13.5%	-3.3%	+7.3%	-2.9%		
	4w vs 4	-9.6%	-10.7%	+8.7%	-5.6%			
H	1	62.2	6,242	2.73	9,308	4.30	1.231	11
	4	58.0	6,298	2.85	9,394	4.26	1.243	23
	4 vs 1	-6.8%	+0.9%	+4.4%	+0.9%	-0.9%	+1.0%	
	4w	55.4	5,885	3.00	9,252	4.51	1.224	126
4w vs 1	-10.9%	-5.7%	+9.9%	-0.6%	+4.9%	-0.6%		
	4w vs 4	-4.5%	-6.6%	+5.3%	-1.5%			
VH	1	78.2	7,101	2.23	8,589	1.85	1.146	11
	4	70.6	7,135	2.36	8,627	1.92	1.151	19
	4 vs 1	-9.7%	+0.5%	+5.8%	+0.4%	+3.8%	+0.4%	
	4w	70.6	6,916	2.38	8,601	1.95	1.147	62
4w vs 1	-9.7%	-2.6%	+6.7%	+0.1%	+5.4%	+0.1%		
	4w vs 4	0.0%	-3.1%	+0.8%	-0.3%			

Algorithm 4w further improves on the results of Algorithm 1 in terms of number of vehicles (-9.7% to -16.6%), vehicle miles (-2.3% to -12.3%), vehicle productivity (+6.7% to +13.3%), passenger miles and ride time ratio. The improvement is more prominent for the L and M scenarios than for the H and VH scenarios. This occurs because the DARP is a heavily constrained problem, and as the problem gets more restricted, the feasible region for improvement becomes more limited. This conclusion is based on scenarios in which vehicles are already heavily loaded. It is expected that if vehicles are less loaded or time windows are wider, the improvement will be greater but would require much more computation time. As the problem gets less constrained (from VH to L), the computation time increases nonlinearly.

Table 3-7 shows the number of vehicles required for the five randomly generated replications. The average and standard deviation are shown in the last two columns. The results show that Algorithm 4 outperforms Algorithm 1 except for the second replication of scenario L. Algorithm 4a uses much fewer vehicles than Algorithm 1. Algorithm 4a outperforms Algorithm 4 except for the fourth replication of scenario VH. The standard deviation of Algorithm 4a increases as the problems get more restrictive. The standard deviations of algorithm 4a for all service scenarios are below 2.2 vehicles. From an operational point of view, this means that the fluctuations in the number of vehicles required for a given level of demand are not evident even if the demand is randomly distributed over time and space. Table 3-7 also reports in parentheses the computation time in second for Algorithm 4w.

Table 3-7. Variability of number of vehicles over five replications

Scenario	Algo.	Replication					Average	Standard deviation
		1	2	3	4	5		
L	1	42	40	40	38	41	40.2	1.48
	4	38	41	38	36	39	38.4	1.82
	4w	33 (2168)	34 (1233)	35 (1661)	34 (1480)	34 (1890)	34 (1686)	0.71 (361)
M	1	47	54	50	47	49	49.4	2.88
	4	45	47	45	44	47	45.6	1.34
	4w	41 (575)	42 (375)	40 (447)	42 (555)	41 (532)	41.2 (497)	0.84 (84)
H	1	61	64	63	63	60	62.2	1.64
	4	57	60	57	58	58	58.0	1.22
	4w	53 (127)	57 (119)	56 (108)	56 (122)	55 (156)	55.4 (126)	1.52 (18)
VH	1	78	77	76	83	77	78.2	2.77
	4	71	71	70	71	70	70.6	0.55
	4w	71 (62)	71 (55)	67 (57)	73 (62)	71 (73)	70.6 (62)	2.19 (7)

*(2) Test the effectiveness of the rejected-reinsertion operator*

Table 3-8 shows the minimum number of vehicles required for all demand that results from six algorithm variations. In Table 3-8, Algorithms 1 and 3, 2 and 4, 2w and 4w are comparison pairs. The rejected-reinsertion operator is implemented in the latter algorithm in each pair. It is found that the rejected-reinsertion operator used in Algorithms 3, 4 and 4w is very effective in reducing the vehicle fleet for all four scenarios.

Table 3-8. Comparison of algorithms with and without rejected-reinsertion operator

Algorithm	# of vehicles required for each scenario			
	L	M	H	VH
1	40.2	49.4	62.2	78.2
2	43.8	50.2	63.4	76.4
2w	37.0	44.0	58.8	75.8
3 (3 vs 1)	37.6 (-6.5%)	45.4 (-8.1%)	58.4 (-6.1%)	72.8 (-6.9%)
4 (4 vs 2)	38.4 (-12.3%)	45.6 (-9.2%)	58.0 (-8.5%)	70.6 (-7.6%)
4w (4w vs 2w)	34.0 (-8.1%)	41.2 (-6.4%)	55.4 (-5.8%)	70.6 (-6.9%)

*(3) Test of fixed vs variable fleet size*

Still in Table 3-8, comparing Algorithm 1 with 2, and 3 with 4, Algorithm 2 (variable fleet size) slightly underperforms Algorithm 1 (fixed fleet size) in the L scenario, but slightly outperforms it in the VH scenario. Algorithm 4 (variable fleet size) performs similarly with Algorithm 3 (fixed fleet size), except that in the VH scenario, Algorithm 4 succeeds with slightly fewer vehicles. While the differences are small, a common trend is that as the problem gets more restricted, algorithms with variable fleet sizes become more preferable. One advantage of the algorithm using variable fleet size over the one using fixed fleet size is that there is no need to run the algorithm repeatedly each time trying a different fleet size and finding the smallest one that satisfies all demand. This implies that the advantage of using variable fleet size becomes more relevant if an algorithm needs much computation time (i.e. an algorithm with a periodical improvement procedure).



*(4) Sensitivity analysis of initial fleet size*

Table 3-9 shows the fleet required to serve all demand by Algorithms 4 and 4w using different initial fleet sizes. It is found the results are quite insensitive to the initial fleet size. In general, using an initial fleet size close to the required fleet size achieves slightly better results, which also requires fewer rejected-reinsertion operations than using a smaller initial fleet size. The required fleet size can be easily estimated by running the algorithm once using any reasonable initial fleet size.

Table 3-9. Effects of initial fleet size on final fleet size with Algorithms 4 and 4w

Scenario	Algo.	Initial fleet size					
		10	20	30	40	50	65
L	4	38.4	38.8	38.4			
	4w	35.0	34.6	34.0			
M	4		45.8	45.6	45.6		
	4w		42.4	42.4	41.2		
H	4			58.4	59.4	58.0	
	4w			57.0	56.8	55.4	
VH	4				71.6	71.8	70.6
	4w				69.8	70.2	70.6

### 3.3.2 Diana and Dessouky's test problems

The main purposes of testing the second set of problems are to check that our algorithms are correctly implemented in software and to compare their performances with other available sources.

#### *(1) Input data characteristics*

The test problems of Diana and Dessouky (2004) include one 500-request problem and one 1000-request problem, each with five replications. The data are randomly generated, but based on data provided by a realistic DAR system run by Access Services, Inc. For example, the distribution from which the pickup times of the samples were drawn was based on the empirical distribution derived from Los Angeles County. Interested readers may find more information on the data generation in Diana and Dessouky (2004) and in Dessouky and Adam (1998). In this paper the 1000-request problem is tested and used to compare algorithms.

The basic operational scenario is summarized as follows:

Total service area: 150 mile  $\times$  150 mile

Vehicle speed: 15 mph (the number has been corrected by the author during our correspondence)

Probability of serving a wheelchair passenger: 0.2

Service time distribution: uniform (1, 3) minutes for wheelchair passengers

30 seconds for others

Simulation period: 0:00 ~ 23:59.

Figure 3-4 shows the demand distributions over time for one replication of the 500-request problem and one replication of the 1000-request problem. Figure 3-5 shows the direct travel time distributions for one of the replications of the 1000-request problem.

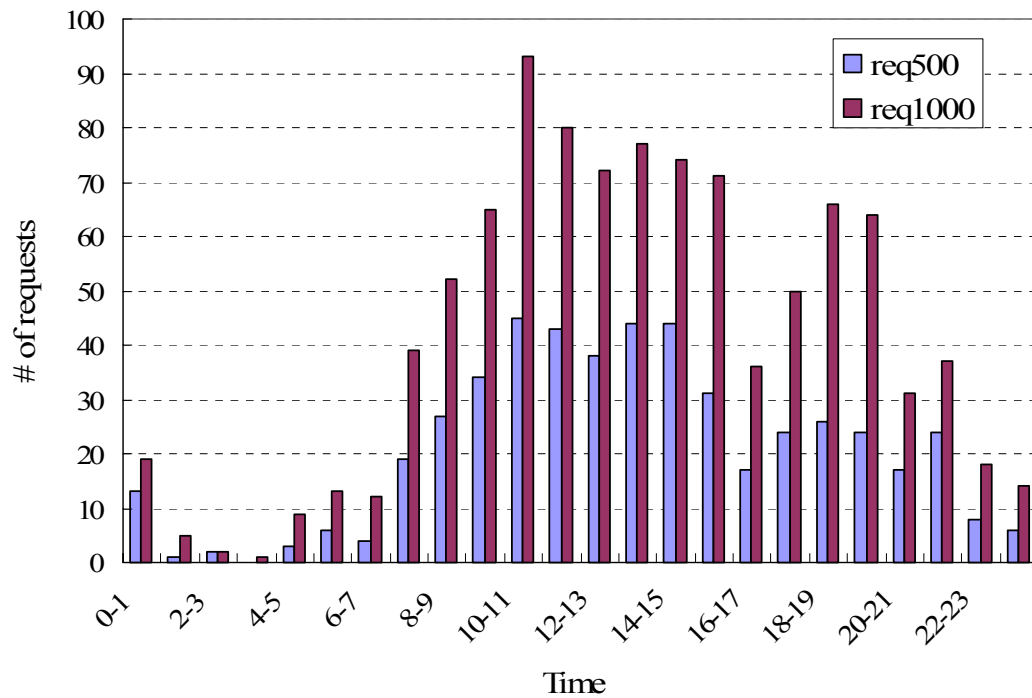


Figure 3-4. Demand distribution over time

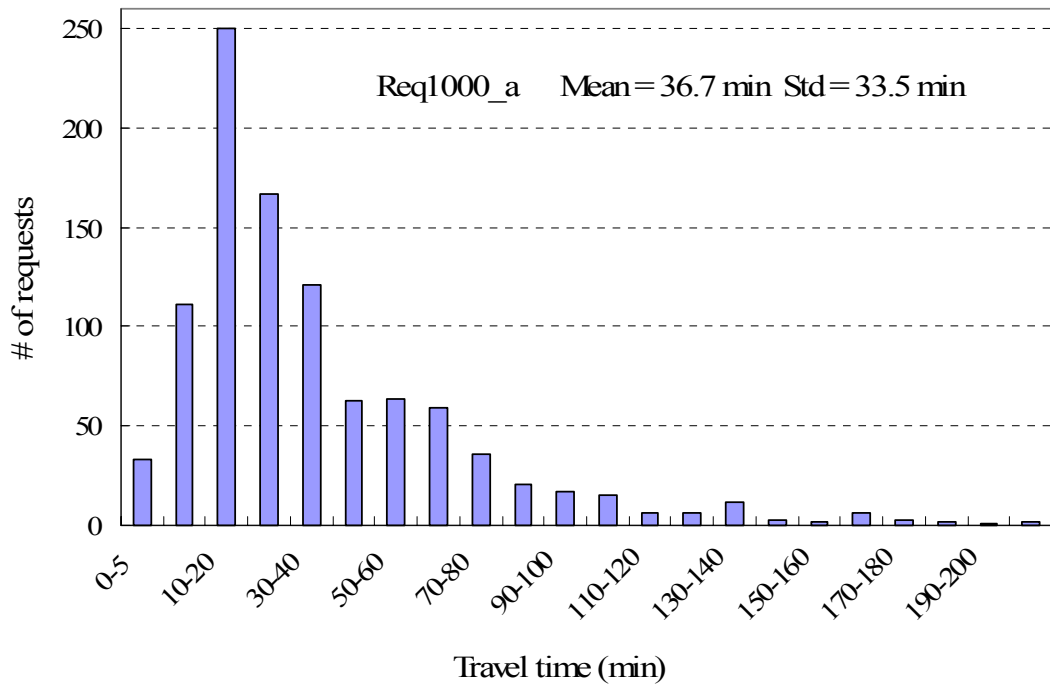


Figure 3-5. Direct travel time distribution

For the 1000-request problem, three scenarios ‘L’, ‘M’ and ‘H’, whose constraint settings are defined in Table 3-10, are tested. (In Diana and Dessouky (2004), one base scenario was tested too, in which the service quality is between M and H.) Note that although the service qualities defined here are very similar to those defined in Table 3-1 for the randomly generated problems, the problems here are more constrained than the randomly generated ones. This occurs because the average direct travel time is longer and the area covered is larger in the problems defined by Diana and Dessouky (2004) than in the randomly generated problems.

Table 3-10. Constraint settings for three service quality scenarios

Scenario	Maximum time deviation (min)	Constant term $a_0$ in Equation (3-10) (min)	Slope $a_1$ in Equation (3-10)
L	30	20	2.0
M	15	10	1.5
H	5	5	1.2

*(2) Computational results*

In Tables 3-11 ~ 3-13, Algorithms Diana1 and Diana5 correspond to Algorithms 1 and 5 in Diana and Dessouky (2004), which represent the basic parallel insertion algorithm (into same schedule block) and their proposed regret insertion algorithm (across multiple schedule blocks and schedule as soon as possible). For comparability, the same objective function is used; thus, the weights for vehicle travel distance, passenger excess ride time and vehicle idle times within the schedule are 0.45, 0.50 and 0.05. The definition of the time windows by Diana and Dessouky includes the stop service time, while ours does not. Their maximum ride time constraint is interpreted as the sum of the actual ride time and of the service times at the pickup and delivery stops must not exceed the maximum ride time, while in ours the service times are not counted in the maximum allowable ride time. Those small discrepancies have been adjusted in the problem definitions to make results comparable.

Tables 3-11 ~ 3-13 shows the computational results for the 1000-request problem under the L, M and H scenarios. ‘# of vehicles’, ‘Vehicle miles’, ‘Ride time ratio’ and ‘Comp.

time' are as previously defined . 'Idle hours' reports the total length of all the vehicle idling times. Based on correspondence with one of the authors, values of vehicle miles for Diana 1 and Diana 5 in Tables 3-11 ~ 3-13 have been adjusted due to a rounding problem

Table 3-11. Computational results for scenario L of the 1000-request problem

Algorithm	# of vehicles	Vehicle miles	Idle hours	Ride time ratio	Comp. time (sec)
Diana 1	63.2	15,675	288	1.395	n/a
Diana 5	58.4	14,820	301	1.476	n/a
1	60.8	13,917	138	1.193	8
(1 vs Diana 1)	(-3.8%)	(-11.2%)	(-52.1%)	(-14.5%)	
4	52.2	13,788	97	1.271	16
(4 vs Diana 5)	(-10.6%)	(-6.8%)	(-67.8%)	(-13.9%)	
4w	51.6	13,402	104	1.214	74
(4w vs Diana 5)	(-11.6%)	(-9.6%)	(-65.4%)	(-17.8%)	

Table 3-12. Computational results for scenario M of the 1000-request problem

Algorithm	# of Vehicles	Vehicle miles	Idle hours	Ride time ratio	Comp. time (sec)
Diana 1	77.2	17,655	350	1.173	n/a
Diana 5	70.0	16,530	374	1.204	n/a
1	72.0	15,811	220	1.102	9
(1 vs Diana 1)	(-6.7%)	(-10.4%)	(-37.1%)	(-6.1%)	
4	66.4	15,771	200	1.105	11
(4 vs Diana 5)	(-5.1%)	(-4.6%)	(-46.5%)	(-8.2%)	
4w	65.6	15,462	200	1.101	32
(4w vs Diana 5)	(-6.3%)	(-6.5%)	(-46.5%)	(-8.6%)	

Table 3-13. Computational results for scenario H of the 1000-request problem

Algorithm	# of Vehicles	Vehicle miles	Idle hours	Ride time ratio	Comp. time (sec)
Diana 1	92.8	20,160	464	1.034	n/a
Diana 5	87.2	19,110	485	1.042	n/a
1	91.6	18,386	368	1.022	8
(1 vs Diana 1)	(-1.3%)	(-8.8%)	(-20.7%)	(-1.2%)	
4	86.8	18,443	346	1.023	15
(4 vs Diana 5)	(-0.5%)	(-3.5%)	(-28.7%)	(-1.8%)	
4w	86.6	18,376	348	1.022	24
(4w vs Diana 5)	(-0.7%)	(-3.8%)	(-28.2%)	(-1.9%)	

Comparing Algorithms 1 and Diana 1, both are basic parallel insertion heuristics but with variable and fixed fleet size. Algorithm 1 uses similar numbers of vehicles for the H scenario but slightly fewer vehicles for the L and M scenarios. However, Algorithm 1 outperforms Diana 1 in terms of total vehicle miles, idle hours and ride time ratio. The big reduction of idle times by Algorithm 1 may be due to the use of a smaller initial fleet size because the demand level is low during the early service period and thus few vehicles are needed.

Algorithm 4 outperforms Algorithm 1, as in the randomly generated test cases described earlier, with the vehicle reduction up to 14.1% for scenario L. Comparing Algorithm 4w (rejected-reinsertion with periodical improvement) with Diana 5 (regret insertion), 4w outperforms Diana 5 with up to 11.6% fewer vehicles used in the L scenario. Its

advantage decreases as the service quality increases (i.e. the problem gets more restricted). In their computational test, Diana and Dessouky (2004) found the regret insertion algorithm to perform better with medium to small time window constraints. Note that the advantage of Algorithm 4w over Algorithm 4 is relatively limited in this set of test problems compared to the set of randomly generated problems, since this set of problems is more restrictive in that 1000 requests are distributed in a very big area (i.e. 150 mile  $\times$  150 mile) which makes the scheduling more difficult. For those more restrictive problems, Algorithm 4 yields similar results to those of Algorithm 4w, but with faster computation.

Note that the ride time ratio in Table 3-13 is very low, even though the constant and slope terms for the maximum ride time (Equation 3-10) for this scenario are 5 minutes and 1.5, respectively. The obtained average ride time ratio is far below half of the maximum ride time ratio. The vehicle occupancy is around 0.51 for Algorithms 1, 4 and 4w. (It is not reported for Algorithm Diana 1 and Diana 5.) As the constraint on deviation from desired time and the maximum ride time constraint get more restrictive, more vehicles are required and vehicle productivity and vehicle occupancy decrease. The indicated tradeoffs between the vehicle resources and service quality (i.e. average time deviation and excess ride time) should be very useful to DAR planners.

The proposed heuristic is very efficient computationally. Without the periodical improvement procedure, Algorithm 4 solves a 1000-request problem within 16 seconds. The computation time for algorithms with the periodical improvement increases as the



problem gets less restricted. The low quality case of the 1000-request problem takes about 74 seconds. (The computational times reported by Diana and Dessouky (2004) are 26 min for the 500-request problems and 195 min for the 1000-request problems on a Pentium III computer.) The computation times of the proposed heuristic are clearly fast enough for practical applications.

### **3.4 Conclusions**

In this chapter, we propose a rejected-reinsertion heuristic for the multi-vehicle DARP with service quality constraints. The main innovation of the heuristic is a rejected-reinsertion operator. Whenever the insertion of a new request is infeasible, this operator persists in inserting it by trying to move previously assigned requests elsewhere. The least cost set of moves is determined and implemented. The insertion process is tested with fixed and variable fleet sizes. A periodical improvement procedure involving trip reinsertion and trip exchange is also tested and implemented to further improve the solution.

Through the computational study, the proposed heuristic is shown to be effective, especially in reducing the number of required vehicles and thus increasing vehicle productivity. The rejected-reinsertion heuristic without periodical improvement can achieve moderately better results than parallel insertion heuristics for all cases studied. The rejected-reinsertion heuristic with periodical improvement outperforms the parallel insertion heuristic by using up to 17% fewer vehicles. Among the problems considered

here, the periodical improvement procedure is more effective for the less constrained ones. The heuristic still maintains the advantages of an insertion-based method whose computational performance is quite good and which can be extended to a dynamic problem. Using a variable fleet size rather than fixed fleet size does not change the results much, but it eliminates the trial-and-error process for obtaining the minimum required fleet size.

Based on its performance on the DARP studied in this research, the proposed rejected-reinsertion operator seems promising for other vehicle routing problems with time windows, especially for heavily time-constrained problems (e.g., PDP or taxi scheduling). This operator alleviates the myopic behavior of an insertion method in an efficient way. The quality and computational efficiency of the heuristic also make it attractive for application in dynamic problems, in which at least some demand arises in real-time. In the next chapter, we will extend the heuristic to the dynamic version of the problem, and test its performance in dynamic applications.

## **Chapter 4 Online Heuristics for the Dynamic Dial-a-Ride**

### **Problem**

DAR services may operate according to one of the following two modes. In the static mode, all requests are known in advance (i.e. typically one day before the service actually takes place). In the dynamic mode, at least part of the requests are revealed and need to be scheduled in real-time. In this chapter, online heuristics for the dynamic DARP are presented and their performances are tested and compared through a computational study.

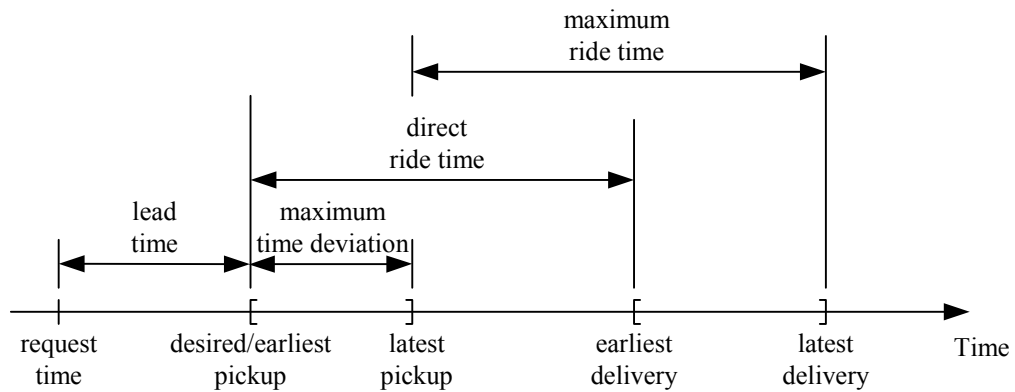
#### **4.1 Operating Scenario of the Dynamic Problem**

In a dynamic problem, it is assumed that the service requests are received throughout the service period. In addition to the operating scenario described for a static DARP in the last chapter, we define a term “lead time” for the dynamic DARP to describe the “dynamic” demand of the problem.

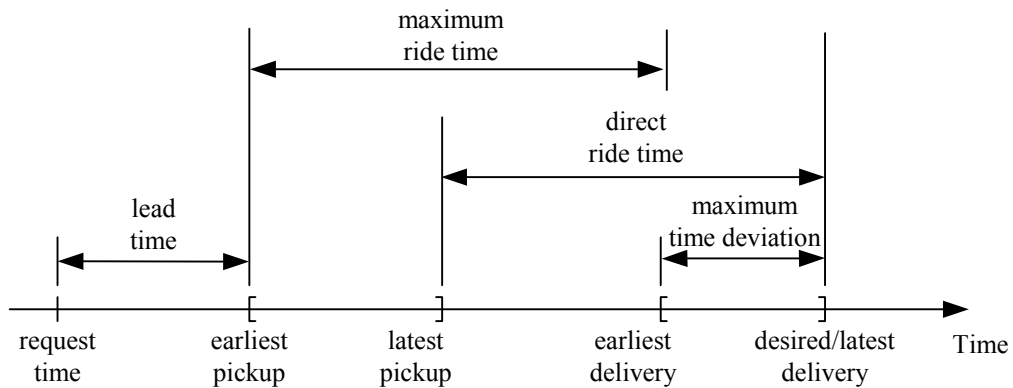
We define that

*Lead Time* is the time elapsed between the passenger’s request (calling) time and the earliest pickup time, no matter the request specifies desired pickup or delivery.

Figure 4-1 illustrates the relation of the lead time to other time components for both pickup- and delivery-specified passengers. The lead time is the measure indicating how far in advance the requests are made. The smaller the value of the lead time, the more immediate (urgent) the request is. If lead times of all the requests are very long (e.g. 24 hours), then the problem reduces to the static problem. In an immediate DAR service, the request should be fulfilled as soon as possible.



(a) Desired Pickup



(b) Desired Delivery

Figure 4-1. Definition of the lead time

## **4.2 Online Insertion Heuristics**

For the DAR service, a straightforward online heuristic is to repeat the static algorithm each time the system is updated (i.e. new call arrives). Thus, a short computation time is required for the algorithm. Based on the DARP algorithms reviewed in Section 2.3, insertion-based heuristics seem to be the most promising candidate for a large-scale DARP. An insertion-based heuristic is computationally efficient, and it could be well adapted to the dynamic version by freezing all the schedules that have already take place and continuously inserting the new requests. Its concept is straightforward and can easily handle many uncertainties involved in the DAR operation, such as vehicle breakdown and cancellation of trips, without reconsidering the whole problem. For example, assume a vehicle breaks down during the service with several passengers still on board to be delivered and several passengers waiting to be picked up. An insertion-based algorithm can sequentially re-insert those demands into other vehicles' schedules in a very short time.

Below, two online insertion-based heuristics for the dynamic DARP are presented.

### **4.2.1 Immediate online insertion heuristic**

As mentioned, a straightforward online heuristic inserts the new request into the vehicle once the request is received, considering that pickup times of passengers already onboard cannot be changed and vehicle locations should be updated. Apparently, passengers delivered before the call time of the new request are no longer considered. Further requests are not predicted, due to the uncertainty in the positions of the stop locations,

and the uncertainty in the widths and starting times of their corresponding time windows.

The immediate online insertion heuristic can be described as follows:

Immediate Insertion Heuristic:

Step 0: Initialize locations of available vehicles.

Set periodical improvement interval  $\Delta$ .

Step 1: Wait until the appearance of the new request.

Step 2: Update the locations of available vehicles. Freeze the route and schedules up to the current time instance.

Step 3: Insert request into the vehicle routes and determine the schedules. Go to Step 1.

Step 4: (Optional) Perform improvement procedure at interval  $\Delta$ .

In Step 2, 'Freeze' means that the pickup times of those passengers on-board cannot be changed and only their delivery times can be adjusted by the new insertions. Both the pickup and delivery times of those passengers still waiting at their origins can be adjusted by the new insertions. It is assumed that positions and status of the vehicles are known at all times (e.g. by automatic vehicle location technology).

#### **4.2.2 Rolling horizon online insertion heuristic**

The immediate online heuristic inserts the requests in the order of their calling times.

However, in a system with requests having different lead times, the deferment of insertion of some requests whose desired time of service is relatively far away from the

current time may results better routing and scheduling decision because of the flexibility reserved to serve more urgent requests which may arise soon. For this purpose, a rolling horizon principle is applied to the dynamic DARP. The rolling horizon principle is illustrated in Figure 4-2. The new online heuristic can be described as follows:

#### Rolling Horizon Insertion Heuristic

Step 0: Initialize locations of available vehicles.

List all known unassigned requests  $P$  in order of earliest pickup time.

Set length of the time horizon  $L$  and the rolling interval  $\alpha$  ( $\alpha \leq L$ ).

Set periodical improvement interval  $\Delta$ .

Set  $k = 1$ ,  $t_k = 0$  ( $k$  is the index of the iteration);

Step 1: Next horizon is  $(t_k, t_k + L)$ .

Form list of requests eligible for insertion  $P'$ ,  $P' \subseteq P$  (all requests in  $P$  whose earliest pickup times are between  $t_k$  and  $t_k + L$ ).

Step 2: If  $P' \neq \phi$ ,

Update vehicle locations at time  $t_k$ .

Freeze routes and schedules up to the current time instance  $t_k$ .

Insert requests from  $P'$  to vehicles and determine the schedules.

Remove assigned requests from the unassigned requests list  $P = P - P'$ .

Step 3: Wait until the appearance of the new request at  $T_i$  or until the rolling horizon time  $t_k + \alpha$  is reached. If the new request appears, go to Step 4; else, go to Step 5.

Step 4: If the new request is an urgent request (i.e.  $EPT_i < T_i + L$ ), schedule the new request immediately; else, insert the new request into the unassigned list  $P$  according to the earliest pickup times. Go to Step 3.

Step 5: Roll time horizon  $t_{k+1} = t_k + \alpha$ . Go to Step 1.

Step 6: (Optional) Perform improvement procedure at interval  $\Delta$ .

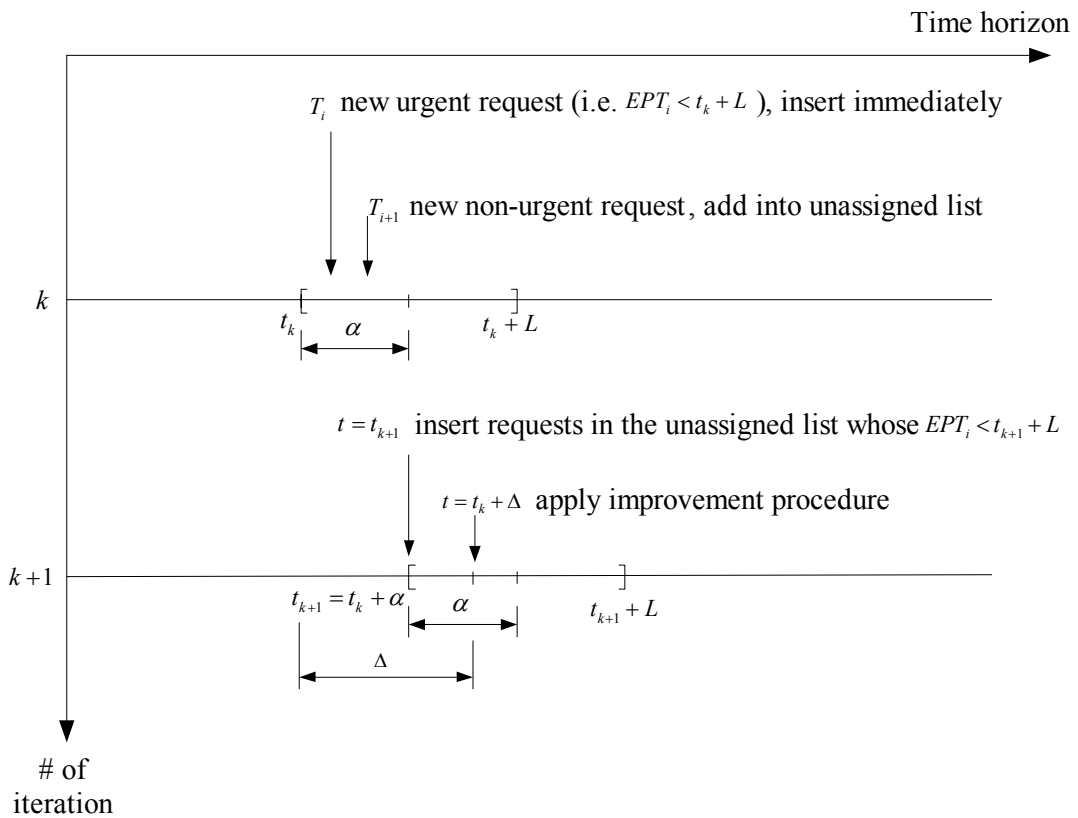


Figure 4-2. Schematic representation of the rolling horizon online insertion principle



### **4.3 Computational Study**

The heuristics are tested on the same randomly generated test problems introduced in Chapter 3 for the static DAR problem except that additional lead time information will be generated for the dynamic version of the problems. Five replications are generated for each problem, to deal with the randomness of the demand, and the statistics reported are the average over the five replications. We use the same objective function as in the static case. The computer program is coded using visual C++ and is run in a Pentium M, 1.6 GHz laptop.

The lead time distribution for all the test problems is generated as follows. It is assumed that half of the total requests are advance demand (i.e. call the service one day ahead). For the remaining half of the requests, the lead time follows a uniform distribution [60,120] minutes.

Throughout this dissertation, the default parameter settings for the rolling horizon online insertion heuristics are as follows, except otherwise specified. The rolling horizon is set as 1 hour and the rolling interval is set as 10 minutes. The interval of the periodical improvement is 30 minutes.

#### **4.3.1 Comparison of rolling horizon and immediate online insertion heuristics**

For each scenario considered, four heuristic variations are implemented and compared. In Tables 4-1 to 4-3, Heuristics D1 and D1w ('D' represents dynamic algorithm) are the immediate insertion online heuristics without and with the periodical improvement.

Heuristics D2 and D2w are the rolling horizon online heuristics without and with the periodical improvement. For insertion, the rejected-reinsertion heuristic with variable fleet size is implemented. The initial fleet sizes are 30, 40 and 50 for scenarios L, M and H, respectively. The statistics reported in the tables of this chapter are the same as those defined in 3.3.1. The last column is the average computation time for each additional request.

Table 4-1. Comparison of rolling horizon vs immediate online insertion heuristics for scenario L

Heuristic	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)	Comp. time per request (sec)
D1	41.8	5,175	3.74	11,460	14.00	1.488	9.4	157	0.1
D2	40.6	4,838	4.00	11,868	14.14	1.538	10.0	68	0.05
D2 vs D1	-2.9%	-6.5%	+7.0%	+3.6%	+1.0%	+3.4%	+6.4%	-57%	
D1w	37.4	4,436	4.12	10,381	14.61	1.356	9.2	8601	6.9
D2w	35.6	4,230	4.46	10,709	14.87	1.401	9.6	1701	1.2
D2w vs D1w	-4.8%	-4.6%	+8.3%	+3.2%	+1.8%	+3.3%	+4.3%	-80%	

Table 4-2. Comparison of rolling horizon vs immediate online insertion heuristics for scenario M

Heuristic	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)	Comp. time per request (sec)
D1	50.4	5,783	3.20	10,516	9.18	1.372	7.4	66	0.05
D2	45.6	5,341	3.52	10,676	8.96	1.397	7.6	26	0.02
D2 vs D1	-9.5%	-7.6%	+10.0%	+1.5%	-2.4%	+1.8%	+2.7%	-61%	
D1w	45.6	5,100	3.44	9,861	9.54	1.293	7.6	2215	1.6
D2w	41.8	4,808	3.82	10,094	9.59	1.325	7.6	597	0.44
D2w vs D1w	-8.3%	-5.7%	+11.0%	+2.3%	+0.5%	+2.5%	0%	-73%	

Table 4-3. Comparison of rolling horizon vs immediate online insertion heuristics for scenario H

Heuristic	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)	Comp. time per request (sec)
D1	66.2	6,942	2.51	9,200	4.22	1.218	6.2	29	0.02
D2	59.6	6,346	2.78	9,347	4.18	1.236	5.8	17	0.01
D2 vs D1	-10.0%	-8.6%	+10.8%	+1.6%	-0.9%	+1.5%	-6.5%	-41%	
D1w	62.4	6,331	2.61	8,958	4.32	1.184	6.0	469	0.34
D2w	56.8	5,936	2.88	9,204	4.36	1.216	6.2	141	0.10
D2w vs D1w	-9.0%	-6.2%	+10.3%	+2.7%	+0.9%	+2.7%	+3.3%	-70%	

Based on Tables 4-1 to 4-3, the rolling horizon online heuristics outperform the immediate online heuristics either without or with the periodical improvement (D2 vs D1, and D2w vs D1w) in terms of number of vehicles required, vehicle miles and vehicle productivity, for all three scenarios. The advantage of rolling horizon online heuristics over the immediate online heuristics increases as the problem gets more restrictive (from L to H). The differences in terms of passenger miles, average passenger time deviation and average passenger ride time ratio are relatively small. Besides, the rolling horizon online heuristics are more computationally efficient, i.e. for scenario L, the computation time of D2 versus D1 is 68 versus 157 seconds, and the computation time of D2w versus D1w is 1701 versus 8601 seconds. This occurs because the rolling horizon insertion method constructs the routes for the requests in the order of their urgency and the routes evolve as the time frames of the requests so that the number of rejected-reinsertion, trip reinsertion and trip exchange operations decreases, thus reducing the computation time.

Table 4-4 shows the number of vehicles required for the five randomly generated replications. The average and standard deviation are shown in the last two columns. The results show that Heuristic D2w always uses the fewest vehicles for all individual replications. The standard deviations of all heuristics for all service scenarios are around one vehicle. From an operational point of view, this means that the fluctuations on the number of vehicles required for a given level of demand are not evident for a dynamic problem even if the demand is randomly distributed over time and space. Table 4-4 also reports in parentheses the computation time in seconds for Heuristic D2w.

Table 4-4. Variability of number of vehicles over five replications

Scenario	Algo.	Replication					Average	Standard deviation
		1	2	3	4	5		
L	D1	42	43	42	40	42	41.8	1.10
	D1w	37	39	37	37	37	37.4	0.89
	D2	41	40	42	40	40	40.6	0.89
	D2w	36 (1670)	36 (1646)	34 (1704)	35 (1696)	37 (1788)	35.6 (1701)	1.14 (54)
M	D1	51	53	48	50	50	50.4	1.82
	D1w	45	46	45	47	45	45.6	0.89
	D2	46	47	46	45	44	45.6	1.14
	D2w	42 (601)	43 (550)	40 (594)	42 (614)	42 (626)	41.8 (597)	1.10 (29)
H	D1	65	68	65	66	67	66.2	1.30
	D1w	62	64	61	63	62	62.4	1.14
	D2	58	60	60	61	59	59.6	1.14
	D2w	57 (151)	57 (134)	55 (150)	58 (128)	57 (142)	56.8 (141)	1.10 (10)

### 4.3.2 Performance of the periodical improvement procedure

The results of Tables 4-1 to 4-3 are rearranged in Tables 4-5 to 4-7 to show the performance of the periodical improvement procedure in the dynamic problem. As shown in the tables, the periodical improvement procedure is effective in reducing the number of vehicles required (up to -12.3%), the total vehicle miles (up to -14.3%), and thus increasing the vehicle productivity (up to 11.5%). The improvement is more prominent

for the less restricted problem (i.e. scenario L) than the more restricted problem (i.e. scenario H). The periodical improvement is consistently effective in both the static and dynamic problems.

Table 4-5. Comparison of online heuristics without and with periodical improvement for scenario L

Heuristic	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)
D1	41.8	5,175	3.74	11,460	14.00	1.488	9.4	157
D1w	37.4	4,436	4.12	10,381	14.61	1.356	9.2	8601
D1w vs D1	-10.5%	-14.3%	+10.2%					
D2	40.6	4,838	4.00	11,868	14.14	1.538	10.0	68
D2w	35.6	4,230	4.46	10,709	14.87	1.401	9.6	1701
D2w vs D2	-12.3%	-12.6%	+11.5%					



Table 4-6. Comparison of online heuristics without and with periodical improvement for scenario M

Heuristic	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)
D1	50.4	5,783	3.20	10,516	9.18	1.372	7.4	66
D1w	45.6	5,100	3.44	9,861	9.54	1.293	7.6	2215
D1w vs D1	-9.5%	-11.8%	+7.5%					
D2	45.6	5,341	3.52	10,676	8.96	1.397	7.6	26
D2w	41.8	4,808	3.82	10,094	9.59	1.325	7.6	597
D2w vs D2	-8.3%	-10.0%	+8.5%					

Table 4-7. Comparison of online heuristics without and with periodical improvement for scenario H

Heuristic	# of vehicles	Vehicle miles	Vehicle prod. (trips/veh hr)	Pass. miles	Avg dev. (min)	Ride time ratio	Max pass. onboard	Comp. time (sec)
D1	66.2	6,942	2.51	9,200	4.22	1.218	6.2	29
D1w	62.4	6,331	2.61	8,958	4.32	1.184	6.0	469
D1w vs D1	-5.7%	-8.8%	+4.0%					
D2	59.6	6,346	2.78	9,347	4.18	1.236	5.8	17
D2w	56.8	5,936	2.88	9,204	4.36	1.216	6.2	141
D2w vs D2	-4.7%	-6.5%	+3.6%					

### 4.3.3 Test of the effectiveness of the rejected-reinsertion operation

To test the performance of the rejected-reinsertion operation in the dynamic version of the problem, we run four additional heuristics corresponding to Heuristics D1, D1w, D2 and D2w, but without using the rejected-reinsertion operator in the insertion process. The minimum number of vehicles required by those heuristics without and with the rejected-reinsertion operation are reported in Table 4-8.

Table 4-8. Comparison of heuristics without and with rejected-reinsertion operation

Heuristic	Rejected-reinsertion	# of vehicles (Comp. time in seconds)					
		L		M		H	
D1	without	49.8	(34)	58.6	(26)	74.0	(19)
	with	41.8	(157)	50.4	(66)	66.2	(29)
	difference	-16.1%		-14.0%		-10.5%	
D1w	without	41.4	(7030)	49.4	(2070)	69.0	(414)
	with	37.4	(9394)	45.6	(2215)	62.4	(469)
	difference	-9.7%		-7.7%		-9.6%	
D2	without	44.8	(25)	49.8	(16)	63.4	(10)
	with	40.6	(68)	45.6	(26)	59.6	(17)
	difference	-9.4%		-8.4%		-6.0%	
D2w	without	36.8	(1606)	44.2	(567)	59.4	(141)
	with	35.6	(1701)	41.8	(597)	56.8	(141)
	difference	-3.3%		-5.4%		-4.4%	

Based on Table 4-8, the rejected-reinsertion operation proposed for the static DARP is also very effective for the dynamic DARP. The results show that heuristics with the rejected-reinsertion operator require smaller vehicle fleets than heuristics without the operator, for all service quality scenarios. The improvement tends to be smaller when the periodical improvement procedure is implemented. For heuristics without the periodical improvement (i.e., D1 or D2), the rejected-reinsertion operation can reduce the number of vehicles by up to 16.1% for immediate online insertion heuristic (D1) and up to 9.4% for rolling horizon online insertion heuristic (D2). For heuristics with the periodical improvement (i.e., D1w or D2w), the rejected-reinsertion operation can reduce the number of vehicles by up to 9.7% for immediate online insertion heuristic (D1) and up to 5.4% for rolling horizon online insertion heuristic (D2). The online rejected-reinsertion heuristic with periodical improvement achieves the best results.

#### **4.3.4 Effect of advance information**

In Section 4.2, a notation “lead time” is defined which is a measurement of how far in advance a passenger calls in for a trip request. In principle, the earlier the passengers make the trip requests, the more flexibility a planner can have to schedule the trips. With all other parameters fixed at their default values, we vary the average lead time  $\bar{T}_{adv}$ . The lead time is uniformly distributed as  $U \sim [0, 2\bar{T}_{adv}]$ . Figure 4-3 to 4-5 shows the number of vehicles required by rolling horizon online heuristics, varying average lead time. ‘Best static result’ is the solution for the corresponding static problem.

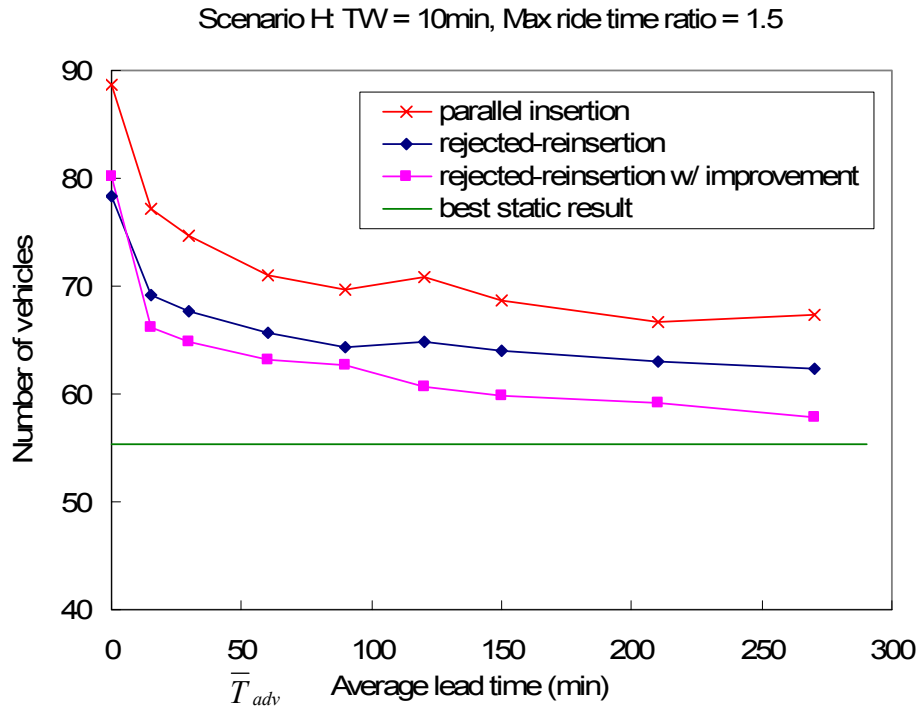


Figure 4-3. Vehicle fleet size requirement vs average lead time for scenario H

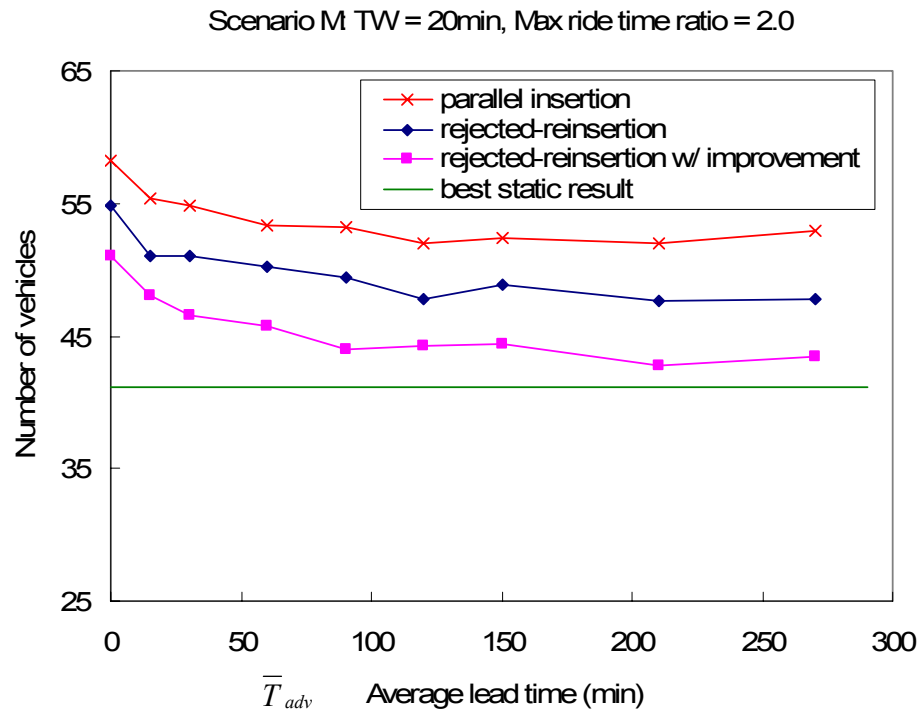


Figure 4-4. Vehicle fleet size requirement vs average lead time for scenario M

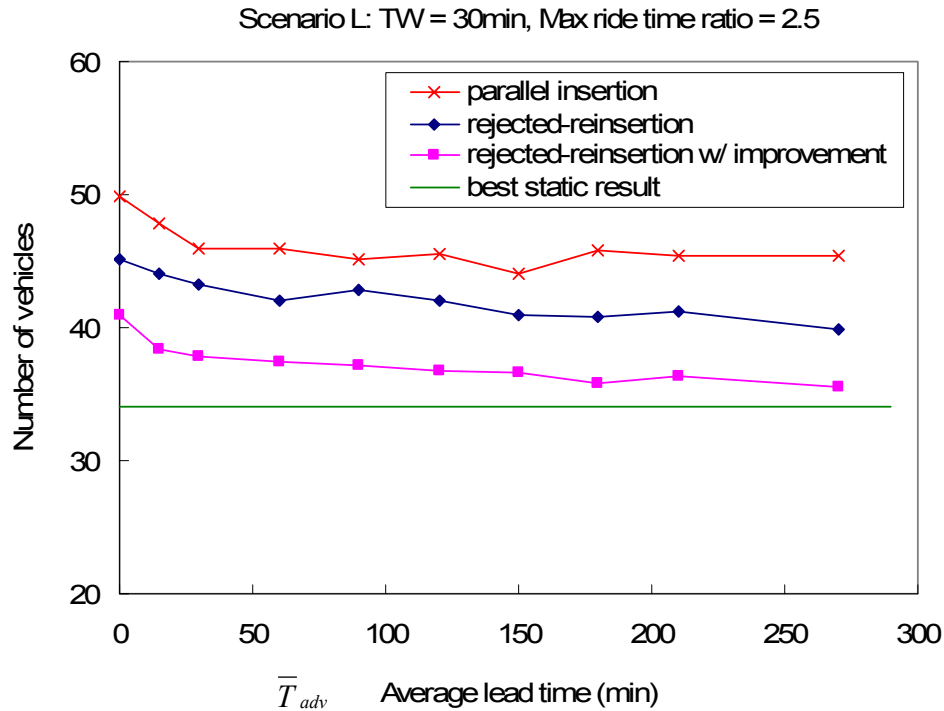


Figure 4-5. Vehicle fleet size requirement vs average lead time for scenario L

The results show that:

- The performance of all three heuristics improves as the average lead time increases.
- The performance is sensitive to a small lead time, especially for high service quality scenario H. This means that requests with very short trip notice will require a much larger vehicle fleet to satisfy the demand, especially for high service quality systems. For those systems, a minimum notice time (lead time) may be required.
- The performance tends to be relatively constant when the average lead time exceeds 1 hour.

- The rejected-reinsertion rolling horizon heuristic with periodical improvement performs the best and its results are close to the static results when average lead time exceeds 1 hour, which shows the effectiveness of the proposed online heuristic in dealing with dynamic demand.

Figures 4-6 ~ 4-8 show the effect of minimum lead time  $T_{adv}^{\min}$  on the heuristic performance. Only the best rejected-reinsertion online heuristic with periodical improvement is tested. The lead time is uniformly distributed as  $U \sim [T_{adv}^{\min}, 2\bar{T}_{adv} - T_{adv}^{\min}]$ . The results show that the minimum lead time constraint can moderately improve the heuristic performance for high and medium service quality scenarios. No great improvement is observed when the minimum lead time increases from 15 to 30 minutes.

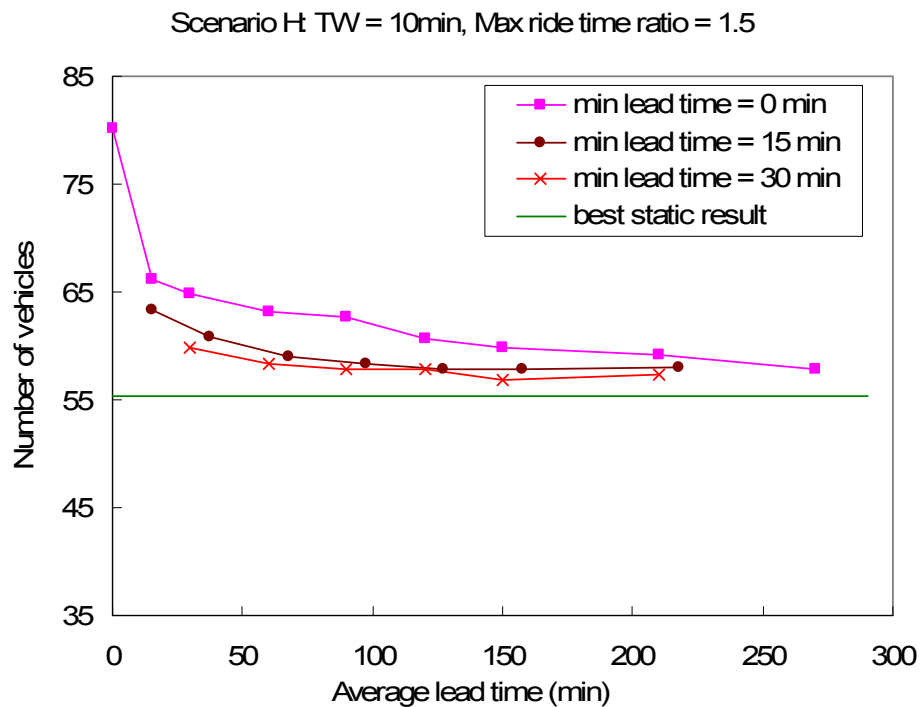


Figure 4-6. Effect of minimum lead time for scenario H

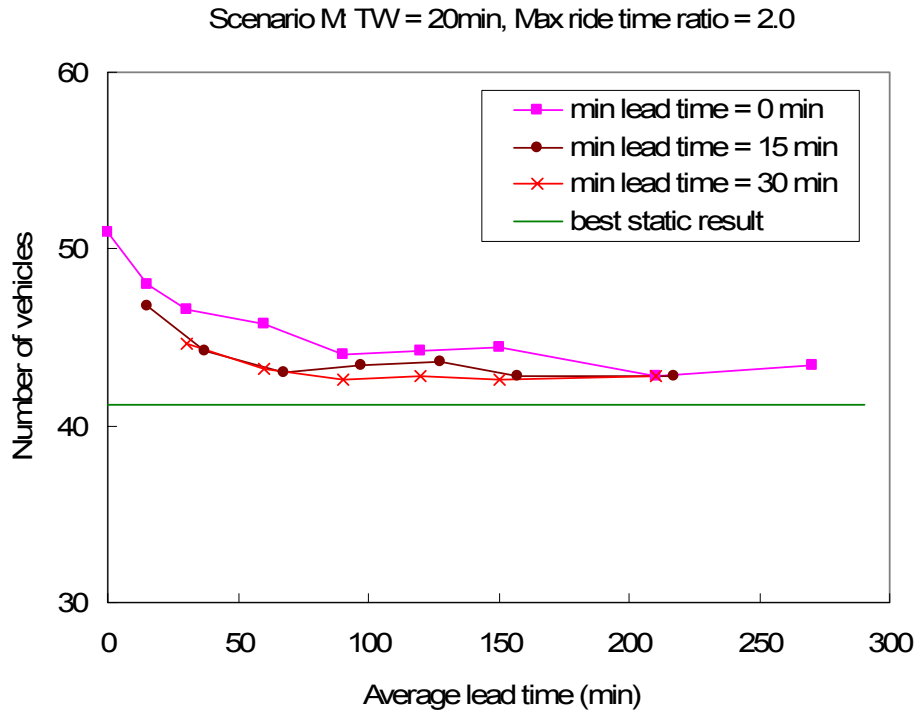


Figure 4-7. Effect of minimum lead time for scenario M

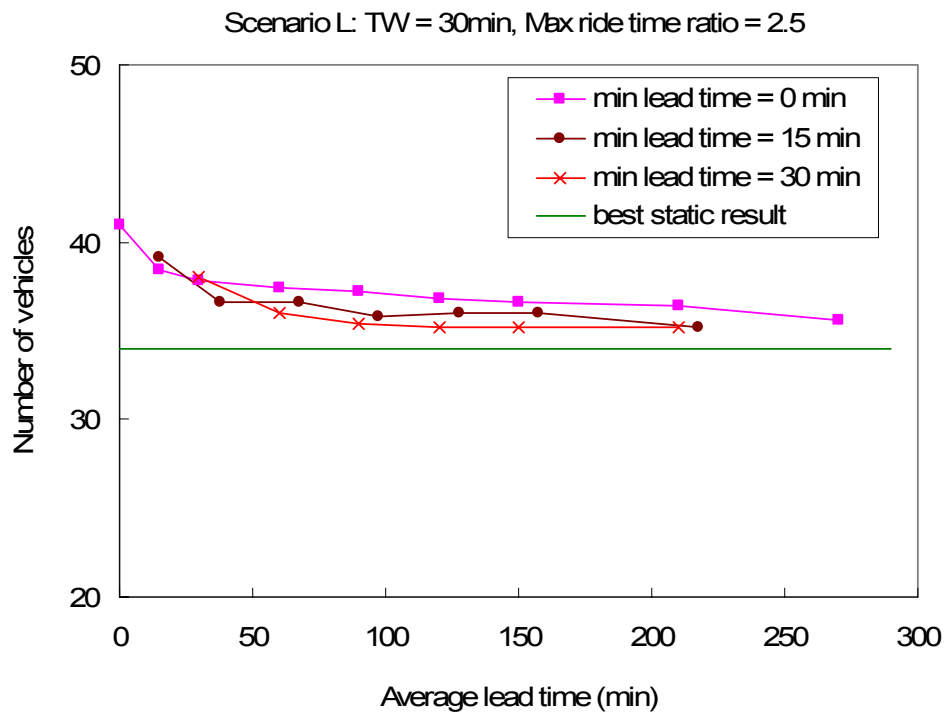


Figure 4-8. Effect of minimum lead time for scenario L

The results in this section are useful for DAR policy makers in determining the minimum trip notice time the passengers should be asked to provide. A minimum notice time of 15-30 minutes is suggested in order to lower the system operating cost.

#### **4.3.5 Sensitivity analysis of parameter settings of the rolling horizon online heuristic**

The rolling horizon online heuristic involves a few heuristic parameters (i.e. the length of the horizon, the rolling interval, and the improvement period if periodical improvement is implemented). This section tests the sensitivity of those parameters.

Table 4-9 shows the number of vehicles required as we only vary the length of the horizon (at 30, 60, and 90 minutes) with other parameters fixed at their default values, using heuristic D2 and D2w (rolling horizon heuristic without and with periodical improvement). It is found that the computation time increases with the horizon. This occurs because the longer the horizon, the more reinsertion or exchange operations may be involved in the computation. The number of vehicles is not very sensitive to the horizon. Slightly more vehicles are needed for a horizon of 30 minutes rather than a horizon of 60 or 90 minutes. Therefore, horizon values are not very critical for the rolling horizon heuristic. A horizon of 60 minutes seems appropriate and is used in this study.



Table 4-9. Sensitivity to the horizon

Heuristic	Horizon (min)	# of vehicles (computation time in seconds)		
		L	M	H
D2	30	41.2 (42)	47.2 (19)	61.2 (11)
	60	40.6 (68)	45.6 (26)	59.6 (17)
	90	40.0 (81)	46.4 (35)	59.6 (19)
D2w	30	36.4 (717)	43.6 (248)	59.4 (54)
	60	35.6 (1,701)	41.8 (597)	56.8 (141)
	90	35.0 (2,975)	42.0 (1011)	57.0 (232)

Table 4-10 shows the number of vehicles required when we only vary the rolling interval  $\alpha$  (at 5, 10, and 20 minutes), while other parameters stay at their default values, using Heuristic D2 and D2w (rolling horizon heuristic without and with periodical improvement ). The results show no effect of the rolling interval on the number of vehicles. The computation time is comparable as well. Again, it is not very critical to set the rolling interval value for the rolling horizon heuristic. A rolling interval of 10 minutes is used in this study.

Table 4-10. Sensitivity to the rolling interval  $\alpha$

Heuristic	Rolling interval $\alpha$ (min)	# of vehicles (computation time in seconds)		
		L	M	H
D2	5	40.4 (69)	46.8 (31)	59.6 (17)
	10	40.6 (68)	45.6 (26)	59.6 (17)
	20	40.2 (62)	46.2 (28)	59.4 (16)
D2w	5	35.4 (2041)	42.6 (581)	57.0 (139)
	10	35.6 (1701)	41.8 (597)	56.8 (141)
	20	35.4 (1991)	42.0 (558)	56.0 (148)

Table 4-11 shows the number of vehicles required when we only vary the time interval (at 15, 30, and 60 minutes) at which the periodical improvement procedure is implemented, while other parameters stay at their default values, using Heuristic D2w (rolling horizon heuristic with periodical improvement). The general trend is that as the time interval decreases, the computation time increases nonlinearly with very limited improvement in the results. An improvement interval of 30 minutes is used in this study.

Table 4-11. Sensitivity to the improvement interval

Heuristic	Interval (min)	# of vehicles (computation time in seconds)		
		L	M	H
D2w	15	34.0 (3225)	42.0 (969)	56.4 (227)
	30	35.6 (1701)	41.8 (597)	56.8 (141)
	60	36.2 (1030)	42.2 (348)	57.4 (96)

#### **4.3.6 Comparison of two vehicle scheduling policies**

In Section 3.2.8, we have mentioned the vehicle scheduling which refers to the determination of the actual pickup and delivery times of the new insertion and the corresponding modification of the actual pickup and delivery times of the affected passengers assigned once the insertion sequence is determined. Two scheduling policies are analyzed and compared. The first policy is to schedule the new request as soon as possible. In the second policy, schedules are sought that minimize the time deviation of the passengers from their desired pickup or delivery times.

Tables 4-12 and 4-13 show the comparison of the number of vehicles required and average passenger time deviation by the two scheduling policies. As expected, the results from the two scheduling policies are quite similar. This occurs because, for a heavily loaded system, there is not much slack time available for shifting the schedule block in which the new request has been inserted. However, the second policy always yields slightly less average passenger time deviation from desired times. In this dissertation, the second scheduling policy is adopted.

Table 4-12. Effects of two scheduling policies on fleet size

Heuristic	Scheduling Policy	Number of vehicles required		
		L	M	H
D1	(1) ASAP	42.2	51.6	65.4
	(2) Min time deviation	41.8	50.4	66.2
D1w	(1) ASAP	37.0	46.0	61.8
	(2) Min time deviation	37.4	45.6	62.4
D2	(1) ASAP	40.0	44.8	58.4
	(2) Min time deviation	40.6	45.6	59.6
D2w	(1) ASAP	36.0	41.6	56.6
	(2) Min time deviation	35.6	41.8	56.8

Table 4-13. Effects of two scheduling policies on average passenger time deviation

Heuristic	Scheduling policy	Average passenger time deviation (min)		
		L	M	H
D1	Min time deviation	14.00	9.18	4.22
	ASAP	14.19	9.41	4.76
D1w	Min time deviation	14.61	9.54	4.32
	ASAP	14.96	10.04	4.50
D2	Min time deviation	14.14	8.96	4.18
	ASAP	14.41	9.30	4.71
D2w	Min time deviation	14.87	9.59	4.36
	ASAP	15.08	10.01	4.84

#### **4.4 Conclusions**

In this chapter, two online insertion heuristics (with four variations) are developed for the dynamic DARP. These are the immediate insertion online heuristic and the rolling horizon online heuristic. The rejected-reinsertion heuristic for the static problem is incorporated in the online heuristics. The performances of the heuristics are tested and compared for a set of randomly generated problems.

The rolling horizon online heuristic outperforms the immediate insertion online heuristic for demand scenario in which different demand lead times exist. The heuristic is computationally efficient. It is simple in concept, and it does not involve complex algorithm parameters which need to be tested for specific problems. The rolling horizon online heuristic with periodical improvement, the best among those heuristic variations developed, is used in the simulation experiments for the development of the performance models.

## **Chapter 5 Development of Performance Metamodels**

The routing and scheduling heuristics developed in the last two chapters to solve the DARP for a given operational scenario can be thought of as a mechanism that turns the settings of a group of experimental factors (i.e. demand, service area, time constraints, vehicle characteristics, etc.) into output performance measures (i.e. number of vehicles required). However, the explicit functional form of the relation of outputs with respect to the input parameters is unknown. Response surface methodology (Box and Draper, 1987; Khuri and Cornell, 1996; Myers and Montgomery, 2002) is a very popular metamodeling technique used to approximate this kind of functional relation. The resulting functions (or models) are usually called metamodels in that they provide a “model of the model” (Kleijnen, 1987). The approximate formula or equations could be used to predict the performance for different number of input parameter combinations.

In this chapter, performance metamodels are developed using the response surface methodology. In Section 5.1, response surface metamodeling technique is introduced. Section 5.2 describes the design of the experiments, in which the main input factors, region of interest of the factors, a face-centered central composite design and generation

of the demand scenarios are discussed in details. Monte Carlo simulation is used to generate demand scenarios. Then for each scenario, the dynamic routing and scheduling algorithm assigns demands to routes and schedules their pickup and delivery times. The output performance measures are collected. Section 5.3 contains the regression analysis for the experiment data and the models development. Validation of the metamodels using a set of new generated data is performed in Section 5.4. Finally, the metamodels are summarized in the last section.

### **5.1 Introduction of Response Surface Methodology**

Response surface methodology is a collection of statistical and mathematical techniques useful for developing, improving, and optimizing processes (Myers and Montgomery, 2002). It is based on the work of Box (1954) and Hunter (1958, 1959a, 1959b), and has been used effectively in other areas (Box & Draper, 1987; Box, Hunter, & Hunter, 1978). The most extensive applications of responsive surface methodology are in situations where several input variables potentially influence some performance measure of quality characteristic of the product or process. This performance measures or quality characteristic is called the *response*. It is typically measured on a continuous scale. The input variables are sometimes called *independent variables*. For a relationship between a response and less than three input variables, the responses for different combinations of input variables constitute a response surface, which has led the term response surface methodology. One typical application of the response surface methodology is to map or approximate a response surface over a particular region of interest.

Response surface methodology is a set of techniques that encompasses (Khuri and Cornell, 1996):

1. Setting up a series of experiments that will yield adequate and reliable measurements of the response of interest.
2. Determining a mathematical model that best fits the data collected from the design.
3. Determining the optimal settings of the experimental factors that produce the optimal value of the response.

The first two techniques are employed in this dissertation to develop the performance metamodels.

The approximate empirical functions or models are usually built using statistical regression methods. The most common models used in response surface methodology are the polynomial first-order and second-order response surface models. Note that response surface methods are additional techniques employed before, while, and after a regression analysis is performed on the data (Khuri and Cornell, 1996). The experiment must be designed, that is, the input parameters must be selected and their value during experimentation must be designated before the regression analysis. After the regression analysis is performed, certain model testing procedures are applied.

The general form of a first-order model in  $k$  input variables  $X_1, X_2, \dots, X_k$  is

$$Y = \beta_0 + \sum_{i=1}^k \beta_i X_i + \varepsilon \quad (5-1)$$



where  $Y$  is an observable response variable,  $\beta_0, \beta_1, \dots, \beta_k$  are unknown parameters, and  $\varepsilon$  is a random error term. The general form of a second-order model is

$$Y = \beta_0 + \sum_{i=1}^k \beta_i X_i + \sum_{i=1}^k \sum_{j=1}^k \beta_{ij} X_i X_j + \varepsilon \quad (5-2)$$

where  $\beta_i$  ( $i = 1, 2, \dots, k$ ),  $\beta_{ij}$  ( $i = 1, 2, \dots, k; j = 1, 2, \dots, k$ ) are unknown parameters. If no interaction term is considered, Equation (5-2) becomes

$$Y = \beta_0 + \sum_{i=1}^k \beta_i X_i + \sum_{i=1}^k \beta_{ii} X_i^2 + \varepsilon \quad (5-3)$$

Note that the following multiplicative model (5-4) is intrinsically linear (Draper and Smith, 1998) and can be transformed into a linear model (5-5) by a logarithmic transformation

$$Y = \beta_0 \prod_{i=1}^k X_i^{\beta_i} \varepsilon \quad (5-4)$$

$$\log_{10} Y = \log_{10} \beta_0 + \sum_{i=1}^k \beta_i \log X_i + \log_{10} \varepsilon \quad (5-5)$$

The multiplicative model is also used in the regression analysis of this study.

## 5.2 Experimental Design

In experimental-design terminology, the input parameters and structural assumptions composing a model are called *factors*, and the output performance measures are called *responses* (Law and Kelton, 2000). The main tasks in the experimental design include:

- Selection of the input factors (parameters) which mostly affect the interested response
- Setting the interested range of the factors
- Determination of the number and values of the experimental points (one point corresponds one combination of the factors)
- (If the experiment includes simulation ), setting the simulation parameters (i.e. length of the simulation, number of replications)

### **5.2.1 Input factors**

In this dissertation, the output performance measure we are mostly concerned with is the vehicle resource requirement given demand and service quality level. Other measures include average passenger time deviation (waiting time if passengers specify desired pickup time) and average passenger ride time ratio.

Since the main purpose of the performance models is to aid in the planning stage and it is the most important to understand the tradeoff relation between the vehicle resource requirement and the level of service provided, we identify the following six factors as the main contributors to the vehicle resource requirement:

- Demand density
- Service area size
- Maximum time deviation
- Maximum ride time ratio
- Vehicle operating speed

- Passenger boarding and alighting time

Note that the road circuitry (ratio of the actual distance to the direct distance) affects the vehicle travel time. However, the effect of increasing road circuitry on the vehicle travel time is equivalent to the effect of decreasing vehicle operating speed on the vehicle travel time. Therefore, road circuitry is not treated as a separate factor. Instead, its effect will be incorporated with the vehicle operating speed into the final models. The speed can then be defined as the average speed based on Euclidean distances rather than actual distances through road networks.

The average passenger time deviation is expected to be mostly affected by the maximum time deviation and the average passenger ride time ratio is expected to be mostly affected by the maximum ride time ratio.

There are other input parameters and assumptions which are considered as fixed aspects of the models:

- Demand distribution in space may also have some effect on the performance measures. However, the demand pattern in space differs considerably in each practical scenario and it is difficult to fully describe it quantitatively (i.e. must specify uniform, Poisson or other distribution qualitatively). A uniform distribution of all origins and destinations is used to represent the most general case. A sensitivity analysis of non-uniform distribution in one direction of the area is performed in Section 6.1.2.

- For the demand distribution in time, the calling time interval between successive passenger requests is assumed to have a negative exponential distribution.
- The service area is assumed to be square. Simulation results by Eilon et al. (1971) suggest that minor variations in the shapes of zones (e.g. square, circle and equilateral triangle) with uniform internal demand do not greatly affect the length of traveling salesman tours within them. A sensitivity analysis of area shape (rectangular with different width length ratio) is provided in Section 6.1.1.
- It is assumed that half of the total requests are advance demand and the lead time for the remaining real-time requests is uniformly distributed as  $U \sim [0, 120]$  minutes. In a rolling horizon scheme, actually the advance demand is equivalent to the demand with lead time more than the rolling horizon. The effect of the advance information analyzed in Section 4.3.4 shows that the performance is not particularly sensitive to the distribution of the lead time for a medium or low service quality system. It also shows that it is costly for a high service quality system to allow short trip notice times for most customers. A mixed demand with some urgent and some non-urgent requests is considered.
- The rejected-reinsertion rolling horizon online heuristic with periodical improvement is used for the routing and scheduling, which is efficient in solving the large-scale dynamic DARP and is the best available.
- The probability that a passenger specifies a desired pick up or delivery time follows a binary distribution with 0.5 probability.
- A 1.15 road circuitry factor is used.

### 5.2.2 Region of interest

Table 5-1 shows the lower and upper values of the six factors considered. They are considered to cover the general region of interest for a DAR service.

Table 5-1. Lower and upper values of the factors

i	Factor	Lower value	Upper value
1	Service area (sq. mi.)	9	81
2	Demand density (trips/hr/sq. mi.)	1	10
3	Maximum time deviation (min)	10	30
4	Maximum ride time ratio	1.5	2.5
5	Vehicle operating speed (mph)	10	40
6	Boarding or alighting time (min)	0.5	1.5

### 5.2.3 Factorial design and face-centered central composite design

Factorial designs are widely used in experiments involving several factors where it is necessary to study the joint effect of the factors on a response (Montgomery, 2001).

Assume that the input variable is coded to take the value -1 when at its low lever and +1

when at its high, a  $2^k$  factorial design is such a design that requires  $2 \times 2 \times \dots \times 2 = 2^k$

observations with each factor chosen at the -1 and +1 levels. The  $2^k$  factorial design is an economic strategy to measure factor interactions and screen out unimportant factors.

Since only two levels are measured for each factor, the  $2^k$  factorial design is one of the first-order designs that are used to estimate first-order models. Similarly, a  $3^k$  factorial

design requires  $3 \times 3 \times \dots \times 3 = 3^k$  observations with each factor chosen at the -1, 0 and +1 levels. The  $3^k$  factorial design is one of the second-order designs that are used to estimate second-order models.  $3^k$  factorial design requires a large number of design points even for moderate value of  $k$ . For example, for  $k = 6$  as in this study,  $3^k$  factorial design requires  $3^6 = 6,561$  design points. If one design point needs 5 replications in a simulation experiment context, the total number of simulation runs would be 32,805, which is computationally expensive.

The class of central composite designs introduced by Box and Wilson (1951) is an alternative class of designs to the  $3^k$  factorial design. A central composite design consists of a  $2^k$  factorial design points augmented with  $2k$  axial points at  $(\pm\alpha, 0, 0, \dots, 0)$ ,  $(0, \pm\alpha, 0, \dots, 0)$ ,  $\dots$ ,  $(0, 0, 0, \dots, \pm\alpha)$  and  $n_c$  ( $n_c \geq 1$ ) center points  $(0, 0, 0, \dots, 0)$ . In Figure 5-1a for  $k = 3$ , a central composite design consists of a  $2^3 = 8$  factorial design points augmented with  $2 \cdot 3 = 6$  axial points and  $n_c$  center points. If the region of interest is cuboidal, a useful variation of the central composite design is the face-centered composite design with  $\alpha = 1$  (Figure 5-1b).

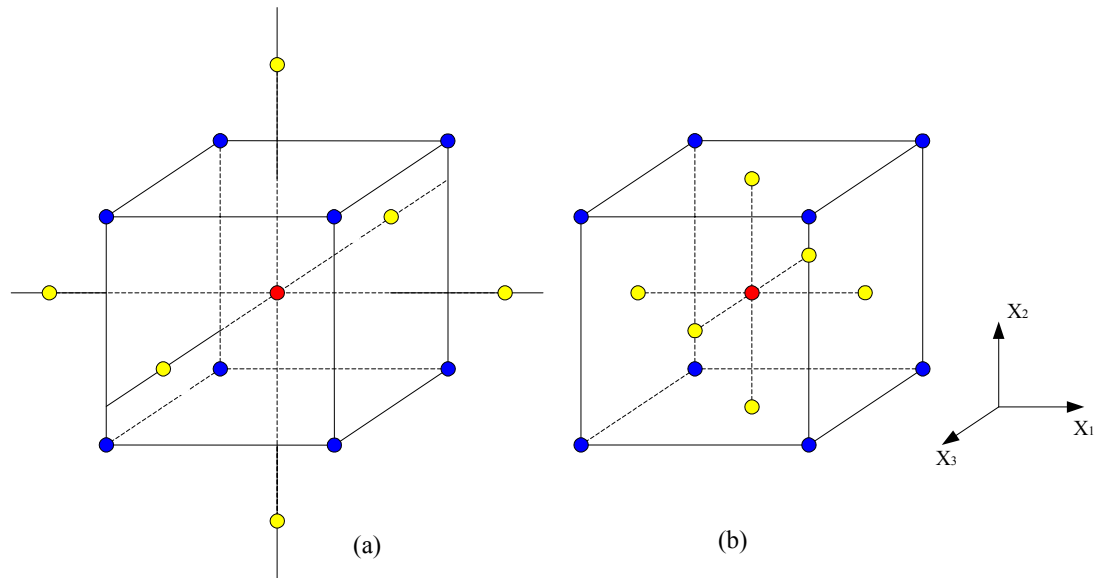


Figure 5-1. Central composite design (CCD) for  $k = 3$

(a) general CCD, (b) face-centered CCD

The face-centered central composite design is chosen for this study. Since the number of factors considered is 6, the design consists of  $2^6 = 64$  factorial design points and  $2 \cdot 6 = 12$  axial points. The number of center points is set as 6. Table 5-2 shows an example of the factor combinations for a face-centered composite design when  $k = 3$  and  $n_c = 2$ .

Table 5-2. Factor combinations for a face-centered CCD for  $k = 3$

<b># of Experiment</b>	<b>Factor 1</b>	<b>Factor 2</b>	<b>Factor 3</b>
1	-1	-1	-1
2	-1	-1	+1
3	-1	+1	-1
4	-1	+1	+1
5	+1	-1	-1
6	+1	-1	+1
7	+1	+1	-1
8	+1	+1	+1
9	-1	0	0
10	+1	0	0
11	0	-1	0
12	0	+1	0
13	0	0	-1
14	0	0	+1
15	0	0	0
16	0	0	0

#### **5.2.4 Generation of demand scenarios**

Monte Carlo simulation is used to generate specific demand attributes such as the origin, destination and calling time of each request. Here Monte Carlo simulation means a scheme employing random numbers to generate scenarios of demand configurations. It is assumed that origins and destinations of requests are uniformly distributed over the service area. The inter-arrival times of calls have a negative exponential distribution. Requests are generated for a three-hour service period, which represents a typical peak



hour period. To deal with the randomness of the demand, each experiment (each of the factor combination) is repeated five times. The average performance over those five replications represents one design point.

### 5.3 Regression Analysis

The notation used for the performance models is defined as follows:

Responses:

- $F$  Minimum number of operating vehicles which satisfy all demand for given time constraints
- $\bar{R}$  Average passenger ride time ratio (actual ride time divided by direct ride time)
- $\bar{T}_{dev}$  Average passenger time deviation from desired time (min) (the absolute value of the difference between the desired pickup/delivery time and the actual pickup/delivery time)

Factors:

- $A$  Service area size (sq. mi.)
- $b$  Total boarding and alighting time per person (min)
- $D$  Demand density (trips/sq. mi./hr)
- $R$  Maximum ride time ratio
- $V$  Vehicle operating speed (mph)
- $W$  Maximum time deviation (min)

### 5.3.1 Vehicle resource requirement model

In the first step, a first-order linear model as in Equation 5-1 is fitted to the response, which is the vehicle fleet size  $F$ . However, residual analysis suggests the transformation of the response  $F$  may result in better fit. A multiplicative form is hypothesized for the vehicle resource requirement model as follows:

$$F = \alpha_0 \frac{A^{\alpha_1} D^{\alpha_2} b^{\alpha_6}}{W^{\alpha_3} R^{\alpha_4} V^{\alpha_5}} \quad (5-6)$$

Equation (5-6) can be transformed into the following linear form:

$$\begin{aligned} \log_{10} F = \log_{10} \alpha_0 + \alpha_1 \log_{10} A + \alpha_2 \log_{10} D - \alpha_3 \log_{10} W - \alpha_4 \log_{10} R - \alpha_5 \log_{10} V \\ + \alpha_6 \log_{10} b \end{aligned} \quad (5-7)$$

Polynomial first-order and second-order models with transformed response  $\log_{10} F$  are also analyzed and their regression results along with the multiplicative model are compared. They are shown in Equations (5-8) and (5-9), respectively. No interaction terms are considered in the second-order model.

$$\log_{10} F = \beta_0 + \beta_1 A + \beta_2 D - \beta_3 W - \beta_4 R - \beta_5 V + \beta_6 b \quad (5-8)$$

$$\begin{aligned} \log_{10} F = \gamma_0 + \gamma_1 A + \gamma_2 D + \gamma_3 W + \gamma_4 R + \gamma_5 V + \gamma_6 b \\ + \gamma_{11} A + \gamma_{22} D + \gamma_{33} W + \gamma_{44} R + \gamma_{55} V + \gamma_{66} b \end{aligned} \quad (5-9)$$

All three models are estimated using linear regression with SPSS software (version 11.0).

The estimated parameters with standard errors in the parentheses, the corresponding adjusted  $R^2$  values,  $F$  values, the normal probability plots (Figures 5-2, 5-6 and 5-10) and the plots of residual against the predicted value (Figures 5-3, 5-7 and 5-11) are shown for each of the three models. The normal probability plot is used to check the

normality assumption of the error term in the least squares regression. The normal probability plot in SPSS plots the cumulative proportion of a single numeric variable against the cumulative proportion expected if the sample were from a normal distribution (SPSS, v11.0). If the sample is from a normal distribution the points will cluster around a straight line. The plots of residual vs the predicted value provide one way of checking the model's adequacy. If the model is adequate, the residual should contain no obvious patterns (Montgomery, 2001). Figures 5-4, 5-8, and 5-12 show the comparison of the observed  $\log_{10} F$  from the simulation experiments with the estimated ones by the regression models. Figures 5-5, 5-9, and 5-13 show the comparison in terms of the number of vehicles instead of taking the logarithm. The comparisons indicate graphically how well the models describe the data.

(1) multiplicative Model F1

$$\begin{aligned} \log_{10} F = & 0.680 + 1.074 \log_{10} A + 0.723 \log_{10} D - 0.287 \log_{10} W - 0.370 \log_{10} R \\ & (0.044) \quad (0.012) \quad (0.011) \quad (0.023) \quad (0.050) \\ & - 0.678 \log_{10} V + 0.205 \log_{10} b \\ & (0.018) \quad (0.014) \end{aligned} \quad (5-10)$$

$$\text{adjusted } R^2 = 0.989, F = 2,442$$

Equation (5-10) can be transformed back to the multiplicative form as

$$F = 4.79 \frac{A^{1.07} \cdot D^{0.72} \cdot b^{0.21}}{W^{0.29} \cdot R^{0.37} \cdot V^{0.68}} \quad (5-11)$$

(2) first-order Model F2 fitted to  $\log_{10} F$

$$\log_{10} F = 0.891 + 0.0136A + 0.0755D - 0.00784W - 0.0913R - 0.0144V + 0.0512b$$

$$(0.056) (0.0003) (0.0024) (0.0011) (0.022) (0.0007) (0.0088)$$

$$\text{adjusted } R^2 = 0.959, F = 606$$

(5-12)

(3) second-order Model F3 fitted to  $\log_{10} F$

$$\log_{10} F = 0.584 + 0.0358A + 0.198D - 0.00684W - 0.0819R - 0.033V + 0.0637b$$

$$(0.047) (0.0024) (0.018) (0.0005) (0.010) (0.006) (0.004)$$

$$- 0.00024A^2 - 0.0107D^2 + 0.000387V^2$$

$$(0.000025) (0.0017) (0.00012)$$

(5-13)

$$\text{adjusted } R^2 = 0.991, F = 1906$$

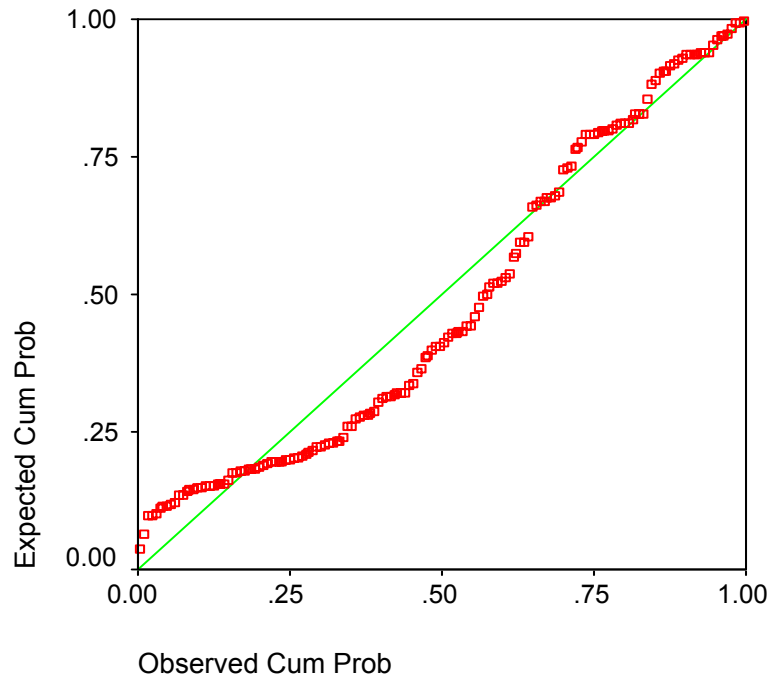


Figure 5-2. Normal probability plot of  $\log_{10} F$  of the multiplicative Model F1

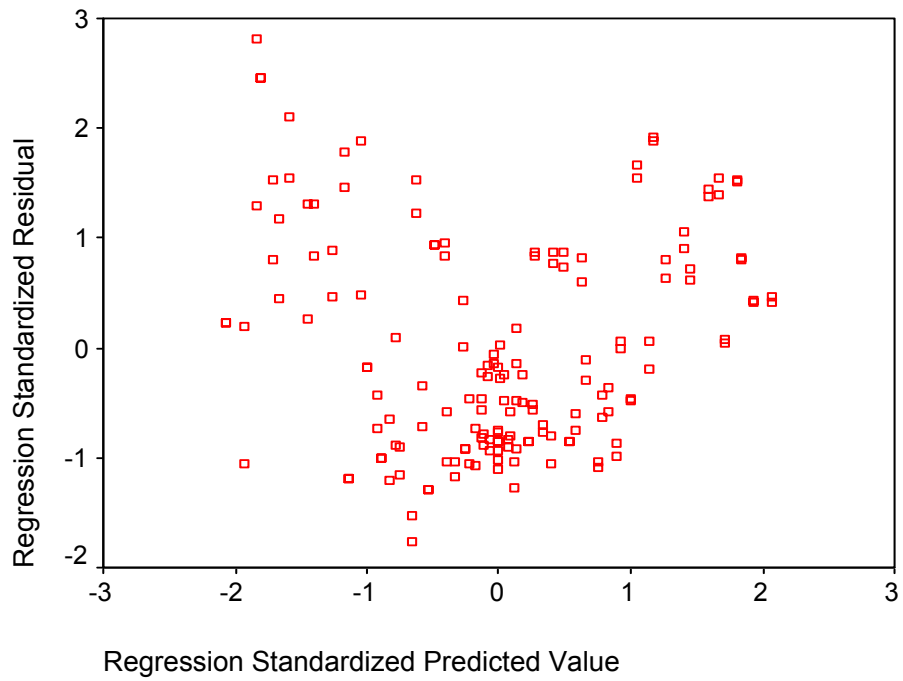


Figure 5-3. Residual vs the predicted  $\log_{10} F$  of the multiplicative Model F1

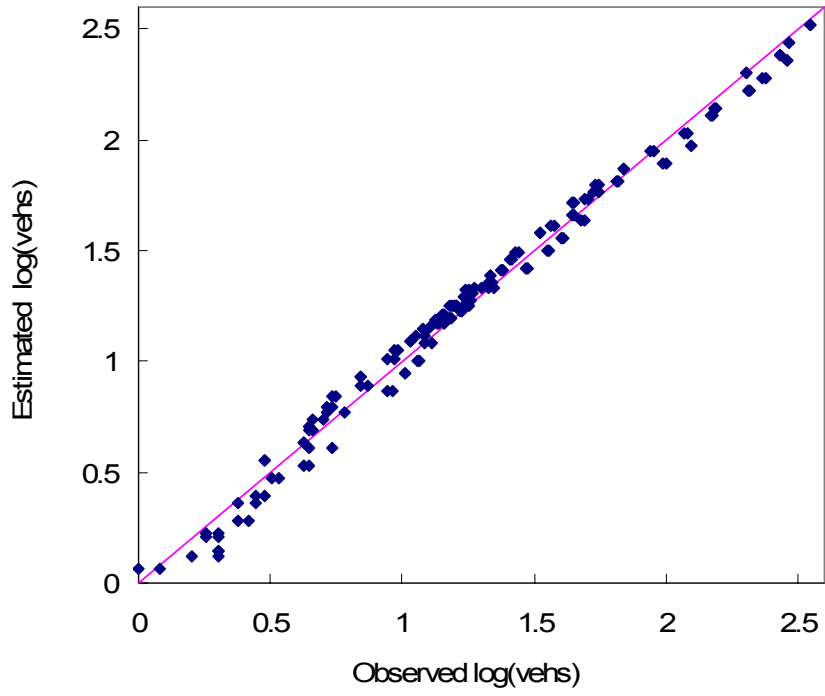


Figure 5-4. Estimated vs observed  $\log_{10} F$  of the multiplicative Model F1

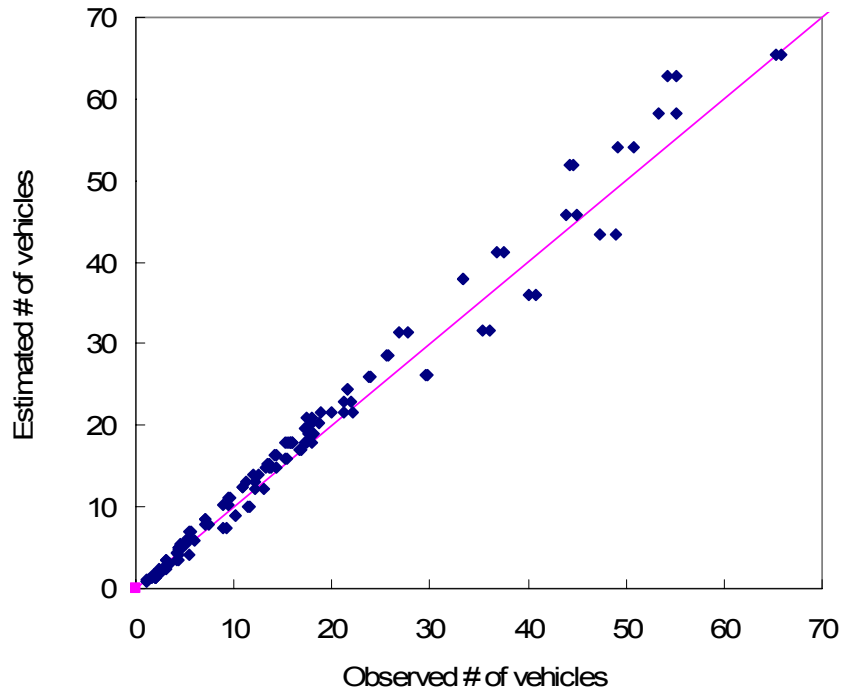


Figure 5-5. Estimated vs observed  $F$  of the multiplicative Model F1

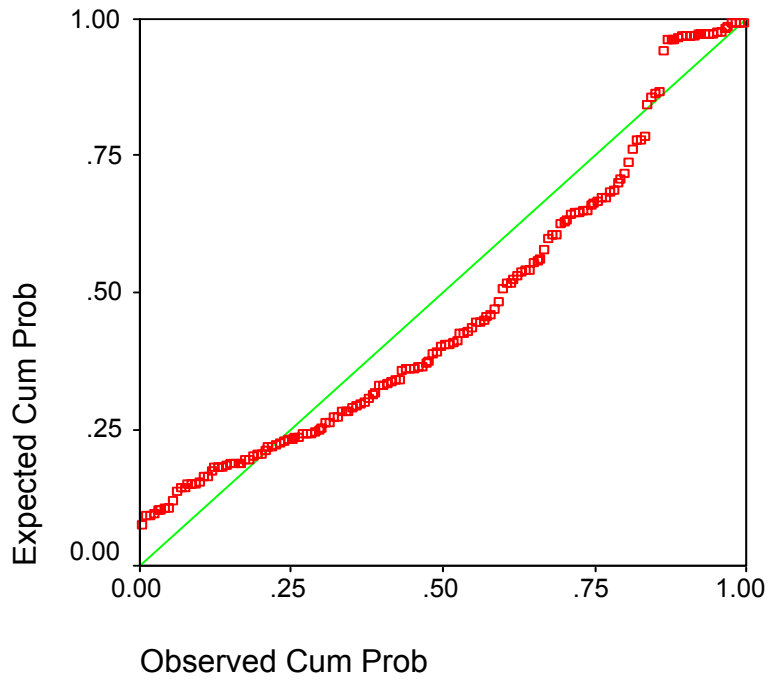


Figure 5-6. Normal probability plot of  $\log_{10} F$  of the first-order Model F2

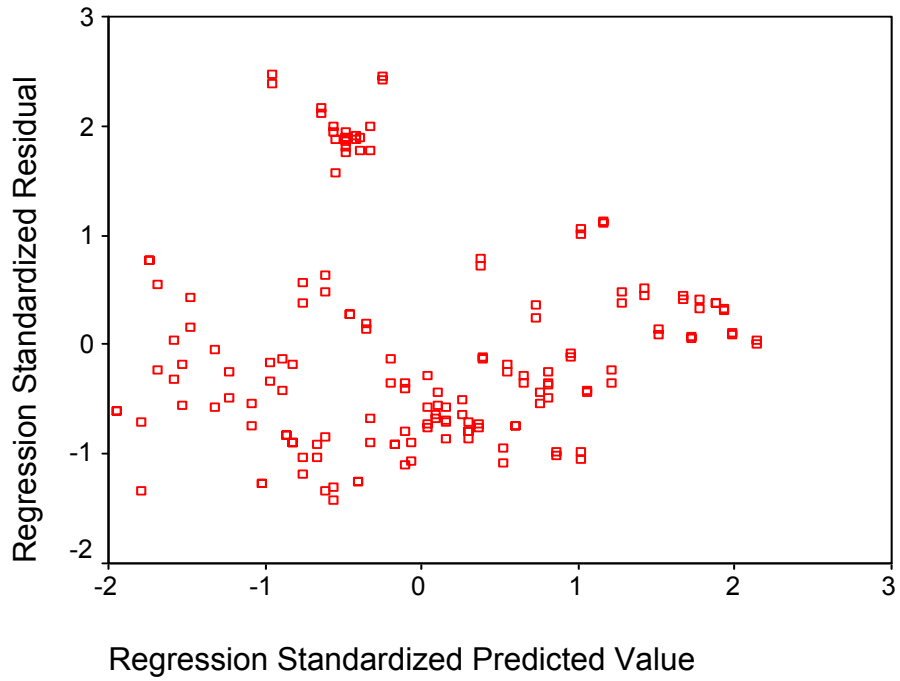


Figure 5-7. Residual vs the predicted  $\log_{10} F$  of the first-order Model F2

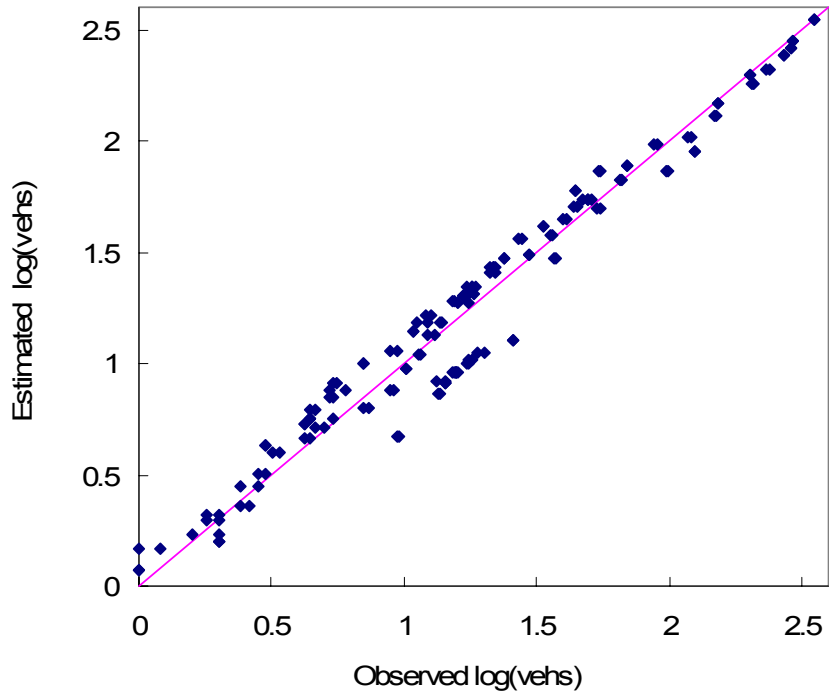


Figure 5-8. Estimated vs observed  $\log_{10} F$  of the first-order Model F2

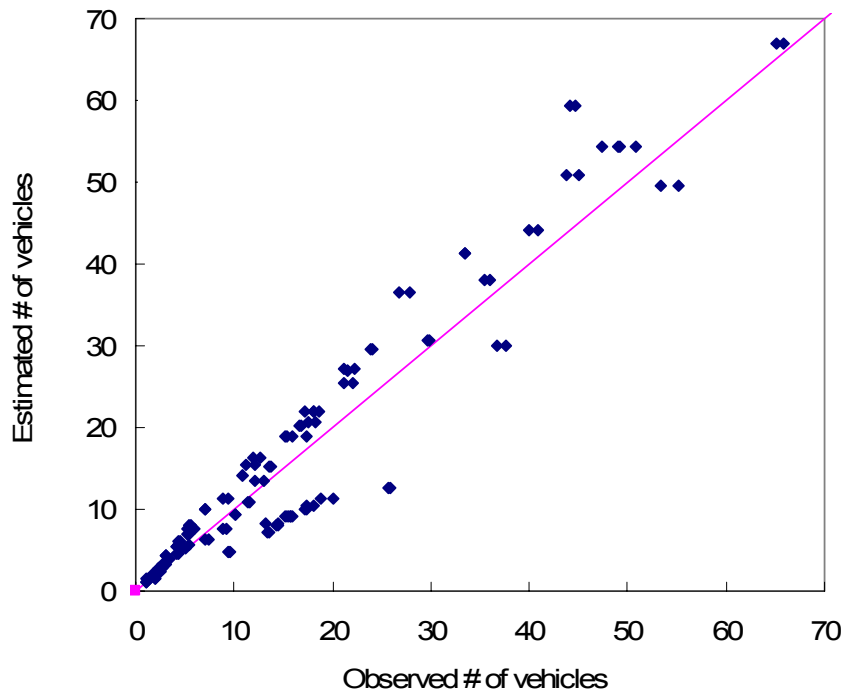


Figure 5-9. Estimated vs observed  $F$  of the first-order Model F2



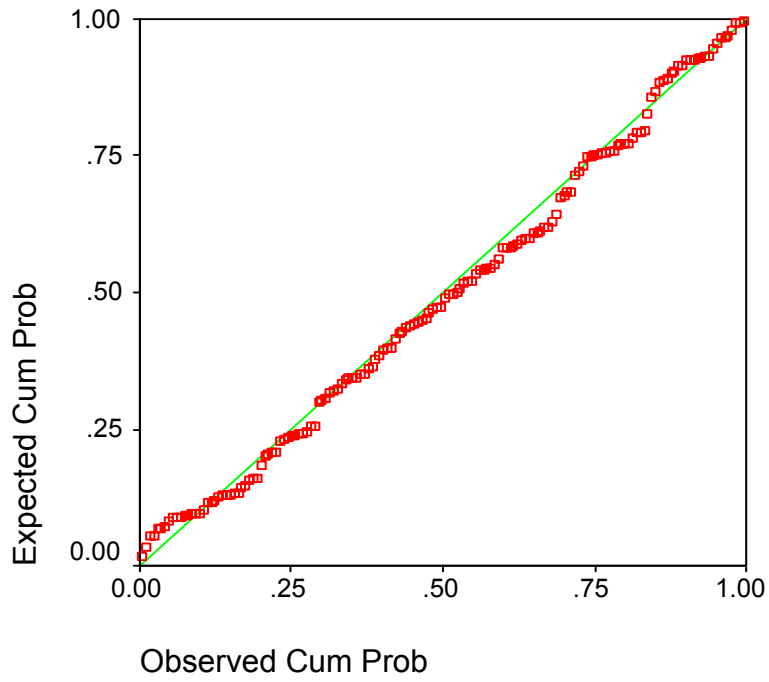


Figure 5-10. Normal probability plot of  $\log_{10} F$  of the second-order Model F3

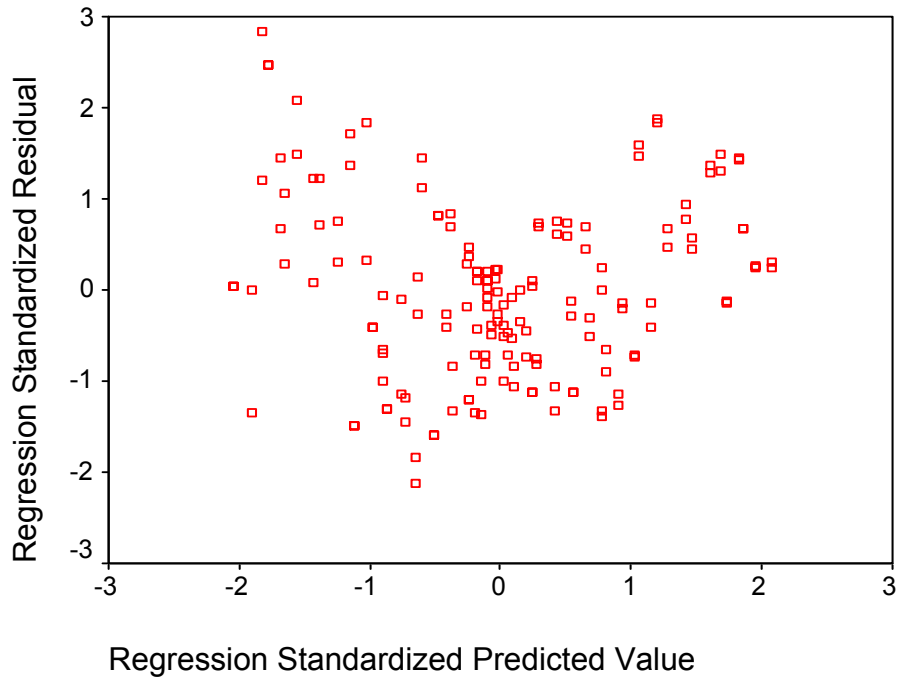


Figure 5-11. Residual vs the predicted  $\log_{10} F$  of the second-order Model F3

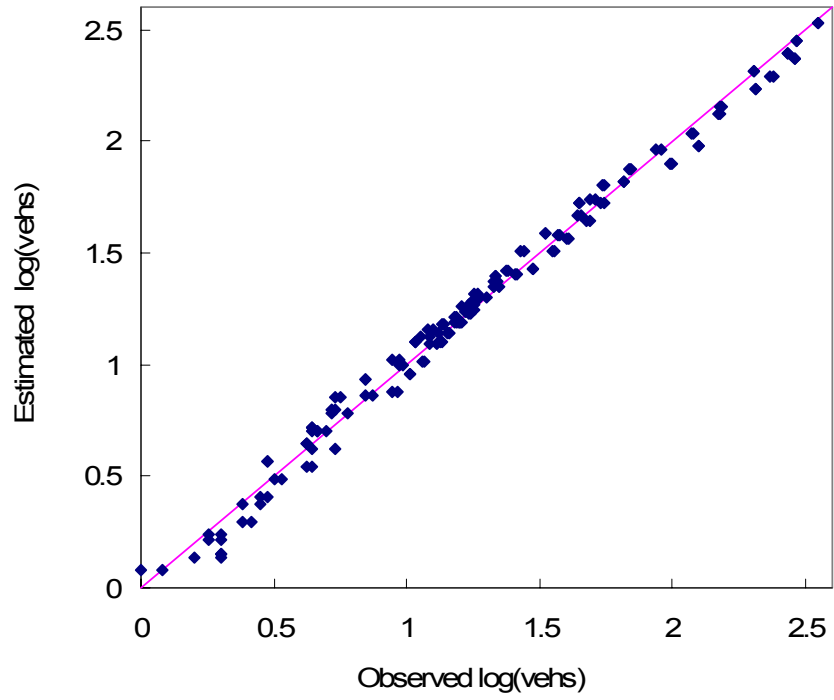


Figure 5-12. Estimated vs observed  $\log_{10} F$  of the second-order Model F3

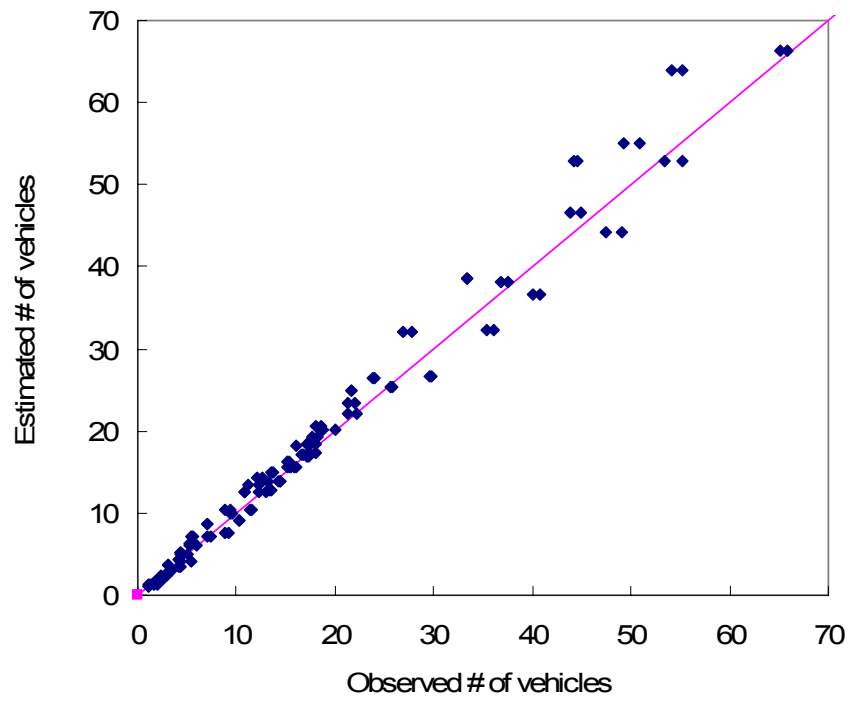


Figure 5-13. Estimated vs observed  $F$  of the second-order Model F3

The observations from the above results are as follows:

- $F$  values of all three models are significant at the  $\alpha = 0.01$  level.
- The probability plot of Model F3 falls very close to the 45-degree line, indicating strong conformity to the normality assumption. The probability plots of Models F1 and F2 deviate somewhat from the 45-degree line. However, no strong indications are observed that the normality assumption is violated.
- There is no clear pattern observed from the plot of residual against the predicted value for each model. Model F2 shows a slightly abnormal pattern with a few points clustered.
- According to the plot of the estimated vs observed values, Models F1 and F3 predict better than Model F2. Models F1 and F3 are comparable. However, Model F1 is preferred to Model F3 because of its relatively simple form and few parameters.

### **5.3.2 Time deviation model**

The time deviation model predicts the average passenger time deviation from their desired pickup or delivery time. Since the maximum time deviation is imposed as a hard constraint in the routing and scheduling algorithm, the output average passenger time deviation is expected to be mostly related with the maximum time deviation set as an operating policy. The experiment results also indicate that the average passenger time deviation is linearly related to the maximum time deviation. Other factors such as demand density and area size have been identified to contribute to the response through regression analysis considering all the six factors, however, their contributions are far less

important than the maximum time deviation. The comparisons of the models considering different combinations of these three factors are shown in Equations (5-14) through (5-16) and Figures 5-14 through 5-22. All three models are first-order polynomial models since regression analysis shows that the first-order polynomial models fit the data well.

(1) Model D1

$$\bar{T}_{dev} = -1.60 + 0.00564A + 0.0942D + 0.50W \quad (5-14)$$

(0.118) (0.001) (0.010) (0.004)

adjusted  $R^2 = 0.988$ ,  $F = 4,200$

(2) Model D2

$$\bar{T}_{dev} = -1.37 + 0.0957D + 0.50W \quad (5-15)$$

(0.115) (0.010) (0.005)

adjusted  $R^2 = 0.986$ ,  $F = 5,564$

(3) Model D3

$$\bar{T}_{dev} = -0.90 + 0.50W \quad (5-16)$$

(0.127) (0.006)

adjusted  $R^2 = 0.979$ ,  $F = 7,189$

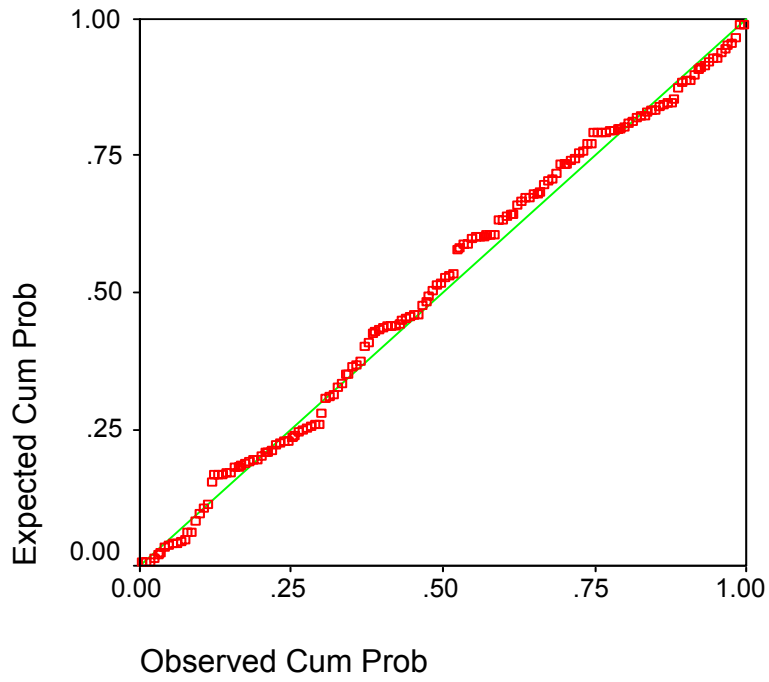


Figure 5-14. Normal probability plot of time deviation of Model D1

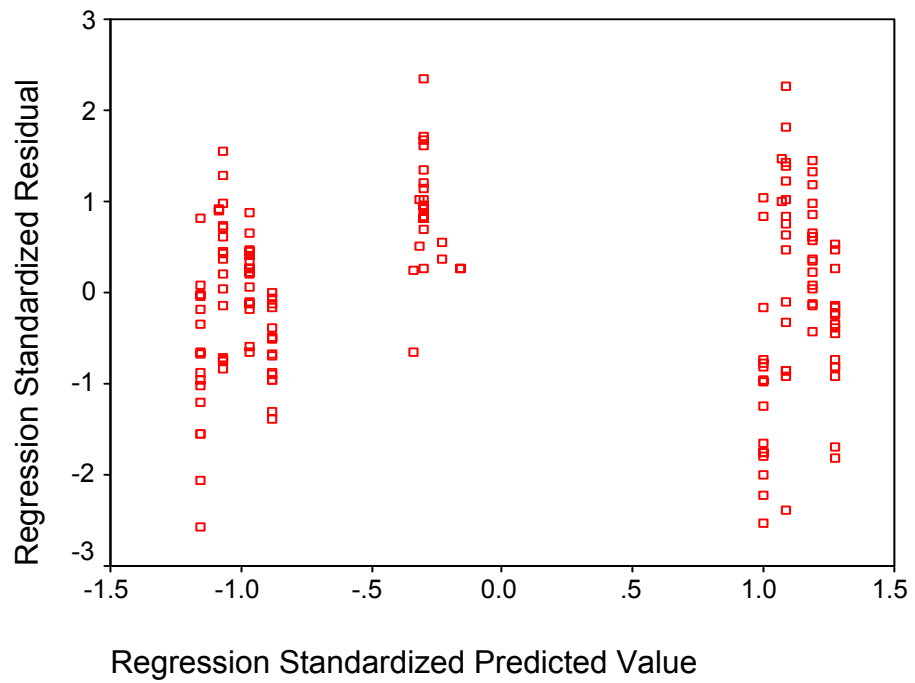


Figure 5-15. Residual vs the predicted time deviation of Model D1

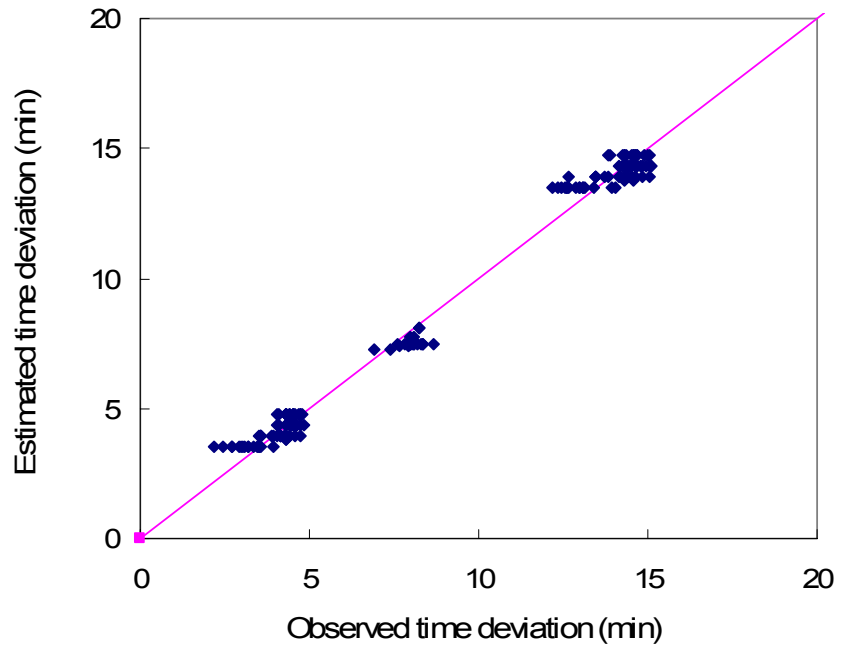


Figure 5-16. Estimated vs observed time deviation of Model D1

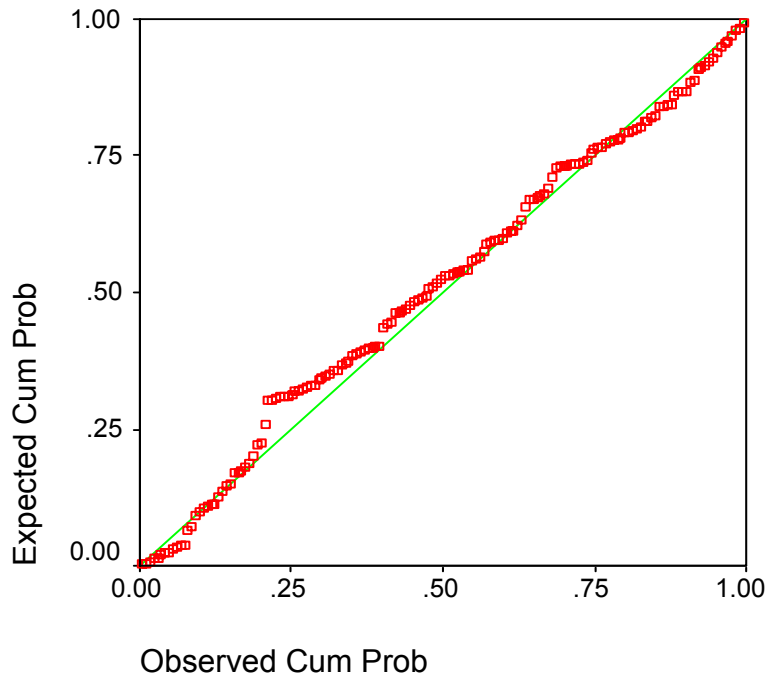


Figure 5-17. Normal probability plot of time deviation of Model D2

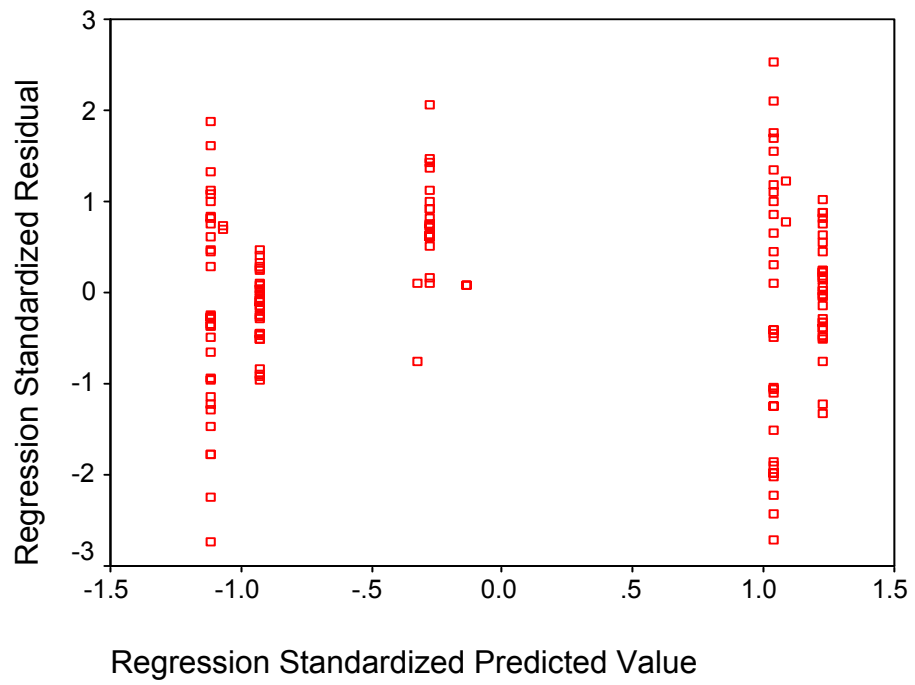


Figure 5-18. Residual vs the predicted time deviation of Model D2

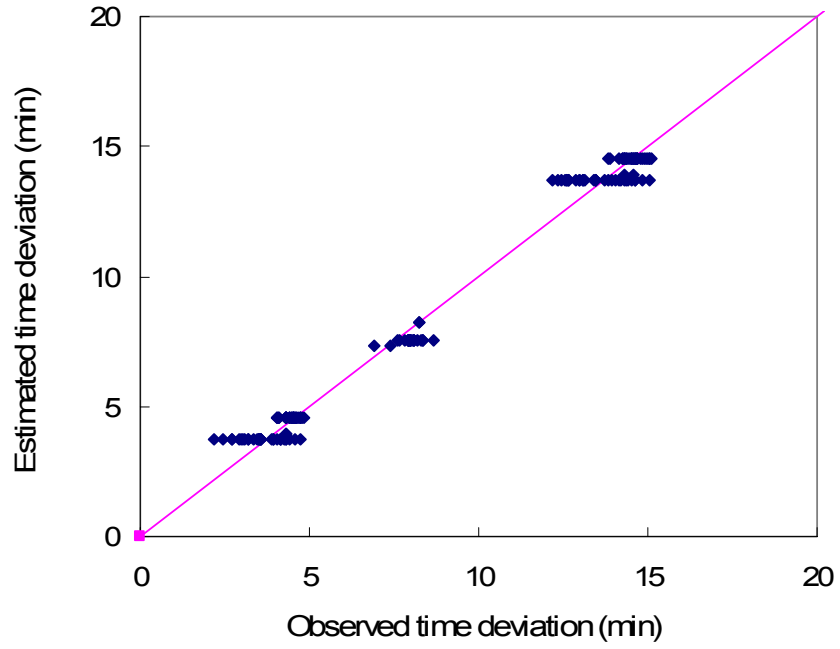


Figure 5-19. Estimated vs observed time deviation of Model D2



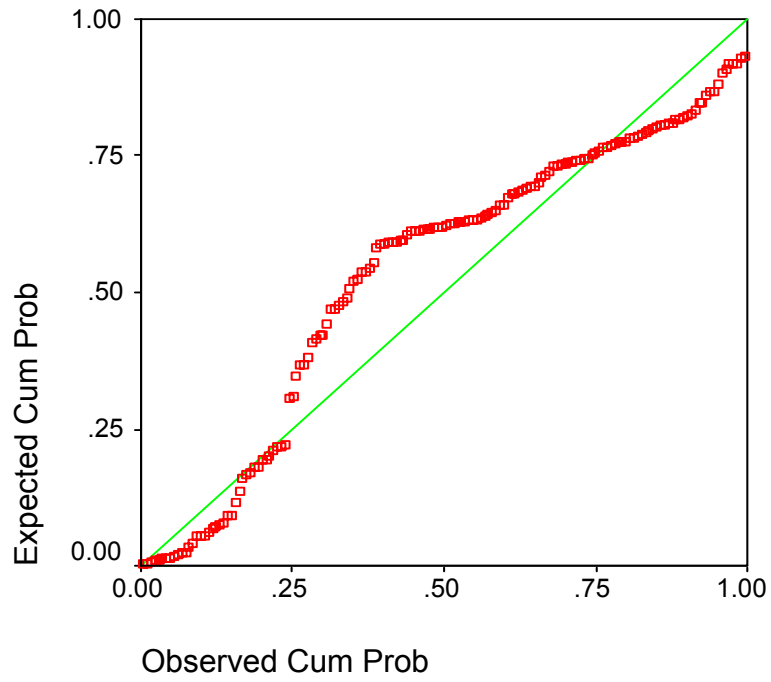


Figure 5-20. Normal probability plot of time deviation of Model D3

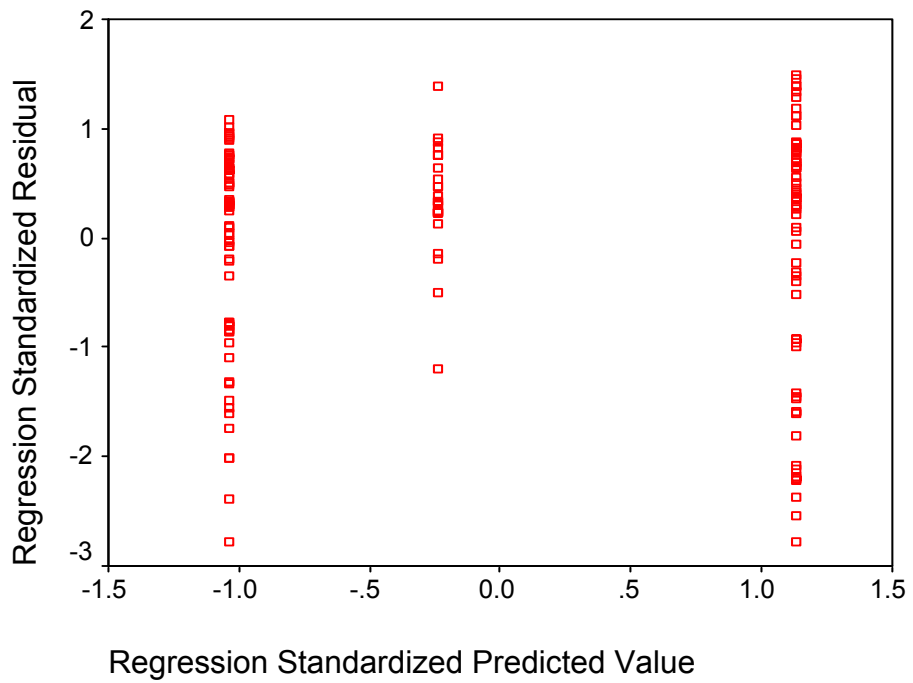


Figure 5-21. Residual vs the predicted time deviation of Model D3

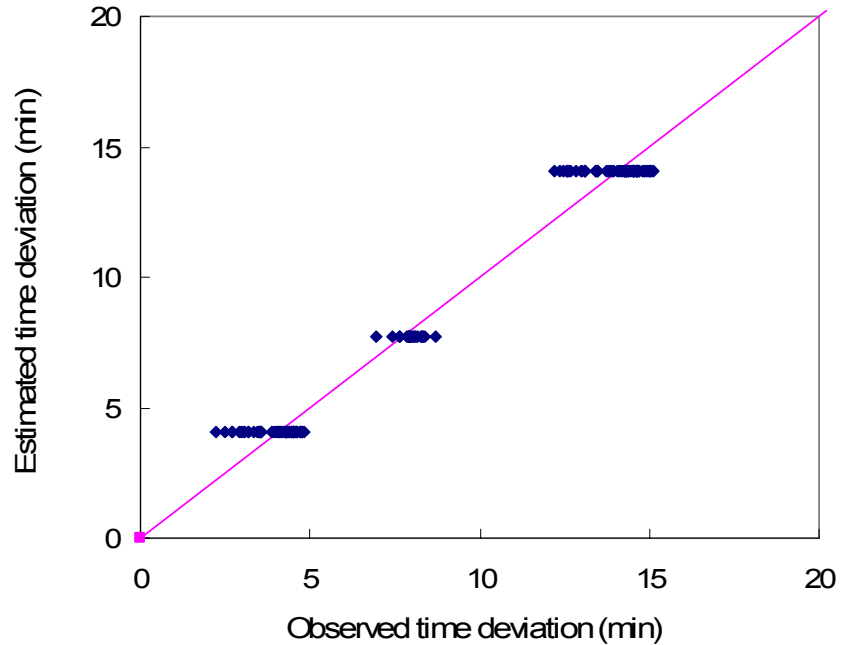


Figure 5-22. Estimated vs observed time deviation of Model D3

Note that in Figure 5-22, the points fall within three horizontal clusters. This occurs because the maximum time deviation is the only factor contributing to the Model D3 (Equation 5-16) and we only use three values for each factor in the simulation experiments.

- $F$  values of all three models are significant at the  $\alpha = 0.01$  level.
- The adjusted  $R^2$  values are close for all three models.
- The probability plot of Models D1 and D2 falls very close to the 45-degree line, indicating strong conformity to the normality assumption. The probability plots of Model D3 deviate somewhat from the 45-degree line. However, no strong indications are observed that the normality assumption is violated.

- The regression standardized residual skews somewhat to the upper part of the figure when the regression standardized predicted value falls within -0.5 to 0.0. (Figures 5-15, 5-18 and 5-21), which indicates that adding second-order terms might improve the model fit. However, a first-order model might still be preferred because of its simplicity. Its prediction accuracy might be sufficient for a particular planning purpose.
- From the plot of the estimated vs observed values, all three models fit the data well. Model D3 is the simplest model with only one factor.
- Average time deviation is a little less than half of the maximum time deviation. This is expected for a tightly constrained DARP with the restriction of no vehicle idling when carrying passengers. The time deviation for each passenger can range from 0 to the maximum limit, thus the average is approximately the half.

### **5.3.3 Ride time ratio model**

The ride time ratio model predicts the average passenger ride time ratio, which is the actual ride time divided by the direct ride time. It is expected that the output average passenger ride time ratio is mostly related with the maximum ride time ratio, which is imposed as a hard constraint in the routing and scheduling algorithm for the DARP. First-order models, second-order models and multiplicative models are all fitted considering all the six factors. The experiment results indicate that the average ride time ratio is mostly related with maximum ride time ratio, demand density and area size, in the order of importance. The most promising models identified during regression analysis are shown in Equations (5-17) to (5-20). The corresponding normal probability plot, residual plot

and the plot of observed versus predicted response value are shown in Figures 5-23 to 5-34.

(1) First-order Model R1

$$\bar{R} = 0.428 + 0.00120A + 0.0120D + 0.424R \quad (5-17)$$

(0.028) (0.0001) (0.001) (0.013)

$$\text{adjusted } R^2 = 0.890, F = 420$$

(2) First-order Model R2

$$\bar{R} = 0.532 + 0.429R \quad (5-18)$$

(0.035) (0.017)

$$\text{adjusted } R^2 = 0.802, F = 630$$

(3) Second-order Model R3

$$\bar{R} = 0.336 + 0.00136A + 0.0783D - 0.00589D^2 + 0.426R \quad (5-19)$$

(0.029) (0.0001) (0.011) (0.001) (0.012)

$$\text{adjusted } R^2 = 0.910, F = 395$$

(4) Multiplicative Model R4

$$\log_{10} \bar{R} = -0.106 + 0.0342 \log_{10} A + 0.0374 \log_{10} D + 0.605 \log_{10} R \quad (5-20)$$

(0.007) (0.004) (0.003) (0.015)

$$\text{adjusted } R^2 = 0.921, F = 606$$

Equation (5-20) can be transformed back to the multiplicative form as

$$\bar{R} = 10^{-0.106} \cdot A^{0.0342} \cdot D^{0.03737} \cdot R^{0.605} \quad (5-21)$$

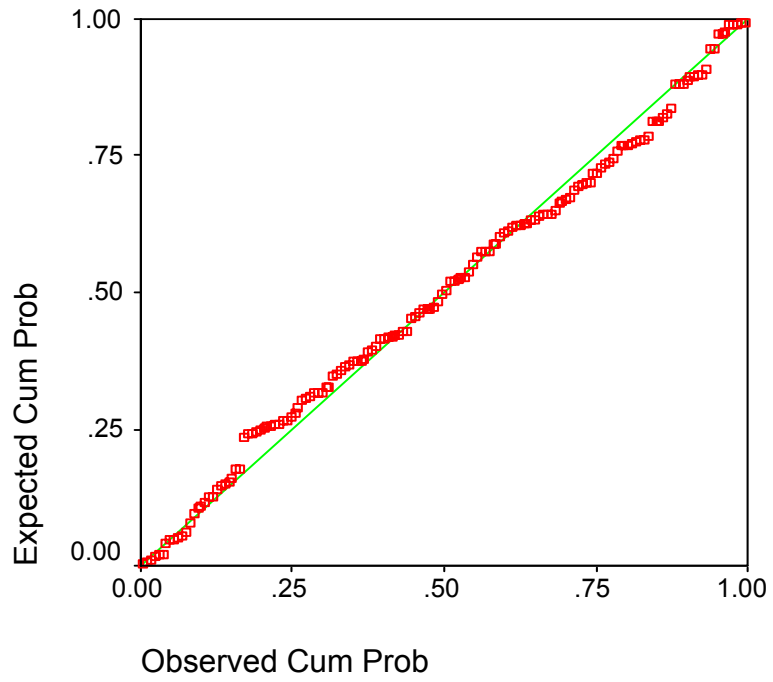


Figure 5-23. Normal probability plot of ride time ratio of the first-order Model R1

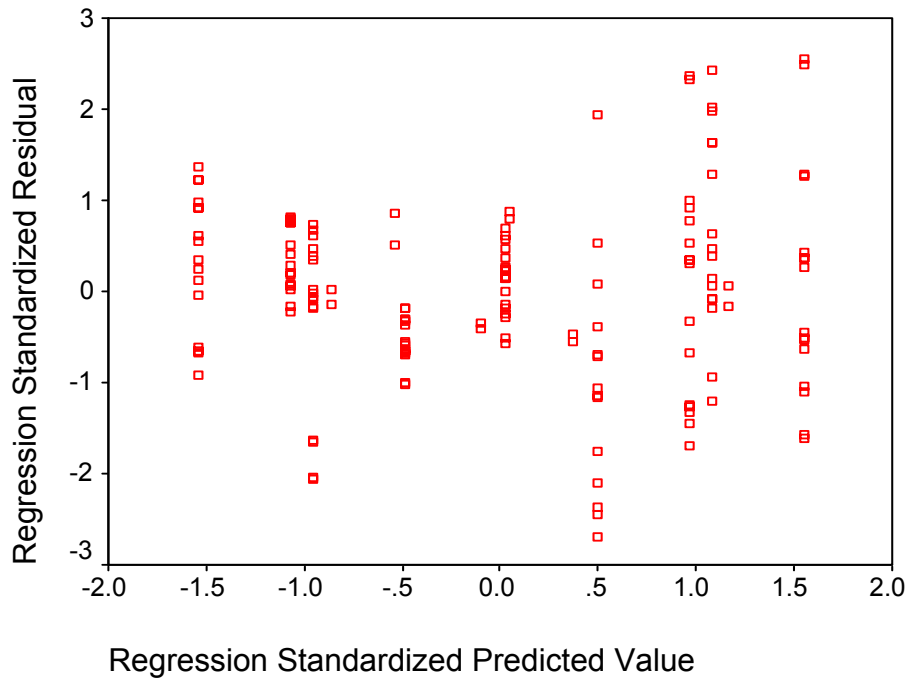


Figure 5-24. Residual vs the predicted ride time ratio of the first-order Model R1

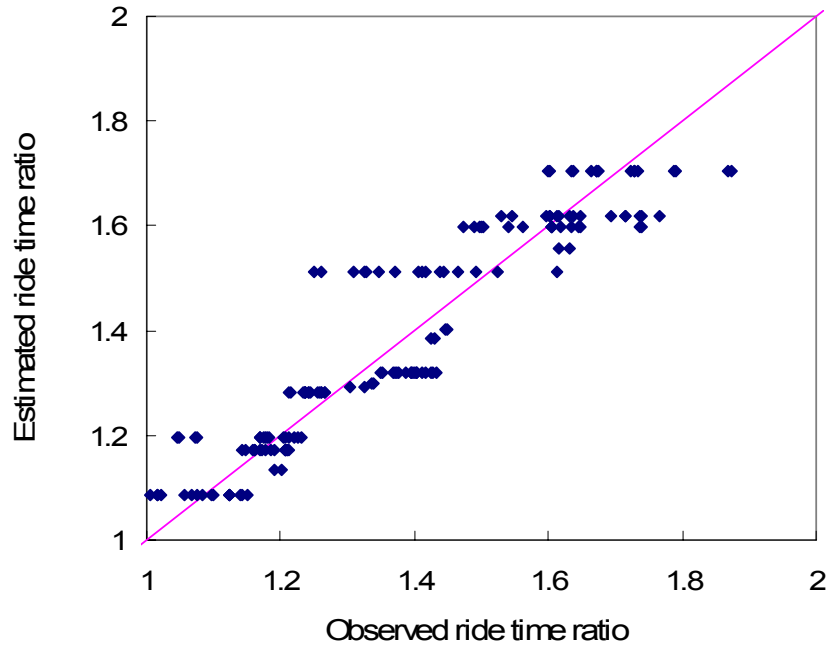


Figure 5-25. Estimated vs observed ride time ratio of the first-order Model R1

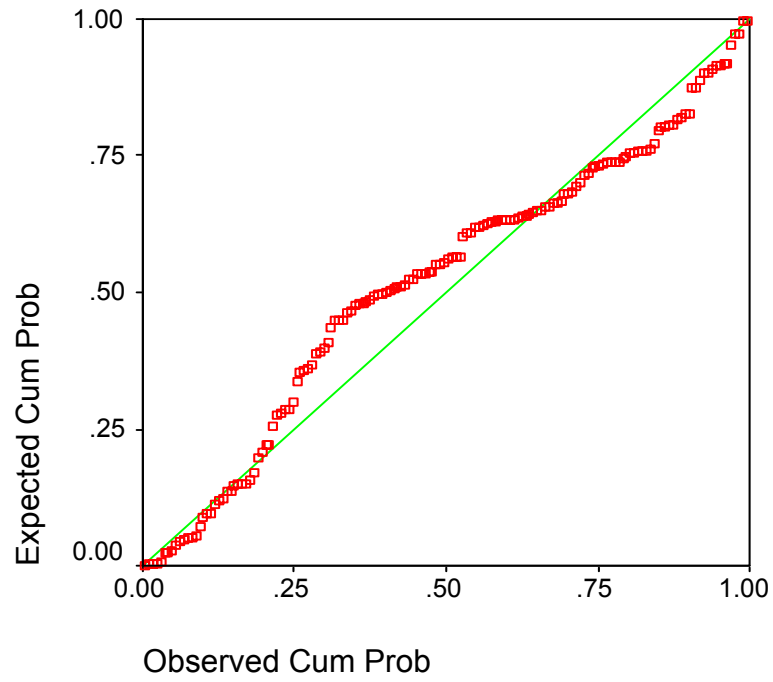


Figure 5-26. Normal probability plot of ride time ratio of the first-order Model R2

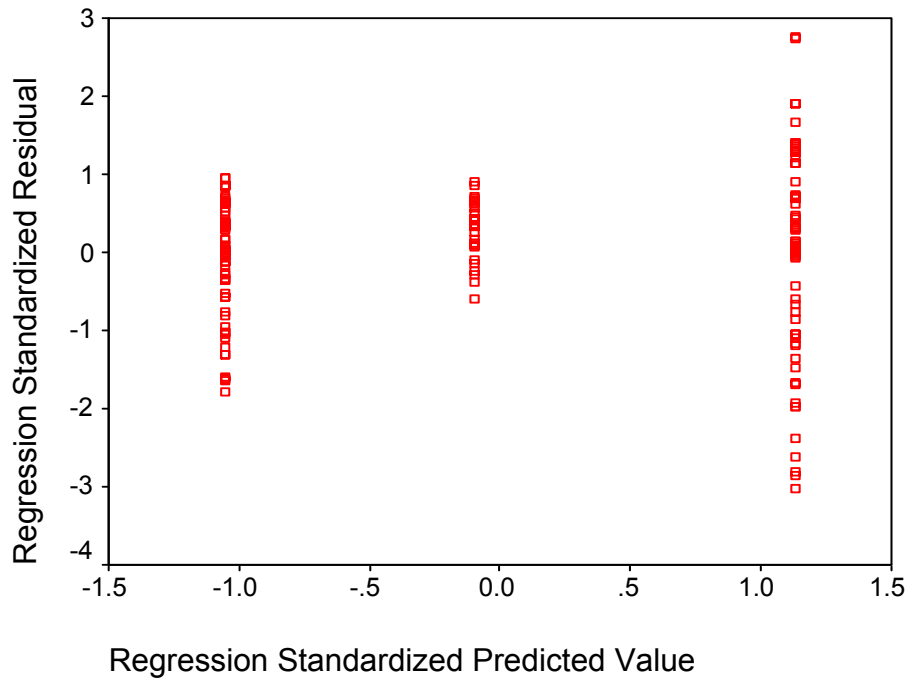


Figure 5-27. Residual vs the predicted ride time ratio of the first-order Model R2

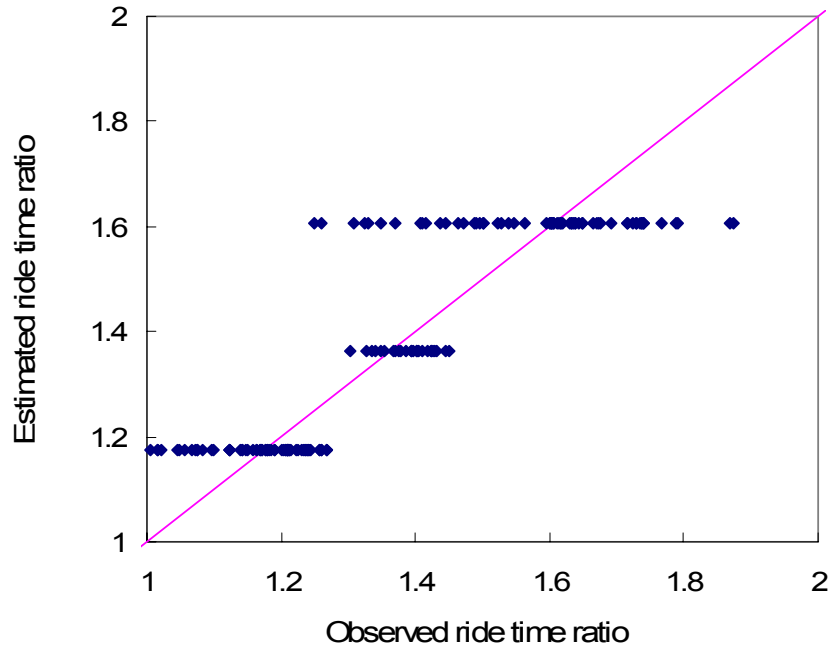


Figure 5-28. Estimated vs observed ride time ratio of the first-order Model R2



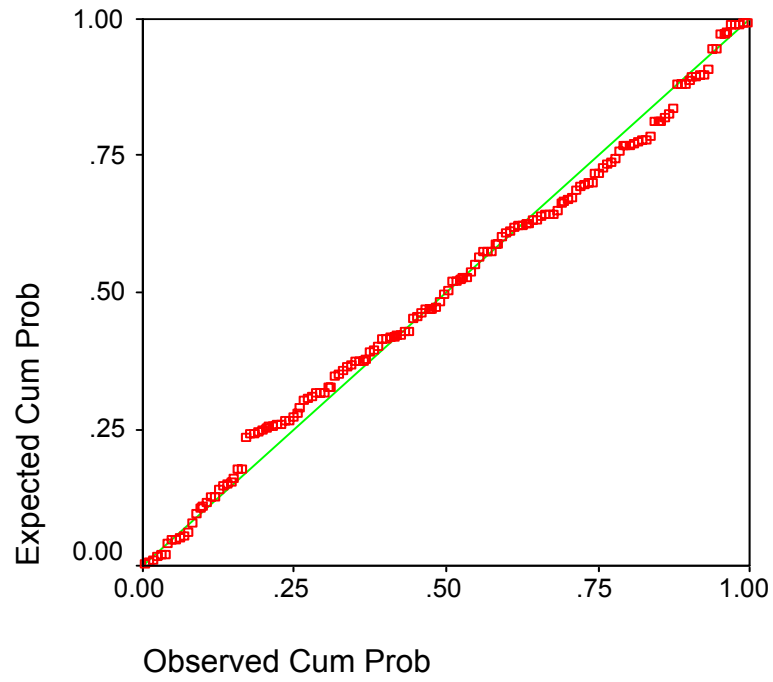


Figure 5-29. Normal probability plot of ride time ratio of the second-order Model R3



Figure 5-30. Residual vs the predicted ride time ratio of the second-order Model R3

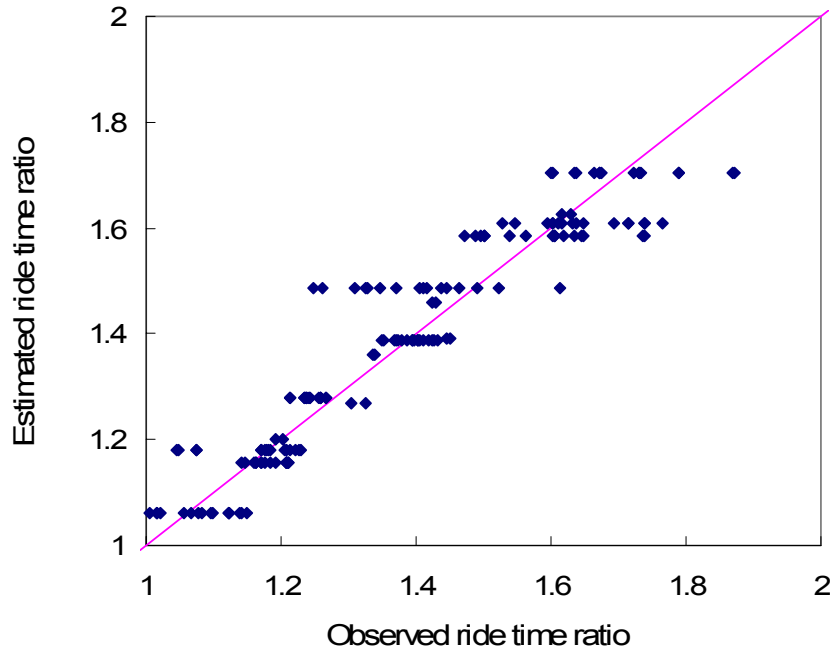


Figure 5-31. Estimated vs observed ride time ratio of the second-order Model R3

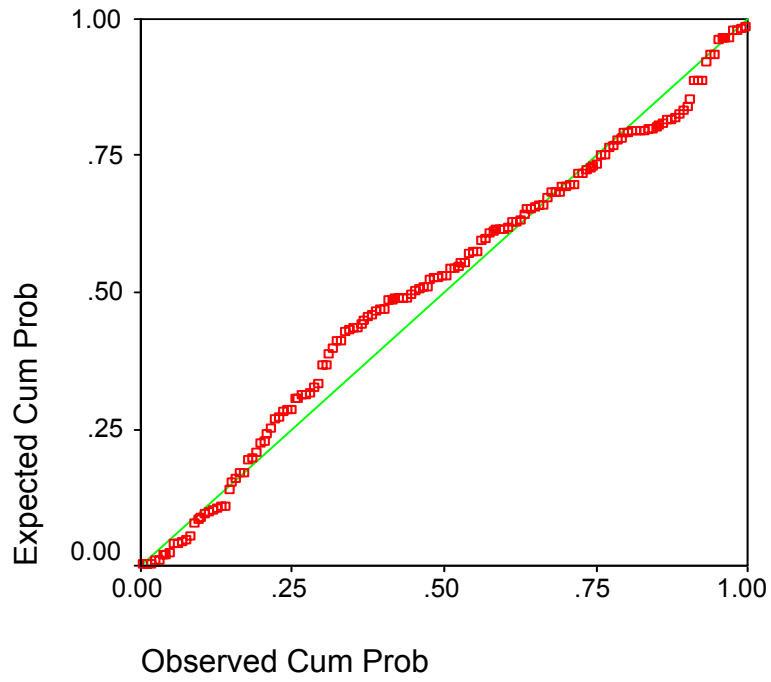


Figure 5-32. Normal probability plot of ride time ratio of the multiplicative Model R4

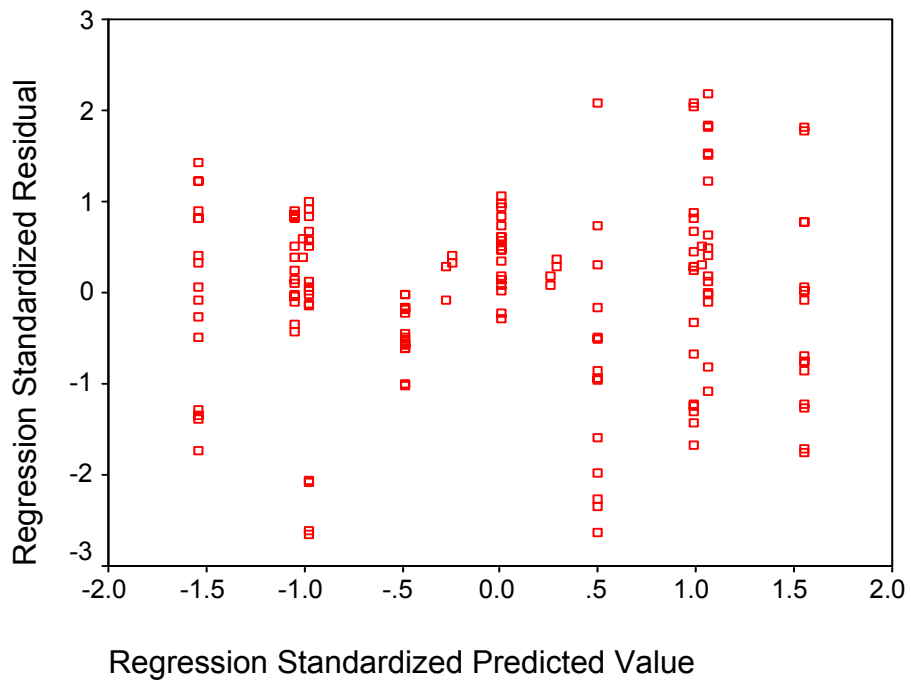


Figure 5-33. Residual vs the predicted ride time ratio of the multiplicative Model R4

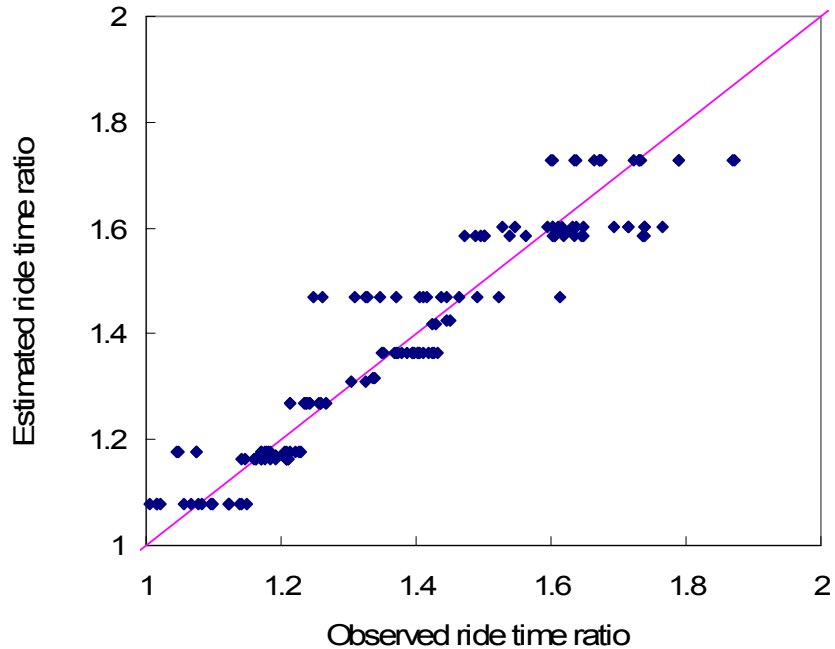


Figure 5-34. Estimated vs observed ride time ratio of the multiplicative Model R4

- $F$  values of all four models are significant at the  $\alpha = 0.01$  level.
- No strong indications are observed for any of these models that the normality assumption is violated.
- No clear pattern is observed from the plot of residual against the predicted value for Models R1, R3 and R4. The regression standardized residual is larger when the predicted value is higher for Model R2 (Figure 5-27), indicating some degree of non-constant variance of the residual.
- Comparing Models R1, R3 and R4, each involving three factors, Model R4 has the highest adjusted  $R^2$  value and the Model R3 has the second highest adjusted  $R^2$  value. However, the values for three models are close and no significant difference has been observed from the plots of estimated versus observed ride

time ratio for the three models. All the models are tested again using the new design points in the model validation (Section 5.4).

- Model R2 has the simplest form with adjusted  $R^2$  value 0.802.

#### **5.4 Metamodel Validation**

Regression analysis, used to develop the general linear metamodel, is very much a data-based technique. It finds the model with the best possible fit to the data. Models thus estimated might not perform well on new data. In this dissertation the metamodels are validated against 30 new design points other than the ones that were used to build the metamodels. The new design points are randomly generated within the region of interest, as shown in Table 5-1. More specifically, values for each factor are generated from uniform distributions bounded by their respective lower and upper values. In the face-centered composite design, most design points used to develop the metamodels are located in the “corner” or “boundary” of the design space. The metamodel validation, in some sense, tests how well the models fit the points that are more internally distributed in the design space.

### 5.4.1 Vehicle resource requirement model

Figure 5-35 shows the estimated versus predicted number of vehicles of the multiplicative Model F1

$$F = 4.79 \frac{A^{1.07} \cdot D^{0.72} \cdot b^{0.21}}{W^{0.29} \cdot R^{0.37} \cdot V^{0.68}} \quad (5-11)$$

The plots falls very close to the 45-degree line, indicating that the above model fits the new randomly generated design points very well.

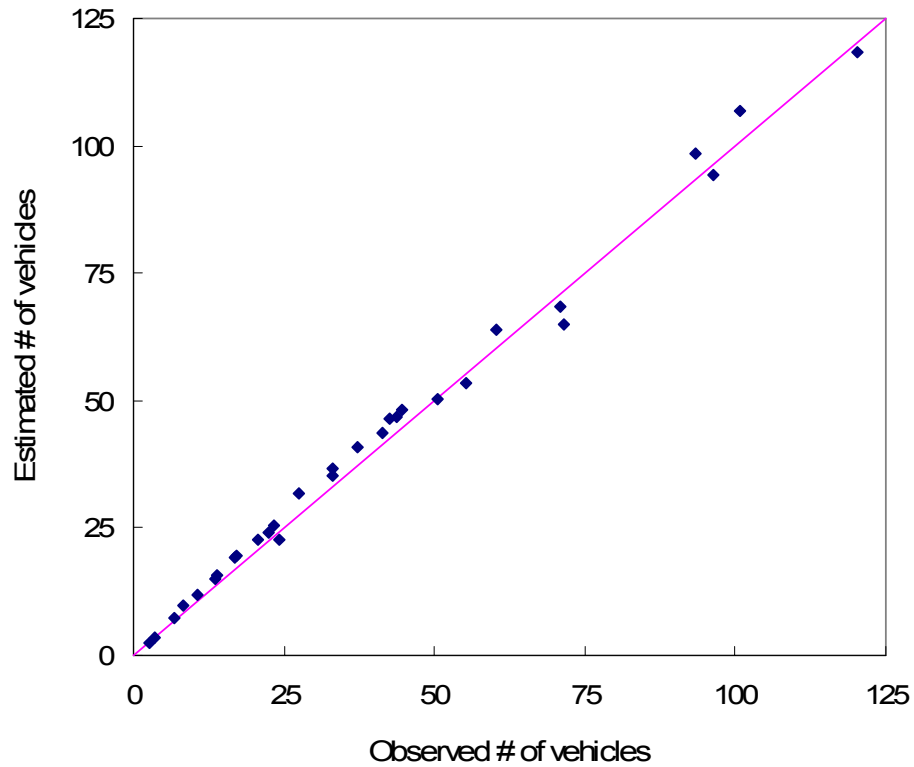


Figure 5-35. Model validation: Estimated vs observed vehicles of the multiplicative Model F1

### 5.4.2 Time deviation model

Figure 5-36 shows the estimated versus predicted average passenger time deviation of Model D3

$$\bar{T}_{dev} = -0.90 + 0.50W \quad (5-16)$$

The plots fall very close to the 45-degree line. However, most of the points fall on the lower side of the line, indicating that the time deviations are slightly underestimated (-4.9% on average) by the metamodel. This underestimation maybe due to the omission of the possible second-order terms in the model. However, the model accuracy should be acceptable for planning purposes.

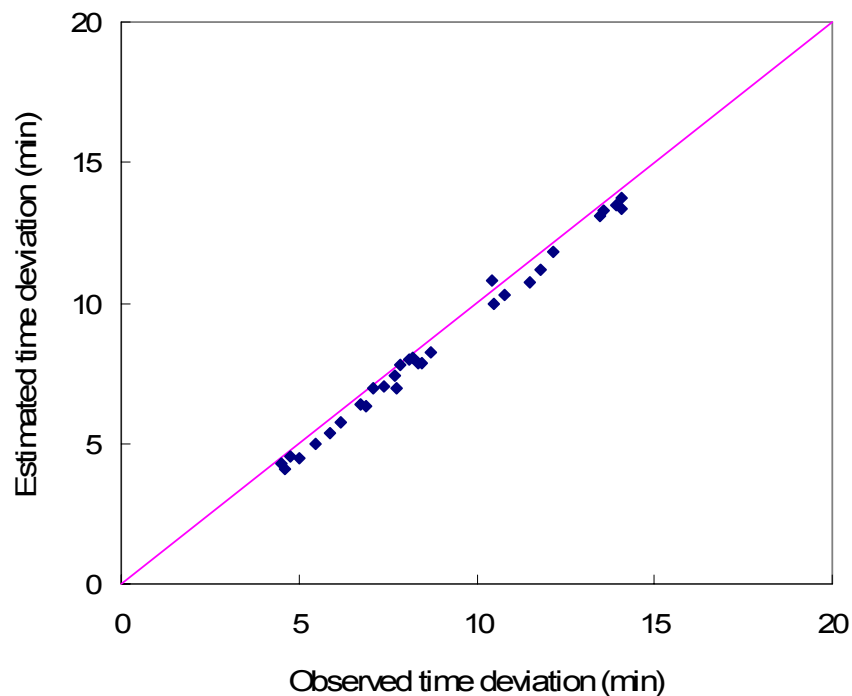


Figure 5-36. Model validation: Estimated vs observed time deviation of Model D3

### 5.4.3 Ride time ratio model

In Section 5.3.3, four metamodels are developed and compared. They are the first-order Model R1 (Equation 5-17), the first-order Model R2 (Equation 5-18), the second-order Model R3 (Equation 5-19), and the multiplicative Model R4 (Equation 5-21).

$$\bar{R} = 0.428 + 0.00120A + 0.0120D + 0.424R \quad (5-17)$$

$$\bar{R} = 0.532 + 0.429R \quad (5-18)$$

$$\bar{R} = 0.336 + 0.00136A + 0.0783D - 0.00589D^2 + 0.426R \quad (5-19)$$

$$\bar{R} = 10^{-0.106} \cdot A^{0.0342} \cdot D^{0.03737} \cdot R^{0.605} \quad (5-21)$$

The statistical performance of Models R1, R3 and R4 are comparable. Model R2 has the simplest form with the little inferior statistical performance. Since there is no clear indication that which model dominates the others, the average passenger ride time ratio predicted by those four models are compared with the observed values using the new data sets. The results are shown in Figures 5-37 through 5-40.

Comparing Models R1, R3 and R4, Model R3 (the second-order model) fits the new data points best, since its data points fall around the 45-degree line while the other two underestimate the response values. The omission of the area size and demand density variables from Model R2 compare to Model R1 does not greatly deteriorate the prediction.



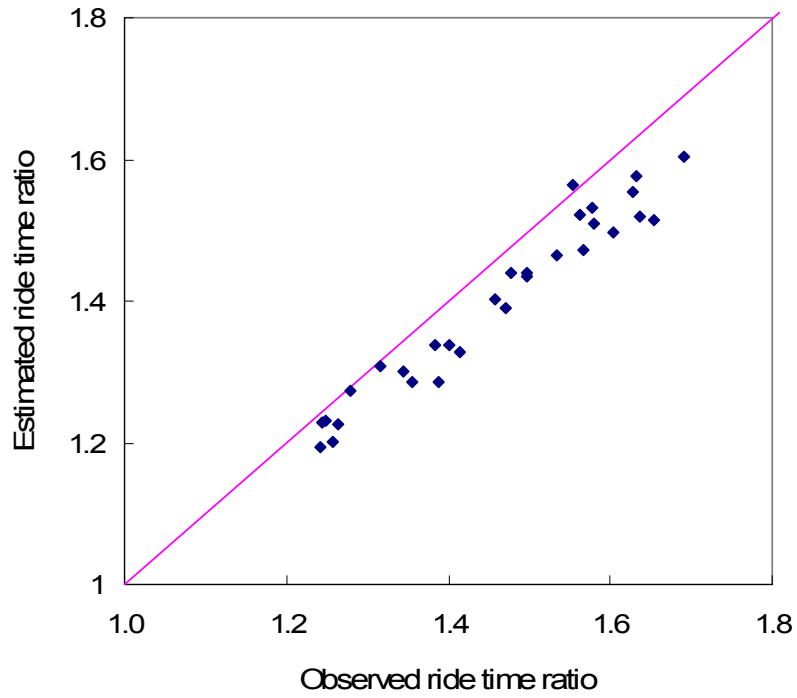


Figure 5-37. Model validation:  
Estimated vs observed ride time ratio of the first-order Model R1

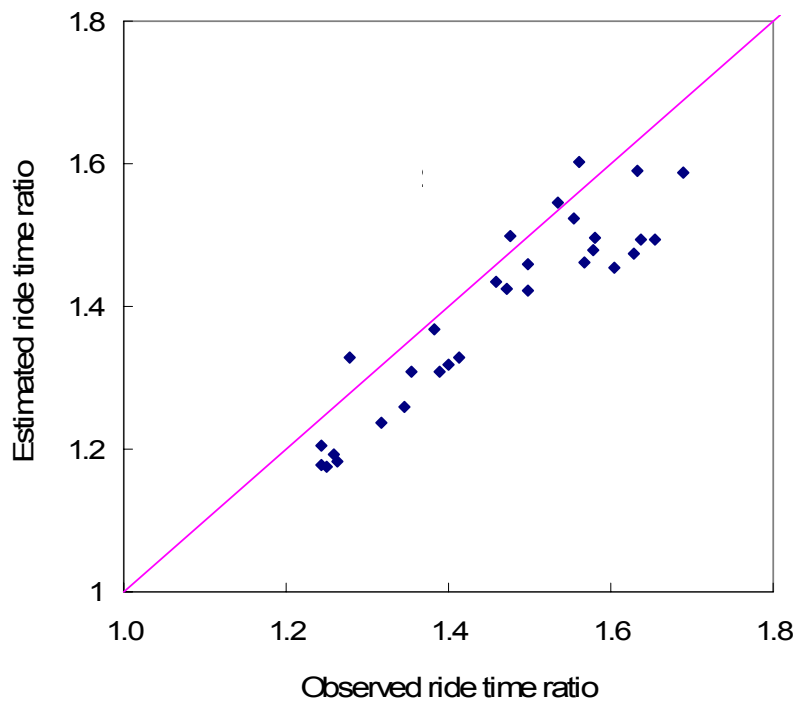


Figure 5-38. Model validation:  
Estimated vs observed ride time ratio of the first-order Model R2

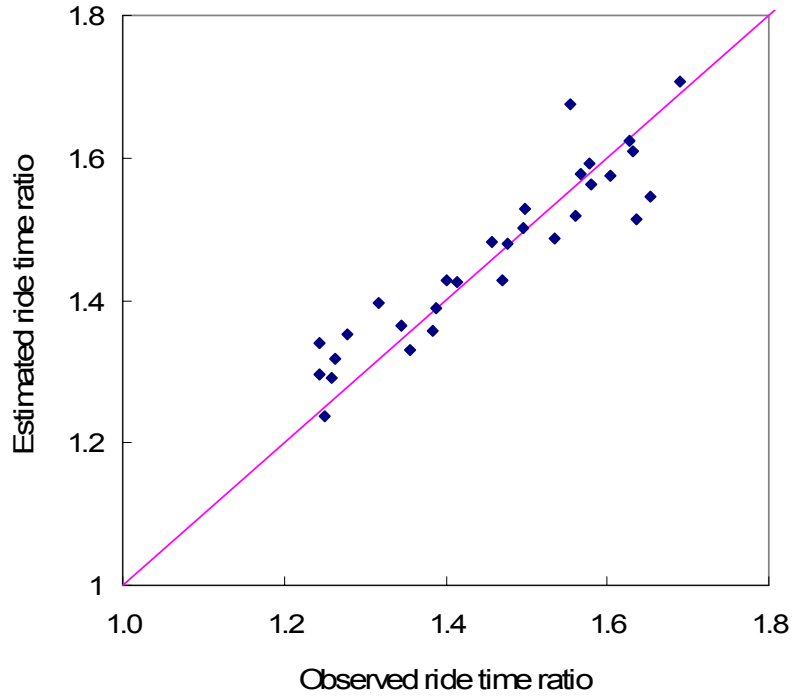


Figure 5-39. Model validation:  
 Estimated vs observed ride time ratio of the second-order Model R3

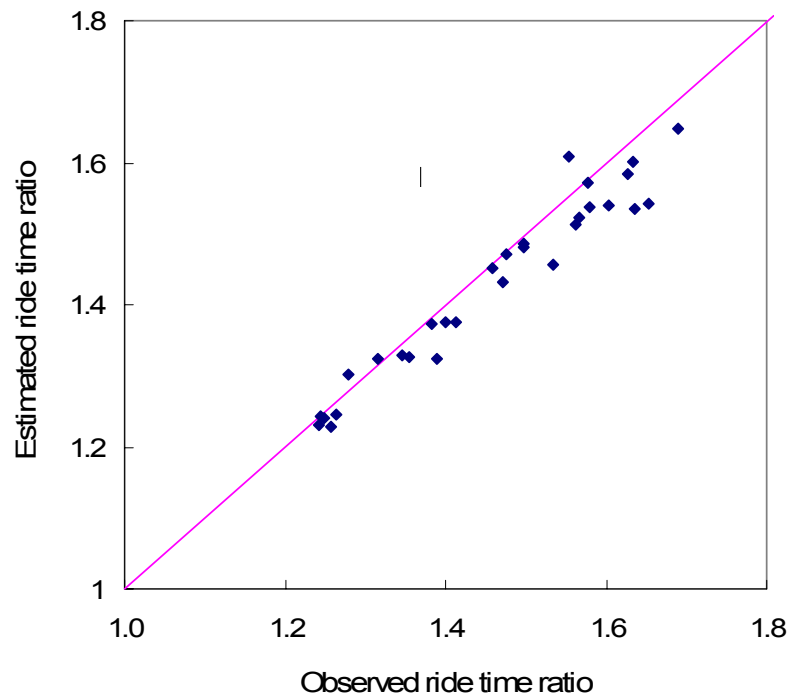


Figure 5-40. Model validation:  
 Estimated vs observed ride time ratio of the multiplicative Model R4

Table 5-3 shows the estimated and observed response values for the multiplicative vehicle requirement Model F1, the time deviation Model D3, and the ride time ratio second-order Model R3 for the 30 experiments with new design points.

Due to the randomness of the demand nature, each design point is replicated five times. From the statistical viewpoint, these repeated runs can be used to estimate the pure error variance  $\sigma^2$ . The pure error represents the error due to the random variation of the experiments such as the randomness of the demand. More information on this is provided in statistics books such as Draper and Smith (1998), and Kleinbaum et al. (1988). The estimated standard deviations of the error  $\sigma$  due to the random variation of the experiments for Models F1, D3, and R3 shown in Table 5-3 are 1.61 vehicles, 0.42 minutes, and 0.018, respectively.

Table 5-3. Observed vs estimated values from the performance models for the 30 validation experiments

Experiment	Number of vehicles			Average time deviation			Average ride time ratio		
	Observed	Model F1	Difference	Observed	Model D3	Difference	Observed	Model R3	Difference
1	16.8	19.3	14.8%	7.7	7.0	-9.8%	1.50	1.50	0.3%
2	20.8	22.6	8.7%	8.7	8.3	-4.9%	1.38	1.36	-1.9%
3	50.6	50.3	-0.6%	10.5	9.9	-5.1%	1.26	1.32	4.3%
4	96.2	94.2	-2.1%	14.1	13.7	-2.3%	1.63	1.62	-0.2%
5	27.4	31.7	15.6%	4.7	4.6	-3.9%	1.46	1.48	1.6%
6	13.6	15.1	10.8%	11.5	10.7	-6.8%	1.64	1.51	-7.5%
7	41.4	43.7	5.5%	5.0	4.5	-10.3%	1.69	1.71	0.9%
8	43.8	46.7	6.6%	7.1	7.0	-1.9%	1.58	1.56	-1.1%
9	23.2	25.6	10.1%	6.2	5.7	-7.5%	1.63	1.61	-1.4%
10	3.6	3.5	-2.2%	10.4	10.8	3.4%	1.56	1.52	-2.7%
11	71.6	65.1	-9.1%	11.8	11.2	-4.9%	1.25	1.24	-1.0%
12	10.6	11.8	11.8%	7.8	7.8	-0.6%	1.48	1.48	0.2%
13	71	68.4	-3.6%	10.8	10.3	-4.8%	1.24	1.34	7.7%
14	8.4	9.9	17.6%	13.5	13.1	-2.9%	1.39	1.39	0.0%
15	120.4	118.3	-1.7%	6.7	6.4	-4.5%	1.55	1.67	7.8%

Table 5-3. Observed vs estimated values from the performance models for the 30 validation experiments (Cont')

Experiment	Number of vehicles			Average time deviation			Average ride time ratio		
	Observed	Model F1	Difference	Observed	Model D3	Difference	Observed	Model R3	Difference
16	2.8	2.4	-14.8%	13.9	13.5	-3.2%	1.53	1.49	-3.1%
17	100.8	106.8	6.0%	4.6	4.1	-10.3%	1.32	1.40	6.0%
18	17.2	19.6	14.0%	8.5	7.9	-7.2%	1.35	1.33	-1.8%
19	42.6	46.4	8.9%	6.9	6.3	-8.2%	1.35	1.36	1.3%
20	24.2	22.5	-6.9%	13.6	13.3	-1.9%	1.65	1.54	-6.6%
21	14	15.7	12.0%	7.4	7.0	-5.2%	1.47	1.43	-2.9%
22	55.4	53.6	-3.3%	5.9	5.4	-8.4%	1.60	1.57	-1.9%
23	33	36.6	11.1%	8.4	7.8	-6.2%	1.50	1.53	2.1%
24	37.2	40.8	9.7%	5.5	5.0	-8.8%	1.57	1.58	0.8%
25	93.4	98.5	5.5%	12.1	11.8	-2.4%	1.58	1.59	1.0%
26	6.8	7.2	6.0%	4.5	4.3	-4.6%	1.28	1.35	5.7%
27	44.6	48.0	7.7%	8.2	8.0	-2.0%	1.41	1.43	0.8%
28	33.2	35.2	5.9%	8.1	8.0	-1.6%	1.24	1.30	4.3%
29	60.4	63.9	5.8%	7.7	7.4	-3.8%	1.40	1.43	2.0%
30	22.4	24.1	7.5%	14.1	13.3	-5.1%	1.26	1.29	2.6%

## 5.5 Model Summary

Based on the results presented and discussed in Section 5.3-5.4, the following performance models are recommended:

- (1) Vehicle resource requirement Model F1

$$F = 4.79 \frac{A^{1.07} \cdot D^{0.72} \cdot b^{0.21}}{W^{0.29} \cdot R^{0.37} \cdot V^{0.68}} \quad (5-11)$$

To incorporate the effect of road circuitry  $f_c$  on the vehicle resource requirement (a 1.15 circuitry factor is used in all experiments for metamodel development), the model (Model F1a) would be

$$F = 4.79 \frac{A^{1.07} \cdot D^{0.72} \cdot b^{0.21}}{W^{0.29} \cdot R^{0.37} \cdot (1.15V / f_c)^{0.68}} \quad (5-22)$$

- (2) Time deviation Model D3

$$\bar{T}_{dev} = -0.90 + 0.50W \quad (5-16)$$

- (3) Ride time ratio model

First-order Model R2

$$\bar{R} = 0.532 + 0.429R \quad (5-18)$$

Second-order Model R3

$$\bar{R} = 0.336 + 0.00136A + 0.0783D - 0.00589D^2 + 0.426R \quad (5-19)$$

For the ride time ratio model, the first-order model is also included because of its simplicity.

Vehicle productivity can be estimated from the vehicle fleet size requirement Model F1a by dividing the hourly demand by the number of vehicles. The passenger in-vehicle travel time can be estimated by the ride time ratio estimated by Model R2 or R3 multiplied by the direct trip time, which might be estimated through demand forecasting analysis.

## **Chapter 6 Sensitivity Analysis and Model Applications**

In Chapter 5, performance metamodels have been developed using the response surface methodology. The models assume a square service area and uniformly distributed demand in the area. It is also assumed that half of the users specify desired pickup time and the remaining half specify desired delivery time. In this chapter, the effects of these assumptions on the performance are investigated. Simulation experiments are performed (1) on rectangular areas with different aspect ratios (defined as the ratio of the length to width of a rectangular area), (2) with linearly distributed demand along one side of the area representing a graduate decreasing demand density, and (3) with different percentages of users specifying desired pickup time.

Two of the model applications have been demonstrated in Section 6.2. Parametric analysis of the model results as a single parameter is varied has been performed to better understand the interrelationships of the system. Questions such as how many additional vehicles are required if the maximum time deviation decreases from 20 minutes to 10 minutes can be answered by such analysis. Tradeoffs between the service quality and vehicle resource requirement can thus be evaluated. The performance models are also



applied in the optimization of the service considering the combined operator and user costs.

## 6.1 Sensitivity Analysis

### 6.1.1 Shape of the service area

As the service area becomes more irregular or elongated in shape, the expected straight line travel distance between random points increases. Thus, one might expect that the increased average travel distance of the passengers may result in more vehicles required.

In this section, the effect of the service area shape on the number of vehicles is investigated. Other assumptions such as the uniformly distributed demand locations defined in Chapter 5 are retained. Denote the length and width of a rectangular area as  $l$  and  $w$  and let aspect ratio  $r = l/w$ . Rectangular areas with various aspect ratios from 1 to 4 are tested.

Test instances of the dynamic DARPs have been generated by Monte Carlo simulation and are similar as those described in Section 3.3.1. An 8 mile  $\times$  8 mile service area with the depot located in the center of the area is studied. The Euclidean distance metric is used with a circuitry factor of 1.3. Vehicle speed is 15 mph. The instances have 9 hours of demand with 120, 120, 160, 200, 200, 160, 160, 120, 120 requests per hour. Half of the requests are advance requests. The lead time for remaining requests is uniformly distributed as  $U \sim [60,120]$  minutes. The boarding and alighting times are not considered.

It is expected that the more the boarding and alighting times contribute to a passenger's total trip time the less the area shape will affect the performance. Each test scenario is replicated five times using different streams of random seeds. The same stream of random seeds is used for scenarios with different aspect ratios.

The expected Euclidean distance  $D$  between two randomly-chosen points uniformly distributed in a rectangular area can be obtained from Lazoff and Sherman (1994). The values of  $D$  for rectangles of constant area 1 with selected aspect ratios are listed in Table 6-1. The expected travel time can then be obtained given the vehicle speed and road circuitry. The expected direct travel times for passengers with origins and destinations uniformly distributed in the 8 mile  $\times$  8 mile area are drawn as the aspect ratio increases in Figure 6-1. The expected travel time increases approximately linearly as the aspect ratio increases from 1 to 4.

Table 6-1. Expected Euclidean distance  $D$  with different aspect ratios  
(Lazoff and Sherman, 1994)

$l/w$	$l$	$w$	$D$
1	1.0000	1.0000	0.5214
2	1.4142	0.7071	0.5691
4	2.0000	0.5000	0.7137
8	2.8284	0.3536	0.9642

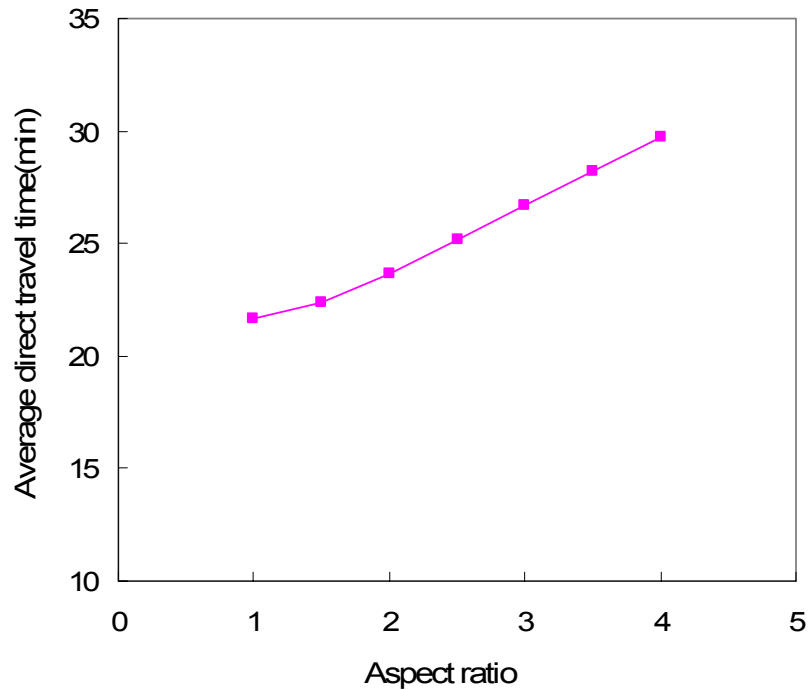


Figure 6-1. Average direct travel time vs aspect ratio

Figures 6-2 through 6-4 show how the area shape in terms of aspect ratio affects the number of vehicles required for three service quality scenarios H, M and L as the time constraints get more restrictive. The results indicate that the number of vehicles is quite insensitive to the aspect ratio of the service area for all three service quality scenarios analyzed with the rolling horizon heuristics. The fluctuation of the results obtained with the rolling horizon heuristic without the improvement procedure is actually caused by the randomness of the demand and it is observed that the heuristic with the improvement procedure can produce results with less variance. The insensitivity to the aspect ratio of the area might be explained as follows: The elongated area might ease the routing and scheduling process for the DARP and more shared rides become available. In the extreme case imagine a narrow stripe area. Most passengers must travel in the elongated

direction. Vehicles may just move in one direction and pickup or delivery passengers if the time constraints are satisfied. Since most passengers traveled in the approximate same direction, more shared rides are then available. Therefore, the increased direct travel distance with elongated area might just be offset by the increase in shared rides, which decreases sensitivity to the aspect ratio of the area.

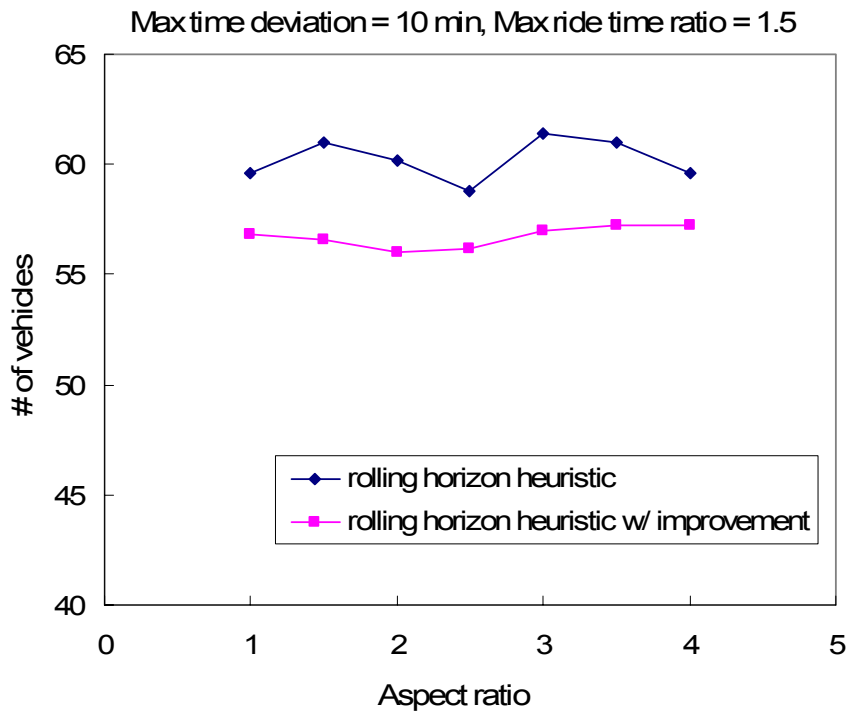


Figure 6-2. Effect of area shape on vehicles needed for service scenario H

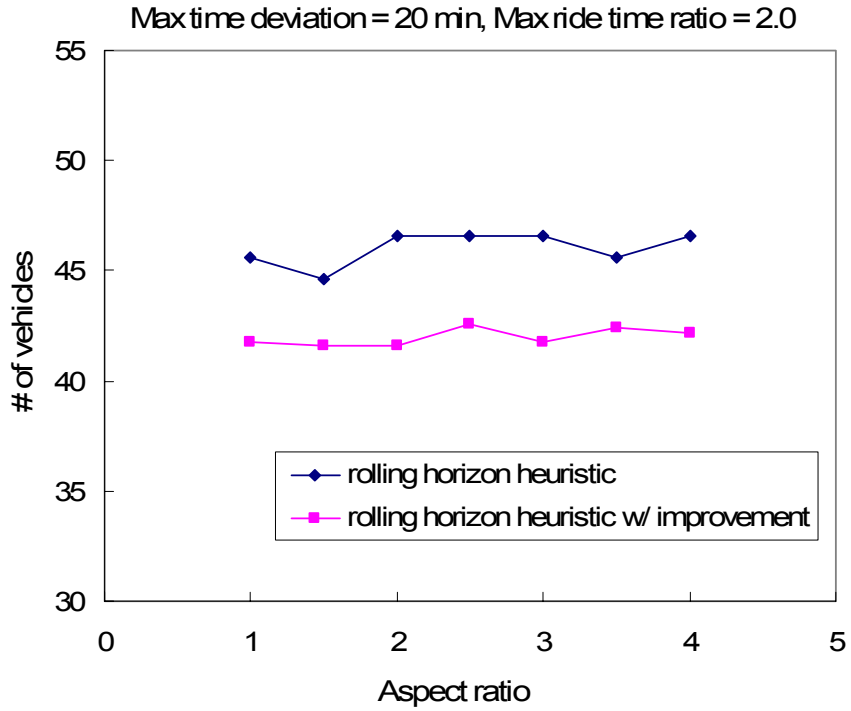


Figure 6-3. Effect of area shape on vehicles needed for service scenario M

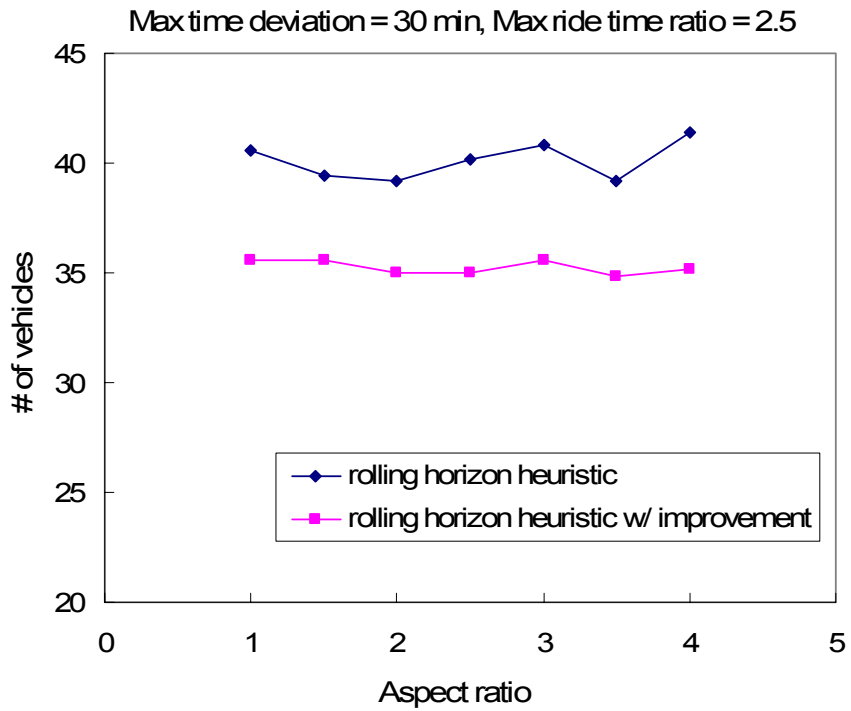


Figure 6-4. Effect of area shape on vehicles needed for service scenario L

### 6.1.2 Demand distribution in space

The metamodels developed in the last chapter rely upon the assumption of uniformly and randomly distributed demand origins and destinations over the service area. In this section, the effect of typical non-uniform demand patterns on the vehicle fleet size requirement is examined. In practice, the most usual non-uniformity of spatial demand consists of declining density as one moves away from the central city. The following experiments use a square service area in which demand density declines in one direction but is uniform in the other direction. Other operation settings are the same as those described in Section 6.1.1.

Figure 6-5 shows the probability density function for the distribution of the demand density along one side of the service area. As one moves away from 0 to  $a$ , the probability density function of the demand density decreases linearly from  $c$  to  $f \cdot c$  ( $0 \leq f \leq 1$ ).  $f = 1$  represents a special case when demand is uniformly distributed. As  $f$  decreases from 1 to 0, the slope of the demand density increases.

Values from 0 to 1 for the  $f$  are tested. Each test scenario is replicated five times using different streams of random seeds. The same stream of random seeds is used for scenarios with different  $f$  values. Figures 6-6 through 6-8 show the results. Just for clarity,  $1 - f$  instead of  $f$  is plotted as  $x$ -axis. The number of vehicles is quite constant as  $1 - f$  rises from 0 to 0.7 and decreases when  $f$  is larger than 0.7.

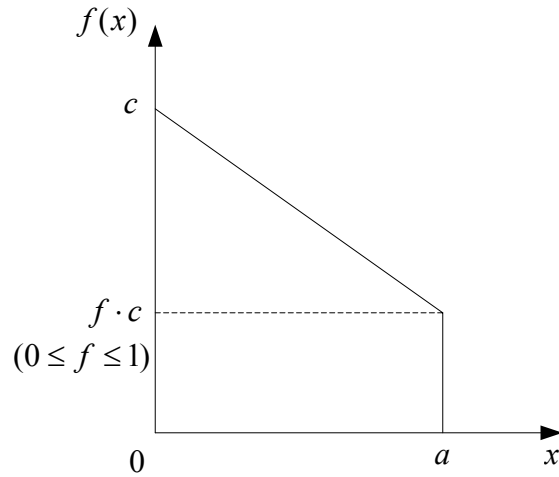


Figure 6-5. Probability density function for linear distribution of demand density along one side of the service area

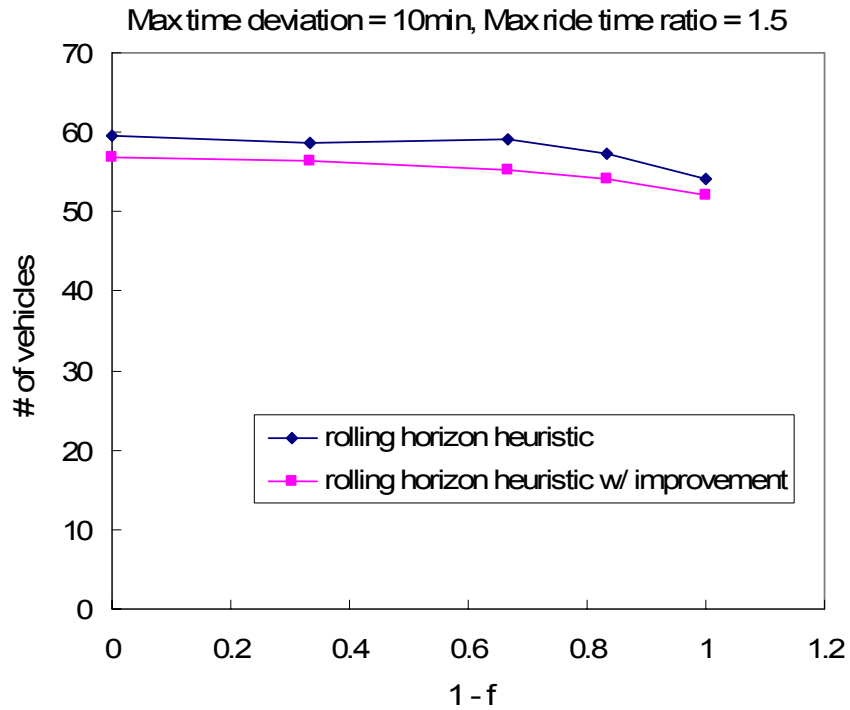


Figure 6-6. Effect of non-uniform demand distribution on vehicles needed for service scenario H

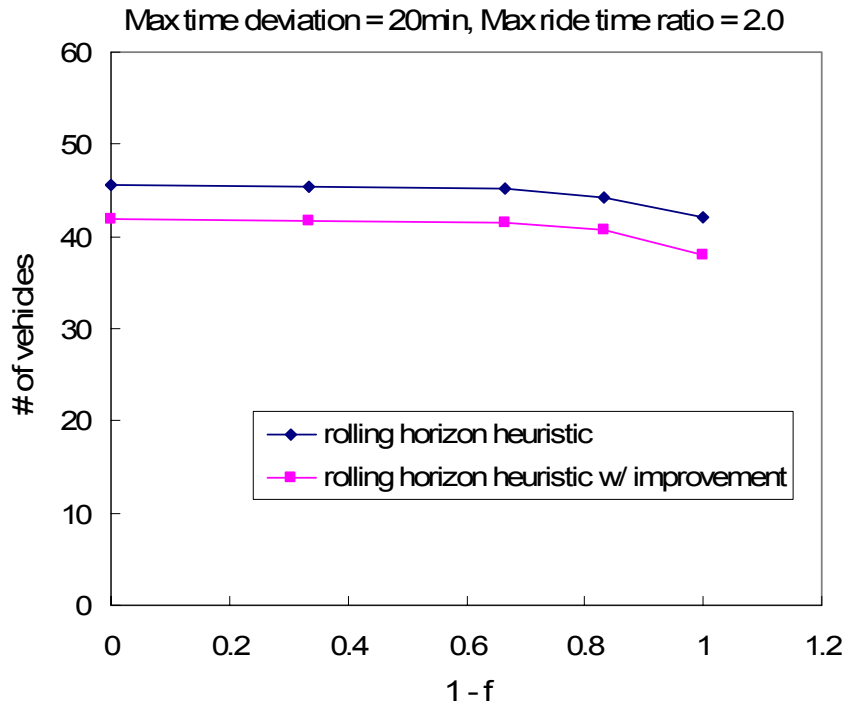


Figure 6-7. Effect of non-uniform demand distribution on vehicles needed for service scenario M

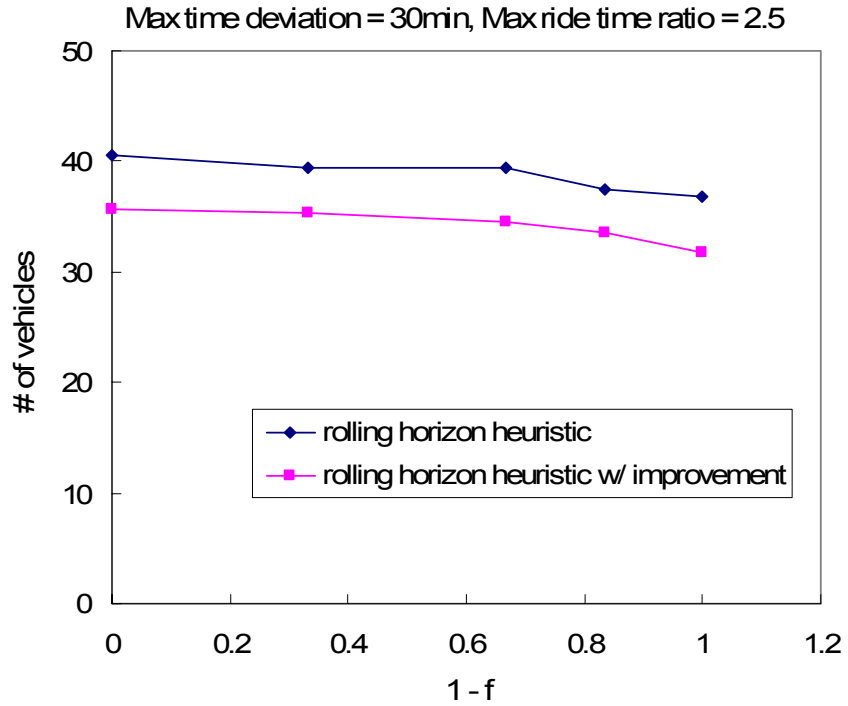


Figure 6-8. Effect of non-uniform demand distribution on vehicles needed for service scenario L



### 6.1.3 Percentage of passengers specifying desired pickup time

The metamodels developed in the last chapter assume that half of the passengers specify desired pickup time and the remaining half specify desired delivery time. In this section, the effects of that assumption on the vehicle fleet size requirement and average time deviation are examined. The percentage of passengers who specify desired pickup time ranges from 0% to 100% in the following tests. The experiments use an 8 mile  $\times$  8 mile service area and a uniform demand distribution with a demand density of 4 trips/sq. mi./hr. The service period is 3 hours. The dwell time for each pickup or delivery stop is 1 minute. The vehicle operating speed is 20 mph and the circuitry factor is 1.15. It is assumed half of the trips are requested in advance and the lead time distribution for the remaining half is uniformly distributed as  $U \sim [0, 120]$  minutes.

Values from 0% to 100% for the percentage of passengers specifying desired pickup time are tested. Each test scenario is replicated five times using different streams of random seeds. The same stream of random seeds is used for scenarios with different  $f$  values. Figure 6-9 shows the results for the number of vehicles required and Figure 6-10 shows the results for the average passenger time deviation.

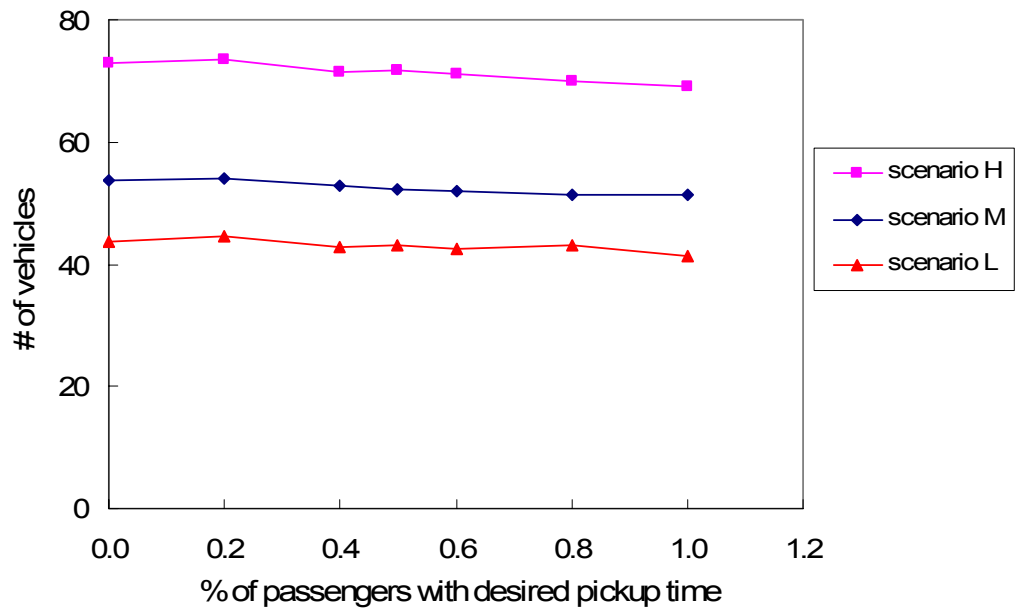


Figure 6-9. Effect of percentage of passengers with desired pickup time on vehicles required

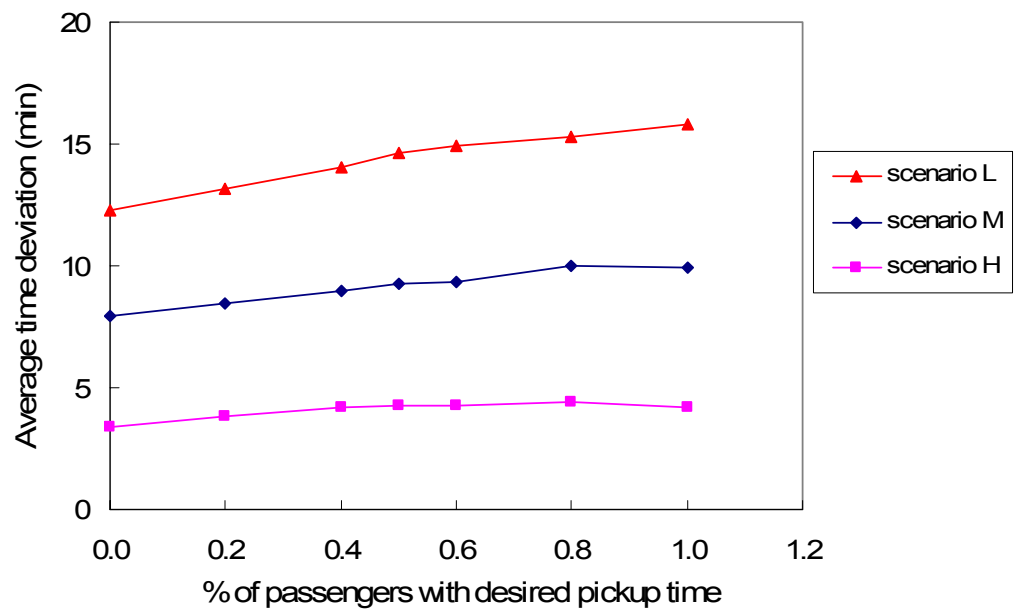


Figure 6-10. Effect of percentage of passengers with desired pickup time on average time deviation

Figure 6-9 shows that the number of vehicles required is not sensitive to the percentage of passengers specifying desired pickup time. For all three service quality scenarios, the number of vehicles decreases very slightly. Figure 6-10 shows that the average time deviation tends to be slightly lower at higher percentages of passengers specifying desired delivery time. In practice, the percentage would most probably be around 0.3 ~ 0.7. Therefore, the average time deviation is not very sensitive to the percentage of passengers specifying desired pickup time within the practical range.

The slightly lower average time deviation when most requests specify a desired delivery time is due to the way the passengers are scheduled. In the experiments, passengers are scheduled to minimize the time deviations. For a system with most trips having a desired delivery time, the service time is scheduled as late as possible to minimize the time deviation to the desired delivery time. In this way, some of the flexibility is lost by postponing the service schedules in a dynamic context. A slightly larger fleet size is required with more idling time left within the schedules, which results in slightly lower time deviation. Therefore, in a dynamic context, the ASAP scheduling policy might be preferred over the policy minimizing the time deviation if most requests specify a desired delivery time.

The results in Section 6.1 indicate that the performance metamodels are fairly robust, in that deviation from the assumptions of square service area, uniform demand distribution and 50% desired pickup-specified passengers would not greatly affect the accuracy of the predictions.



- Vehicle speed                                      20 mph
- Boarding and alighting time                2 min
- Road circuitry factor                            1.2

(1) Demand density

Figure 6-11 shows the number of vehicles required with varying demand density and other parameters fixed at default values. The number of vehicles increases with the demand density, at a decreasing rate. This implies that as the demand density increases, the opportunity for shared rides increases and fewer vehicles are required for additional trips. However, that saving is limited due to the difficulty of combining the trips with different time and geographical constraints.

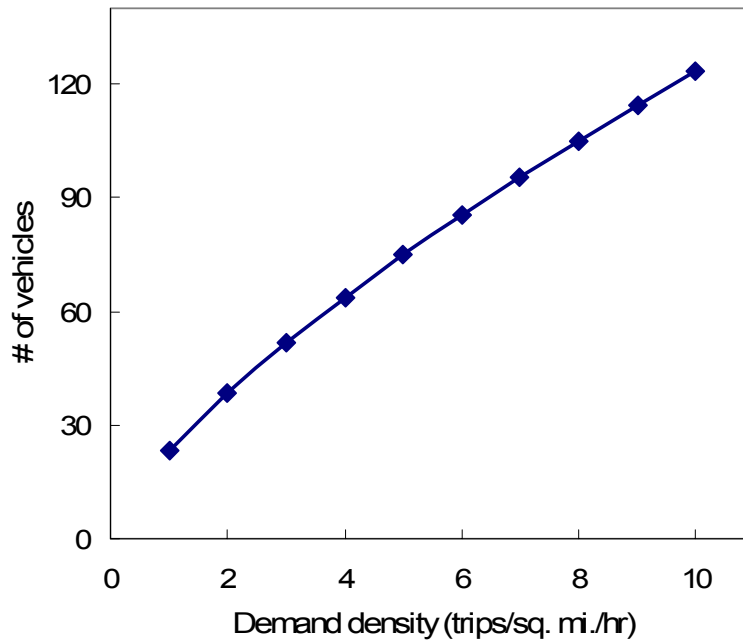


Figure 6-11. Effect of service demand density on vehicles required

Figure 6-12 shows vehicle productivity instead of the number of vehicles required by dividing the number of trips per hour by the number of vehicles for an area of 64 square miles. Vehicle productivity increases with increasing demand density, at a decreasing rate. Taxis show a relatively constant vehicle productivity since they usually can only carry one passenger party at any time. Conventional fixed-route buses, conversely, are well suited to take advantage of the economies of scale and their vehicle productivity would continue rising with increasing demand level. The results suggest that the DAR would be more suitable for a service area with low demand density. Once the demand density reaches a certain level, the conventional bus is more productive.

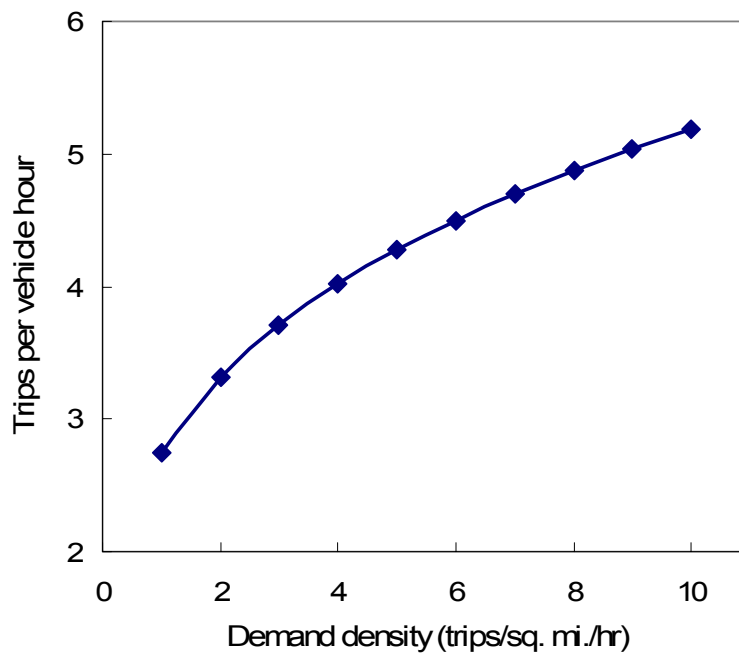


Figure 6-12. Effect of service demand density on vehicle productivity

(2) Area size

The second series of tests investigates the effect of area size, holding the demand density and service constraints at their default values. Figure 6-13 shows the results in terms of number of vehicles required. It illustrates that the required number of vehicles increases approximately linearly with increasing service area size

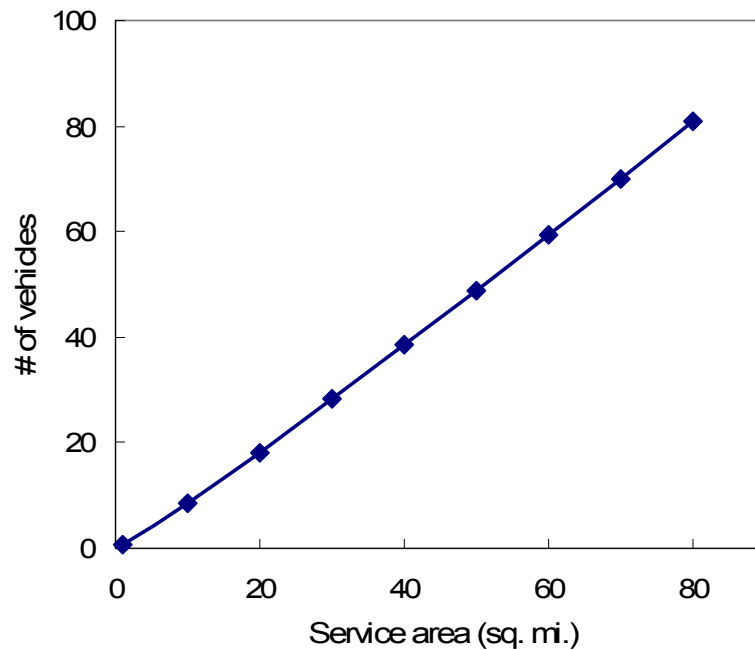


Figure 6-13. Effect of service area size on vehicles required

(3) Maximum time deviation

Figures 6-14 and 6-15 investigate the effect of one of the time constraints, maximum time deviation, while holding the demand density and service area fixed at 4 trips/sq. mi./hr and 64 sq. mi., respectively. Figure 6-14 shows the tradeoff relation between the operator cost in terms of vehicle fleet size requirement and the user cost in terms of maximum time deviation. Figure 6-15 shows the results in terms of vehicle productivity instead of number of vehicles. Figure 6-14 indicates a nonlinear relation between the

number of vehicles and the maximum time deviation. The decrease in number of vehicles when maximum time deviation increases from 10 to 15 minutes is larger than that when maximum time deviation increases from 25 to 30 minutes. And the corresponding vehicle productivity gained when maximum time deviation increases from 10 to 15 minutes is larger than that when maximum time deviation increases from 25 to 30 minutes. The results are more sensitive to shorter than longer maximum time deviations.

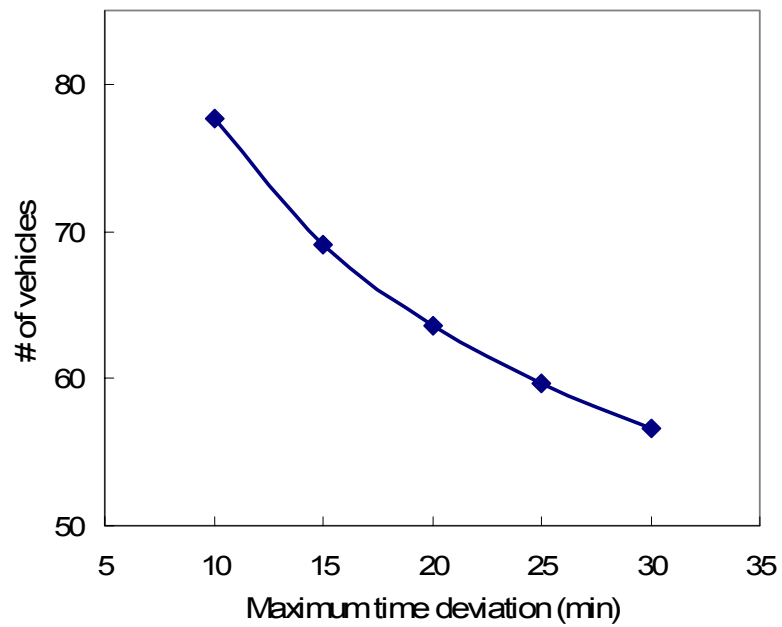


Figure 6-14. Effect of maximum time deviation on vehicles required



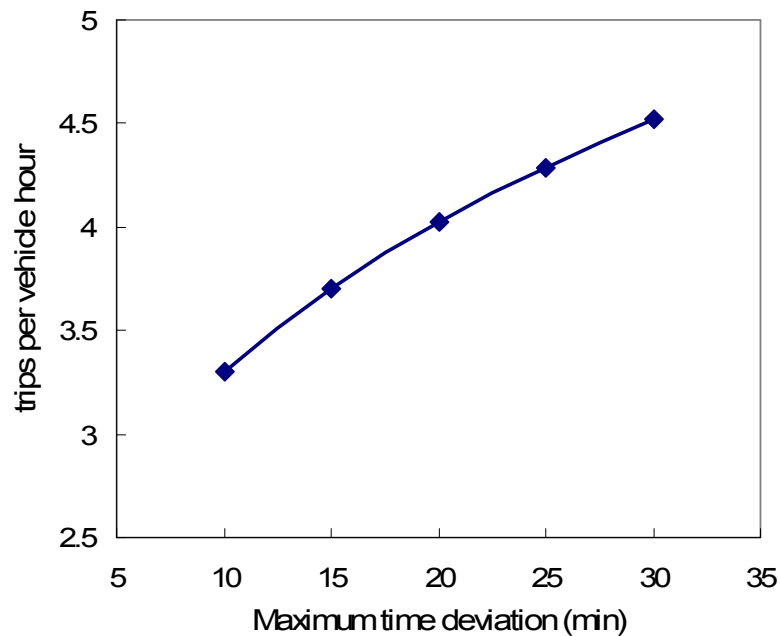


Figure 6-15. Effect of maximum time deviation on vehicle productivity

#### (4) Maximum ride time ratio

Figures 6-16 and 6-17 investigate the effect of maximum ride time ratio, holding the demand density and service area fixed at 4 trips/sq. mi./hr and 64 sq. mi. respectively.

Figure 6-16 shows the results in terms of number of vehicles required while Figure 6-17

shows the results in terms of vehicle productivity. Results in Figure 6-16 indicate an approximately linear relationship between the number of vehicles and the maximum ride time ratio. The vehicle productivity in Figure 6-17 increases approximately linearly as the maximum ride time ratio increases. Figure 6-18 shows the vehicle productivity as the maximum time deviation and maximum ride time ratio vary simultaneously. The vehicle productivity increases from about 3 to 5 trips per vehicle hour as both time constraints get less restrictive and service quality decreases.

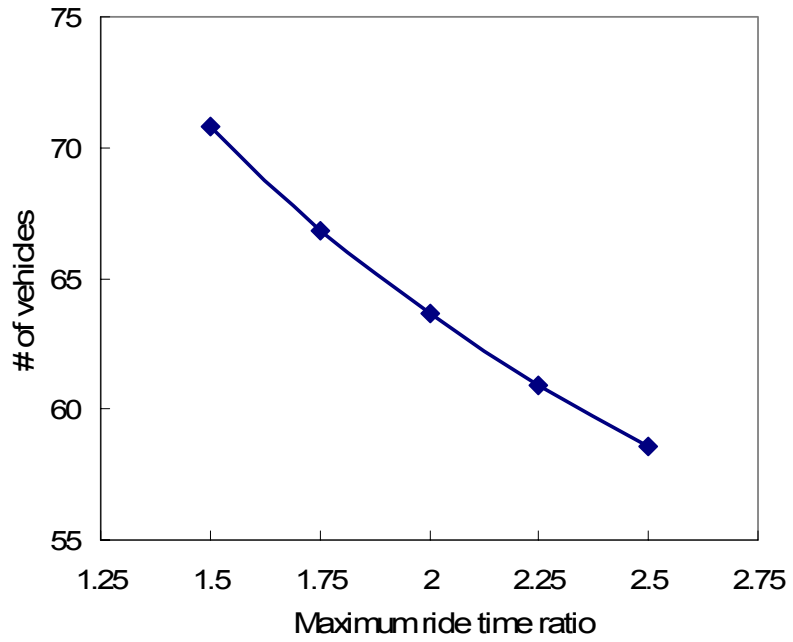


Figure 6-16. Effect of maximum ride time ratio on vehicles required

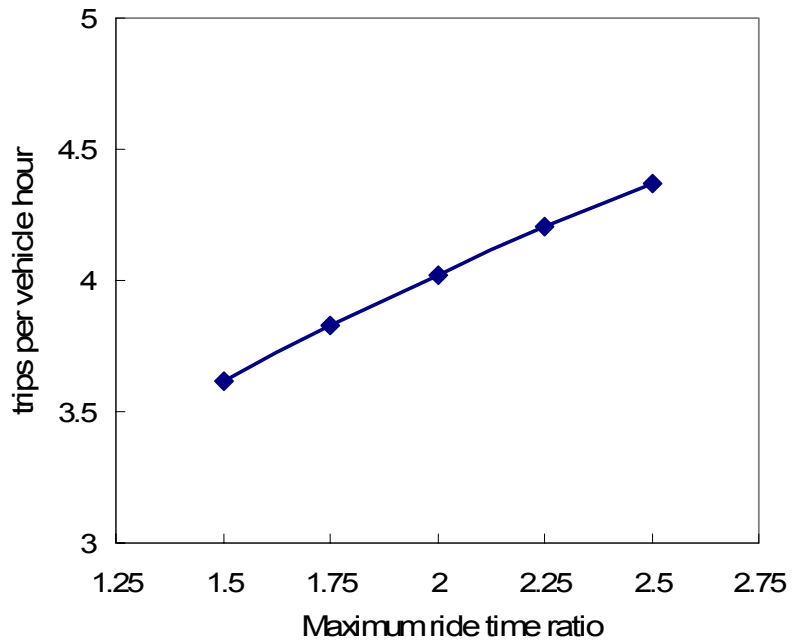


Figure 6-17. Effect of maximum ride time ratio on vehicle productivity

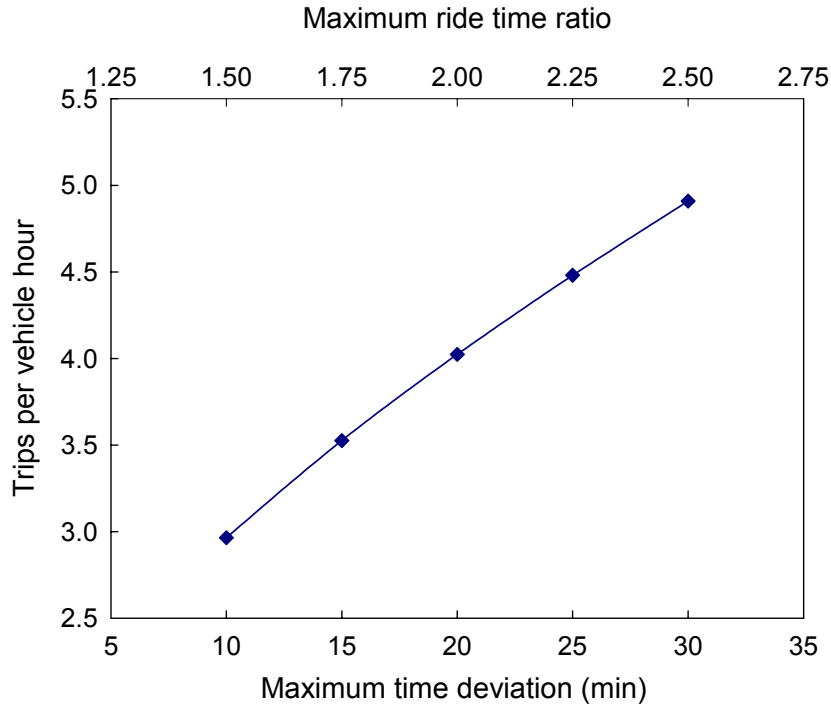


Figure 6-18. Effect of both time constraints on vehicle productivity

### 6.2.2 Optimization of the dial-a-ride service

This section illustrates the application of the developed performance models in the optimization of the DAR service considering the tradeoffs between the service quality and operating cost. In the planning stage of a DAR, given a demand level, service area characteristics and vehicle operating characteristics, decisions to be made include the determination of the fleet size and service level provided which is constrained and/or measured by the maximum time deviation and maximum ride time ratio. As the service level increases, the passenger time deviation and passenger ride time decrease while the vehicle resource requirement and operating cost increases. The tradeoffs between the service quality and operating cost should be well balanced in a public transit system. In the following case study, a total system cost is minimized, which takes into account both

the operator cost and user cost. The total cost function can be interpreted as a weighted sum of operator cost and disutility to the system's customers due to the time deviation from users' desired time and the in-vehicle travel time.

The following notation is used and baseline values are provided after the definitions for the case study:

- $A$  Area size,  $8 \times 8$  sq. mi.
- $B$ : Bus operating cost, 60 \$/hr
- $b$  Total boarding and alighting time, 2 min
- $C_o$ : Operator cost, in \$/hr
- $C_t$ : Total cost, in \$/hr
- $C_{uv}$ : User in-vehicle cost, in \$/hr
- $C_{uw}$ : User time deviation cost, in \$/hr
- $D$  Demand density, 4 trips/sq. mi./hr
- $f_c$  Roadway circuitry, 1.2
- $R$  Maximum ride time ratio
- $V$  Vehicle operating speed, 20 mph
- $v_{in}$ : Value of passenger in-vehicle time, 12 \$/passenger/hr
- $v_w$ : Value of passenger time deviation, 20 \$/passenger/hr
- $W$  Maximum time deviation, in min

Note that all of the cost calculations are on a hourly basis. The operator cost  $C_o$  is the fleet size  $F$  multiplied by the hourly operating cost  $B$  (which can incorporate the vehicle depreciation or rental cost). The fleet size  $F$  is estimated with Equation (5-11).

$$C_o = F \cdot B \quad (6-1)$$

$$F = 4.79 \frac{A^{1.07} \cdot D^{0.72} \cdot b^{0.21}}{W^{0.29} \cdot R^{0.37} \cdot (1.15V / f_c)^{0.68}} \quad (5-22)$$

The user cost consists of passenger in-vehicle cost  $C_{uv}$  (disutility due to in-vehicle travel time) and passenger time deviation cost  $C_{tw}$  (disutility due to time deviation from desired time). The total passenger in-vehicle cost  $C_{uv}$  in \$/hr can be estimated as the total passenger in-vehicle travel time per hourly demand  $T_{in}$ , multiplied by the value of in-vehicle time  $v_{in}$ . The total passenger in-vehicle travel time  $T_{in}$  can be estimated as the average ride time ratio  $\bar{R}$  multiplied by the total direct travel time per hourly demand

$$\sum_i T_{d,i} \cdot$$

$$C_{uv} = v_{in} T_{in} \quad (6-2)$$

$$T_{in} = \bar{R} \cdot \sum_i T_{d,i} \quad (6-3)$$

$$\bar{R} = 0.336 + 0.00136A + 0.0783D - 0.00589D^2 + 0.426R \quad (5-19)$$

This total direct travel time may usually be estimated from the demand analysis. In this case study, the total direct travel time is approximately estimated by using the average direct distance for two randomly generated points in a square area which is estimated to

be  $0.512l$  ( $l$  is length of area side). The second-order ride time Model R3 (Equation 5-19) is employed in this study.

The passenger time deviation cost  $C_{uw}$  is the average passenger time deviation, multiplied by hourly demand and the value of time deviation  $v_w$ .

$$C_{uw} = v_w D \bar{T}_{dev} \quad (6-4)$$

$$\bar{T}_{dev} = -0.90 + 0.50W \quad (5-16)$$

The total cost is the sum of the three cost components defined above.

$$C_t = C_o + C_{uv} + C_{uw} \quad (6-5)$$

Figure 6-19 shows the cost components (in \$/trip), which are obtained by dividing the costs in \$/hr by hourly demand, when both the maximum time deviation and maximum ride time vary accordingly. The maximum ride time ratio varies from 1.5 to 2.5 linearly as the maximum time deviation varies from 10 min to 30 min to provide consistent level of service. (Note that the maximum time deviation and ride time ratio can also vary independently and be optimized as two decision variables.) From the Figure 6-19, user cost increases and the operator cost decreases as the level of service decreases. (Both the maximum time deviation and maximum ride time ratio increase.) Figure 6-20 shows the total cost per trip as summing up three cost components. For this case study, the optimized maximum time deviation is 28 minutes and the optimal maximum ride time ratio is 2.4. The corresponding number of vehicles is 54.

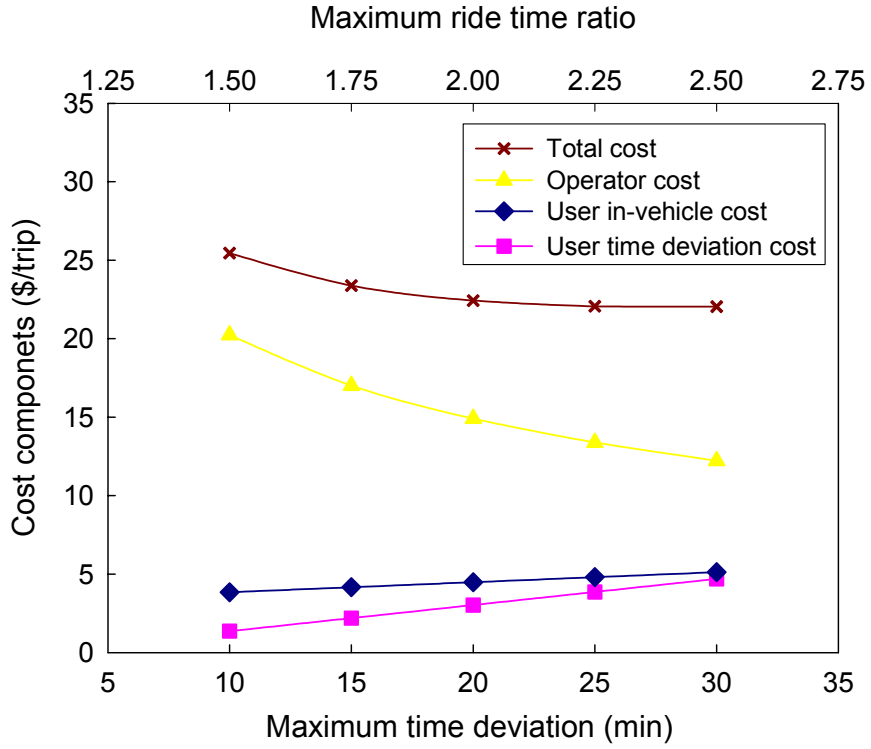


Figure 6-19. Cost components of the dial-a-ride service in the case study

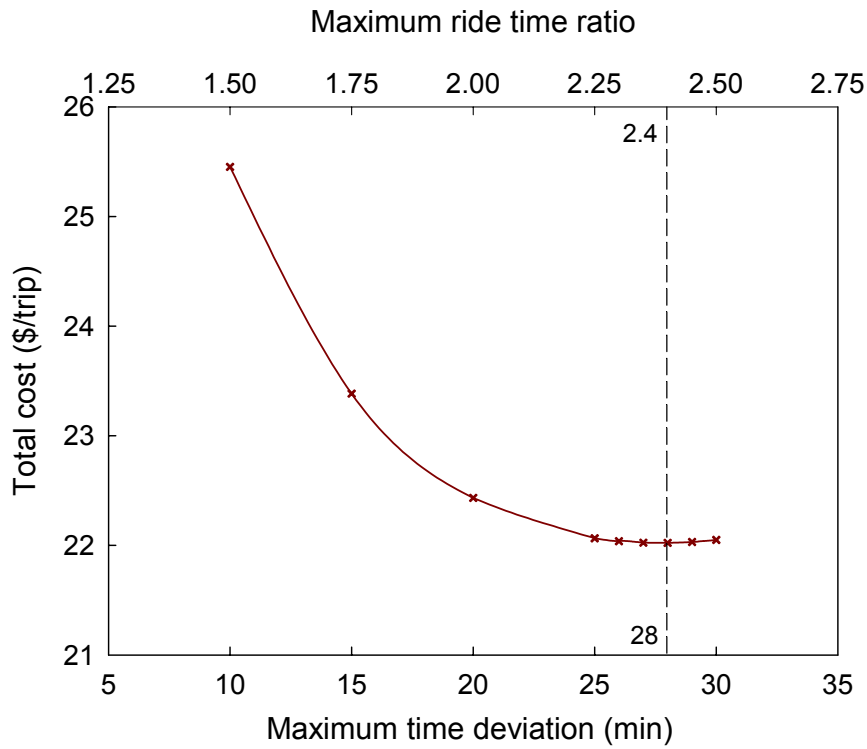


Figure 6-20. Total cost of the dial-a-ride service in the case study

Figure 6-21 plots the optimized cost components versus demand density using the baseline parameter values defined early in this section. The total cost per trip decreases as the demand density increases from 1 to 10 trips/sq. mi./hr. The decrease is steeper when the demand density is low (i.e. 1-3 trips/sq. mi./hr).

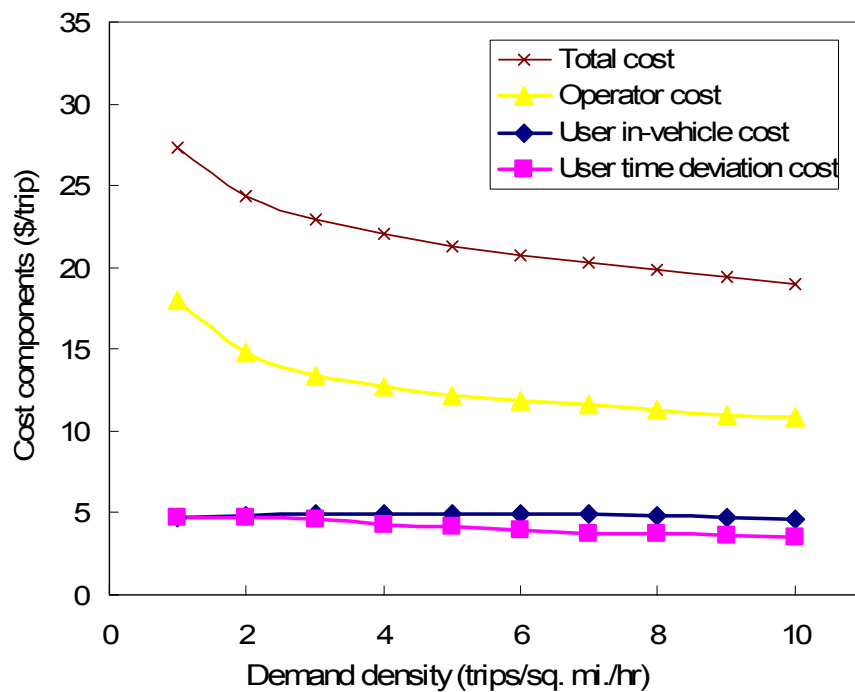


Figure 6-21. Costs vs demand density of the dial-a-ride service in the case study

Unlike a DAR service with its flexible route and schedule, fixed route conventional bus services are characterized by their fixed routes and schedules. They can provide relatively high passenger-carrying capacities at relatively low average costs to system operators. However, their service quality is limited since passengers must somehow reach some predetermined stations, wait for a vehicle, possibly transfer several times, and then travel



from their exit stations to their destinations. Thus, conventional transit services are least disadvantaged in areas and time periods with high demand densities, which can sustain high network densities and service frequencies. Information such as in Figure 6-21 provides very useful insights in determining the system operating configuration at the planning stage (i.e. whether fixed route conventional bus service or flexible route and schedule DAR service should be provided). If combined with similar information for a fixed route conventional bus service, a threshold analysis (as in Chang and Schonfeld, 1991) can be used to determine which service type is preferable under given circumstances.

## **Chapter 7 Conclusions and Future Research**

### **7.1 Conclusions**

In this dissertation, three performance metamodels have been developed using the response surface metamodeling approach for dynamic many-to-many DARP. The models predict, respectively, the minimum vehicle fleet size requirement, the average passenger time deviation from desired time, and the average passenger ride time ratio.

The metamodeling approach can incorporate in its simulation experiments detailed vehicle routing algorithm and passenger time constraints, which are oversimplified or omitted by an analytical approach. The technique used for developing the performance models is summarized as follows:

- Develop an online routing and scheduling heuristic for the dynamic DARP
- Design simulation experiments (which include the determination of input factors, their ranges of interest and selection of design points)
- Execute experiments (apply heuristic to solve simulated scenarios)
- Collect data from experiments

- Develop relations among performance and input factors through statistical estimation
- Validate the metamodels

This work also contributes to the development of heuristics for the static and dynamic DARPs with time constraints. A new heuristic, which is named a rejected-reinsertion heuristic, has been developed for the static multi-vehicle DARP. This method improves the conventional parallel insertion heuristic with a new rejected-reinsertion operation and a periodical improvement procedure involving trip reinsertion and trip exchange operations. Tables 7-1 and 7-2 show the vehicle reduction due to rejected-reinsertion heuristics compared with parallel insertion and/or regret insertion of Diana and Dessouky (2004). The proposed heuristics are very efficient computationally.

Table 7-1. Vehicle reductions due to rejected-reinsertion heuristics for static problem

	Parallel insertion	Diana 5
Rejected-reinsertion	-5% ~ -10%	-1% ~ -11%
Rejected-reinsertion with improvement	-10% ~ -17%	-1% ~ -12%

Table 7-2. Vehicle reductions due to rejected-reinsertion rolling horizon heuristics for dynamic problem

	Parallel insertion
Rejected-reinsertion	-6% ~ -9%
Rejected-reinsertion with improvement	-10% ~ -21%

The static heuristic has been extended to two online heuristics for the dynamic large-scale DARP, namely, the immediate-insertion online heuristic and the rolling horizon online heuristic. The immediate-insertion heuristic re-solves the static problem upon the appearance of the new request, while the rolling horizon heuristic uses a rolling horizon scheme which defers the insertion of the non-urgent requests in order to reserve more flexibility for future requests. The rolling horizon heuristic outperforms the immediate insertion heuristic for demand scenario in which different lead times for demand exist. The heuristic is computationally efficient, which makes it usable in real dynamic applications. It is simple in concept, and it does not involve complex algorithm parameters which must be tested for specific problems. The rolling horizon online heuristic with periodical improvement, the best among those heuristic variations developed here, is employed in the simulation experiments for the development of the performance models. Table 7-3 shows the vehicle reduction due to rolling horizon heuristics compared with immediate insertion heuristics.

Table 7-3. Vehicle reductions due to rolling horizon heuristics for dynamic problem

	Immediate insertion	Immediate insertion with improvement
Rolling horizon	-3% ~ -10%	
Rolling horizon with improvement		-5% ~ -9%

The response surface metamodeling approach is applied in the development of the performance model. The functional relation between an output (i.e. number of vehicles)

of the DAR operation process and its input factors is modeled through well designed simulation experiments and post statistical analysis based on data collected from the experiments. A face-centered central composite design is used in this study to determine the design points (the value of input factors). Data collected from the simulation experiments are fitted through linear regression with SPSS software. Polynomial first-order, second-order and multiplicative models are estimated and their statistical results are analyzed and compared. The best models in terms of both statistical properties and simplicity of the model form are suggested. The metamodels are validated using an additional set of randomly generated data.

The developed metamodels are as the follows:

- (1) Vehicle resource requirement Model F1a

$$F = 4.79 \frac{A^{1.07} \cdot D^{0.72} \cdot b^{0.21}}{W^{0.29} \cdot R^{0.37} \cdot (1.15V / f_c)^{0.68}} \quad (5-22)$$

- (2) Time deviation Model D3

$$\bar{T}_{dev} = -0.90 + 0.50W \quad (5-16)$$

- (3) Ride time ratio Model R3

$$\bar{R} = 0.336 + 0.00136A + 0.0783D - 0.00589D^2 + 0.426R \quad (5-19)$$

The variables in the above equations are defined in Section 5.3. The resulting models are relatively simple in structure and inexpensive to use. Sensitivity analysis also indicates that the performance metamodels are fairly robust, in that deviation from the assumptions of square service area, uniform demand distribution and percentage of passengers

specifying desired pickup time in the practical range would not affect much the accuracy of the predictions. They are approximate in nature, and mostly suited for use at the high-level planning stage of a system. The applications of the performance models are illustrated through the parametric analysis and optimization of a DAR service considering the tradeoff between operator cost and user cost.

## **7.2 Future Research**

The developed performance models might be applied to optimize an integrated system including both flexibly and fixed route transit services. Fixed conventional bus services are least disadvantaged in areas and time periods with high demand densities, which can sustain high network densities and service frequencies, while flexible route DAR services are suitable for suburban areas or time periods with low demand densities. When operated separately both services suffer from the variability of demand over time. In an integrated system, the entire fleet might be used to provide conventional bus service during peak hours and the excess fleet is used during off-peak to provide DAR service with higher service quality to low-density surrounding areas. The vehicle resource allocation can be optimized to obtain the best combination of cost and service quality based on the performance of the two systems.

The performance models are developed for many-to-many DAR service. When some of the origins and/or destinations coincide and when the requests of pickup and/or delivery at the same place are within a certain time period, the system can then accommodate

multiple pickups and/or deliveries at one place and is expected to operate more efficiently. The research can be extended to investigate the appropriate measurements for the cluster of the demand in both space and time and its effect on the performance.

Similar performance models can also be developed for PDP (e.g. pickup and delivery mails or packages), which usually has more applications than DARP.

Field operating data from similar systems should be collected, if they become available, in order to further compare and evaluate the developed models. Note that comparison with one single real system from a specific location might not mean much since the models are developed for the high-level planning purpose and are based on some general assumptions such as the square area.

The rejected-reinsertion operation, developed to accommodate those requests that are infeasible by direct insertion, improves the parallel insertion with very little additional computational cost. It can be applied and further tested in other related vehicle routing problems, especially for the dynamic problems because of its computational efficiency.

## Notation

$A$	Service area size (sq. mi.)
$a_0$	Constant term in maximum ride time Equation (3-10) (min)
$a_1$	Slope in maximum ride time Equation (3-10)
$AT_i$	Actual pickup or delivery time for stop $i$
$AUP_i (ADOWN_i)$	Maximum amount of time by which stop $i$ and all its following stops can be advanced (delayed) without violating the time window constraints.
$B$	Bus operating cost (\$/hr)
$BUP_i (BDOWN_i)$	Maximum amount of time by which stop $i$ and all its preceding stops on the same <i>vehicle route</i> can be advanced (delayed) without violating the time window constraints
$b$	Total boarding and alighting time per person (min)
$C_o$	Operator cost (\$/hr)
$C_t$	Total cost (\$/hr)
$C_{uv}$	User in-vehicle cost (\$/hr)
$C_{uw}$	User time deviation cost (\$/hr)
$D$	Demand density (trips/sq. mi./hr)
$EDT_i (LDT_i)$	Earliest (latest) delivery time for request $i$
$EPT_i (LPT_i)$	Earliest (latest) pickup time for request $i$



$ET_i$ ( $LT_i$ )	Earliest (latest) pickup or delivery time for stop $i$
$F$	Minimum number of operating vehicles
$f_c$	Roadway circuitry
$Idle_k$	Idling time before schedule block $k$
$l$	Length of a rectangular area
$MRT_i$	Maximum ride time for request $i$
$R$	Maximum ride time ratio
$\bar{R}$	Average passenger ride time ratio
$r$	Aspect ratio $r = l / w$
$\bar{T}_{dev}$	Average passenger time deviation from desired time (min)
$T_i$	Scheduled time for stop $i$
$T_{detour}^i$	Additional travel time due to inserting both stops $+i$ and $-i$
$T_{i,j}$	Direct ride time from stop $i$ to stop $j$
$TW_i$	Maximum deviation from desired time for request $i$
$V$	Vehicle operating speed (mph)
$v_{in}$	Value of passenger in-vehicle time (\$/passenger/hr)
$v_w$	Value of passenger time deviation (\$/passenger/hr)
$W$	Maximum time deviation (min)
$w$	Width of a rectangular area

## References

- Arrillaga, B. and Medville, D. M. (1974). Demand, supply, and cost modeling framework for demand-responsive transportation systems. *Demand-Responsive Transportation Special Report 147*, 32-48.
- Atanasio, A., Cordeau, J. F., Ghiani, G. and Laporte, G. (2004). Parallel tabu search heuristics for the dynamic multi-vehicle dial-a-ride problem. *Parallel Computing*, 30, 377-387.
- Barton, R. R. (1998). Simulation metamodels. *Proceedings of the 1998 Winter Simulation Conference*, eds Medeiros, D. J., Watson, E. F., Carson, J. S. and Manivannan, M.S., 167-174, Washington D.C.
- Baugh, J. W., Kakivaya, G. K. R. and Stone, J. R. (1998). Intractability of the dial-a-ride problem and a multiobjective solution using simulated annealing. *Engineering Optimization*, 30, 91-123.
- Beardwood, J., Halton, J. H. and Hammersley, J. M. (1959). The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55, 299-328.
- Bodin, L., Golden, B., Assad, A. and Ball, M. (1983). Routing and scheduling of vehicles and crews: the state of the art. *Computers and Operations Research*, 10(2), 63-211.
- Bodin, L. D. and Sexton, T. (1986). The multi-vehicle subscriber dial-a-ride problem. *TIMS Studies in the Management Sciences*, 22, 73-86.

- Borndorfer, R., Grottschel, M., Klostermeier, F. and Kuttner, C. (1997). Telebus Berlin: vehicle scheduling in a dial-a-ride system. *Computer-Aided Transit Scheduling, Lecture Notes in Economics and Mathematical Systems 471*, ed Wilson, N. H. M., 391-422, Springer, Berlin.
- Box, G. E. P. (1954). The exploitation and exploration of response surfaces: some general considerations and examples. *Biometrics*, 10(1), 16-60.
- Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*, Wiley, New York.
- Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*, Wiley, New York.
- Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society*, B13, 1-38.
- Campbell, A. M. and Savelsbergh, M. (2004). Efficient insertion heuristics for vehicle routing and scheduling problems. *Transportation Science*, 38(3), 369-378.
- Chang, S. K. (1990). *Analytic optimization of bus systems in heterogeneous environments*. Ph.D. Thesis, Department of Civil and Environmental Engineering, University of Maryland, College Park.
- Chang, S. K. and Schonfeld, P. (1991). Optimization models for comparing conventional and subscription bus feeder services, *Transportation Science*, 25(4), 281-298.
- Cordeau, J. F. and Laporte, G. (2003). The dial-a-ride problem: variants, modeling issues and algorithms. *Quarterly Journal of the Belgian, French and Italian Operations Research Societies*, 4OR, 1, 89-101.
- Cordeau, J. F. and Laporte, G. (2003). A tabu search heuristic for the static multi-vehicle dial-a-ride problem. *Transportation Research*, 37B, 579-594.
- Daganzo, C. F., Hendrickson, C. T. and Wilson, N. H. M. (1977). An approximate analytic model of many-to-one demand responsive transportation systems. In: *Proceedings of the Seventh International Symposium on Transportation and Traffic Theory*, Kyoto Daigaku, 743-772.
- Daganzo, C. F. (1978). An approximate analytic model of many-to-many demand responsive transportation systems. *Transportation Research*, 12, 325-333.

- Daganzo, C. F. (1984a). The length of tours in zones of different shapes. *Transportation Research*, 18B(2), 135-145.
- Daganzo, C. F. (1984b). The distance traveled to visit N points with a maximum of C stops per vehicle: an analytic model and an application. *Transportation Science*, 18(4), 331-350.
- Desaulniers, G., Desrosiers, J., Erdmann, A., Solomon, M. M. and Soumis, F. (2002). VRP with pickup and delivery. *The Vehicle Routing Problem, SIAM Monographs on Discrete Mathematics and Applications*, eds Toth, P. and Vigo, D., 225-242, Philadelphia.
- Desrosiers, J., Dumas, Y., Solomon, M. M. and Soumis, F. (1995). Time constrained routing and scheduling. *Network Routing, Handbooks in Operations Research and Management Science*, eds Ball, M. O., Magnanti, T. L., Monma, C. L. and Nemhauser, G. L., 8, 35-139. North-Holland, Amsterdam.
- 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 vehicle dial-a-ride problem. *Computer-Aided Transit Scheduling, Lecture Notes in Economics and Mathematical System 308*, eds Daduna, J. R. and Wren, A., 15-27. Springer, Berlin.
- Dessouky, M. and Adam, S. (1998). *Real-Time Scheduling of Demand Responsive Transit Service – Final Report*. University of Southern California, Department of Industrial and Systems Engineering, Los Angeles.
- Dessouky, M., Rahimi, M. and Weidner, M. (2003). Jointly optimizing cost, service, and environmental performance in demand-responsive transit scheduling. *Transportation Research*, 8D, 433-465.
- Diana, M. and Dessouky, M. M. (2004). A new regret insertion heuristic for solving large-scale dial-a-ride problems with time windows. *Transportation Research*, 38B(6), 539-557.
- Draper, D. R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd edition, Wiley, New York.
- Dumas, Y., Desrosiers, J. and Soumis, F. (1991) The pickup and delivery problem with time window. *European Journal of Operational Research*, 54, 7-22.

- Eilon, S., Watson-Gandy, C. D. T. and Christofides, N. (1971). *Distribution Management: Mathematical Modelling and Practical Analysis*, Hafner, New York.
- Flusberg, M. and Wilson, N. H. M. (1976). A descriptive supply model for demand responsive transportation system planning. *Proceedings of the 17th Annual Meeting Transportation Research Forum*, 425-431.
- Fridman, L. W. (1996). *The Simulation Metamodel*. Kluwer Academic Publishers, Norwell, Massachusetts.
- Fridman, L. W. and Pressman, I. (1988). The metamodel in simulation analysis: can it be trusted? *Journal of the Operational Research Society*, 39(10), 939-948.
- Fu, L. (2002a). A simulation model for evaluating advanced dial-a-ride paratransit system. *Transportation Research*, 36A, 291-307.
- Fu, L. (2002b). Scheduling dial-a-ride paratransit under time-varying, stochastic congestion. *Transportation Research*, 36B, 485-506.
- Fu, L. (2003). Analytical model for paratransit capacity and quality-of-service analysis. *Transportation Research Record*, 1841, 81-89.
- Gendreau, M., Hertz, A. and Laporte, G. (1992). New insertion and postoptimization procedures for the traveling salesman problem. *Operations Research*, 40, 1086-1094.
- Hart, S. M. (1996). The modeling and solution of a class of dial-a-ride problems using simulated annealing. *Control and Cybernetics*, 25(1), 131-157.
- Heathington, K. W., Miller, J., Knox, R. R., Hoff, G. C. and Bruggeman, J. (1968). Computer simulation of a demand-scheduled bus system offering door-to-door service. *Highway Research Record 251*, 26-40.
- Hunter, J. S. (1958). Determination of optimal operating conditions by experimental methods. *Industrial Quality Control*, 15, Part II-1.
- Hunter, J. S. (1959a). Determination of optimal operating conditions by experimental methods. *Industrial Quality Control*, 16, Part II-2.
- Hunter, J. S. (1959b) Determination of optimal operating conditions by experimental methods. *Industrial Quality Control*, 16, Part II-3.

- Ioachim, I., Desrosiers, J., Dumas, Y., Solomon, M. M and Villeneuve, D. (1995). A request clustering algorithm for door-to-door handicapped transportation. *Transportation Science*, 29(1), 63-78.
- Jaw, J. J. (1984). *Heuristic Algorithms for Multi-vehicle, Advance-Request Dial-a-Ride Problems*. Ph.D. Thesis, Department of Aeronautics and Astronautics, M.I.T., Cambridge, MA.
- 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.
- Johnson, D. S., McGeoch, L. A. and Rothberg, E. E. (1996). Asymptotic experimental analysis for the Held-Karp traveling salesman bound. In: *Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms*, 341-350.
- Khuri, A. I. and Cornell J. A. (1996). *Response Surfaces Designs and Analyses*, 2nd edition, Marcel Dekker, New York.
- Kikuchi, S. and Rhee, J. H. (1989). Scheduling method for demand-responsive transportation system. *Journal of Transportation Engineering*, 115(6), 630-645.
- Kleijnen, J. P. C. (1987). *Statistical tools for simulation practitioners*, Marcel Dekker, New York.
- Kleinbaum, D. G., Kupper, L. L. and Muller, K. E. (1988). *Applied Regression Analysis and Other Multivariable Methods*, PWS-Kent Publishing Company, Boston.
- Lazoff and Sherman (1994). *An Exact Formula for the Expected Wire Length between Two Randomly Chosen Terminals*, Technical Report TR CS-94-08, Computer Science Department, University of Maryland Baltimore County.
- Law and Kelton (2000). *Simulation Modeling and Analysis*, 3rd edition, McGraw-Hill, Boston.
- Lerman, S., Flusberg, M., Pecknold, W. and Wilson, N. H. M. (1977). *A Method for Estimating Patronage of Demand Responsive Transportation Systems*. U. S. Department of Transportation Report DOT-TSC-977.
- Lerman, S. and Wilson, N. H. M. (1974). Analytic model for predicting dial-a-ride system performance. *Transportation Research Board Special Report 147*, 48-53.

- Lin, S. and Kernighan, B. (1973). An effective heuristic algorithm for the traveling salesman problem. *Operations Research*, 21, 498-516.
- Madsen, O. B. G., Rawn, H. F. and Rygaard, J. M. (1995). A heuristic algorithm for the dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Annals of Operations Research*, 60, 193-208.
- Madu, C. N. and Kuei, C. (1994). Regression metamodeling in computer simulation – The state of the art. *Simulation Practice and Theory*, 2, 27-41.
- Mason, F. J. and Mumford, J. R. (1972). Computer models for designing dial-a-ride systems. Society of Automotive Engineers, Automotive Engineering Congress, 720216.
- Myers, R. H. and Montgomery, D. C. (2002). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. 2nd edition, Wiley, New York.
- Menhard, H. R., Flusberg, M. and Englisher, L. S. (1978). *Modeling Demand-Responsive Feeder Systems in the UTPS Framework: Final Report*. Multisystems, Inc., UMTA-MA-06-0049-78-9.
- Mitrovic-Minic, S. (1998). *Pickup and Delivery Problem with Time Windows: A Survey*. SFU CMPT TR 1998-12, <ftp://fas.sfu.ca/pub/cs/techreports/1998>.
- Mitrovic-Minic, S. (2001). *The Dynamic Pickup and Delivery Problem with Time Windows*. Ph.D. Thesis, School of Computing Science, Simon Fraser University, Burnaby, Canada.
- Mitrovic-Minic, S., Krishnamurti, R. and Laporte, G. (2004). Double-horizon based heuristics for the dynamic pickup and delivery problem with time windows. *Transportation Research*, 38B, 669-685.
- Montgomery, D. C. (2001). *Design and Analysis Experiments*, 5th edition, Wiley, New York.
- Potvin, J. Y. and Rousseau, J. M. (1992). Constraint-directed search for the advanced request dial-a-ride problem with service quality constraint. *Computer Science and Operations Research: New Developments in Their Interfaces*, eds Balci, O., Sharda, R. and Zenios, S. A., 457-474, Pergamon Press, Oxford, England.

- Psaraftis, H. N. (1980). A dynamic programming solution to the single vehicle many-to-many immediate request dial-a-ride problem. *Transportation Science*, 14(2), 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(3), 351-357.
- Psaraftis, H. N. (1986). Scheduling large-scale advance-request dial-a-ride systems, *American Journal of Mathematical and Management Science*, 6, 327-367.
- Psaraftis, H. N. (1988). Dynamic vehicle routing problems. *Vehicle Routing: Methods and Studies*, eds Golden, B. L. and Assad, A. A., 223-248, Elsevier Science Publishers B. V., North-Holland, Amsterdam.
- Psaraftis, H. N. (1995). Dynamic vehicle routing: status and prospects. *Annals of Operations Research*, 61, 143-164.
- Savelsbergh, M. W. P. (1995). The general pickup and delivery problem. *Transportation Science*, 29(1), 17-29.
- Sexton, T. and Bodin, L. D. (1985a). Optimizing single vehicle many-to-many operations with desired delivery times: I. scheduling. *Transportation Science*, 19, 378-410.
- Sexton, T. and Bodin, L. D. (1985b). Optimizing single vehicle many-to-many operations with desired delivery times: II. routing. *Transportation Science*, 19, 411-435.
- Solomon, M. (1987). Algorithms for the vehicle routing and scheduling problems with time window constraints. *Operations Research*, 35(2), 254-265.
- The SPSS homepage, [www.spss.com](http://www.spss.com).
- Stein, D. M. (1978a). An asymptotic, probabilistic analysis of a routing problem. *Mathematics of Operations Research*, 3(2), 89-101.
- Stein, D. M. (1978b). Scheduling dial-a-ride paratransit systems. *Transportation Science*, 12(3), 232-249.
- Teodorovic, D. and Radivojevic, G. (2000). A fuzzy logic approach to dynamic dial-a-ride problem. *Fuzzy Sets and Systems*, 116, 23-33.
- Toth, P. and Vigo, D. (1996). Fast local search algorithms for the handicapped persons transportation problem. *Meta-Heuristics: Theory and Applications*, eds Osman, I. H. and Kelly, J. P., 677-690, Kluwer, Boston.



- Toth, P. and Vigo, D. (1997). Heuristic algorithms for the handicapped persons transportation problem. *Transportation Science*, 31(1), 60-71.
- Van Der Bruggen, L. J. J., Lenstra, J. K. and Schuur, P. C. (1993). Variable-depth search for the single-vehicle pickup and delivery problem with time windows, *Transportation Science*, 27(3), 298-311.
- Wilson, N. H. M. and Colvin, N. H. (1977). *Computer Control of the Rochester Dial-a-Ride System*. Technical Report R77-31, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA.
- 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., Higonnet, B. T. and Goodman, L. A. (1970). Simulation of a computer-aided routing system (CARS). *Highway Research Record* 318, 66-76.
- Wilson, N. H. M., Sussman, J. M., Wong, H. and Higonnet, B. T. (1971). *Scheduling Algorithms for Dial-a-Ride Systems*. Technical Report USL TR-70-13, Massachusetts Institute of Technology, Cambridge, MA.
- Wilson, N. H. M. and Weissberg, H. (1976). *Advanced Dial-a-Ride Algorithms Research Project: Final Report*. Technical Report R76-20, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA.