ON SPARE PARTS SUPPLY CHAINS WITH FORWARD STOCKING LOCATION RECOURSE

A Thesis
Presented to
The Academic Faculty

by

Brian Joseph Kues

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial & Systems Engineering

Georgia Institute of Technology
August 2016

ON SPARE PARTS SUPPLY CHAINS WITH FORWARD STOCKING LOCATION RECOURSE

Approved by:

Professor Alan Erera, Advisor
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Professor Atalay Atasu
Scheller College of Business
Georgia Institute of Technology

Professor Martin Savelsbergh
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Professor Julie Swann
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Professor Chelsea White
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Date Approved: 31 March 2016

To my parents, who have blessed me with their love and support my entire life.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Alan Erera, for his guidance and financial support throughout the course of my Ph.D. studies at Georgia Tech. His insightful questioning forced me to grow as a researcher, improving my ability to step back from the intimate details of a problem and synthesize the big picture. I would also like to thank Professor Chip White for advising and supporting me as a research assistant for numerous semesters as well as serving on my thesis committee. Professors Atalay Atasu, Martin Savelsbergh, and Julie Swann gave generously of their time to serve on my thesis committee and provided helpful feeback and direction, for which I am grateful. Thank you to Professor Gary Parker for admitting me to the Ph.D. program and believing in my ability.

One of the benefits of the ISyE department at Georgia Tech being so large is that there are so many good people affiliated with it, and I have had the great fortune of getting to know some of them very well in the classroom, in the office, and on the intramural field during my six years as a Ph.D. student. I would like to extend my gratitude to James Bailey, Anthony Bonifonte, Vinod Cheriyan, Jin Lee, Camilo Ortiz, Jeff Pavelka, Daniel Silva, and Rodrigue Ngueyep Tzoumpe for being especially good friends and making my experience a memorable and enjoyable one. I am also glad to have met and spent time with all of my ISyE colleagues, Josh Hale, Andres Iroume, Ben Johnson, Mathias Klapp, Todd Levin, Alvaro Lorca, Josh McDonald, Carl Morris, MinKyoung Kang, Alborz Parcham-Kashani, Matt Plumlee, Aurko Roy, Kevin Ryan, Mallory Soldner, Tim Sprock, Stefania Stefansdottir, Monica Villareal, Daniel Zink, and many others whose names I am forgetting as I write this. A special thank you to Yanling Chang for being such a good collaborator on our risk assessment project. Thank you to my fellow Tech students from other departments and Atlanta friends I met through them, Carlos Campo, Felipe and Stephanie Castrillon, Gina and Edward Kim, Barbara Piña, Maria Restrepo, Tatiana Restrepo, and Alejandro Suarez, for the summer barbecues, boat rides, and trips to Charleston.

The sport of volleyball has provided me with an outlet to get some exercise, take my mind off work, and meet many great people. Thank you to Karol Chudy, Hunter Evans, Jake Heinrichs, Ryan Hojnacki, James Mauro, David Mustard, Eric Topper, Marin Zaimov, and others in the Atlanta volleyball community for so often inviting me to play in tournaments and for hanging out after the matches were over. The Georgia Tech volleyball program made me feel like a member of the family ever since I first volunteered for them. Thanks to Tonya Johnson, Craig Bere, Ed Tolentino, Chuck Crawford, Carla Gilson, Scott McDonald, Christen Steele, and Danny Karnik for being so welcoming.

Having lived at GLC, Tenth and Home, 1075 McMillan, State Street, Walton River, and Alexander on Ponce, I have had the pleasure of many awesome roommates. Thank you to Jin Lee, Karol Chudy, Dan Chudy, Mandy Miller, Jeff Pavelka, James Bailey, Marin Zaimov, Dimitar Kostov, Alex Samarchi, Brandon Makinson, Nick Magina, Jake Heinrichs, Bo Hatchett, and Matt Vaughan for helping me stretch my stipend, engaging in thought-provoking conversation, and being such cool people.

Of course, I could have never completed this long journey without the unwavering and ever-present support of my parents, brothers, and extended family back home in Ohio. Thanks, Mom, Dad, Derek, and Matthew for everything. Thanks, Aunt Carol and Uncle Mick, Aunt Jan and Uncle Bob, Aunt Eileen, Coutts family, Taylor family, Collins family, Schneider family, Arnold family, for the thoughts, prayers, and cards over the years. I am tearing up as I write this, thinking about how lucky I am to have you all in my corner.

Finally, thank you to the National Center for Food Protection and Defense for their financial support during the later years of my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis investigates how a service provider responsible for completing repairs of durable goods deployed throughout a region can improve profitability by reconfiguring its existing supply chain. This supply chain stocks spare parts inventory both in kits that technicians take to the repair sites, and at forward stocking locations (FSLs) distributed throughout the region. The investigation was conducted in three parts.

The first part examines the decision of which spare parts to stock in the kit for a single technician who must complete multiple repairs within a single period. The technician has access at a single FSL to all parts not stocked in the kit but at a time penalty which decreases the likelihood of completing a satisfactory number of repairs by the end of the period. This decision is modeled with a binary optimization problem which has a single probabilistic constraint, namely, that the allocation of spare parts between the kit and the FSL allows the technician to complete the required number of repairs in expectation. Formulated as such, the decision problem is $\mathcal{NP}$-hard. Six heuristics are proposed to quickly find spare parts allocations that have low cost. A method to find a lower bound on the optimal value is developed and then used to demonstrate that many of the heuristics can generate solutions whose costs are within ten percent of optimal. The heuristics are also compared against one another on a test suite of problem instances with important parameters varied over ranges of values. Finally, the first part concludes with a case study in which the heuristics are analyzed for effectiveness using instances drawn from real-world data from an industry collaborator.

The second part of the thesis broadens the scope of the first part to include the decision of how to sequence the repairs for the technician when the geographic locations of the customers are explicitly incorporated into the model. The total travel time, which depends on the sequence in which customers are visited, affects the expected number of completed repairs, which again is constrained to be above a certain required level in expectation. Travel time, however, does not directly impact the objective to minimize total inventory cost. Given a method for sequencing customers, the decision problem is identical to that in part one. The fact that the customers change from one period to the next is accounted for by evaluating the expected fill rate for a repair kit on average across multiple customer instances. Heuristics for sequencing customers are proposed, and heuristics for determining inventory are reused from the first part of the thesis. Computational results show that random routing leads to inventory costs twenty percent higher than those for all other proposed methods of routing. Furthermore, it is shown that almost all benefits of smarter routing can be generated even when using a simple greedy heuristic for routing decisions.

The third and final part of the thesis again broadens the scope of the first part, but in a different direction to include the decisions of how many technicians to employ and how many FSLs to operate. Each technician must complete the repairs assigned to him or her, the number of which depends on the total number of technicians employed, and shares access to inventory at all of the FSLs, the number of which influences the time delay needed to retrieve a part not in the repair kit. The objective of the decision problem is to minimize the average total cost of repair kit inventory, FSL inventory, and technician labor per customer repair job in a single period. A straightforward algorithm to find a good solution, completely specified by a kit-or-FSL decision for all part types, a number of technicians, and a number of FSLs, is motivated by the fact that the inventory-setting algorithms developed in the first part of the thesis can be reused for a given number of technicians and number of FSLs. Computational tests reveal that the best solutions have a single FSL and either fewer but busy technicians with close-to-full kits or more but less busy technicians with empty kits. When inventory is four times more costly than labor, the solutions fall somewhere between these two extremes. Such an arrangement offers savings of 10% to 30% over a supply chain without inventory recourse at an FSL.

CHAPTER I

INTRODUCTION

In his book *The Rational Optimist: How prosperity evolves*, Matt Ridley argues that, at the moment, human beings are better off in almost all ways than they have ever been due to human innovation over the past 100,000 years, the source of which is "collective intelligence evolving by trial and error resulting from the sharing of ideas through exchange and specialization" [1]. On a daily basis, human beings who are fortunate enough to live in countries with advanced economies are consumers of goods and services whose creation required the efforts of hundreds and sometimes thousands of other people and often natural resources which first began developing millenia before their eventual harvest. If these consumers were required to subsist entirely on the fruits of their own labors, their standards of living would be incomparably lower than those they enjoy now. As Ridley succinctly puts it, "Self-sufficiency is another word for poverty" [2].

Not only are almost all users of high technology incapable of creating that technology themselves, but also they cannot repair it when it inevitably ceases functioning properly. (Being able to use and benefit from technology whose inner workings one does not completely understand is of course better than not having access to such technology at all. In fact, it is inherent in the nature of the prosperity we enjoy that this will almost always be the case, according to Ridley's reasoning. Nevertheless, the dilemma remains.) A purchaser of a high technology device expects to be able to use the device for a certain length of time. More often than not, the device fails during this expected lifetime. In an effort to satisfy their customers, some manufacturers take up the role of repairing their products to working order with their own fleet of mobile technicians. The timely and cost-effective completion of such repairs is the focus of this thesis.

1.1 Background

After-sales field service can account for a significant proportion of total revenue and expenses for a business that manufactures and sells high technology equipment. Often at the time of purchase, customers enter service-level agreements with the manufacturer which stipulate that the machine must be available for a certain percentage of the year or that repairs must be completed within a certain period of time after the customer notifies the manufacturer of a problem. Customers who do not purchase their equipment but rather lease it will usually have such service-level terms stipulated in the leasing contract that they sign.

Depending on the importance and priority of the equipment to the customer's business (or personal) operations, the responsiveness with which the manufacturer must act may vary from a few hours to a week. For example, a laundry rental agency can schedule a dryer repair for a few days after the customer calls to report a problem. On the other hand, a factory which produces millions of dollars worth of goods in a single day would not settle for that level of service if one of its machines broke down and halted production.

The repairs which must be completed by the manufacturer (or agents acting on its behalf) usually require working spare parts to be installed in the failed machine, replacing parts which stopped functioning properly. The proximity of the necessary spare parts to the machines is a key determinant of down time or service level. However, inventory that is held closer to the customers, for example in a technician's kit, leads to a larger system-wide repair parts inventory that is held at a larger cost to the provider. In real-world problems, the number of parts which might need to be replaced can run into the thousands, due to the variety of machine types and versions installed within a given region that a technician would cover. Stocking all such parts as close to the customer as possible, especially when many of them are used rarely, leads to bloated inventory costs and service levels that exceed expectation or contractual obligation. The fundamental trade-off in spare parts problems is low inventory cost versus high service level.

1.2 Concepts

The design and operation of spare parts supply chains requires decisions to be made across all lengths of time horizon. Strategic decisions must be made, for example, about how many warehouses and FSLs to have in the supply chain and where to locate each of those facilities. Tactical decisions include the stocking levels for each inventory facility, given their total number and locations. Operational decisions about how to schedule repairs, and how to dispatch parts both to technicians or from one facility to another are required as well. This thesis discusses decisions of all types, but devotes the most attention to the tactical decision of setting stocking levels.

Machines that are repaired with spare parts are complicated systems that require numerous parts to work together effectively. When such a machine breaks down, a skilled technician diagnoses the problem and often removes a malfunctioning assembly or part of the machine. The technician might be able to repair the malfunctioning part and then return it to the machine in simple cases. In more difficult cases, the technician ships the part to a central location, where more specialized staff repair it, and either immediately installs a working spare if one is stocked where the machine is located or places an order for a replacement part with the higher-level supplying location. The machine with a removed part can also be put back into service immediately if there is another machine at the same location that is out-of-service due to a different part. This type of spare parts replacement is known as cannibalization. If there are no working parts of the needed type available on-site, the machine will wait until the technician repairs the malfunctioning part or a replacement arrives from the supplying facility.

Resupply is the process of shipping parts from the supplying higher echelon to the demanding lower echelon when a needed part is not on-hand. Resupply events may occur at a discrete set of epochs, or at any point in continuous time. Not all spare parts models, however, incorporate the repair of parts at a higher echelon facility and subsequent shipping of spare parts from one echelon to another. A special case of the spare parts supply chain is one that does not have resupply. Supply chains with periodic resupply from a source with essentially infinite inventory can also be modeled this way. Examples of this special case include flyaway kits used in the military or repair kits for service technicians where a fixed inventory must support operations for a period of time without replenishment, the interest of this thesis.

A final point about these supply chains is that the approach taken to study them can be a system level or item level approach. The latter examines each part individually and optimizes the stocking level without taking into account any information about the other parts that must be supported. The advantage of such an approach is the simpler nature of its calculations. The disadvantage is that its solutions are often either suboptimal with respect to cost, or simply not feasible given that the model does not properly capture the true system. Decision makers are concerned with maintaining a certain level of service, which is a function of all the spare parts under their purview, and would like to consider the problem from this integrated perspective, which is best accomplished with a system level approach.

1.3 Model

This thesis presents a model for a spare parts problem with periodic resupply and inventory available in two different types of locations. One type of location, the repair kit, has a higher holding cost, but the parts are available for immediate use at the site of the repair. The second type, the forward stocking location (FSL), has a lower holding cost, but the parts must be fetched by the technician and brought to the repair site, which adds time to the process. The work was originally motivated by an industry project with a collaborator that wanted to improve its after-sales service performance without enlarging its fleet of technicians. In order to accomplish this goal, we needed to develop inventory policies that took advantage of a modified network design.

Our partner manufactures and sells point-of-sale devices and other equipment that is installed on-site at their customers' locations, often retail settings. The structure of the supply chain is similar to that of multi-echelon spares supply chains (as well as the special case of those which are repair kit problems) but that framework does not completely capture the trade-offs in this setting.

In this supply chain, there is one central warehouse that receives and repairs all the removed spare parts from the technicians. This warehouse resupplies both the repair kit and the FSL via overnight shipments.The technicians visit multiple customer sites in a single day, and must both diagnose the problem and complete the repair by the end of the day for job to be considered a success, which means that resupply from the central warehouse does not affect the service level. The repair kit contains a subset of the part types that might be needed to complete a repair, and the remainder are kept at the FSL. When the technician does not have the required spare part in his or her kit, he or she travels to the FSL to retrieve the part. This extra trip might cause later jobs to become service failures even though they require part types stocked in the repair kit. The inventory at the FSLs (which are less numerous than the technicians) is shared by all technicians. Hence, it is cheaper to stock a part at the FSL than to stock it in the repair kit due to pooling.

While keeping parts at the FSL decreases inventory costs, it also decreases technician productivity, because retrieving a part which otherwise would have been in the kit forces the technician to spend time traveling from the customer site to the FSL and back. Models for multi-echelon supply chains with resupply do not capture this trade-off of low inventory cost versus high rate of technician productivity. In fact, such existing models do not incorporate the availability of technicians at all. Once the working spare part is at the right location, the repair is considered complete.

In response to this shortcoming in the literature, this thesis answers the question of how to manage a spare parts supply chain with forward stocking location recourse.

1.4 Overview

This thesis is divided into three chapters, each of which addresses a different aspect of the spare parts supply chain with forward stocking location recourse.

1.4.1 The Multiple-Job Repair Kit Problem with Forward Stocking Location Recourse

The first chapter investigates the question of how to determine optimal stocking levels of parts in the technician's repair kit when the technician must visit multiple customers between replenishments but has access to recourse inventory at an FSL. We assume that every customer job requires exactly one spare part to successfully complete and that these demands are independent of one another. If this part is not stocked in the repair kit, then the technician will immediately make a trip to the FSL, where he or she can retrieve the needed part which is guaranteed to be stocked there. The technician does not run out of parts stocked in the kit. Hence, the probability of a job requiring a trip to the FSL is identical for all customers over the course of a day.

Previous work on the repair kit problem assumes that jobs that require parts not stocked in the repair kit are service failures, and ignores how these jobs are ultimately resolved. Since all parts are available to the technician (at the FSL if not in the kit) and the cost of not having a part is a decrease in productivity, we constrain the total time available for the technician to complete repairs. In our simple model, the technician is assigned m jobs to complete by the end of the business day, which lasts d time. A kit job, where all necessary parts are in the kit brought on-site by the technician, requires α time while an FSL job takes β ($> \alpha$) time. We assume that a technician will never abandon a job once he or she discovers that it requires a part not in the repair kit and will instead go retrieve the part from the FSL immediately.

We propose and evaluate six algorithms of varying complexity and effectiveness to find repair kits that meet the service requirement at low cost. Four of the six perform significantly better than the other two and more or less equivalently to one another. We demonstrate that two of the top-performing four are not necessarily optimal, and furthermore that the worst case performance ratio of one of those two cannot be bounded. However, all of the four best algorithms produce solutions with costs about 10% greater than a developed lower bound over the test suite of problem instances.

1.4.2 The Multiple-Job Repair Kit and Technician Routing Problem with Forward Stocking Location Recourse

In the second chapter we extend our work to take into account the geographic relationships between the customers. The amount of time for a kit job or FSL job at a given customer depends on its predecessor in the technician's route as well as the customer's distance from the FSL. The time needed to successfully complete a customer repair is the sum of two components, travel time and service time. The travel time consists of the time it takes the technician to travel from the previous customer site once work there is finished to the current customer, and the round trip time to the FSL should such a visit be warranted. The service time represents the time needed to diagnose the malfunction, remove the inoperable part, and replace it with a working spare. The travel time from the previous customer and the service time require the same duration for kit and FSL jobs. The extra travel time for a customer is zero for kit jobs and the round trip travel time to the FSL for FSL jobs.

This new, richer model requires the decision maker to choose which parts to stock in the repair kit as well as how to route the technician to complete the jobs in a given day. The route chosen has a direct impact on the productivity of the technician and the number of jobs that can be filled in a day, which itself has a direct impact on the number of parts needed in the kit to meet the service level requirements. We show through computational testing that the indirect impact of routing on repair kit inventory is significant.

We propose six methods of varying complexity and effectiveness for routing the technicians to customer job sites and reuse the four best inventory algorithms from the first part to create twenty-four ways for designing a repair kit. We compare the solutions to determine how much cheaper inventory costs are for good routing policies than for bad ones. Our results suggest that repair kits chosen with random routing are at least twenty percent more expensive than those chosen with any non-random routing method. Somewhat surprisingly, the simplest routing heuristic captures almost all of the benefits of non-random routing, and more sophisticated heuristics do not bring much if any further payoff in terms of inventory savings.

1.4.3 Design of a Spare Parts Supply Chain with Forward Stocking Location Recourse

The third part of the thesis examines the strategic design of a spare parts supply chain with inventory held in technician repair kits as well as in FSLs. The decision maker must choose the number of FSLs, the number of technicians, and the allocation of part types between repair kit and FSL with the objective to meet the required level of service at minimum total cost per repair job. It is again assumed that every part type is stocked either in the repair kit or at the FSL. The supply chain must be capable of handling a certain average number of repair jobs on a per-period basis. Each technician is assigned the same number of jobs per period, which depends on the total number of technicians employed. The cost of stocking a part type in the repair kit depends on the number of technicians employed and likewise the FSL cost depends on the number of FSLs. For a given combination of number of technicians and number of FSLs, the problem of deciding where to stock part types can be solved with the algorithms from part one of the thesis.

The motivation for the problem studied in this chapter is the situation faced by a service provider whose existing supply chain structure does not contain any supplementary inventory locations, but wants to evaluate whether to implement an FSL strategy. In reality, the customers requiring repair change on a daily basis but modeling this transience exactly is too granular for a first attempt at representing the problem. Instead, we approximate the times required for the two repair job outcomes with constant parameter values in order to make closed-form analysis of system performance measures, *i.e.*, expected fill rate, possible.

The feasible set of decision alternatives is similar to the corresponding feasible set from the first part of the thesis but has two more dimensions, those for the number of technicians and number of FSLs. Given a rough target number of repair jobs to be completed in a single period and the average productivity for a technician to complete the two types of jobs, the feasible set of the number of technicians, which must be integer, is quite small. The number of FSLs, also an integer variable, is similarly small, covering the integers from zero to the number of technicians employed. Taking these observations into account, an exhaustive algorithm is proposed that reuses algorithms from the first part to find good inventory policies for all feasible combinations of technicians and FSLs, and chooses the supply chain configuration with the lowest total cost per job.

The algorithms were tested on a suite of problem instances with key problem parameters varied over reasonable ranges. The test results showed that the best spare parts supply chain with FSL recourse has either fewer but busy technicians with close-to-full kits or more but less busy technicians with empty kits, the former when inventory is cheaper relative to labor and the latter otherwise. When inventory is around four times more expensive than labor, the best supply chain configuration falls somewhere between the two extremes. The latter structure may be practically infeasible, but it offers savings of 10% to 30% over a no-FSL strategy.

CHAPTER II

THE MULTIPLE-JOB REPAIR KIT PROBLEM WITH FORWARD STOCKING LOCATION RECOURSE

2.1 Introduction

There are many industries which utilize electronic and mechanical equipment installed at various locations throughout a geographic region. The proper functioning of such machinery is vital to the business operations of its purchaser. When a machine ceases working as intended, its owner requests that a technician, employed by the manufacturer or a third-party repair orgainzation, come repair the machine to working order. In many cases these repairs require replacement of defective parts with spare ones, although only the technician can determine in-person which parts to switch out. Since the machines are located at customer sites, the technician would like to diagnose the problem and complete the repair in a single trip, which necessitates carrying around an array of commonly used spare parts. If the repair requires a part not carried by the technician, he or she must retrieve the part from a central warehouse or order it to be shipped overnight and then return to the customer site at a later date to complete the repair. This latter outcome is less than ideal because it results in worse service for the customer, whose equipment remains inoperable for a longer period of time, and more time spent on the repair by the technician, reducing time available to service repair jobs for other customers in the region.

One strategy that some providers adopt is to store extra inventory in fixed locations, *e.g.*, self-storage rental units, throughout the region so that there is recourse inventory available close by to technicians completing repairs at customer sites. These locations allow a technician to complete a job (time-permitting) for which he or she does not have the necessary parts in the vehicle on a same-day schedule, and in less additional time than it would take to travel to the regional warehouse to retrieve a part. All technicians who work in the region have access to the inventory at these locations, so storing parts there is less costly than including them on every service vehicle. The benefits of this strategy are improved service for the customers and less time spent by technicians to complete repairs. The costs clearly are the additional holding costs for inventory stored at the forward stocking locations (FSLs).

In this chapter, we introduce a new model to represent this service strategy and provide six algorithms to solve the decision problem of which part types to stock in the repair kit and which in the FSL. Four algorithms generate good but not necessarily optimal solutions to the problem, as confirmed by comparing their solution values to a lower bound developed in the chapter. In Section 4.2, we review the previous work on the repair kit problem and in Section 4.3 we develop our new model of the repair process as well as a lower bound on the true optimal solution. We describe six algorithms to determine inventory policy in Section 4.4 with some theoretical results about two algorithms following in Section 2.5. In Section 4.5, we demonstrate the effectiveness of the four best performing algorithms using both hypothetical problem instances and also instances derived from real-world data. Furthermore, we report the average cost savings from using the FSL strategy before finally concluding in Section 4.6.

2.2 Related Literature

Although our contribution extends previous research into the repair kit problem, part of what we have done is closely related to the area of order fill rates in multi-item, base-stock inventory systems. First, we briefly address the analysis of order fill rates for multi-item, base-stock inventory systems and argue why our model necessitates something distinct from existing results in this area. Then we review the history of the repair kit problem and explain how our work adds to the literature.

2.2.1 Order fill rate literature

As part of our model we must determine the job fill rate in a multi-item, base-stock inventory system with replenishment at constant time points. There is previous work by other authors that at first glance seems adaptable to our needs but we argue to the contrary.

Song [67] derives the immediate order fill rate for a continuous-time, multi-item, base-stock inventory system in which replenishment lead times are constant. Song and Yao [68] extends the analysis by relaxing the constant replenishment lead time assumption with stochastic lead times. These papers model demand for groups of items (kits of parts) as Poisson processes and determine the probabilities of each type of demand being satisfied immediately from on-hand inventory.

These demand-type fill rates could be aggregated to form an overall job fill rate using the probabilities that a job is of a certain demand type. However, the fill rates depend only on part availability and are not affected by the available time of a technician, which is essential to the model developed in this chapter. This difference is enough to make the order fill rates developed in Song [67] and Song and Yao [68] unusable for our purposes.

The job fill rate we derive is distinct from anything we are aware of but may be limited in applicability to the special type of model we develop in this chapter.

2.2.2 Repair kit literature

Smith et al. [66] initiated the modern approach to thinking about the repair kit problem. The authors were the first to model job fill rate as opposed to part fill rate. They consider a single-period problem with part types that fail independently and no more than one per job. The authors minimize an unconstrained cost function that includes carrying cost and penalty cost for failed jobs and show that the optimal policy takes a certain nested form. Hence, enumeration can be used to solve the problem efficiently to optimality.

Graves [42] makes the same assumptions as Smith et al. [66] about the problem setting, but minimizes holding cost subject to satisfying a specified job fill rate. The structure of the probabilistic constraint allows the problem to be transformed into a 0/1 knapsack problem, which can be solved with efficient and effective heuristic algorithms. To its advantage, this model does not require a penalty cost parameter and can generate a wider range of solutions.

Mamer and Smith [56] relaxes the assumption of one part failure per job for its model and shows that unconstrained minimization of a total expected yearly cost function is analogous to a selection problem, which, as Balinski [6] demonstrates, can formulated as a network problem. Thus, the optimal stocking policy can be found with a max flow/min cut algorithm. (A selection problem seeks to maximize the profit of a chosen set of subsets of nodes from a graph, where each subset of nodes has a positive profit but the individual nodes themselves have costs.)

Heeremans and Gelders [45] is the first to consider the setting in which the repair technician must visit multiple job sites between inventory replenishments. The authors assume, in addition to independent failures, that at most one part of each type is needed on a single job. While minimizing total cost, however, they constrain the probability of completing an entire tour without running out of stock (tour fill rate) rather than a job fill rate as perceived by the customer. Because their formulation is quite complex, the authors propose a simple knapsack heuristic to solve it.

Teunter [71] extends the work of Heeremans and Gelders [45] by modifying its knapsack heuristic to indeed use job fill rate in determining service level. Such calculations are difficult, especially in more general settings, so Teunter [71] also proposes a second heuristic that uses part fill rate in place of job fill rate.

Bijvank et al. [13] derive a closed-form expression for expected job fill rate in a general multi-period setting where one or more units of a part type may fail and inventory from the kit is not set aside for a job that cannot be completed. The authors also provide a modified greedy algorithm with a local search-type improving step that comes very close to optimal in numerical experiments.

Our contribution to the repair kit literature is to redefine what it means for a job to be a service failure (which requires altering the problem setting slightly) in an effort to evaluate the usefulness of an alternative service strategy. Instead of considering a job failed if at least one required part is not in the kit, we sometimes allow the repair technician to complete such a job successfully but with a time penalty corresponding to a round trip from the job site to an FSL. Clearly, these modifications reframe the problem and introduce additional parameters, but the resulting models provide insight into the value of the FSL strategy.

2.3 The Multiple-Job Repair Kit Problem with Forward Stocking Location Recourse

The decision maker responsible for the spare parts supply chain must choose a stocking policy, which specifies which part types will be stocked in the repair kit (represented by $x_i = 1$ for part types $i = 1, 2, \ldots, n$). We assume that all part types not stocked in the repair kit are stocked in the FSL.

There are two measures of primary interest for a stocking policy in the repair kit problem: the expected job fill rate and the total inventory holding cost. In our model, we take the total inventory holding cost as the objective function and constrain the job fill rate. This type of formulation is not the only arrangement used to model the repair kit problem, but it has the benefit of being intuitive and reflects how the problem is viewed in practice. The goal is to minimize total inventory holding cost of the service parts stocking policy, while supporting a job fill rate no less than a required minimum:

$$\begin{aligned} \min \quad & \text{total inventory holding cost} \\ \text{s.t.} \quad & \text{job fill rate} \geq \text{required fill rate} \end{aligned}$$

It remains to define total inventory holding cost and job fill rate precisely in the setting with FSL recourse.

Let n be the number of distinct part types installed in customer machines located throughout the region of interest. If the technician does not have the part type in the kit necessary to complete a job, he or she will make a round-trip to the nearest FSL to retrieve it. The important question to answer at this point is how to classify a job that requires a trip to the FSL. If we consider it always a service failure, then our problem is no different than ones studied previously by other authors. On the other hand, if we consider it always a success, then the optimal solution is to stock all part types at the FSL (since stocking at the FSL is cheaper than stocking in the repair kit), which is not realistic because the jobs that need to be resolved in a day will take too long.

We introduce an element of time into the model to represent this aspect of the real situation in practice. Let m be the number of jobs assigned for the technician to complete in a single day, which we say is d time units long. A job for which all necessary parts are in the repair kit (*i.e.*, a kit job) takes the technician α time to complete, while one that requires a round-trip to the FSL (*i.e.*, an FSL job) takes β time. We consider a job a service failure if it is not completed by time d. In order for the problem to be interesting, we require $m\alpha < d < m\beta$ so that all jobs would be successful if they are kit jobs but not so if all jobs are FSL jobs. In this scenario, jobs (kit or FSL) are only successes if they can be completed prior to d, which depends on the prior jobs serviced during the technician's day.

We assume that exactly one part type fails on each job (with p_i the probability

Table 1: Model notation for deterministic parameters

Time for kit job	α
Time for FSL job	β
Number of jobs per day	m
Length of day	d
Minimum fill rate	γ
Cost to stock part i in kit	c_i^k
Cost to stock part i in FSL	c_i^f
Probability part i is needed for a job	p_i
Binary decision to stock part i in the kit	x_i

Table 2: Model notation for random variables

Number of service failures	F
Time to complete job j	T_j
Elapsed time at which job j is completed	$\mathcal{T}_j$

that part type i is the one that does so), independent of the outcomes of previous jobs. We also assume that the technician will never suffer inventory stock-outs within either the kit or the FSL, which means the probabilities of kit or FSL for the jobs in a replenishment cycle are independent and identically distributed.

2.3.1 Job fill rate

A job will be considered a failure if it is not completed by time d. Observe that the probabilities of individual jobs failing are not independent of one another; in fact they, are almost entirely dependent. If a job fails, then all subsequent jobs scheduled after it fail as well. If a job is completed before but close to time d, there is a chance that the next job (and thus all remaining jobs) will fail. The expected job fill rate,

$$\phi(x_1, x_2, \ldots, x_n) = 1 - \frac{E[F]}{m},$$

is the complement of the expected job failure rate, where $E[F]$ is the expected number of failed jobs. Note that we assume technicians will complete all jobs in the scheduled order, whether they require trips to the FSL or not. More advanced models where technicians could choose to skip jobs (as failures) when they are revealed as FSL jobs are not considered here.

The number of failed jobs, F, is a random variable that takes its value in $\{0, 1, 2, \ldots, m\}$ and hence we can calculate its expected value, $E[F]$, as follows in Equations 1, 2, and 3.

$$E[F] = \sum_{j=0}^{\infty} P(F > j) \tag{1}$$

$$= \sum_{j=0}^{m-1} P(F > j) \tag{2}$$

$$= m - \sum_{j=0}^{m-1} P(F \leq j) \tag{3}$$

When we use this expression in the expected fill rate expression, we obtain Equation 4.

$$\begin{aligned} \phi(x_1, x_2, \ldots, x_n) &= 1 - \frac{m - \sum_{j=0}^{m-1} P(F \leq j)}{m} \\ &= \frac{\sum_{j=0}^{m-1} P(F \leq j)}{m} \end{aligned} \tag{4}$$

The event that no more than j jobs fail during a period is also the event that the $(m-j)$th job is completed by the deadline, and hence we have Equation 5, where $\mathcal{T}_{m-j}$ is the elapsed time from the start of the period at which job $m-j$ is completed.

$$P(F \leq j) = P(\mathcal{T}_{m-j} \leq d) \tag{5}$$

The expected job fill rate is then what can be seen below in Equation 6.

$$\phi(x_1, x_2, \ldots, x_n) = \frac{1}{m} \sum_{j=0}^{m-1} P(\mathcal{T}_{m-j} \leq d) \tag{6}$$

The time $\mathcal{T}_i$ is a binomial-type random variable that takes one of at most $i+1$ distinct values from the set $\{i\alpha, (i-1)\alpha+\beta, \ldots, i\beta\}$, depending on the particular combination of kit jobs and FSL jobs among the first i. The probability mass function of $\mathcal{T}_i$ can be seen in Table 3 and its cumulative distribution function in Table 4.

Table 3: Probability mass function for $\mathcal{T}_i$

t	$P(\mathcal{T}_i = t)$
$i\alpha$	$P(\text{Kit})^i$
$(i-1)\alpha + \beta$	$\binom{i}{1} P(\text{Kit})^{i-1}(1 - P(\text{Kit}))^1$
$(i-2)\alpha + 2\beta$	$\binom{i}{2} P(\text{Kit})^{i-2}(1 - P(\text{Kit}))^2$
$\vdots$	$\vdots$
$\alpha + (i-1)\beta$	$\binom{i}{i-1} P(\text{Kit})(1 - P(\text{Kit}))^{i-1}$
$i\beta$	$(1 - P(\text{Kit}))^i$

Table 4: Cumulative distribution function for $\mathcal{T}_i$

$t \in$	$P(\mathcal{T}_i \leq t)$
$(-\infty, i\alpha)$	0
$[i\alpha, (i-1)\alpha + \beta)$	$P(\text{Kit})^i$
$[(i-1)\alpha + \beta, (i-2)\alpha + 2\beta)$	$P(\text{Kit})^i + \binom{i}{1} P(\text{Kit})^{i-1}(1 - P(\text{Kit}))^1$
$\vdots$	$\vdots$
$[\alpha + (i-1)\beta, i\beta)$	$\sum_{j=0}^{i-1} \binom{i}{j} P(\text{Kit})^{i-j}(1 - P(\text{Kit}))^j$
$[i\beta, \infty)$	1

Note that $\sum_{i=1}^{n} p_i = 1$ since we assume exactly one part type fails per job. A simple change of index to promote clarity gives us Equation 7 as the expected job fill rate.

$$\phi(x_1, x_2, \ldots, x_n) = \frac{1}{m} \sum_{j=1}^{m} P(\mathcal{T}_j \leq d) \tag{7}$$

2.3.2 Problem formulation

We assume without loss of generality that each part has an associated cost c_i of stocking in the kit and that there is no cost to stocking the part at the FSL. If there were nonzero costs for stocking at the FSL, it would be reasonable to assume that they would be less than the corresponding kit costs due to inventory pooling. Thus, we can transform the problem to an equivalent one with no FSL costs by subtracting the FSL cost from the kit cost for each part. Let c_i^f be the total FSL cost and c_i^k be the kit cost. Our objective function is

$$\sum_i c_i^f (1 - x_i) + \sum_i c_i^k x_i,$$

which is equivalent to

$$\sum_i c_i^f + \sum_i c_i x_i,$$

where $c_i = c_i^k - c_i^f$. If we let γ be the minimum required job fill rate, then we define the optimization model for determining the kit versus FSL stocking policy as follows.

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} c_i x_i \\ \text{s.t.} \quad & \phi(x_1, \ldots, x_n) = \frac{1}{m} \sum_{j=1}^{m} P(\mathcal{T}_j \leq d) \geq \gamma \\ & x_i \in \{0, 1\},\ i = 1, 2, \ldots, n \end{aligned}$$

2.3.3 Lower bound

Unfortunately for the purpose of solving our optimization problem, the expected job fill rate we developed is not a smooth function. Thus we are unable to use a general-purpose math programming algorithm or verify any claims about the optimality of any given solution. On top of that, realistic repair kit problems can have on the order of thousands of part types, which means that brute force enumeration to find the optimal solution is not an efficient strategy. In order to evaluate the effectiveness of the solution methods we propose, it would be nice to have a tractable lower bound on the optimal objective value. We would like to replace the table look-up necessary to calculate the expected number of service failures with a smooth functional form, and we can accomplish this by examining the expected time needed by the technician to complete all assigned jobs.

Let T_i represent the length of time it takes a technician to service the ith job on his docket. Under our assumptions, which imply kit job probabilities are i.i.d. for all jobs, T_i takes the following distribution for $i = 1, 2, \ldots, m$ where $P(\text{Kit}) = \sum_{j=1}^{n} p_j x_j$.

$$P(T_i = t) = \begin{cases} P(\text{Kit}) & t = \alpha \\ 1 - P(\text{Kit}) & t = \beta \\ 0 & \text{otherwise} \end{cases}$$

Hence, we can calculate the expected time it takes a technician to service all customers as follows.

$$\begin{aligned} E[\mathcal{T}_m] &= E\left[\sum_{i=1}^{m} T_i\right] \\ &= m((\alpha - \beta)P(\text{Kit}) + \beta) \end{aligned}$$

We will hold on to this result and use it shortly.

The total amount of additional time a technician needs to complete all assigned jobs successfully is $\sum_{i=1}^{m} T_i - d$. In the worst case for the service provider, this needed time means that $\lceil(\sum_{i=1}^{m} T_i - d)/\alpha\rceil$ jobs fail, *i.e.*, the technician spent much time fixing FSL jobs early in the day and had a backlog of consecutive kit jobs to close the day. In the best case for the service provider, $\lceil(\sum_{i=1}^{m} T_i - d)/\beta\rceil$ jobs fail; this gives us a lower bound without needing to know anything about the order of kit jobs and FSL jobs throughout the day. We can remove the ceiling function to get a more calculation-friendly lower bound on the number of failed jobs.

$$F \geq \frac{\sum_{i=1}^{m} T_i - d}{\beta} = F'$$

Taking the expectation, we have the following lower bound, which is affine in the decision variables.

$$\begin{aligned} E[F] &\geq E[F'] \\ &= E\left[\frac{\sum_{i=1}^{m} T_i - d}{\beta}\right] \\ &= \frac{m\left((\alpha - \beta)\sum_{j=1}^{n} p_j x_j + \beta\right) - d}{\beta} \end{aligned}$$

Plugging this lower bound into the job fill rate constraint in place of the exact calculation for the expected number of failed jobs gives us a relaxation of our original constraint and that relaxation is affine in the decision variables.

$$\begin{aligned}
\phi(x_1, \ldots, x_n) &= 1 - \frac{E[F]}{m} \\
&\leq 1 - \frac{E[F']}{m} \\
&= \frac{\beta - \alpha}{\beta} \sum_{j=1}^{n} p_j x_j + \frac{d}{\beta m} = \phi'(x_1, \ldots, x_n)
\end{aligned}$$

The relaxed problem now takes the following form, after we rearrange the modified constraint, $\phi'(x_1, \ldots, x_n) \geq \gamma$.

$$\begin{aligned}
\min \quad & \sum_{i=1}^{n} c_i x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} p_i x_i \geq \frac{m\beta\gamma - d}{m(\beta - \alpha)} \\
& x_i \in \{0, 1\}, \ i = 1, 2, \ldots, n
\end{aligned}$$

Observe that the relaxation is a traditional 0/1 knapsack problem, which we can solve to optimality with off-the-shelf software for problem instances of realistic size.

2.4 Solution Algorithms

Our exact optimization problem is a 0/1 knapsack program in the sense that the objective function is linear in the decision variables and there is a single constraint. However, we cannot use traditional knapsack algorithms to solve it because the constraint is not a smooth function. Instead, we propose four different ways to find a feasible solution to the problem. The first way is motivated in a similar fashion to the lower bound presented in the previous section and uses a smooth function that is tighter than the expected fill rate constraint. The second method is a greedy marginal-analysis algorithm which takes one part out of the kit at a time until any further removal would take the expected job fill rate below the minimum requirement.

The third approach is a class of algorithms that simplifies the marginal-analysis approach by calculating values for and sorting the part types only at initialization. The fourth method takes advantage of the fact that only the probability of a kit job (regardless of individual parts within that repair kit) is needed to calculate the expected fill rate to solve the problem with traditional knapsack algorithms. In Section 4.5, we show that the proposed greedy algorithm, two algorithms from the third approach, and the algorithm from the fourth method are on par with one another and better than the other two algorithms.

2.4.1 Smooth constraint algorithm (SCA)

Recall from Section 2.3.3 that in the worst case for the service provider, $\lceil(\sum_{i=1}^{m} T_i - d)/\alpha\rceil$ jobs fail in a single period.

$$F \leq \left\lceil \frac{\sum_{i=1}^{m} T_i - d}{\alpha} \right\rceil$$

If we remove the ceiling function we are not guaranteed that the right hand side of the above inequality will be no less than the actual number of failures. However, if we remove the ceiling function and add one then our guarantee remains and we have an upper bound on the number of failures.

$$F \leq \frac{\sum_{i=1}^{m} T_i - d}{\alpha} + 1$$

Now we have a smooth functional form and we can get an upper bound on the expected number of failures as follows.

$$\begin{aligned} E[F] &\leq E\left[\frac{\sum_{i=1}^{m} T_i - d}{\alpha} + 1\right] \\ &= \frac{m\left((\alpha - \beta)\sum_{j=1}^{n} p_j x_j + \beta\right) - d}{\alpha} + 1 \end{aligned}$$

We can use this expression in the fill rate expression to get a lower bound on the expected fill rate.

$$\begin{aligned}\phi(x_1,\ldots,x_n) &= 1-\frac{E[F]}{m}\\ &\geq 1-\frac{m\left((\alpha-\beta)\sum_{j=1}^{n}p_jx_j+\beta\right)-d}{\alpha m}-\frac{1}{m}\\ &= 1+\frac{\beta-\alpha}{\alpha}\sum_{j=1}^{n}p_jx_j-\frac{\beta}{\alpha}+\frac{d}{\alpha m}-\frac{1}{m}\end{aligned}$$

After substituting this lower bound in place of the expected fill rate for our optimization problem constraint and rearranging, we get the following 0-1 knapsack problem.

$$\begin{aligned}\min \quad & \sum_{i=1}^{n} c_i x_i\\ \text{s.t.} \quad & \sum_{i=1}^{n} p_i x_i \geq \frac{\alpha(m\gamma+1)-d}{m(\beta-\alpha)}+1\\ & x_i \in \{0,1\},\ i=1,2,\ldots,n\end{aligned}$$

We can solve this problem quickly for instances of interest to practitioners to get a feasible inventory policy and hence an upper bound on the true optimal cost.

2.4.2 Subtractive greedy algorithm (SGA)

The aim of a greedy marginal-analysis algorithm is to get the most "bang for the buck" at each move and iterate until feasibility is reached (or breached, as in our case). In our problem, the "bang" corresponds to the decrease in total cost between two solutions or inventory policies while the "buck" is the decrease in job fill rate.

We propose what we call a subtractive greedy algorithm wherein we initialize the algorithm with a solution that stocks all parts in the kit. At each iteration we calculate (for the relevant parts) the decrease in job fill rate from moving the part out of the kit to the FSL and, so long as this drop does not exceed the slack between the current job fill rate and the minimum fill rate, divide this into the decrease in holding cost from such a move. We pivot to the solution with the greatest such ratio in an iteration and repeat the process until there are no more feasible moves. Please see Algorithm 3 for pseudocode of the subtractive greedy algorithm.

Algorithm 1 Subtractive greedy algorithm

```
all parts in the kit
repeat
  clear list
  for each part in the kit do
    if moving part from kit to FSL is feasible then
      calculate ratio and add move to list
    end if
  end for
  pivot solution by move on list with highest ratio
until no feasible moves on list
```

2.4.3 Threshold algorithms

The SGA calculates a score for each part still remaining in the kit at each iteration and removes the part with the best score. Implicit in the algorithm is a sorting process based on the calculated score. Once a part is removed the scores must be recalculated and the parts sorted again by their new score values. This recalculation process requires additional computational effort but the marginal value of that effort has not been assessed. We also propose an alternative algorithm structure that sorts the parts one time at the initialization and then iterates through the list, removing or adding parts as desired without recalculating the part scores after every step but instead evaluating the parts in their original sorted order. The parts can be added until the fill rate is met, or alternatively sorted in the opposite order and removed until just before the fill rate is violated. Thus, in the worst case, n fill rate calculations are needed for a threshold algorithm, which is a stark contrast from the SGA where $n(n+1)/2$ may be used in the worst case.

The criterion used to sort the parts can be anything the supply chain manager thinks may be a reasonable ordering. In this chapter, three threshold algorithms are evaluated, with the following initial sorting criteria:

1. failure probability,

2. ratio calculated in the first iteration of the SGA, and

3. ratio of failure probability to kit minus FSL cost differential.

The first criterion approximates the total throughput of a part, which a service provider might consider using to partition parts into kit and FSL locations. We refer to its algorithm as the Failure Probability Threshold Algorithm (FPTA). The second criterion is a simplified version of the SGA and can be compared to those solutions on the same instances to see the extra inventory cost that can be saved, *i.e.*, value added, by recalculating the ratios of interest at every iteration of the SGA. We refer to the threshold algorithm using this criterion as the Initial Ratio Threshold Algorithm (IRTA). Finally, the third criterion was motivated by examining two-dimensional scatterplots of which parts are in the kit and which in the FSL in an SGA solution with failure probability on one axis and kit minus FSL cost differential on the other. On such plots it seems possible to draw a line through the origin that partitions the parts into kit and FSL classification. Achieving such a solution with fewer calculations than the SGA is possible with a threshold algorithm where the parts are sorted based on the ratio of their two values for the parameters on the axes, that is, kit minus FSL cost differential and failure probability. We refer to this final threshold algorithm as the Probability Cost Threshold Algorithm (PCTA). Please see Algorithm 2 for pseudocode of an additive threshold algorithm.

Algorithm 2 Additive threshold algorithm

```
no parts in the kit
sort all parts by criterion with best-criterion parts at top of list
while job fill rate less than minimum required rate do
   add part on top of list to kit and remove part from list
end while
```

2.4.4 Kit Probability Search plus Knapsack (KPSK)

The final method of finding a solution that we will investigate in this chapter has two phases. In the first phase, the minimal kit job probability needed to satisfy the minimum job fill rate is determined with a binary search, and in the second phase this probability is used as a the right-hand-side value of a minimum "weight" constraint in a 0/1 knapsack problem which can be solved by commerical software packages. Such an algorithm is made possible by the fact that the expected job fill rate of a given repair kit depends only on the the probability of a needed part being in the repair kit, *i.e.*, the sum of failure probabilities for parts in the kit, and not on the individual parts themselves that make up the repair kit. Once this critical value is found (which can be effectively accomplished with binary search) a standard 0/1 knapsack problem can be formulated and then solved with state-of-the-art commercial software. Let p_{min} represent the critical kit job probability value and the knapsack problem is as follows.

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} c_i x_i \\ \text{s.t.} \quad & p_i x_i \geq p_{min} \\ & x_i \in \{0,1\},\ i = 1, 2, \ldots, n \end{aligned}$$

2.5 Theoretical Algorithm Performance

There are three claims that we can make and verify about the algorithms proposed in the previous section.

Proposition 1. *The Subtractive Greedy Algorithm does not always return the optimal solution.*

Proposition 2. *The Subtractive Greedy Algorithm has an infinite worst-case bound.*

Proposition 3. *The Probability Cost Threshold Algorithm does not always return the optimal solution.*

The proofs of these three propositions follow.

2.5.1 SGA is not always optimal

We show that the subtractive greedy algorithm does not always reach the optimal solution via a counterexample for the smallest of problem settings, with two part types and two jobs. In this setting, there are four possible solutions and the algorithm iterates at most twice, which means there is exactly one way to miss the optimal solution. If the only feasible solution is to stock both parts in the kit, then the algorithm will reach this trivial solution. If it is feasible to stock neither part in the kit, then the algorithm will terminate at this cheapest solution because it will select the move with the best ratio in the first step and then it will select to move the other part in the second step. If only one of the intermediate solutions, *i.e.*, where one part is stocked in the kit but the other not, is feasible, then the algorithm will also terminate at this optimal solution after reaching it in the first iteration. However, if both intermediate solutions are feasible but the empty solution is infeasible, then it is possible for the algorithm to pivot to the intermediate solution with higher cost and subsequently terminate at this solution. Note that there is no way for the algorithm to put a part back into the kit after taking it out in a preceding step.

2.5.1.1 Conditions for incorrectness

To put our explanation about the subtractive greedy algorithm's failure to reach the optimal solution into mathematical terms, we have the following necessary and sufficient conditions for incorrectness (without loss of generality we will assume that $\text{cost}(1,0) < \text{cost}(0,1)$ or $c_1 < c_2$):

1. $\text{fill rate}(0,0) < \gamma \leq \min\{\text{fill rate}(1,0), \text{fill rate}(0,1)\}$

2. $\text{ratio}((1,1) \to (1,0)) < \text{ratio}((1,1) \to (0,1))$

When these conditions are satisfied the algorithm will choose to move to $(0,1)$ in the first iteration and then terminate in the second iteration because there are no more

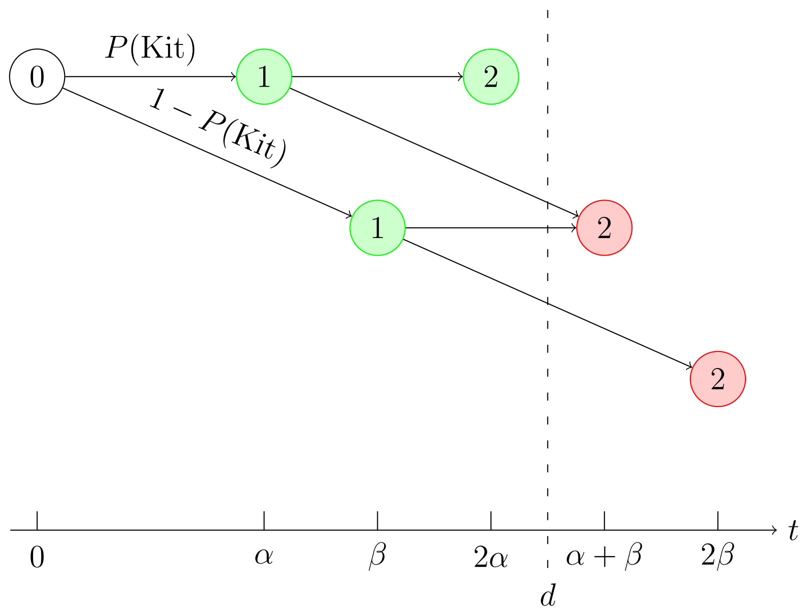

Figure 1: Representation of two-job setting where exactly one job fails when at least one job requires a part not stocked in the kit

feasible moves at that point. The optimal solution, however, is $(1, 0)$, which will not be reached because its ratio during the first iteration is smaller than that of $(0, 1)$.

The two-job setting is only of interest to study when exactly one job is at risk of failing. Clearly, when it is impossible for a single job to fail the solution is trivial. In order for both jobs to fail, the first job must fail and the second job fails as a consequence, but such a situation is impossible given our assumptions. With $2\alpha < d$, the time for an FSL job, β cannot exceed d without violating our assumption of no abandoment and hence the first job will never fail.

There are two cases, depending on the parameters α, β, and d, for exactly one possible job failure. When $\alpha + \beta > d$, one job will fail when at least one job requires a trip to the FSL. See Figure 1 for a representation of this case. When $\alpha + \beta \leq d$ but $2\beta > d$, one job will fail when both jobs require trips to the FSL. We will derive the conditions for incorrectness in the first case and omit the second case due to its similarity.

Table 5: Characteristics of possible solutions for first interesting case of two-part, two-job setting

Solution	Fill Rate	Cost
$(1,1)$	1	$c_1 + c_2$
$(1,0)$	$1 - \frac{1-p_1^2}{2}$	c_1
$(0,1)$	$1 - \frac{1-p_2^2}{2}$	c_2
$(0,0)$	$\frac{1}{2}$	0

The expected number of service failures in the first case is

$$\begin{aligned} E[F] &= \sum_{i=0}^{2} iP(i \text{ jobs fail}) \\ &= P(1 \text{ job fails}) \\ &= 1 - P(\text{Kit, Kit}) \\ &= 1 - P(\text{Kit})^2 \\ &= 1 - (p_1x_1 + p_2x_2)^2 \end{aligned}$$

and thus the job fill rate is

$$1 - \frac{E[F]}{2} = 1 - \frac{1 - (p_1x_1 + p_2x_2)^2}{2}.$$

We summarize the characteristics of the four possible solutions in Table 5. The ratio for the move $(1,1) \to (1,0)$ in the first iteration is

$$\frac{\Delta \text{ cost}}{\Delta \text{ fill rate}} = \frac{c_1 - (c_1 + c_2)}{1 - \frac{1-p_1^2}{2} - 1} = \frac{2c_2}{1 - p_1^2}.$$

The ratio for the move $(1,1) \to (0,1)$ in the first iteration is

$$\frac{\Delta \text{ cost}}{\Delta \text{ fill rate}} = \frac{c_2 - (c_1 + c_2)}{1 - \frac{1-p_2^2}{2} - 1} = \frac{2c_1}{1 - p_2^2}.$$

The two conditions in terms of the problem parameters for this case are:

1. $\frac{1}{2} < \gamma \leq \min\left\{1 - \frac{1-p_1^2}{2}, 1 - \frac{1-p_2^2}{2}\right\}$
2. $\frac{2c_2}{1-p_1^2} < \frac{2c_1}{1-p_2^2}$.

For example, when $p_1 = 0.25$, $p_2 = 0.75$, $c_1 = 1$, $c_2 = 1.5$, and $\gamma \in (0.5, 0.53125]$ the subtractive greedy algorithm willl not reach the optimal solution. The class of problem instances whose parameters satisfy the above conditions provides a counterexample to the claim that the subtractive greedy algorithm always terminates at the optimal solution.

2.5.2 SGA worst-case bound is infinite

Unfortunately from a theoretical perspective, the percent by which the SGA misses the optimal solution can be arbitrarily large, as we now detail. The ratio of the SGA solution cost to the optimal solution value is c_2/c_1. The first condition does not provide any more information when rearranged, assuming that γ can be arbitrarily close to 1/2.

$$\frac{1}{2} < 1 - \frac{1-p_i^2}{2} \Rightarrow p_i^2 > 0, i = 1, 2$$

The second condition for incorrectness can be rearranged as follows.

$$\frac{2c_2}{1-p_1^2} < \frac{2c_1}{1-p_2^2} \Rightarrow \frac{c_2}{c_1} < \frac{1-p_1^2}{1-p_2^2}$$

It is possible to find parameters c_1, c_2, p_1, p_2, and γ such that $p_1 + p_2 = 1$ and both conditions for incorrectness are satisfied and $c_2/c_1 > N$ for any $N \in \mathbb{R}$. After substitution we see that the upper bound on c_2/c_1 can go to infinity in the limit while p_1 and p_2 stay within $(0, 1)$.

$$\lim_{p_2 \to 1} \frac{1-(1-p_2)^2}{1-p_2^2} = \infty$$
$$\lim_{p_1 \to 0} \frac{1-p_1^2}{1-(1-p_1)^2} = \infty$$

The good news is we have a theoretical worst-case bound on the performance of our heuristic, SGA. The bad news is that the bound is not finite.

Table 6: Part parameters

Part	Failure Probability	Kit Minus FSL Cost
1	0.4	1
2	0.2	2
3	0.4	5

2.5.3 PCTA is not always optimal

A threshold algorithm that sorts parts by decreasing ratio of failure probability to kit minus FSL cost difference is not necessarily guaranteed to produce the optimal solution. Consider the following counterexample. A technician is assigned $m = 2$ jobs and has $d = 2.2$ time units in which to complete them. A kit job requires $\alpha = 1$ time unit and an FSL job $\beta = 1.5$. Thus, two kit jobs means two successful jobs and any other outcome means one successful job. There are $n = 3$ parts in this small problem, with parameters as seen in Table 6. The parts must be put into the kit in order of decreasing failure probability to cost difference ratio, which corresponds to increasing numerical order in this example, as illustrated in Figure 2. The minimum required job fill rate is $\gamma = 0.75$. With parts one and two in the kit, the probability of a kit job is 0.6 and the expected number of successful jobs is

$$2(0.6 \cdot 0.6) + 1(1 - 0.6 \cdot 0.6) = 1.36$$

for an expected fill rate of $1.36/2 = 0.68$, which is not sufficient. The threshold algorithm then will not stop until it adds all three parts to the kit, at which the point expected fill rate of 1.0 finally exceeds the minimum required fill rate. However, a satisfactory solution with lower cost would be to include only parts one and three in the repair kit. Such an arrangement has a kit job probability of 0.8 and expected number of successes equal to

$$2(0.8 \cdot 0.8) + 1(1 - 0.8 \cdot 0.8) = 1.64$$

and expected fill rate of 0.82. Hence, we conclude that the threshold algorithm with parts sorted in decreasing order on failure probability to kit minus FSL cost differential does not always provide the optimal solution.

2.6 Computational Testing

Forward stocking locations provide flexibility to spare parts supply chains, which, at worst, is not harmful to the cost-effectiveness of service performance. By how much such flexibility can reduce inventory costs is the major question that we attempted to answer with a computational testing regimen. In order to answer that question, it is necessary to specify how the inventory decisions will be made in the spare parts supply chain with FSL recourse.

2.6.1 Algorithm performance and comparison

In this chapter we have proposed six algorithms (SCA, SGA, FPTA, IRTA, PCTA, and KPSK) to make inventory decisions in spare parts supply chain with FSL recourse. We fed each algorithm the same problem instances to discover which method generates the inventory solutions with lowest cost as well as how much cheaper those lowest-cost solutions are than inventory solutions needed were the FSL not available.

It seems reasonable to speculate that the solutions generated by the proposed algorithms may be sensitive to the values of the parameters, numerous as they are for this model. To investigate this thought, we identified seven problem parameters that we suspected would influence the algorithm solutions, specified a range of levels for these factors, and conducted a full factorial experiment, with ten runs at each factor level. The factors and their levels can be seen in Table 7.

See Figure 3 for boxplots of the solution values for all algorithms relative to the SGA. The value of comparison is the solution cost for the heuristic named in the graph title divided by the solution cost for the SGA on a given problem instance. The SCA was far and away the worst-performing heuristic while the FPTA was clearly worse than the SGA, albeit by around 5% on average. From the latter part we could conclude that selecting parts only based on usage without regard for cost does

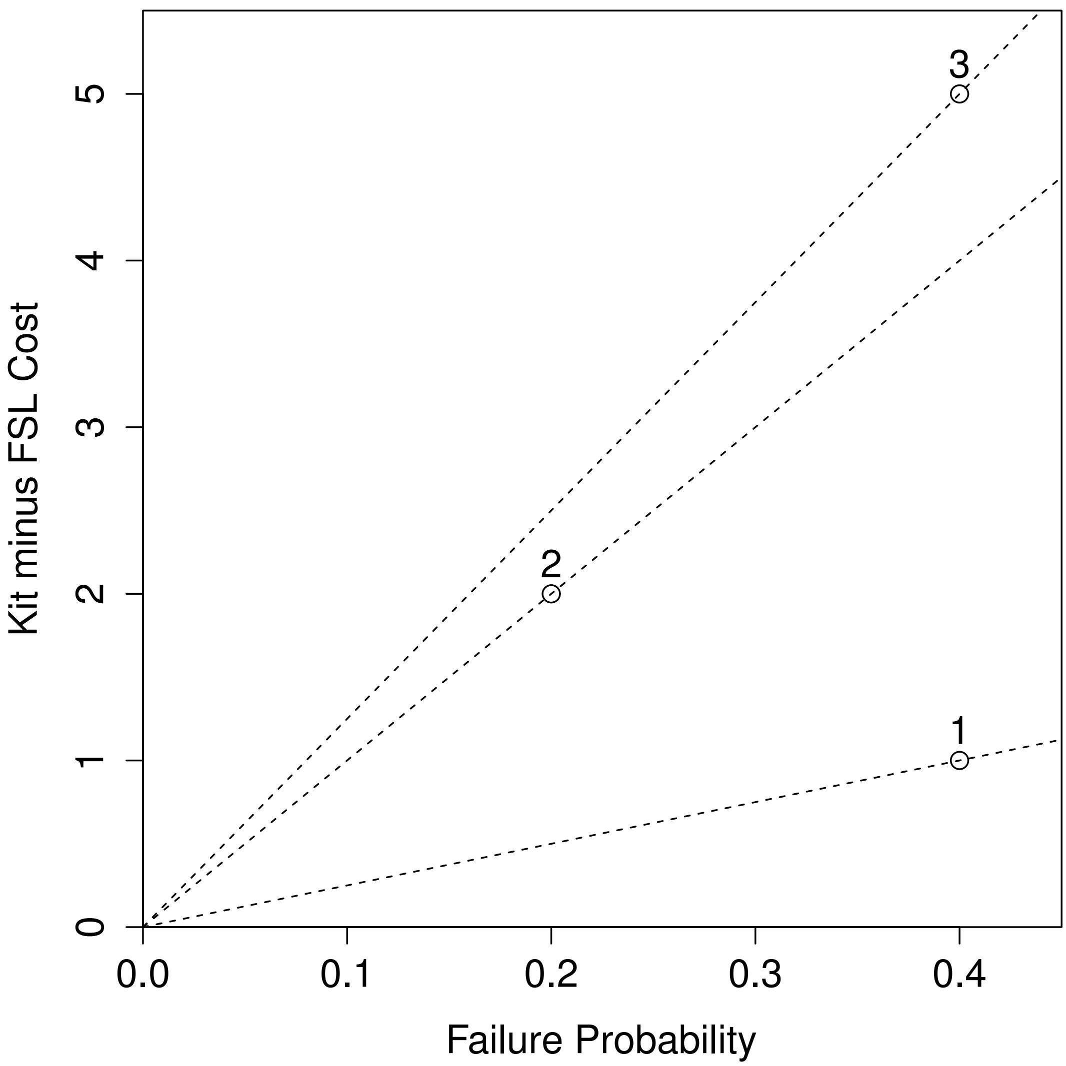

Figure 2: Illustration of counterexample that PCTA is not always optimal

Table 7: Parameters suspected to influence algorithm solutions

Factor	Notation	Levels
Relative time delay for trip to FSL	β/α	$\{1.5, 1.9\}$
Ratio of kit to FSL holding costs	c_i^k/c_i^f	$\{2, 10\}$
Maximum jobs per day	$\lfloor d/\alpha \rfloor$	$\{4, 8\}$
Length of day	d	$\{7, 10\}$
Number of part types	n	$\{100, 2000\}$
Fraction of maximum jobs assigned	$m/\lfloor d/\alpha \rfloor$	$\{0.9, 1.0\}$
Minimum fill rate	γ	$\{0.9, 0.95\}$

not lead to the best solutions available to the decision maker. The remaining three algorithms, IRTA, PCTA, and KPSK, perform quite similarly to the SGA. The IRTA has numerous outlying instances where the SGA outperformed it by a nontrivial percent, which lends support to the argument that reevaluating the criterion ratio at every iteration of the SGA improves the chance of finding a better solution. The PCTA and KPSK give almost identical results to the SGA in all instances tested for this chapter. The lower bound reaches around 90% of the SGA solutions (and hence PCTA and KPSK) on average.

2.6.2 FSL cost savings

We must be precise when comparing inventory policies between the FSL strategy and the conventional strategy because the job fill rate calculation is different between the two strategies. The novel rate we developed for this work corresponds to a setting where technicians can retrieve needed parts from an FSL whereas in the conventional strategy, such recourse is not available and the job fill rate, under our assumption that technicians do not experience stock-outs of parts in the kit, is merely the kit fill rate.

If a job requires parts not in the kit, *i.e.*, is an FSL job in the proposed model, it is considered a service failure under the conventional strategy. Since the probabilities of jobs requiring FSL roundtrips are i.i.d., we have the expected number of failures to be $m \cdot P(\text{FSL})$ and the job fill rate as follows.

$$1 - \frac{m \cdot P(\text{FSL})}{m} = 1 - P(\text{FSL}) = P(\text{Kit}) = \sum_{i=1}^{n} p_i x_i$$

We can solve a similar optimization problem to the one we presented in our formulation where the job fill rate constraint is replaced with a kit job probability constraint and the objective function coefficients are the kit costs as opposed to the difference between the kit costs and the FSL costs. This problem is a 0-1 knapsack problem and can be solved with off-the-shelf software just like the relaxation from Section 2.3.3.

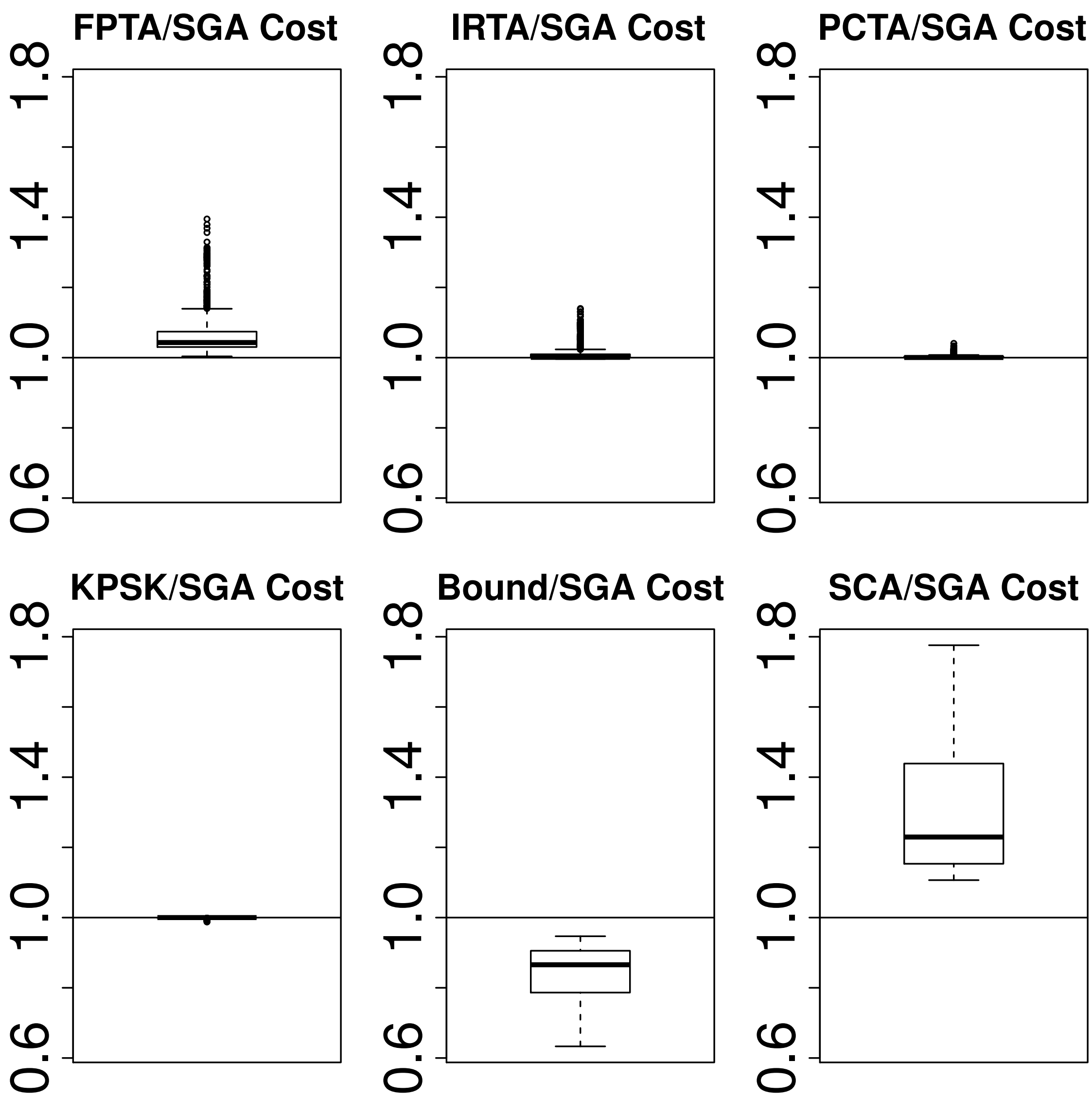

Figure 3: Comparison of algorithm solutions

We must also note that in order to compare the results fairly we must limit ourselves to situations in which the technician cannot reasonably schedule any more jobs in his day. When recourse inventory is available at the FSL the technician must have a little bit of buffer time to make a trip to retrieve it. Otherwise it is pointless to keep this extra stock on hand. When there is no recourse inventory available the technician does not need any buffer time because each customer will take α in the case that the needed part is in the repair kit and less than α if that is not the case. We consider the length of the buffer time (call it b) relative to α as a factor to vary in our testing and do not let it exceed one.

If we examine instances where b is much greater than one (or, equivalently, the fraction of maximum jobs assigned is much less than one) then the FSL side can benefit greatly because there is plenty of time for round trips and hence more parts can be stocked at the lower cost location whereas the no-FSL side will complete its assignment, albeit perhaps missing on some jobs, with plenty of time left over that could have been better used to service more jobs. Implicit in this explanation is the fact that a better picture of cost savings due to the FSL would emerge from a model where the decision maker can choose both the inventory and the number of technicians and number of jobs to assign to them. That is the topic of Chapter 4. For now, we will compromise by looking only at the case where the technicians are scheduled for full utilization. The six factors, including the buffer time, and their levels used for the FSL cost savings tests can be seen in Table 8. A boxplot for the cost savings across all runs can be seen in Figure 4. Boxplots for the cost savings broken down by all factor levels can be seen in Figure 5. Cost savings is defined as

$$1 - \frac{\text{SGA solution cost}}{\text{no-FSL optimal solution cost}}.$$

As Figure 4 illustrates, there are cost savings to the FSL configuration of a spare parts supply chain. It should be noted that the savings in these computational results err on the conservative side because the supply chain manager is limited to kit or

Table 8: Factors and levels for FSL cost savings tests

Factor	Notation	Levels
Length of day	d	$\{7, 10\}$
Jobs per technician	m	$\{4, 8\}$
Buffer relative to kit job time	b/α	$\{0.25, 0.75\}$
Relative time delay for trip to FSL	β/α	$\{1.5, 1.9\}$
Ratio of kit to FSL holding costs	c_i^k/c_i^f	$\{2, 10\}$
Minimum fill rate	γ	$\{0.9, 0.95\}$

FSL stocking for all part types. A true comparison would look at kit, FSL, or neither against kit or neither. However, we limit ourselves to the kit-or-FSL only case for this thesis and do not propose any algorithms for the expanded situation with three possible locations for part types.

The factors with the most significant effects on cost savings were the relative buffer time and relative time delay for FSL trips. The more buffer time available throughout the course of a day and the less extra time needed to retrieve a part from the FSL, the more cost savings can be achieved in a spare parts supply chain by adopting an FSL strategy.

2.6.3 Case Study

We have examined the results of our proposed algorithms in the preceding subsections and shown that our heuristics can work fairly well in generating useful solutions for a wide variety of parameter values. However, we also showed earlier in the chapter that the neither the SGA nor the PCTA were always optimal and that there is no finite cap on how bad the SGA can be in the worst case. It must also be noted that the part instances generated for use in computational testing had parts homogeneously distributed across both failure probability and cost, which may not exactly be the case in a real-world spare parts supply chain. (It is quite possible that certain outlier parts would be excluded in practice from an algorithmic management like the one proposed in this chapter because it is clear where they should be held and therefore

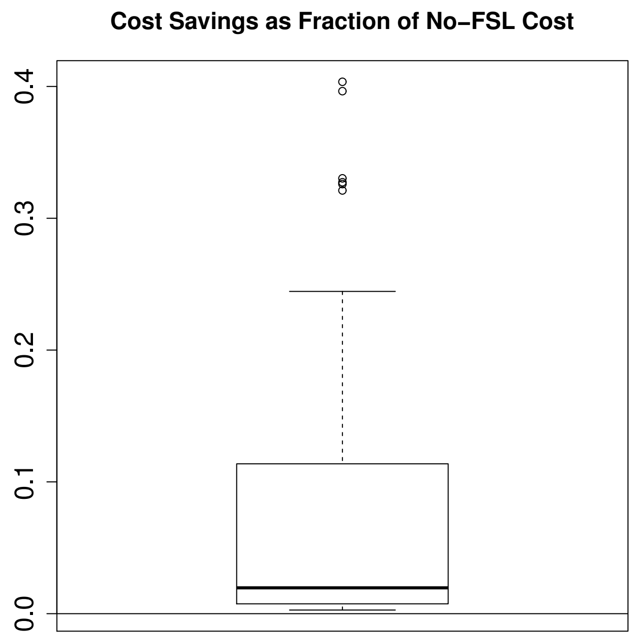

Figure 4: Fraction of inventory cost saved by using FSL strategy

the parts whose best locations are uncertain would be distributed homogeneously, but we digress.) To that end, we also tested our proposed algorithms on real-world spare parts usage data obtained through collaboration with an industry partner.

The information we received from our collaborator contained all resolved customer service jobs from a year-long period in a metropolitan area. Each job entry listed one part number, the one for which a spare version was used to complete the repair, as well as the cost for a part of that type. To approximate the failure probability of a part number, we divided the number of jobs that required that part number by the total number of resolved jobs in the file. These two values, failure probability and unit cost, are enough to define the characteristics of the spare part population for our purposes. Figure 6 shows a scatterplot of the characteristics for all 4636 part types. Of course, there are other parameters outside of the part characteristics needed to define a complete problem instance, and those parameters and the values over which we varied them can be seen in Table 9.

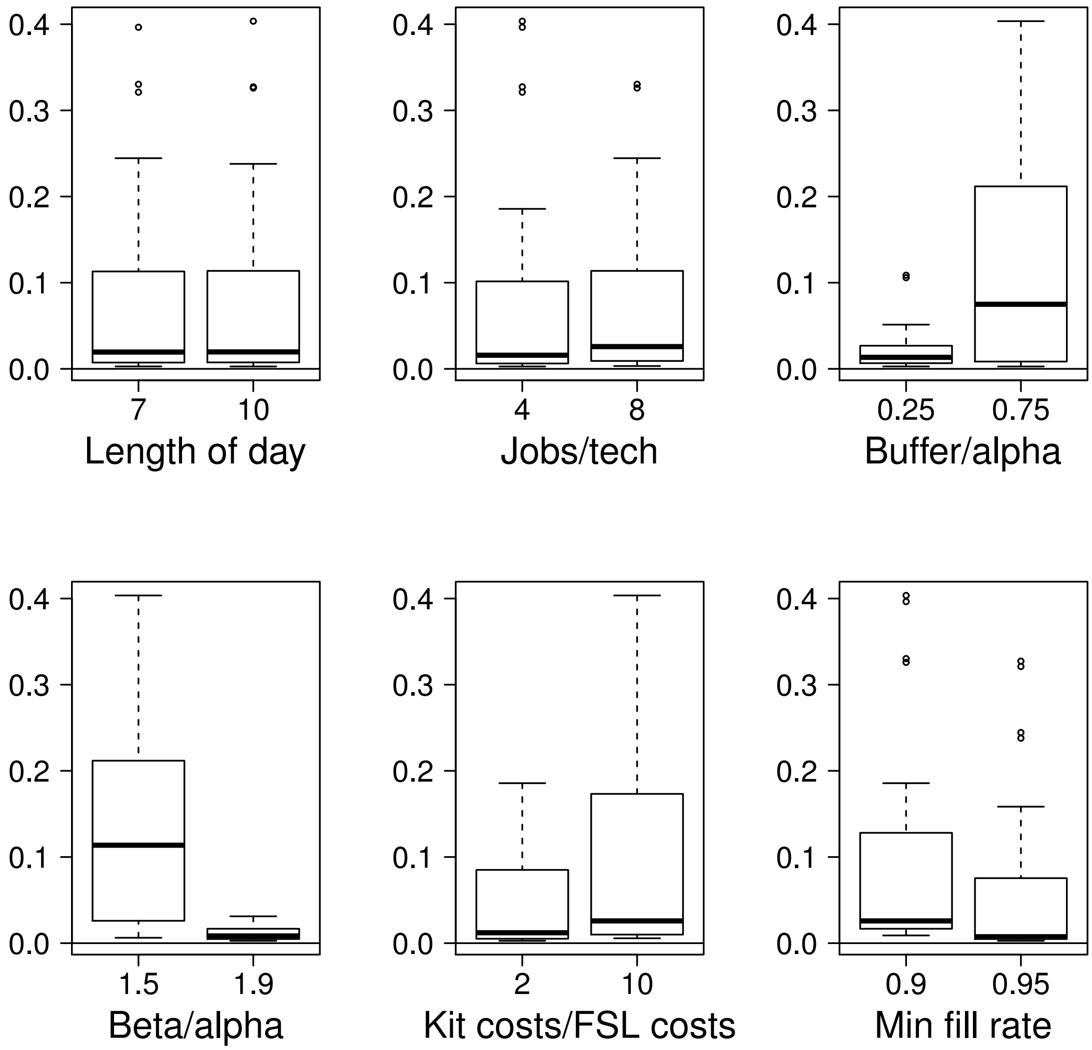

Figure 5: Fraction of inventory cost saved by using FSL strategy, broken down by factor level

Table 9: Parameters used in case study to compare algorithm solutions

Factor	Notation	Levels
Relative time delay for trip to FSL	β/α	$\{1.5, 1.9\}$
Ratio of kit to FSL holding costs	c_i^k/c_i^f	$\{2, 10\}$
Maximum jobs per day	$\lfloor d/\alpha \rfloor$	$\{4, 8\}$
Length of day	d	$\{7, 10\}$
Fraction of maximum jobs assigned	$m/\lfloor d/\alpha \rfloor$	$\{0.9, 1.0\}$
Minimum fill rate	γ	$\{0.9, 0.95\}$

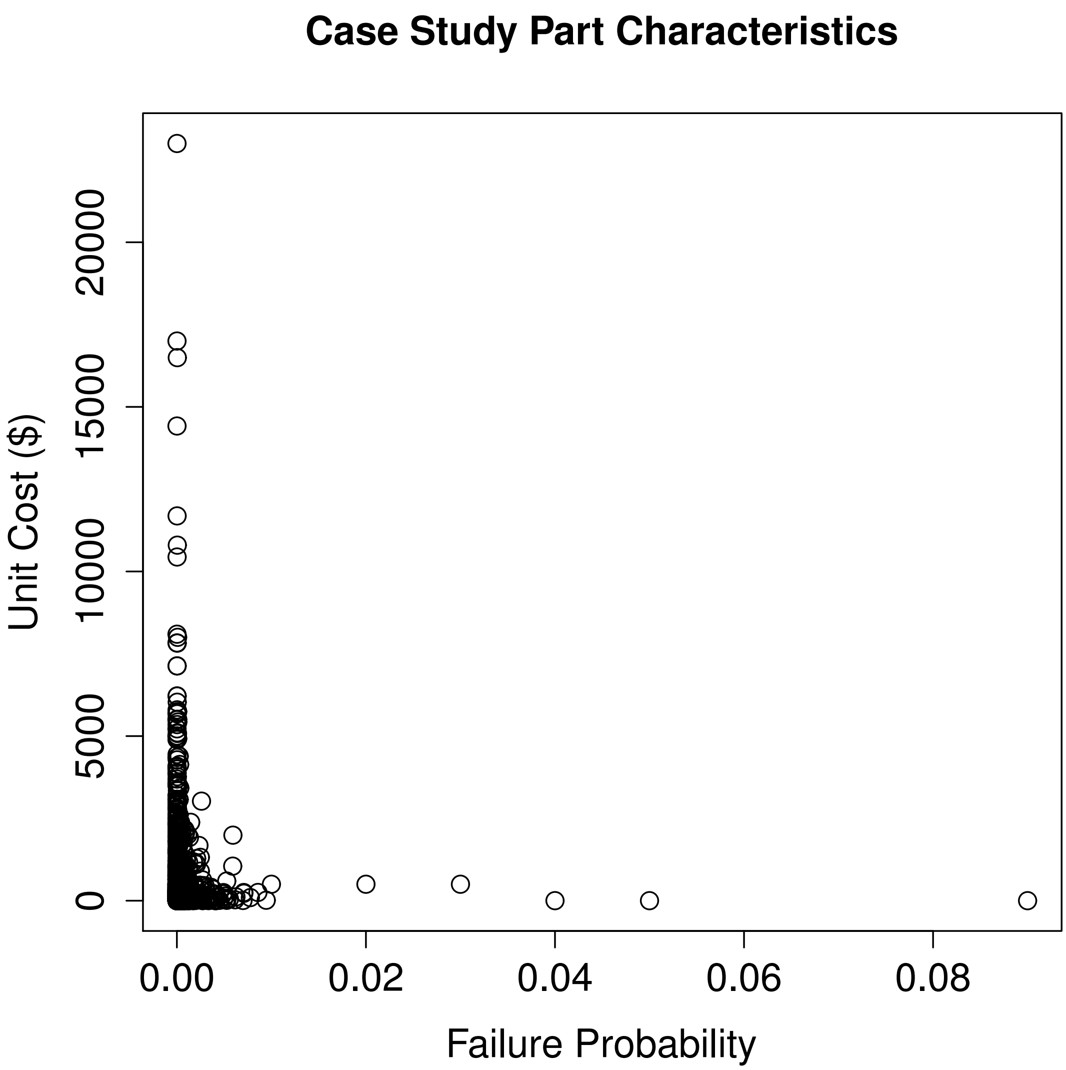

Figure 6: Characteristics of spare parts in case study

The results of the tests can be seen in Figure 7, where the quantity of interest is the solution value of one algorithm divided by the solution value of the SGA. It is clear that the SGA, IRTA, PCTA, and KPSK are quite comparable in their solution quality while the FPTA lags behind, which further corroborates the conclusions we drew about the effectiveness of the algorithms in Section 2.6.1.

2.7 Concluding Remarks

We developed a model for a modified version of the multiple-job repair kit problem in which the technician has access to recourse inventory for use in servicing jobs that require parts not stocked in the kit. We derived an exact measure for the expected job fill rate in this setting and constrain it to be no less than a given rate while minimizing total inventory holding costs. We proposed six algorithms, four of which we showed can produce satisfactory inventory policies at lower costs than the other two via computational testing on both hypothetical and real-world data. However, we showed that two of those algorithms are not always correct and one of those two can be infinitely bad in the worst case. In addition, we presented an upper bound on the expected job fill rate which we used to make a knapsack relaxation of our exact problem. We can solve the relaxed problem to optimality with off-the-shelf commercial software to obtain a lower bound on the optimal solution to the exact problem. Doing so in numerical experiments shows that the four best algorithms were often within 25% of the optimal value.

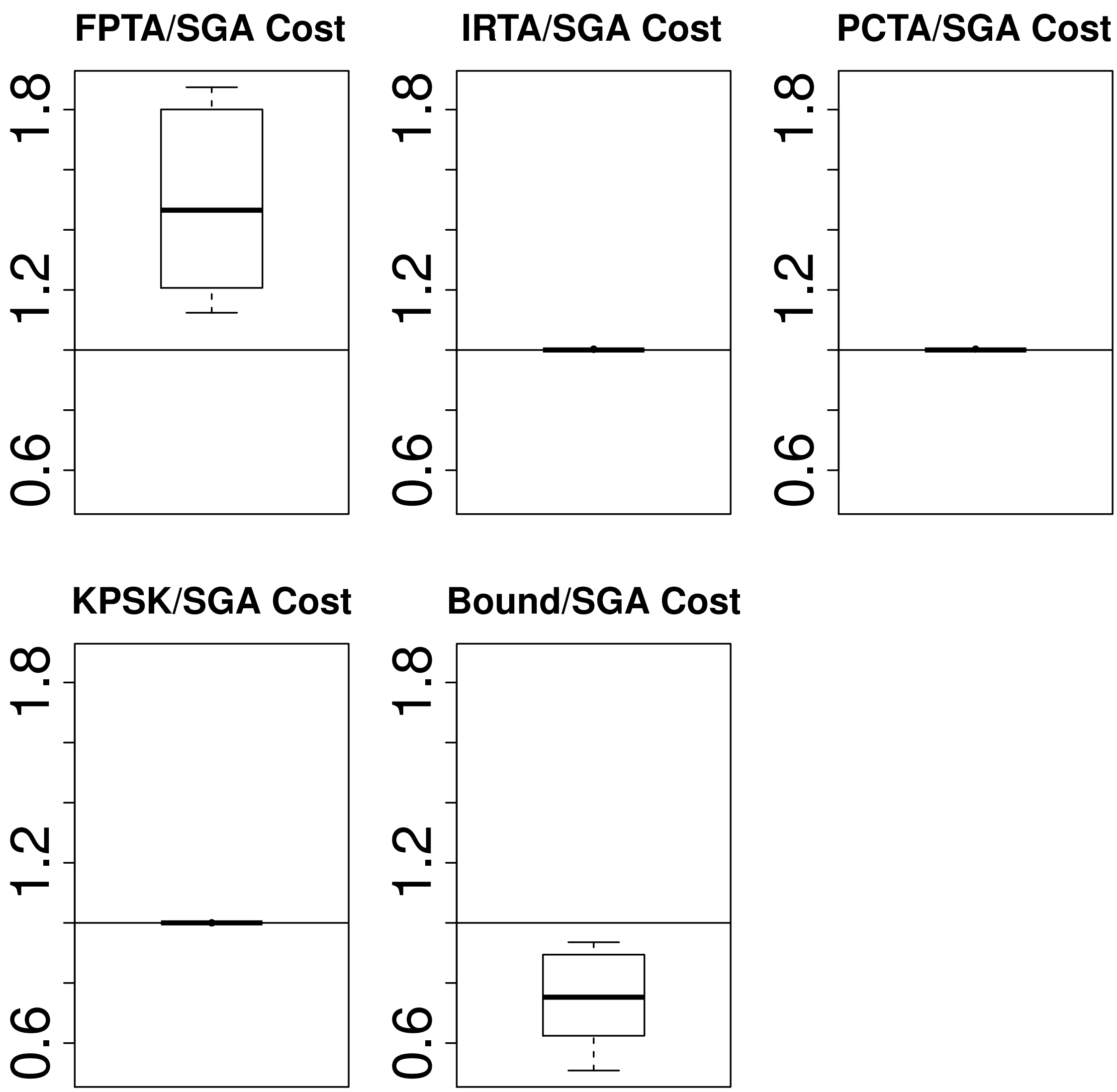

Figure 7: Algorithm comparison for case study

CHAPTER III

THE MULTIPLE-JOB REPAIR KIT AND TECHNICIAN ROUTING PROBLEM WITH FORWARD STOCKING LOCATION RECOURSE

3.1 Introduction

Companies which manufacture and sell high technology mechanical and electrical equipment often also, as part of their business, are repsonsible for repairing that machinery when it malfunctions at their customers' locations. At the time of purchase, customers enter service-level agreements that specify how quickly the manufacturer must resolve problems the user reports. Violations of this agreement have negative consequences for a business, *e.g.*, penalty fees paid to the customer, loss of goodwill, or a lowered perception of customer service.

In most cases, the manufacturer dispatches a technician to bring the machine back on-line by replacing one or more faulty parts with working spare parts. However, which parts are not functioning properly is not known until the technician diagnoses the problem on-site. Carrying in the repair kit every part that might possibly be needed is quite expensive given the number of parts installed in various machines in the field and the number of technicians employed. To address this dilemma, the manufacturer can store spare parts at centralized positions in the field known as forward stocking locations (FSLs). Multiple technicians share access to the inventory at the FSLs and hence the pooling of spare parts decreases total system-wide inventory cost.

This chapter demonstrates that, for a spare parts supply chain with FSL recourse, the operational decision of how to route the technician to visit customer sites for repair jobs significantly impacts the tactical decision of which spare parts to stock in the technician's repair kit. For this purpose, this chapter provides computational test results showing that the service provider can maintain the minimum required level of customer service with more parts at the FSL (instead of in the repair kit) if the technician's route for the period maximizes expected fill rate given the inventory stocking level. With more parts at the FSL the total system-wide inventory cost is lower due to pooling. This chapter presents an approximation of a repair kit's expected fill rate based on sample realizations of customer periods and technician routes as well as multiple algorithms to route the technician in a period given a repair kit and multiple equally-effective algorithms to choose inventory stocking given an expected fill rate black box.

Section 4.2 reviews the literature on the repair kit problem and relevant vehicle routing problems. Section 4.3 contains formulations for the inventory and routing decision problems and illustrates how they relate to one another. Section 4.4 describes solution algorithms for both problems. Section 4.5 presents results from computational experiments and Section 4.6 concludes.

3.2 Related Literature

The main objective of this chapter is to determine the least-cost repair kit inventory composition for a technician who must service multiple jobs in a single period with access to recourse inventory at an FSL. As detailed in Chapter 2, the existence of this recourse inventory affects one customer service measure of interest, the expected job fill rate. This chapter extends that work so that the route a technician takes to visit customers affects the expected job fill rate as well.

3.2.1 Repair kit literature

The seminal repair kit paper by Smith et al. [66] develops a model that minimizes an unconstrained objective function that includes both holding costs and penalty costs for a single-job problem with no more than one part type used on a given job. For the same problem setting, Graves [42] develops a model that minimizes holding cost subject to satisfying a specified job fill rate and thus avoids the determination of a penalty cost parameter. Mamer and Smith [56] also develops a model that minimizes an unconstrained total expected yearly cost function for a single-job repair kit problem but relaxes the assumption of only one part type failure.

The model of Heeremans and Gelders [45] minimizes holding cost for a multiple-job problem but constrains tour fill probability rather than job fill rate as perceived by the customer. Teunter [71] extends the model of Heeremans and Gelders [45] by indeed constraining job fill rate. Bijvank et al. [13] derives a closed-form expression for expected job fill rate in a general multiple-job setting where one or more units of a part type may fail and inventory from the kit is not set aside for a job that cannot be completed. Naturally, all of the authors propose algorithms to solve the repair kit problems they have formulated.

Chapter 2 modifies the repair kit problem framework slightly to represent an inventory strategy observed in practice and shows that such a strategy can benefit service providers by allowing them to operate at the required level of customer service for a lower inventory cost. We assumed that the order in which the customers were served had no effect on the expected job fill rate. This chapter relaxes that assumption and solves the problem of which parts to stock in the repair kit (and which to leave at the FSL) given a technician routing algorithm for various types of such algorithms.

3.2.2 Routing literature

In the literature, there are four types of problems that relate to the routing decision problem studied as part of this chapter. The inventory routing problem (IRP), the traveling repairman problem (TRP), and the technician routing and scheduling problem (TRSP) are similar enough that we review work done on them but argue how our problem differs. The stochastic orienteering problem is a generalization of what we seek to do regarding routing.

3.2.2.1 Inventory Routing Problem

The IRP is a decision problem model for vendor-managed resupply, a newer trend in which the supplier manages the inventory replenishment of its customers. This arrangement combines inventory management and transportation to create value for both the customer and the supplier. Vendors can better coordinate their distribution to save cost and customers are released from the responsibility of managing their own inventory.

Campbell et al. [17] presents and discusses the IRP as well as reviews a representative sample of prior work done on the problem, including Federgruen and Zipkin [33], Golden et al. [38], Dror et al. [28], and Chien et al. [24] among others. Jaillet et al. [48] presents incremental cost approximations to be used in a rolling horizon framework for minimizing total expected annual delivery costs. Campbell and Savelsbergh [19] proposes a two-phase solution approach for the IRP. The first phase creates a delivery schedule via integer programming, and the second phase makes use of routing and scheduling heuristics to create a set of delivery routes. Coelho et al. [25] reviews the first thirty years of IRP literature.

The IRP does not match up exactly with the problem we study in this chapter because it concerns only one type of product whereas our "products" are spare parts with thousands of possible types. Also, the IRP is usually tasked with creating multiple routes whereas we want to create a single route and have no control over the number of customers on that single route. Nevertheless, there is variation of the IRP that modifies the usual framework in a fashion similar to the way we modify the usual repair kit framework, *i.e.*, with satellite facilities. Bard et al. [9] examines an IRP with satellite facilities, where vehicles can refill during the course of the period without needing to return to the central depot should they run out of the product. The authors decompose the problem over the planning horizon in order to solve a daily problem rather than multi-day vehicle routing problems (VRPs). They also develop and test three heuristics for the VRP with satellite facilities (VRPSF), which is the second part of their decomposition scheme. Bard et al. [8] presents a branch-and-cut algorithm for the VRPSF. The same concerns about a single product type and multiple routes apply to these references as well, in addition to the fact that the locations of the satellite facilities are within the decision scope of the problem, which is not the case in this chapter.

3.2.2.2 Traveling Repairman Problem

The TRP is a version of the traveling salesman problem (TSP) in which the objective is to minimize the sum of the waiting times for the customers on the route rather than the total time it takes to execute all the jobs as in the case of the TSP. This problem is also known in the literature by the name minimum latency problem (MLP) and traveling deliveryman problem (TDP).

Minieka [60] explores various characteristics of the TDP and proposes a pseudo-polynomial time algorithm to solve the problem. Lucena [55] proposes a scheme to derive lower bounds for the time-dependent TSP and creates a branch-and-bound algorithm based on that scheme to solve the TDP. Bianco et al. [12] proposes two exact algorithms for the TRP that incorporate lower bounds provided by a Lagrangean relaxation and a heuristic procedure derived from dynamic programming. Blum et al. [15] gives a constant-factor approximation algorithm for the MLP whenver the distance matrix for the customer nodes satisfies the triangle inequality. Goemans and Kleinberg [37] improves the approximation ratio of Blum et al. [15] and Archer et al. [5] improves further upon that. Méndez-Díaz et al. [59] proposes a new integer programming formulation for the TDP along with a cutting plane algorithm that uses a number of valid inequalities shown to be facet-defining for the convex hull of feasible solutions. Mladenović et al. [62] proposes an effective variable neighborhood search (VNS) algorithm for the TDP.

The name TRP makes the problem and the work done on it sound like what we do in this chapter but we argue to the contrary. The objective of our routing problem is to maximize the number of jobs completed before a deadline, which is different enough from the sum of waiting times for all jobs to make these previous investigations of little use to us.

3.2.2.3 Technician Routing and Scheduling Problem

In the TRSP, a set of technicians must be scheduled to serve a set of customers and routed appropriately. The technicians are allowed to differ by skill level, which affects the length of time to complete a service call, as well as spare parts carried or types of service performed, which affects ability to assign certain calls to certain technicians.

Dutot et al. [29] introduces the TRSP in the context of scheduling interventions for telecommunications services offered by France Telecom. Bostel et al. [16] presents an approach for planning and routing technician visits over multiple periods, updating the plan daily on a rolling-horizon basis. Kovacs et al. [52] proposes and evaluates an adaptive large neighborhood search algorithm for the TRSP with time windows at the service sites. Tricoire et al. [72] models the TRSP as a set covering problem and uses both exact and hybrid methods to solve it. Pillac et al. [63] describes a matheuristic composed of a constructive heuristic, a parallel adaptive large neighborhood search, and a post-optimization procedure to solve the TRSP. Binart et al. [14] considers a variant of the TRSP with mandatory and optional customers, and proposes a two-stage method for the solving the problem. Chen et al. [23] expands the traditional TRSP model to explicitly incorporate individualized, experience-based learning and shows that capturing changes in productivity over time due to learning leads to significantly better and different solutions than ignoring heterogeneity. Castillo-Salazar et al. [20] reviews the literature on workforce scheduling and routing problems.

The TRSP is, for the most part, a significant generalization of what we do for routing in this chapter. We create a route for a single technician, thus ignoring any assignment of jobs to technicians or heterogeneity of repair capabilities. Furthermore, our time horizon is a single period. In this chapter, we do not use any algorithms designed for the TRSP since doing so would be overkill. We do, however, model one of our routing algorithms after one used in the literature to solve the stochastic orienteering problem, as detailed in the Section 3.2.2.4.

3.2.2.4 Stochastic Orienteering Problem

Orienteering is a sport where competitors must navigate through a forest and visit a number of 'control points' armed only with a compass and a map that shows the locations of these points. In one version of the game, the competitors must visit all locations and the winner is the one who does so in the shortest amount of time. In another, the controls have score values assigned to them and the goal is to collect as high a score as possible within a given time limit.

Tsiligirides [73] creates a mathematical model for the decision problem of how to proceed, proposes two heuristics to find solutions, and then compares the two versions of the sport against one another. Golden et al. [40] proposes a more effective center-of-gravity heuristic for the orienteering problem (OP). Chao et al. [21] proposes a fast and effective heuristic for the OP. Chao et al. [22] studies the team orienteering problem (TOP) where $M > 1$ routes must be designed, one for each of the M members of the team, with the goal of maximizing total score within the time limit. Tang and Miller-Hooks [70] presents a tabu search heuristic for the TOP. Vansteenwegen [74] reviews the OP literature.

The problem of routing a technician to maximize expected fill rate as defined in this chapter is equivalent to a special case of the orienteering problem with stochastic service times. Maximizing expected fill rate for a given set of customers is equivalent to maximizing the expected number of customers visited when defined to mean the technician has completed the repair. The expected number of customers visited is equivalent to the expected profit under the conditions that visiting a customer earns a reward of one and not visiting a customer incurs a penalty of zero. In a spare parts supply chain with FSL recourse, the service time at a customer is a random variable that depends on which part needs replaced, *i.e.*, whether a trip to the FSL must be made to retrieve a part.

Gupta et al [43] describes a constant-factor approximation algorithm for the best non-adaptive policy for the orienteering problem in which service times at the nodes are stochastic, where non-adaptive means the route must be specified *a priori* and followed without deviation. Evers et al. [32] studies what it calls the OP with stochastic weights, which represent uncertainty in travel or service times. The authors introduce a linearization for the expected profit and develop a heuristic to solve the problem fast than it could be done with sample average approximation. Campbell et al. [18] proposes a variant of the orienteering problem in which travel and service times are stochastic. The travel times between customers and the service times at the customer sites are uncertain but their distributions are known in advance. The objective is to find a tour with maximum expected profit, where a tour is a subset of the customers. Completing service at a customer on the tour before the known deadline earns a reward while not doing so incurs a penalty. The authors use pieces from existing VNS heuristics to solve the general version of the problem that they propose.

One routing algorithm in this chapter builds off of the framework from Campbell et al. [18]. However, the proposed VNS algorithm uses some other neighborhoods for the shaking and local search steps because many of the ones in Campbell et al. [18] rely on the fact that not all customers are included on the tour.

3.3 The Multiple-Job Repair Kit and Technician Routing Problem with Forward Stocking Location Recourse

As mentioned above, we extend the work of Chapter 2 by relaxing the assumption that the times for kit and FSL jobs at all customers are identically distributed; instead those times depend on the geography of the customers and the order in which they are visited within the period. Calculating the expected fill rate in this setting is more complicated and itself requires a decision problem be solved, that of how to route the technicians.

The overall objective of the decision maker is to minimize total inventory cost subject to meeting the contracted minimum fill rate. Additionally, the decision maker can set the routing policy, which affects the fill rate calculation and thus indirectly affects the inventory decisions. We can think of our optimization problem as follows, the same as in Chapter 2.

$$\begin{aligned} \text{min} \quad & \text{total inventory holding cost} \\ \text{s.t.} \quad & \text{job fill rate} \geq \text{required fill rate} \end{aligned}$$

It remains to define precisely the total inventory holding cost and job fill rate for our problem setting. We assume that all part types not carried in the repair kit are held at the FSL and that exactly one part type is required to complete the repair at each customer. In this model, a job is considered filled if the repair is completed by the end of the period. If the necessary parts for a repair are not stocked in the kit then the technician must retrieve them from the nearest FSL and the job is not filled until the technician returns.

3.3.1 Inventory holding cost

Let n be the number of distinct part types installed in customer machines located throughout the region of interest. Let c_i^f be the cost to hold part i in at the FSL and c_i^k be the cost to hold part i in the kit. The total inventory holding cost is

$$\sum_{i=1}^{n} c_i^k x_i + c_i^f (1 - x_i)$$

where $x_i = 1$ if part i is stocked in the kit and $x_i = 0$ otherwise. We claim that minimizing the above objective function is equivalent to minimizing

$$\sum_{i=1}^{n} c_i x_i$$

where $c_i = c_i^k - c_i^f$.

$$\sum_{i=1}^{n} c_i^k x_i + c_i^f (1 - x_i) = \sum_{i=1}^{n} (c_i^k - c_i^f) x_i + c_i^f$$

We can remove the constant term $\sum_{i=1}^{n} c_i^f$ from the objective function and thus we have shown that we can assume the FSL costs are zero.

3.3.2 Job fill rate

We will begin by showing how to calculate the expected fill rate for a certain repair kit given a route over a single instance of customers with uncertain part demands. In reality, the customers that require service will be different every day and new routes will need to be determined every day as well. The expected fill rate of a repair kit will vary based on the relative locations of the customers that require service in a given period. We assume the decision maker's objective is some type of fill rate aggregated over the long run, spanning multiple service periods, and a given repair kit may not necessarily need to meet the expected service rate for every single possible customer instance. To approximate an aggregated fill rate, we will repeat the calculation of expected fill rate on a single customer instance for a number of scenarios and then take the average of those values as the service measure value for a given repair kit. More details follow on the approximated aggregation in Section 3.3.2.2.

3.3.2.1 Expected job fill rate for a single customer instance

We consider a job successful if it is completed by the end of the period, which lasts d time, and we consider a job to be a service failure otherwise. However, in our new setting where the order of the customers matters, we discard our assumption that each type of job takes the same amount of time for every customer. Now we will let each customer's kit job and FSL job times depend on the customer which directly precedes it in the route, reflecting the varying travel times between customer sites.

The expected job fill rate is the complement of the expected job failure rate.

$$\phi(x_1, \ldots, x_n; \rho) = 1 - \frac{E[F]}{m}$$

Let F be the number of job failures in a single period for one customer instance. Clearly, F takes its value in $\{0, 1, 2, \ldots, m\}$ where m is the number of jobs assigned to a technician. The expected number of failures takes the following form.

$$\begin{aligned} E[F] &= \sum_{j=0}^{\infty} P(F > j) \\ &= \sum_{j=0}^{m-1} P(F > j) \\ &= m - \sum_{j=0}^{m-1} P(F \leq j) \end{aligned}$$

When we use this expression in the expected fill rate expression we get the following.

$$\begin{aligned} \phi(x_1, \ldots, x_n; \rho) &= 1 - \frac{E[F]}{m} \\ &= 1 - \frac{m - \sum_{j=0}^{m-1} P(F \leq j)}{m} \\ &= \frac{\sum_{j=0}^{m-1} P(F \leq j)}{m} \end{aligned}$$

The event that no more than j jobs fail during a period is exactly the same as the event that the $(m-j)$th job is completed by the deadline, and hence we have

$$P(F \leq j) = P(\mathcal{T}_{m-j} \leq d)$$

where $\mathcal{T}_{m-j}$ is the elapsed time from the start of the period to the point at which the $(m-j)$th job is completed. Our expected job fill rate becomes the following.

$$\phi(x_1, \ldots, x_n; \rho) = \frac{1}{m} \sum_{j=0}^{m-1} P(\mathcal{T}_{m-j} \leq d) = \frac{1}{m} \sum_{j=1}^{m} P(\mathcal{T}_j \leq d)$$

It remains to derive the distribution of $\mathcal{T}_j$ for $j = 1, 2, \ldots, m$.

Let t_{ij} be the travel time between customers i and j and let customer 0 be the FSL. The kit job time, α_{ij}, is equal to the travel time to the customer (t_{ij} when customer j is preceded by customer i) plus the service time (s_j at customer j). The FSL job time, β_{ij}, is equal to the travel time to the customer plus the service time plus the travel time to go back and forth from the FSL ($t_{ij} + s_j + t_{j0} + t_{0j}$). Now, we assume that the technician can start the period at the site of the first customer and hence $\alpha_{0j} = s_j$ and $\beta_{0j} = s_j + t_{j0} + t_{0j}$ are different for all j because they do not include a travel time from a predecessor customer.

Once we have a route for the technician we have determined the α and β values that represent the two possible times completing service at a customer might take. From these values we can determine the necessary cumulative distribution functions for the expected number of failed jobs and then calculate the expected fill rate for the given inventory and route decision.

Let (i) represent the customer number in the ith position on a route. If the ith job is a kit job, it takes time $\alpha_{(i-1),(i)}$; otherwise it takes time $\beta_{(i-1),(i)}$. The probability mass function (p.m.f.) for the time at which job j is completed, $\mathcal{T}_j$, is created by sorting all possible completion times in ascending order and assigning each one its appropriate probability. From this p.m.f. it is then trivial to create the cumulative distribution function. A probability mass function might look something like what is shown in Table 10.

Table 10: Example probability mass function for $\mathcal{T}_j$

t	$P(\mathcal{T}_j = t)$
$\alpha_{0,(1)} + \alpha_{(1),(2)} + \cdots + \alpha_{(j-1),(j)}$	$P(\text{Kit})^j$
$\beta_{0,(1)} + \alpha_{(1),(2)} + \cdots + \alpha_{(j-1),(j)}$	$P(\text{Kit})^{j-1}(1 - P(\text{Kit})$
$\alpha_{0,(1)} + \beta_{(1),(2)} + \cdots + \alpha_{(j-1),(j)}$	$P(\text{Kit})^{j-1}(1 - P(\text{Kit})$
$\vdots$	$\vdots$
$\beta_{0,(1)} + \beta_{(1),(2)} + \cdots + \beta_{(j-1),(j)}$	$(1 - P(\text{Kit}))^j$

3.3.2.2 Approximation of expected aggregate fill rate

As mentioned, the customers will vary in both number and location from one service period to the next in practice. The fill rate of the chosen repair kit and FSL part type allocation will be achieved over a number of service periods as the inventory decision is made for an extended time frame, at minimum a few months. In order to approximate this fill rate achieved across multiple service periods with multiple customer routes calculated, we measure the expected fill rate on a sample of instances with randomly-generated customer locations. Sampling as many customer instances as service periods in the inventory decision timeframe would be ideal but presents too much computational difficulty for the scope of this chapter. Therefore, we calculate the expected fill rate for an inventory allocation on five customer instances and take the average of those values to be the expected fill rate. See below, where $\mathcal{T}_j^i$ is the completion time for the customer in the jth position on the route in the ith instance and ℓ is the number of instances, five in our case.

$$\begin{aligned}\phi(x_1, \dots, x_n; \rho) &= \frac{1}{\ell} \sum_{i=1}^{\ell} \phi_i(x_1, \dots, x_n; \rho) \\ &= \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{1}{m_i} \sum_{j=1}^{m_i} P(\mathcal{T}_{ij} \leq d)\end{aligned}$$

This number gives us a nice balance between realism of the model and computational tractability.

3.3.3 Routing sub-problem

In practice, the customers that require service may be different every day. The fill rate sought by the service provider takes place over a period of many days. There will be days when the fraction of customer jobs completed successfully is lower than required and days when it is higher. To approximate this heterogeneity and variation among customers we will use multiple customer instances to calculate expected fill rate.

As mentioned previously, the inventory and routing decisions are made at different frequencies. Ideally, the service provider would like to find a collection of part types for the repair kit that gives the satisfactory fill rate in practice and, barring any significant changes to the customer and machine demographics, maintain that inventory policy indefinitely. On the other hand, the routing decisions must be made on a daily basis because the customers that need service will not be the same. The customers that need service on a given day will only be a small subset of all the customers within the region.

We approximate the true fill rate over time as follows. First we generate Monte Carlo samples of customers and route the technicians based only on the locations of the customers and the probability of a kit job, which we assume is the same for all customers. For any route on a single instance, we are able to calculate the expected fill rate as described in the previous section. We route the technicians to maximize expected fill rate and take the average of the expected fill rates for each customer scenario as the overall expected fill rate. The sample customer instances are stored in memory and used for every fill rate calculation during a single run of an algorithm and can be kept for runs of multiple algorithms as well for purposes of comparison.

The objective of the routing problem is to find a complete path starting at a customer node that maximizes expected fill rate for a given repair kit.

3.3.4 Complete formulation

Our optimization problem with all the parameters completely specified follows.

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} c_i x_i \\ \text{s.t.} \quad & \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{1}{m_i} \sum_{j=1}^{m_i} P(\mathcal{T}_{ij} \leq d) \geq \gamma \\ & x_i \in \{0,1\} \forall i = 1, 2, \ldots, n \end{aligned}$$

We have that the fill rate is a function of both the inventory decisions $(x_1, \ldots, x_n)$ and the routing policy, ρ.

3.4 Solution Algorithms

Due to the expected fill rate, which constrains our inventory problem, not taking a smooth functional form, we cannot utilize any information about its structure in the design of algorithms to solve the inventory problem. Hence, the expected fill rate constraint is taken as a black box when solving the overall inventory problem, which means we can use all combinations of inventory problem algorithms and technician routing algorithms. This will allow us to determine whether random routing significantly impacts total inventory cost by negatively influencing customer service.

We have two optimization problems to solve and need a total of at least two algorithms to do so, one for each problem. The overall problem of finding the cheapest repair kit to satisfy the contracted customer service level is identical to that of Chapter 2, except the expected fill rate now depends on a routing procedure. The algorithms proposed to solve the repair kit problem with homogeneous customers relied only upon the value of the expected fill rate for a given repair kit and were completely independent of the manner in which those values were calculated. Hence, we can reuse the algorithms proposed for the inventory problem in Chapter 2. We propose methods to the solve the routing problem, which was not studied previously.

3.4.1 Inventory problem

For multiple-job repair kit problems with FSL recourse but no routing involved, there are four algorithms from Chapter 2 that work similarly well:

1. Subtractive Greedy Algorithm (SGA),
2. Initial Ratio Threshold Algorithm (IRTA),
3. Probability Cost Threshold Algorithm (PCTA), and
4. Kit Probability Search plus Knapsack (KPSK).

We review all four of the algorithms here.

3.4.1.1 Subtractive Greedy Algorithm

The SGA is a greedy marginal analysis heuristic that evaluates all possible moves and pivots the solution by the best feasible move at each iteration. Best in this case is defined to be the one with the highest ratio of cost decrease to fill rate decrease due to removing a part from the kit and feasible means taking the part out of the repair kit does not drop the fill rate below the minimum required value. We initialize the repair kit to contain all part types and iterate until we cannot remove a single part type more without leaving the feasible region of the optimization problem. Please see Algorithm 3 for pseudocode of the subtractive greedy algorithm.

Algorithm 3 Subtractive Greedy Algorithm

```
all parts in the kit
repeat
  clear list
  for each part in the kit do
    if moving part from kit to FSL is feasible then
      calculate ratio and add move to list
    end if
  end for
  pivot solution by move on list with highest ratio
until no feasible moves on list
```

3.4.1.2 Threshold Algorithms

A threshold algorithm sorts the parts for which stocking decisions must be made by a single measure and then iterates through the list until a stopping criterion is satisfied. An additive threshold algorithm is initialized with an empty kit and adds the parts as it iterates from most desirable to least desirable until a feasible kit with sufficient probability of a kit job is reached. The fill rate of the kit is recalculated after each addition step. See Algorithm 4 for pseudocode of an additive threshold algorithm.

Algorithm 4 Additive threshold algorithm

```
no parts in the kit
sort all parts by criterion with parts most desirable for the kit at top of list
while job fill rate less than minimum required rate do
  add part on top of list to kit and remove part from list
end while
```

A subtractive threshold algorithm is initialized with a full kit and subtracts the parts as it iterates until the fill rate of the kit is no longer greater than or equal to the minimum required fill rate. At that point the last part removed is added back to the kit to return it to feasibility. See Algorithm 5 for pseudocode of a subtractive threshold algorithm.

Algorithm 5 Subtractive threshold algorithm

```
all parts in the kit
sort all parts by criterion with parts least desirable for the kit at top of list
while job fill rate greater than or equal to minimum required rate do
  remove part on top of list from kit and remove part from list
end while
add part removed last back to kit
```

The IRTA is a subtractive algorithm that sorts the parts in the descending order of the initial ratio that would be calculated in the first iteration of the SGA. This approach was motivated because the SGA works well but requires recalculation of all ratios at every iteration. The PCTA is an additive threshold algorithm that sorts the parts in descending order of ratio of failure probability to difference between kit and

FSL costs, *i.e.*,

$$\frac{p_i}{c_i^k - c_i^f}.$$

This approach was motivated by visual inspection of solutions that revealed many good solutions appeared to have parts with high values of this ratio in the kit and low values in the FSL.

3.4.1.3 Kit Probability Search plus Knapsack (KPSK)

The KPSK finds the minimal kit job probability needed to satisfy the expected fill rate constraint and then chooses the lowest-cost assortment of part types to meet or exceed the minimal kit probability as a knapsack decision problem. The value of $p_{\min}$ can found quickly with binary search and then used in the following formulation.

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} c_i x_i \\ \text{s.t.} \quad & p_i x_i \geq p_{\min} \\ & x_i \in \{0,1\},\ i = 1, 2, \ldots, n \end{aligned}$$

This 0/1 knapsack problem can be solved efficiently with commerical software packages. This approach was motivated by the observation that the specific part types in the repair kit themselves do not directly influence the fill rate, only the sum of their probabilities, *i.e.*, the probability of a kit job.

3.4.2 Routing sub-problem

In this chapter, we want to explore the impact of routing on the fill rate with the intent to gain insight into how much effort is justified for good routing which requires decreased inventory levels and hence cost. Thus, we need to define different routing policies that we will evaluate at varying levels of goodness and computational effort, which we expect to correlate positively with one another.

3.4.2.1 Random routing

The simplest routing policy that takes almost no effort is random routing. The name is self-explanatory; random routing puts the customers in a random order. This is a conservative tactic and might better represent a situation where customer requests come in over the course of the day and must be added to the queue for that same day.

3.4.2.2 Nearest neighbor style algorithms

Another simple routing strategy that takes minimal computational effort is based on the intuition that it is desirable to serve customers close to the FSL first since their penalty times are small if a needed part is not in the kit. We propose two similar algorithms that work off of this intuition. Both start with the customer closest to the FSL leading off the route and construct the rest of the order in standard nearest neighbor greedy fashion. The first version, called Earliest Completion Neighbor (ECN), uses the travel time to a customer plus the service time, *i.e.*, $t_{ij} + s_j$, as the "distance" and the second, called Earliest Expected Completion Neighbor (EECN) uses travel time to a customer plus service time plus the expected travel time to the FSL and back, *i.e.*, $t_{ij} + s_j + P(\text{Kit})(t_{j0} + t_{0j})$. A sketch of the algorithm structure can be seen in Algorithm 6.

Algorithm 6 Nearest neighbor style algorithm

```
Empty route
Customer closest to FSL gets first spot on route
while Not all customers have a spot on the route do
  Append customer with least "distance" criterion value to route
end while
```

3.4.2.3 Nearest neighbor style plus local search routing

We build off of the simple nearest neighbor style algorithms by adding a first-improving local search that we seed with either the ECN or EECN route. We look for the first route with a higher expected fill rate in a compound neighborhood of what we call reinsert and swap. Reinsert neighbors remove a customer for its current position in the route and put the customer back in at any of the other $m - 1$ slots that would give a different route. Swap neighbors takes two customers on the route and switch them with each other. Whenever an improving solution is found, that solution is used as the seed for a new local search with the same compound neighborhood structure. The algorithm iterates like this until no improving solution is found in the entire compound neighborhood.

3.4.2.4 Variable neighborhood search algorithm

The "smartest" routing method we employ (besides finding the optimal route by brute force) is an implementation of a variable neighborhood search (VNS) algorithm. The concept of variable neighborhood search was first introduced by Mladenović and Hansen [61] and a review is available in Hansen et al. [44]. VNS is a metaheuristic for solving combinatorial optimization problems that uses random perturbations and changing search neighborhoods to escape local minima. There are two phases to VNS, a shaking phase and a local search phase. In the shaking phase, a random solution is returned from a neighborhood of a given solution and then in the local search phase this solution is improved. The routing problem we are solving with a VNS algorithm has as its objective to maximize expected fill rate. An outline of the VNS procedure can be seen in Algorithm 7.

We used a $t_{\max}$ value of 100 because that is what [18] used. The shaking phase chooses a random solution from neighborhood $k \in N_s$, where N_s is an ordered set of neighborhoods. The four neighborhoods we used in N_s are:

1. moving a customer to the end of the route,

2. moving a customer to the front of the route,

3. removing a customer from the route and reinserting elsewhere, and

Algorithm 7 Variable neighborhood search

Initial feasible solution τ
$t = 1$, $k = 1$
repeat
  $\tau' \leftarrow \mathit{Shake}(\tau, k)$
  $\tau'' \leftarrow \mathit{LocalSearch}(\tau')$
  if τ'' has expected fill rate no greater than that of τ **then**
    if $k < |N_s|$ **then**
      $k \leftarrow k + 1$
    else
      $k \leftarrow 1$
    end if
  else
    $\tau \leftarrow \tau''$
  end if
  $t \leftarrow t + 1$
until $t > t_{\max}$

4. swapping the positions of two customers on the route.

Because our problem is a special case that does not penalize the technician for failing to service customers, it is never better to exclude a customer from the route. Hence, our route decision space is smaller since no customers are ever left out and consequently the neighborhoods we could use are also less interesting.

For the local search phase, we implement a form of variable neighborhood descent (VND) where we search a neighborhood of a solution until no improving solution can be found, at which point we search a different neighborhood. We use a steepest descent criteria to search the neighborhood. Our neighborhood set, N_d, is the same as the neighborhood set in the VNS algorithm, N_s. Please see Algorithm 8 for an outline of the method.

3.5 Structure of Optimal Route

It is not immediately clear what a fill-rate-maximizing route looks like in this setting where the technician might have to make an extra trip to the FSL node from any customer node. Because of this probabilistic aspect of additional travel, the route with

Algorithm 8 Variable neighborhood descent

Initial feasible solution τ
repeat
  $\tau' \leftarrow \tau$
  $k \leftarrow 1$
  repeat
    $\tau'' \leftarrow \mathit{BestImproving}(\tau, k)$
    if τ'' has expected fill rate no greater than that of τ **then**
      $k \leftarrow k + 1$
    else
      $\tau \leftarrow \tau''$
    end if
  until $k > |N_d|$
until $\tau = \tau''$

shortest total distance, *i.e.*, the optimal open-TSP route, may not be optimal when maximizing fill rate. The distances of the customer nodes from the FSL influence the expected fill rate of a given route in addition to the distances between consecutive customer nodes on the route. One might suspect that a route which starts at the customer closest to the FSL and proceeds progressively farther away would be optimal since more service jobs could be completed due to the relatively short delay for parts retrieval from the FSL for the intial customers. However, this hypothesis does not hold for customers in a one-dimensional region, *i.e.*, a line out from the FSL, as we detail in the following counterexample.

Consider a region with three customers and one FSL, as shown in Figure 8, and problem setting with the parameters as seen in Table 11. The route $(1, 2, 3)$ (Figure 9) is the nearest neighbor route, which we might suspect to be optimal. We will show, however, that the route $(3, 2, 1)$ (Figure 10) has a higher expected fill rate and thus our hypothesis does not hold for customers in a one-dimensional region. The outcome trees in Figures 11 and 12 capture all the possible sample paths and associated job successes for the two routes of interest. Recall that the time for a kit job is the travel time from the preceding customer (0 for the first job of the day) plus the service time (0 in this example) and the time for an FSL job is the kit job time plus the

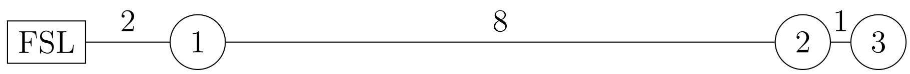

Figure 8: One-dimensional region with three customers and one FSL (distances shown)

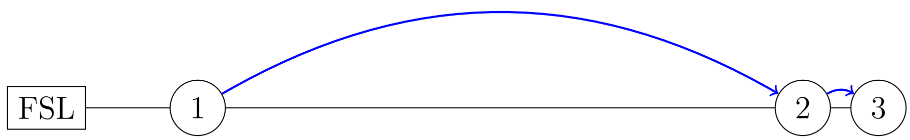

Figure 9: Route $(1, 2, 3)$

roundtrip travel time to the FSL. Note that each of the eight sample paths has the same probability, 1/8, because the probability of a kit job and the probability of an FSL job are identically 1/2.

We can observe that the expected number of successes for route $(1, 2, 3)$ is 1.75 while for route $(3, 2, 1)$ it is 2.125. Hence, the latter route has a higher expected fill rate although it starts farthest away from the FSL. We can conclude that not only does distance from the FSL affect the optimal selection of nodes for a route but also the distance from other customer nodes as well. It is not so simple to say what the fill-rate maximizing route is with the probabilistic "threat" of additional travel (dependent on customer location) that we have in this problem setting.

3.6 Computational Results

The main objective of the computational testing of the model presented in this chapter was to determine how much inventory cost can be saved through good routing. If an optimal (or near optimal) route can allow the technician to achieve the minimum

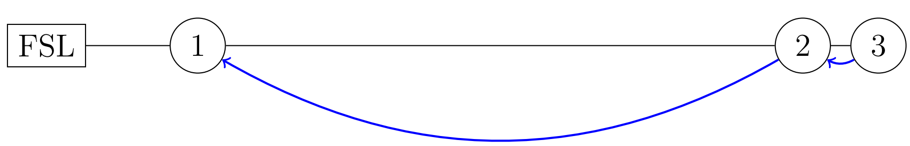

Figure 10: Route $(3, 2, 1)$

Table 11: Problem setting parameters

Service time at customer i	s_i	0
Probability of a kit job	$\sum_{i=1}^{n} p_i x_i$	0.5
Length of day	d	25

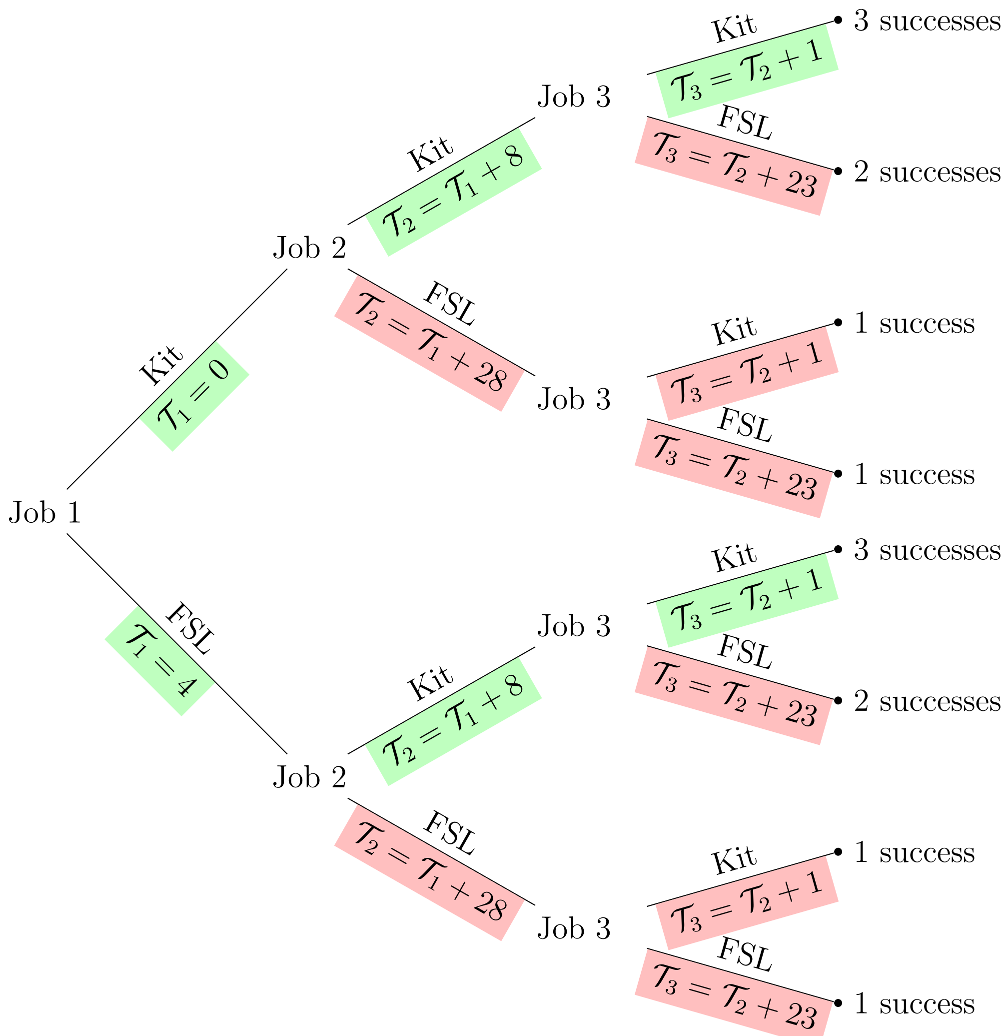

Figure 11: Outcome tree for route $(1, 2, 3)$

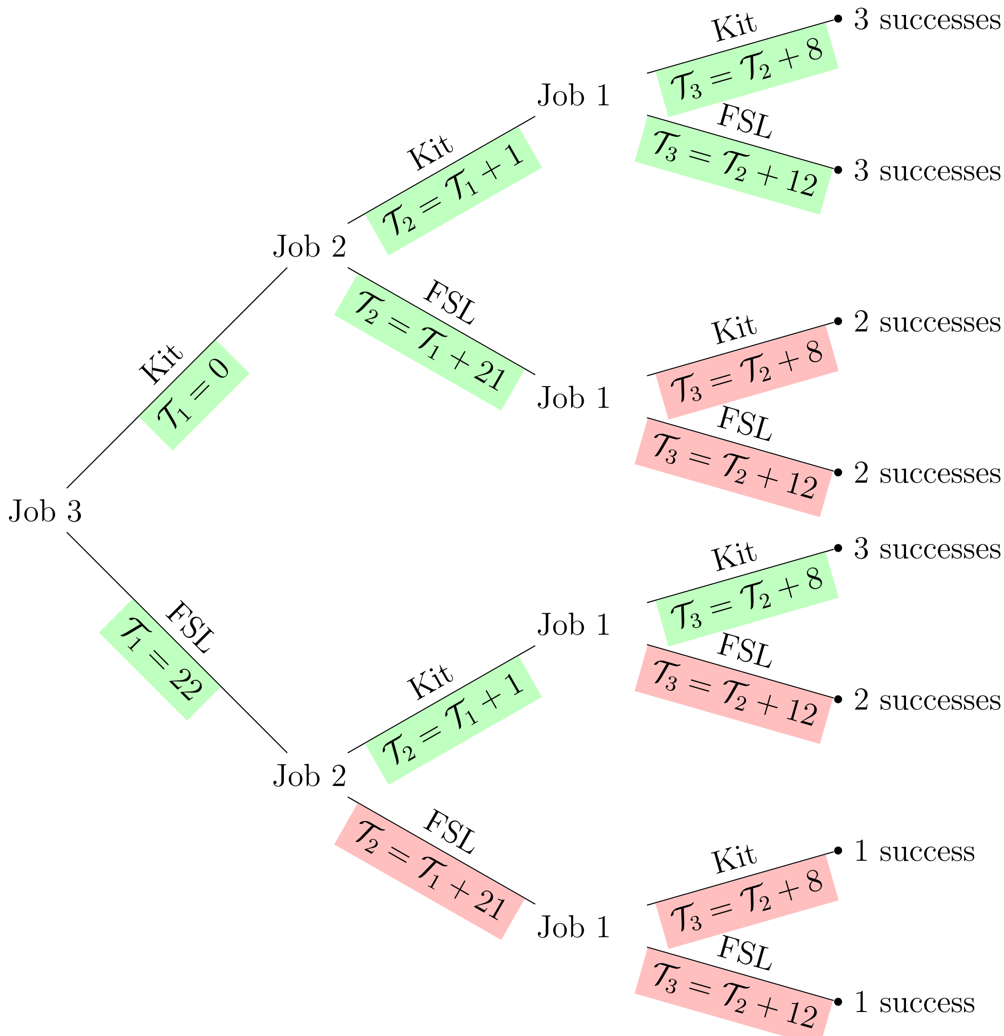

Figure 12: Outcome tree for route $(3, 2, 1)$

Table 12: Factors and level for computational testing

Factor	Notation	Level
Radius of region served by FSL at center (min)	–	$\{30, 60\}$
Length of day (hours)	d	$\{7, 9\}$
Service time at customer i (hours)	s_i	$\{0.75, 1\}$
Number of part types	n	$\{100, 2000\}$
Ratio of kit holding cost to FSL holding cost	c_i^k/c_i^f	$\{5, 10\}$
Minimum fill rate	γ	$\{0.9, 0.95\}$

required fill rate with significantly less repair kit inventory then a service provider has an incentive to invest resources in developing and implementing good routing policies in practice.

Another interpretation of this incentive would be that the service provider would want to structure its service-level agreements so that it can know all its customer locations for a period ahead of time to route them optimally. Random routing could be considered an approximation of the setting where jobs arrive to the service provider dynamically over the course of the period. If non-random routing is better than random routing by a nontrivial margin then the service provider may want to structure its contractual obligations so that it has the ability to route deterministically *a priori*.

To reach this objective we developed an extensive set of part test instances with important characteristics varied over ranges of reasonable values that might be confronted in practice, as summarized in Table 12. The number of customers in an instance (and thus to be routed) under these parameter settings ranged from four to eight, excluding seven. In total, twenty-four repair kits were found for each part instance, one for each of the twenty-four inventory-routing algorithm combinations. We then compared the kits obtained on these same part and customer instances with different routing methods to each other along the lines of total cost and kit job probability. Of course, the total cost is the measure of greatest interest but the kit job probability may provide us a little intuition about the structure of a good repair kit in the FSL setting.

3.6.1 Routing takes time

Before we get into the details of the results it must be noted that finding a route on a single customer instance can take a non-negligible amount of time, especially with VNS (as well as complete enumeration) for the eight customers. Of course, the routing algorithms that take longer to run usually provide higher expected fill rates for the same repair kits than their shorter-running competitors. Figures 13, 14, 15, and 16 show the average run times and average expected fill rates over three customer instances for the six routing methods from this chaptper plus the complete enumeration optimal route.

VNS and complete enumeration provide slightly better fill rates but take much longer to run, with VNS actually running longer than enumeration for the four- and five-customer instances. In an ideal situation, the routing algorithm specified for a test would be run for every instance in the customer sample every time the fill rate is calculated. However, given the number of fill rate calculations performed by our proposed inventory algorithms in some settings, we have chosen not to re-route for every single function call. For example, when the number of part types is 2000 and SGA removes 200 parts from the kit to get to its solution, it will have called the fill rate function around $380,000$ times. With five customer instances in the sample, that is 1.9 million routes that need to be determined just for a part of a run at one factor level, which equates to 130 days of computation time with six seconds per route (VNS in eight-customer case). Naturally, this is a worst-case scenario for what we want to investigate but it demonstrates that re-calculating the routes for every fill rate function call is not practical for all of our factor levels.

To save computation time, we generate routes for each customer instance according to its specified routing algorithm at initialization and use those routes for every fill rate calculation. After running all twenty-four inventory-routing combinations with these "static" fill rate calculations, we also explore some smaller scenarios with the

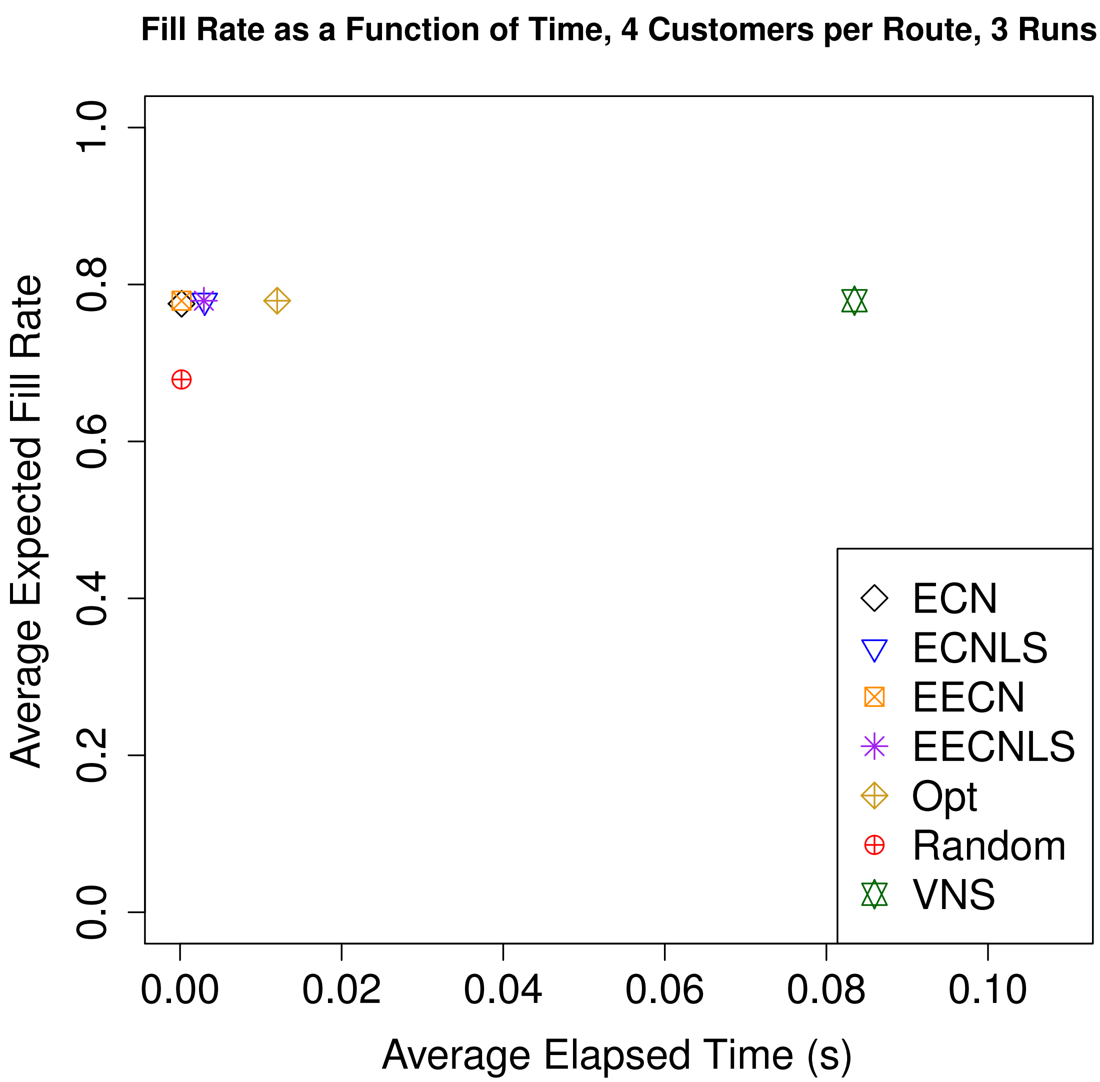

Figure 13: Average expected fill rate as a function of average run time

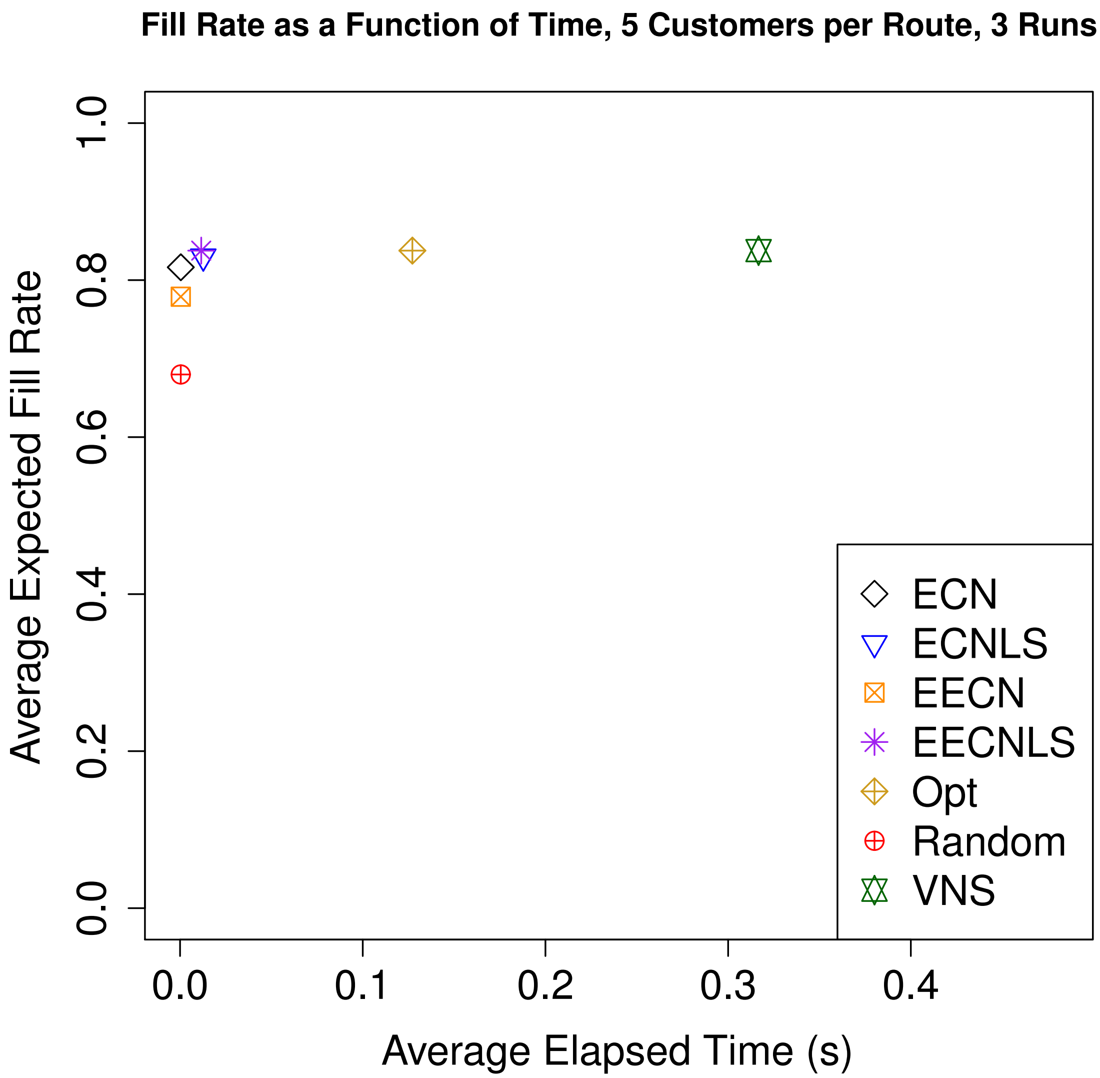

Figure 14: Average expected fill rate as a function of average run time

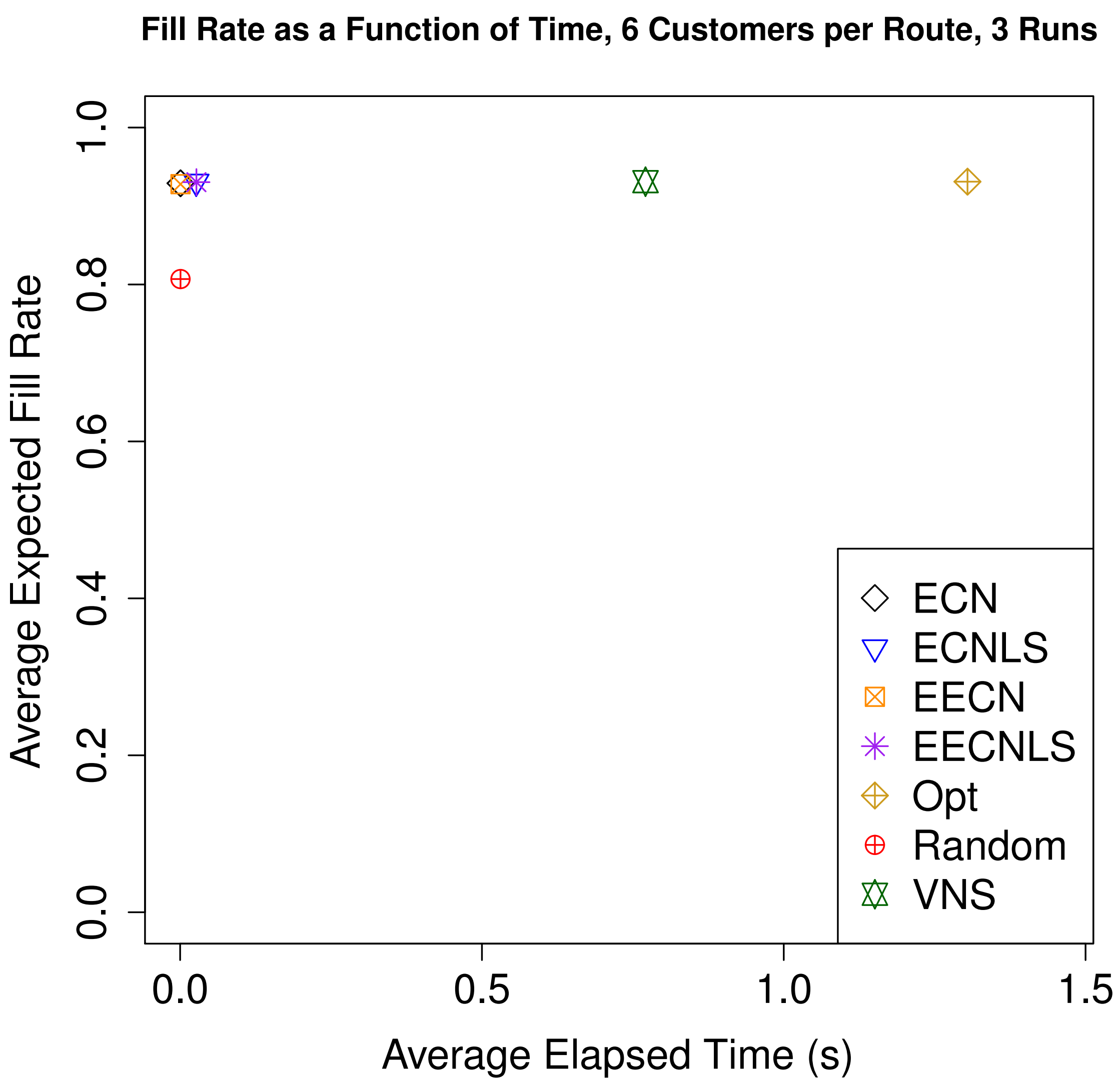

Figure 15: Average expected fill rate as a function of average run time

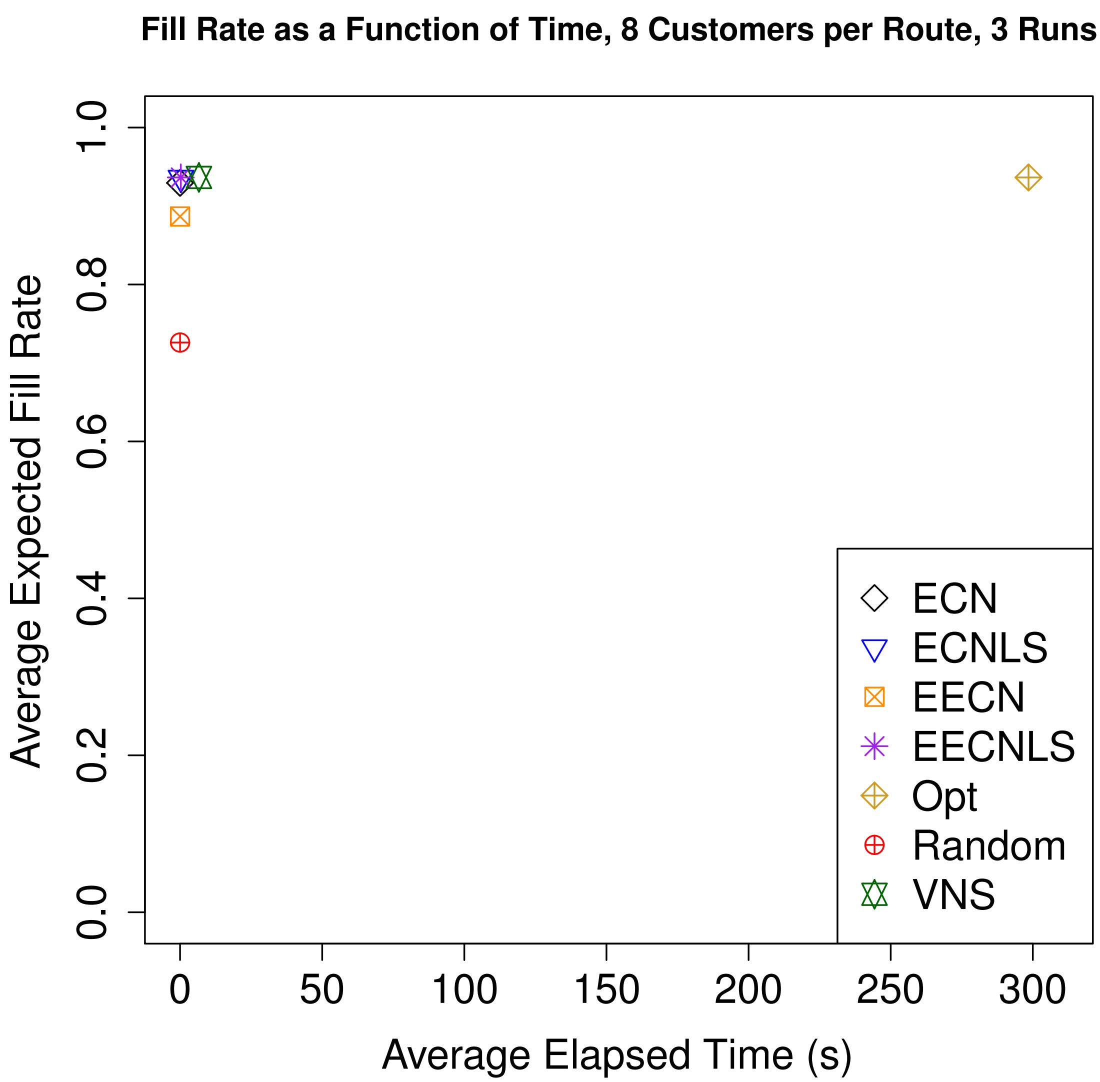

Figure 16: Average expected fill rate as a function of average run time

fill rate calculated “dynamically” (and optimally) to get an idea of what we miss out on by not re-calculating fill rates every time in our full factorial test.

3.6.2 Static fill rate

The major conclusion to draw from the test results where the fill rate is calculated with routes set at initialization is that random routing is an impediment to low cost solutions. Total inventory costs when customers are routed randomly are twenty percent or more higher than when any of other proposed routing algorithms are used, as illustrated in Figures 17, 18, 19, and 20. We show in the next section that this conclusion holds for smaller problem instances when routes are re-calculated for every fill rate calculation as well.

3.6.3 Dynamic fill rate

Instances with four customers and 100 or 2000 part types and instances with five customers and 100 part types are small enough for us to run them through our twenty-four inventory-routing combinations while re-calculating routes for every fill rate call. These instances also have few enough customers that we can find the optimal routes for each fill rate calculation in a reasonable amount of time. The solution cost ratios with optimal routing as the denominator can be seen in Figures 21, 22, 23, and 24.

The same major conclusion holds in this dynamic fill rate setting, *i.e.*, random routing is twenty percent more expensive. A corollary of this conclusion is that all non-random methods perform equally well, which is perhaps a little suprising because they are varied in their levels of sophistication. One of them, ECN, does not even take the probability of a kit job into account.

We speculate that ECN is good because it combines two factors which we think are important to good routes: proximity to FSL and proximity to other customers. The ideal customers, in the service provider’s eyes, are those ones that are very close to the FSL, so roundtrips to retrieve parts do not take much time, and are close to

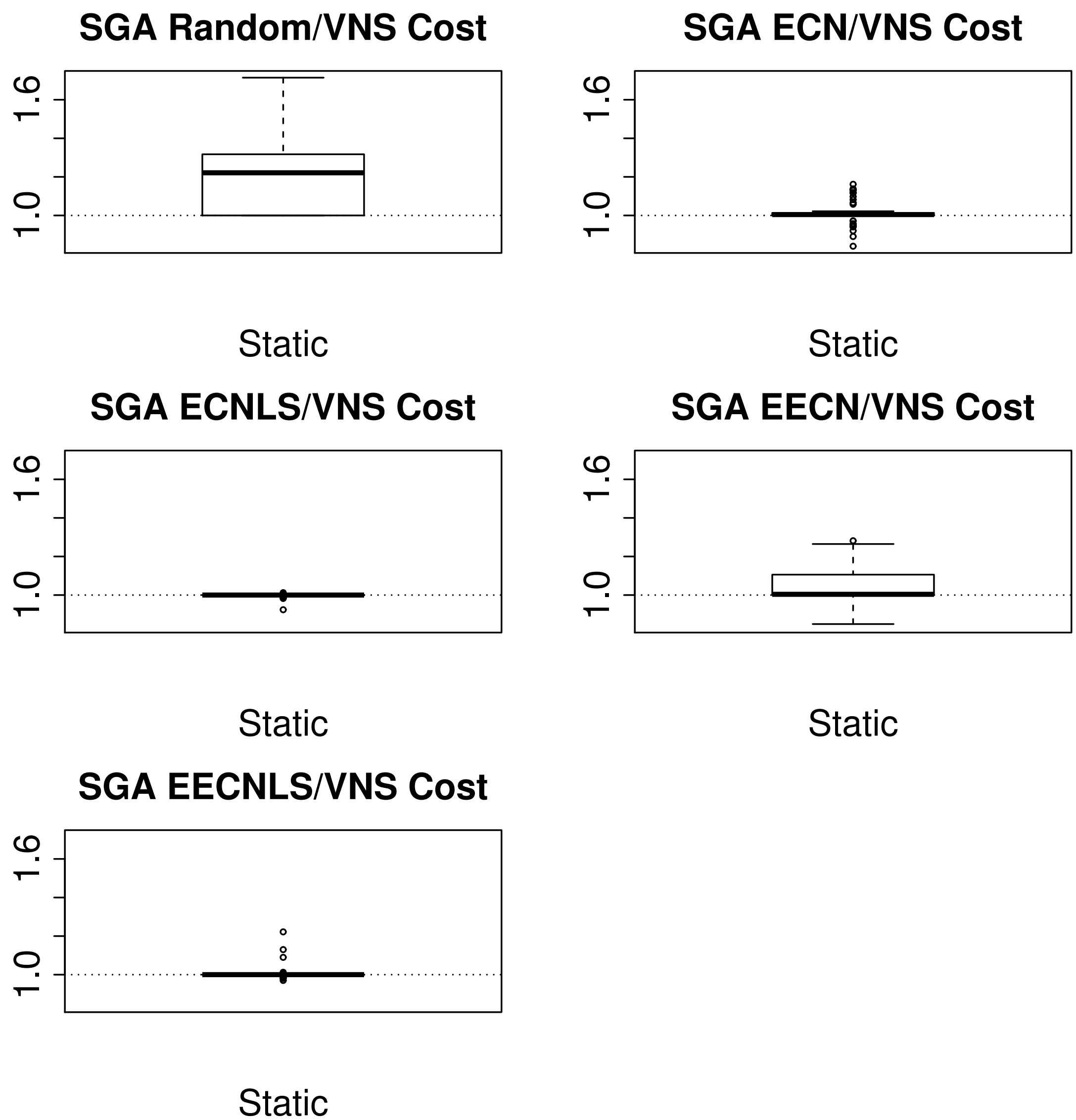

Figure 17: Boxplots of solution cost ratios for SGA over all test runs with static fill rate calculations

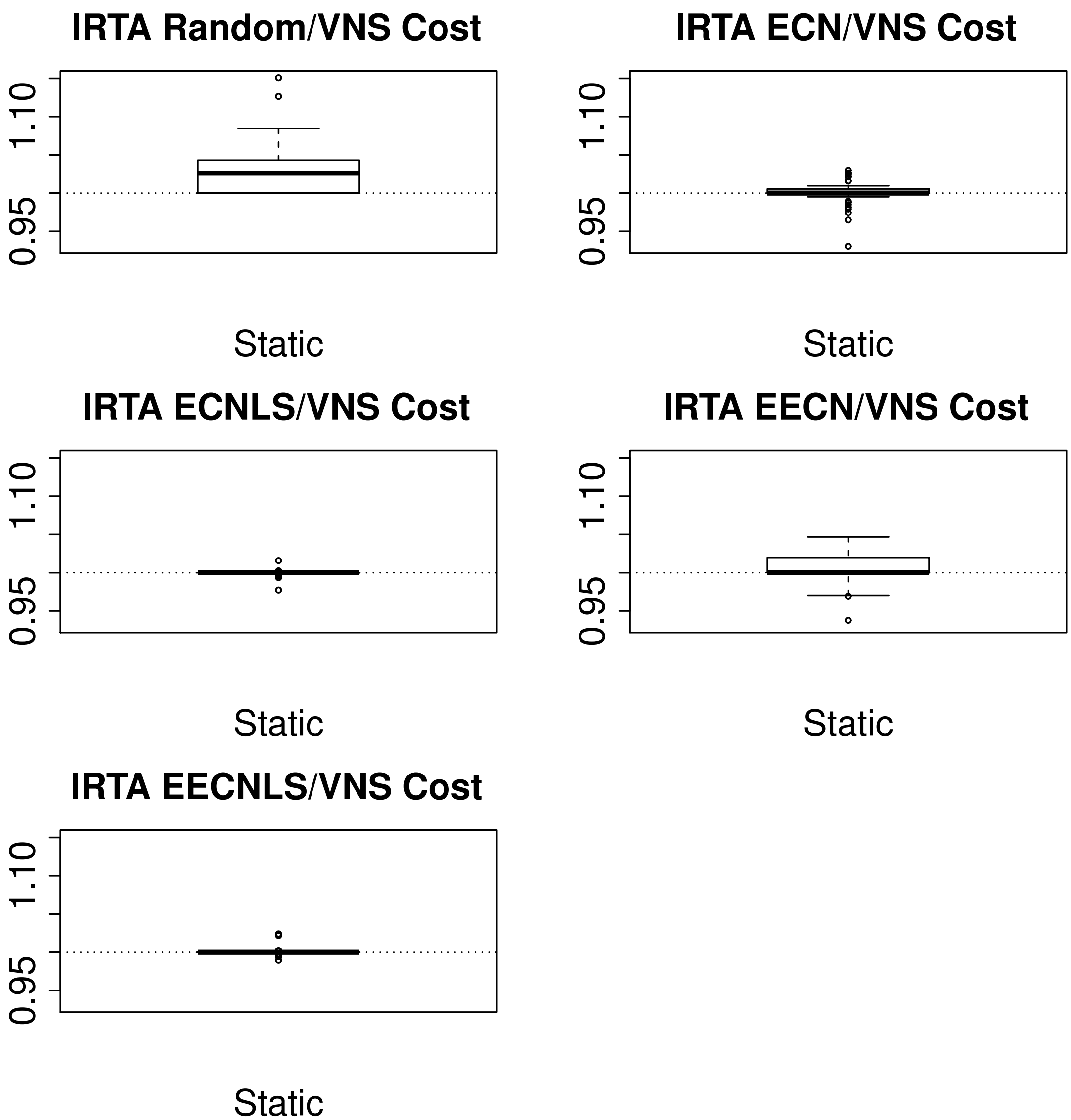

Figure 18: Boxplots of solution cost ratios for IRTA over all test runs with static fill rate calculations

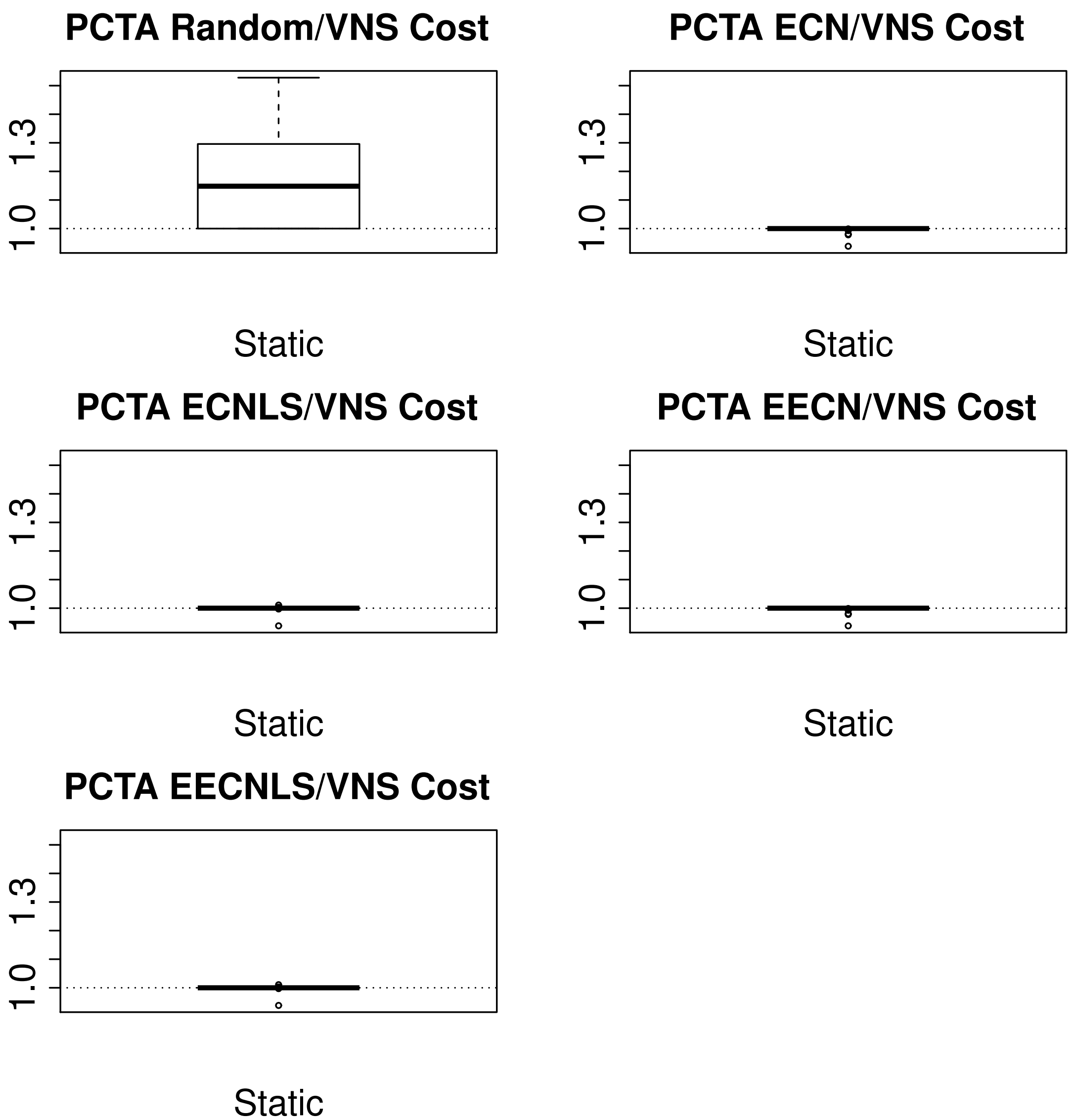

Figure 19: Boxplots of solution cost ratios for PCTA over all test runs with static fill rate calculations

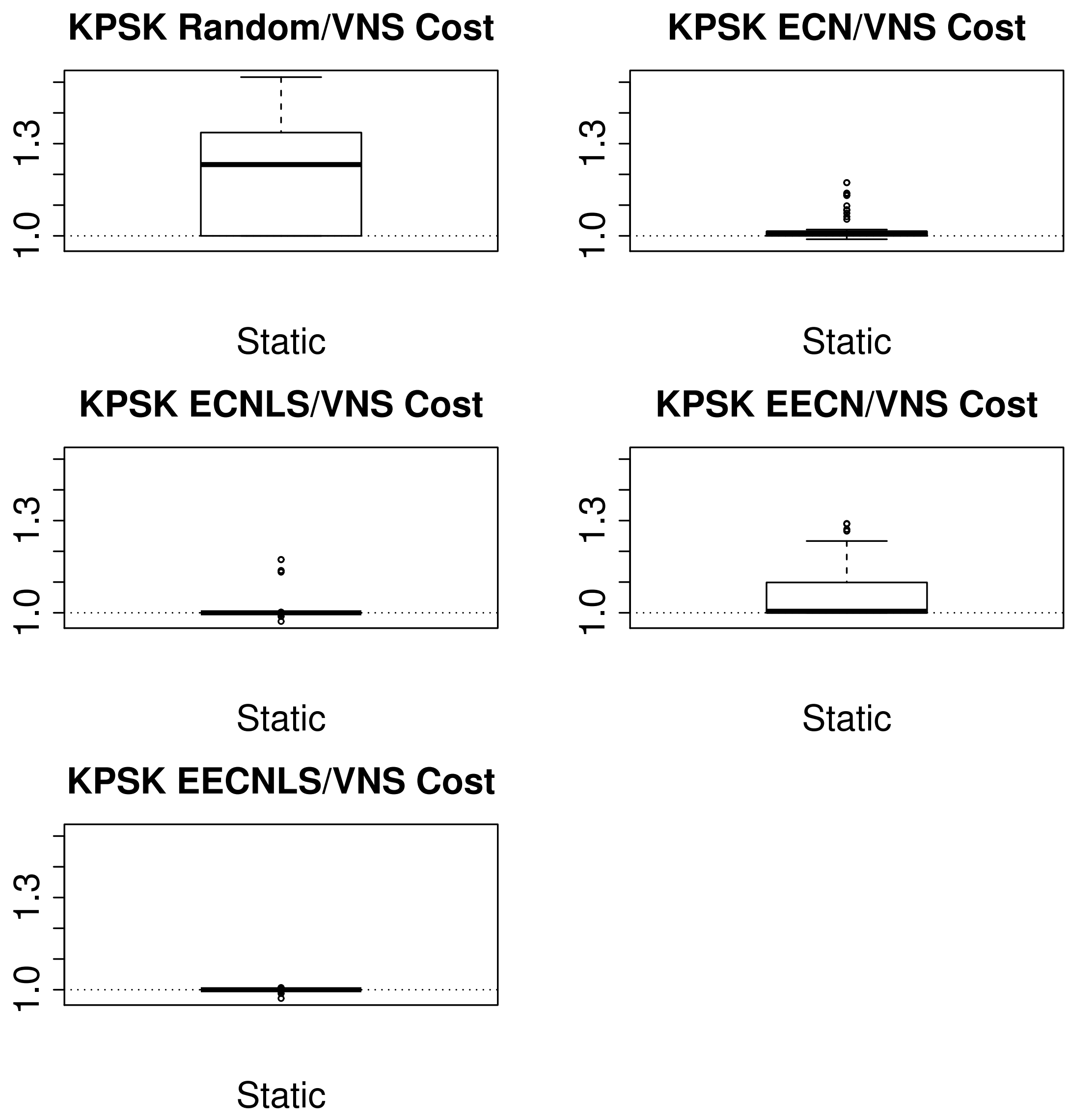

Figure 20: Boxplots of solution cost ratios for KPSK over all test runs with static fill rate calculations

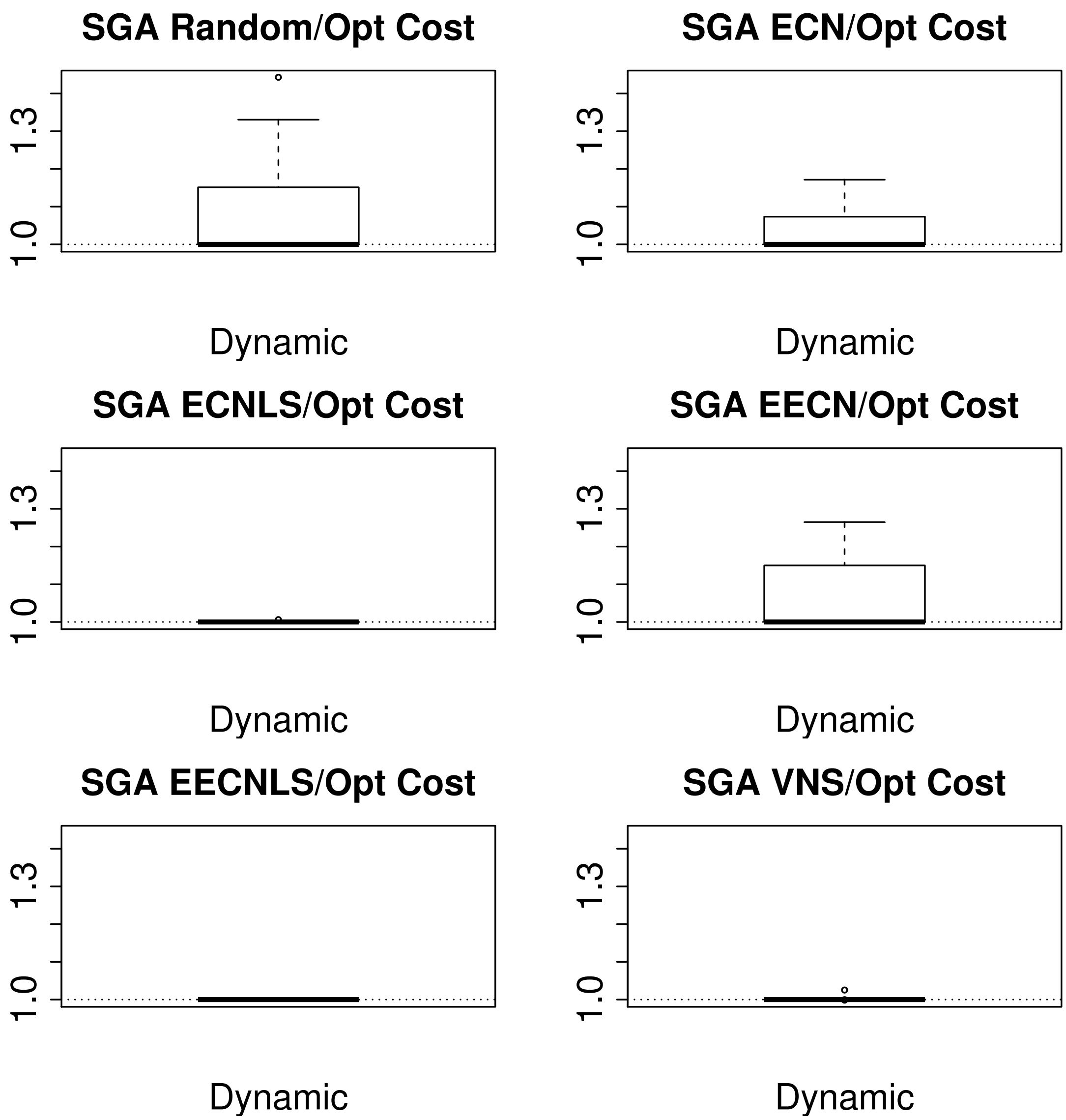

Figure 21: Boxplots of solution cost ratios for SGA over all test runs with dynamic fill rate calculations

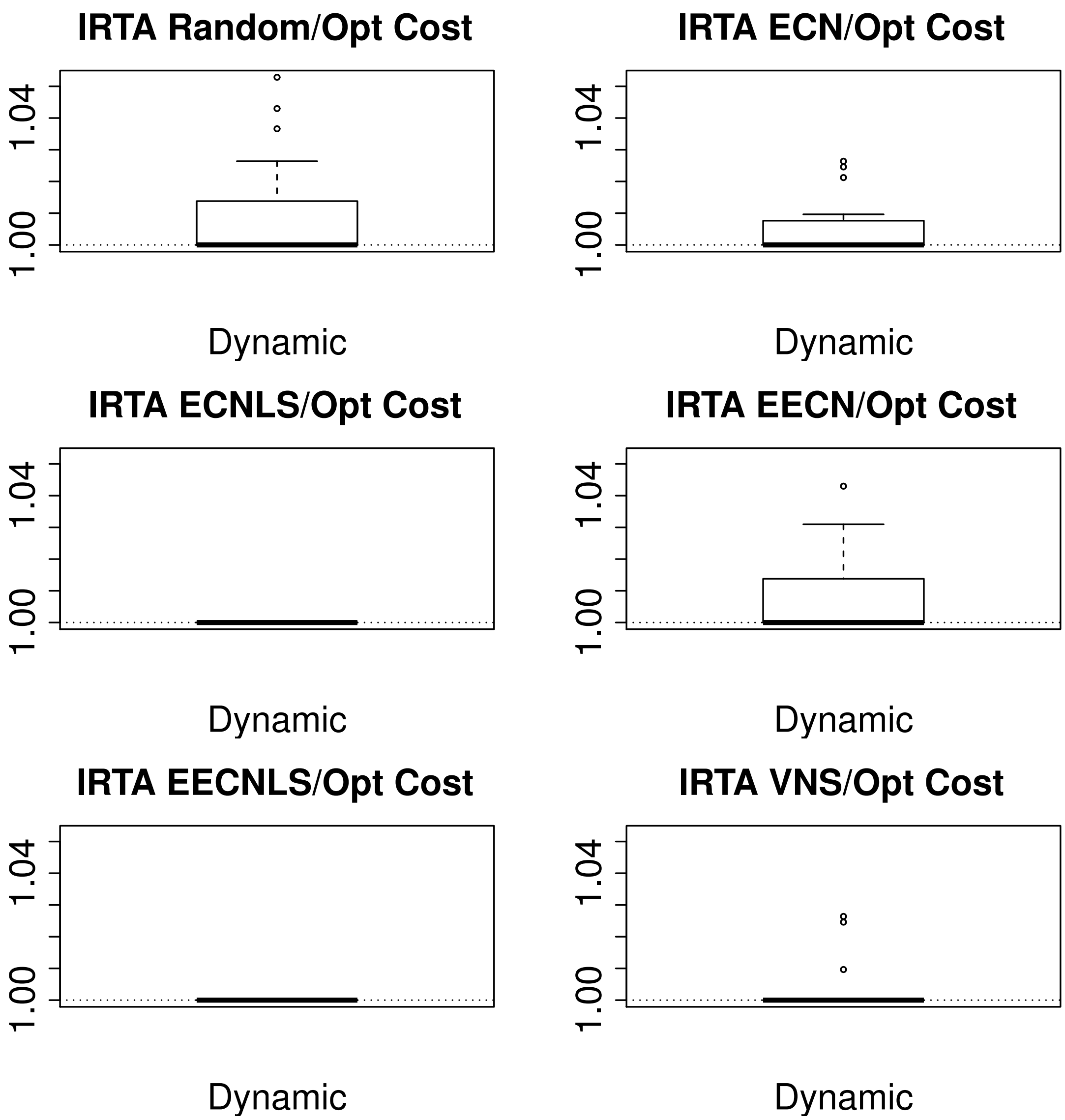

Figure 22: Boxplots of solution cost ratios for IRTA over all test runs with dynamic fill rate calculations

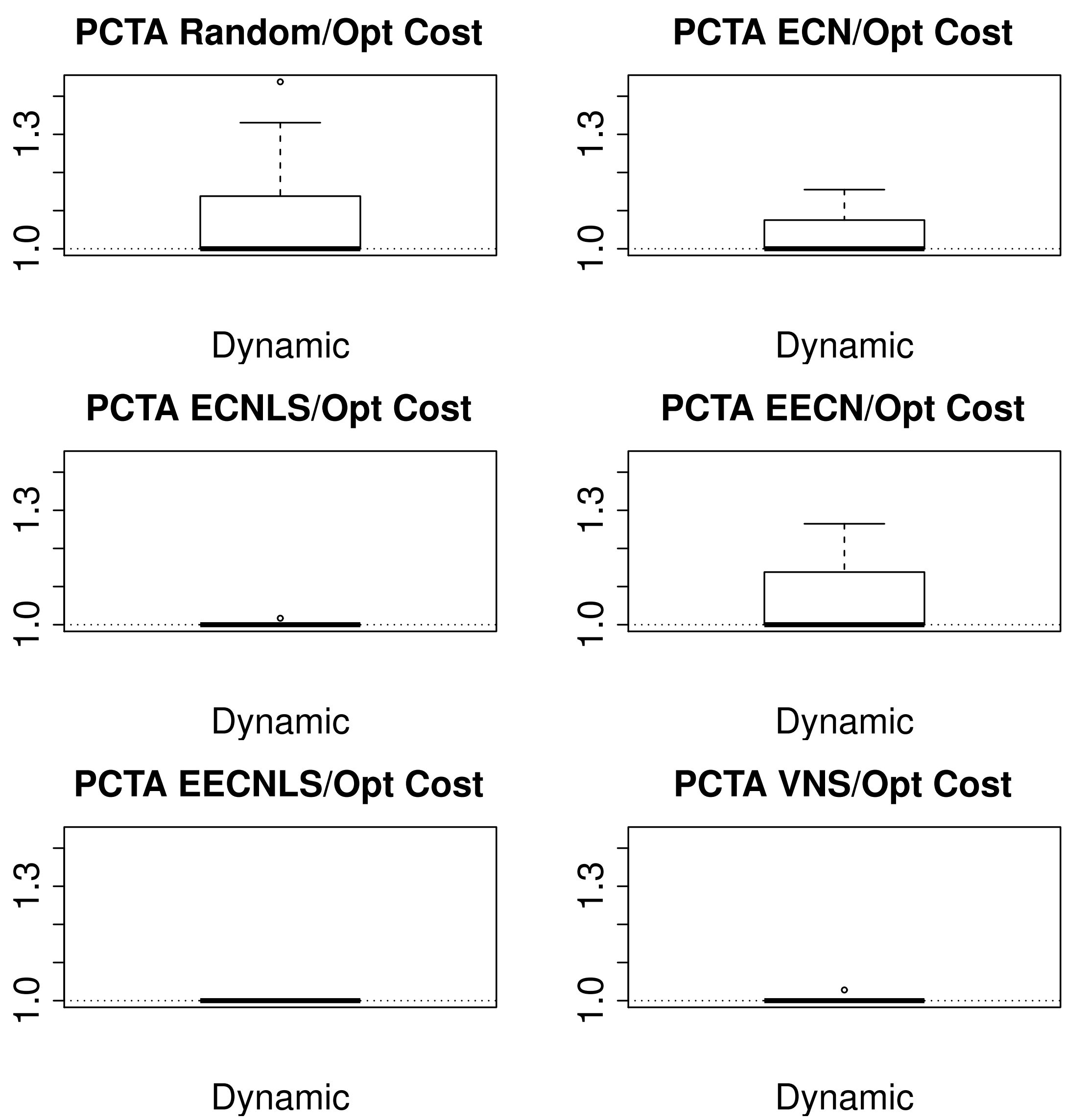

Figure 23: Boxplots of solution cost ratios for PCTA over all test runs with dynamic fill rate calculations

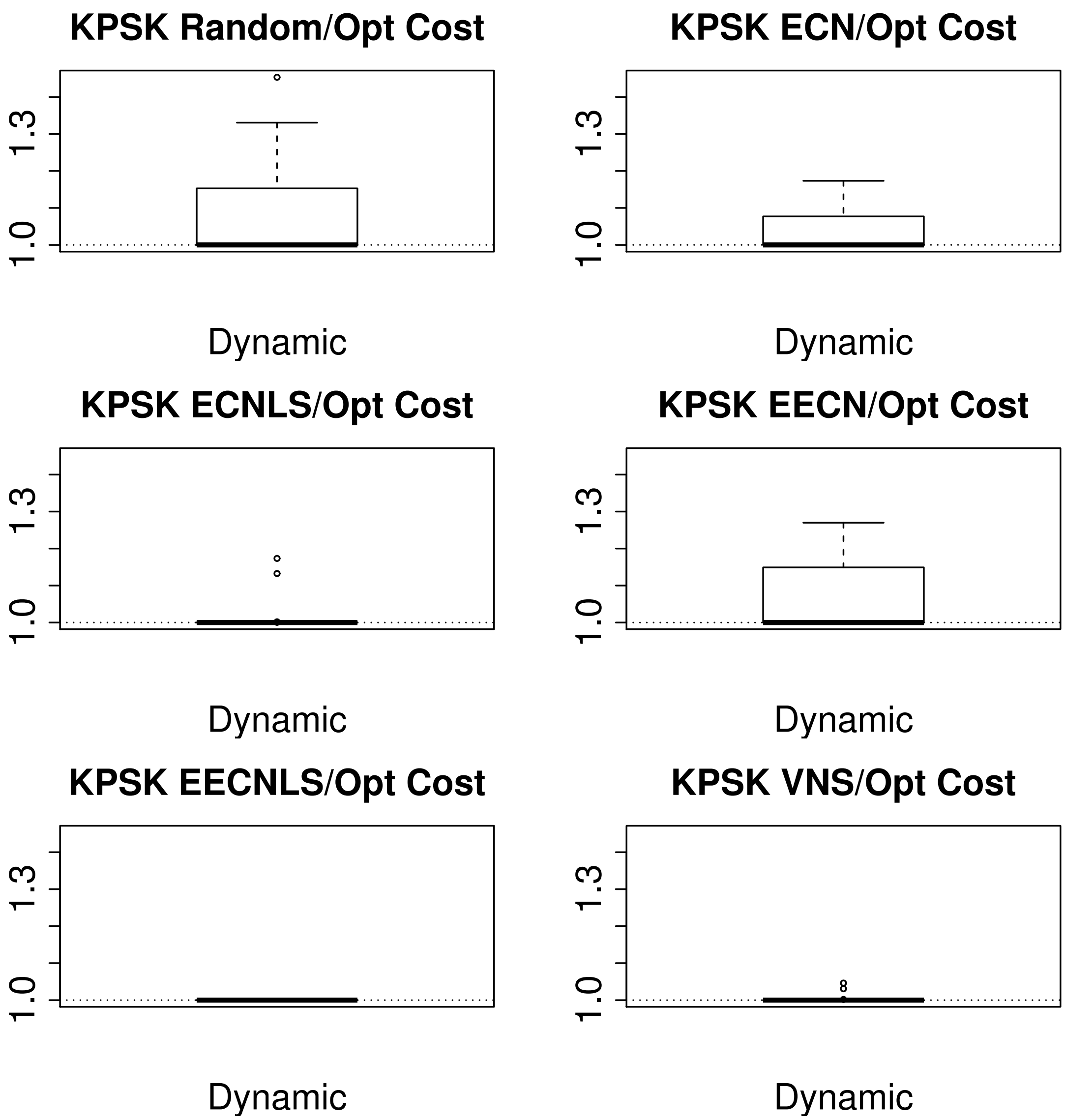

Figure 24: Boxplots of solution cost ratios for KPSK over all test runs with dynamic fill rate calculations

other customers, so travel between customers, an activity that is non-value-add, is minimized. Those are the two big asks of a route and ECN does a decent job of filling them, as borne out by the computational test results. Of course, as we have seen even in two dimensions in Section 3.5, ECN is not always optimal but it is pretty good on the whole and is much simpler to understand and implement than the other non-random routing algorithms.

3.7 Concluding Remarks

This chapter extends the work of Chapter 2 by introducing the element of customer geography into the model of a spare parts supply chain with FSL recourse. The adapted model put forth in this chapter melds two decision problems together, an inventory problem and a routing problem, to satisfy the overall objective of generating a low-cost spare parts inventory stocking policy when technicians have access to recourse inventory at an FSL. To solve the inventory problem we reuse four algorithms from Chapter 2, SGA, IRTA, PCTA, and KPSK, since none of them use any information about the manner in which the expected fill rate of a solution is calculated or structured. To solve the routing problem (which is a part of the fill rate calculation) we propose six heuristics, Random, ECN, ECNLS, EECN, EECNLS, and VNS. In addition, we provide a result that the optimal route for customers in a one-dimensional region is not straight out from the FSL starting at the customer nearest the FSL, which we interpret to mean that both distance from the FSL and distance from other customers affect the construction of the fill-rate-maximizing route ordering of customers.

After a thorough regimen of computational testing documented in this chapter, we concluded that random routing is significantly more expensive than any of the other routing methods, costing twenty percent more in many cases. A service provider has a strong incentive to route all of its customers *a priori*, whether that be through negotiating its contractual obligations or some other means. However, the service provider need not implement any complicated routing decision support systems. It can achieve almost all benefits of routing with a simple heuristic that does require any fill rate calculations, *i.e.*, ECN.

CHAPTER IV

DESIGN OF A SPARE PARTS SUPPLY CHAIN WITH FORWARD STOCKING LOCATION RECOURSE

4.1 Introduction

Many businesses in the modern economy rely on high-technology equipment to provide competitive products and services to their customers. These organizations invest capital in machinery they do not necessarily know how to create or maintain, yet it is essential to their competitiveness that these capital investments work properly. Depending on how they use these machines, such companies have varying levels of urgency for resolutions when problems arise. Companies usually enter into service-level agreements at the time of purchase, which specify how quickly the original manufacturer of the device or other repair agency must diagnose and repair a malfunctioning machine once it is reported. With such an agreement in place, users are free to benefit from machine capabilities with controlled expectations regarding possible downtime.

From the equipment manufacturer's perspective, the set of possible machine failure resolutions may be quite vast since a repair often involves replacing a malfunctioning part with a working spare part, and the number of part types can be large. In most settings, the technician does not know which part will need to be replaced before arriving on-site to the customer. However, carrying every part is costly, and many parts are used infrequently and some others very rarely. One compromise that allows access to more parts at lower cost is pooling inventory between technicians by holding at forward stocking locations (FSLs), which are spread throughout a region and close to the customer sites. These locations provide flexibility to the service provider for inventory storage, and can potentially reduce the total cost of meeting service-level agreements.

This chapter develops a model to assist a decision maker in configuring a supply chain where multiple technicians must complete multiple jobs in a single day and have access to recourse inventory (*i.e.*, inventory not carried on site in a mobile repair kit) at multiple FSLs. This chapter also proposes an algorithm structure to find good solutions to the model; a solution specifies how many FSLs to operate, how many technicians to hire, and which parts to equip the technicians with in their repair kits. Through computational tests, we show that the best strategy is usually one of two extremes: fewer but busy technicians with close-to-full kits or more but less busy technicians with empty kits. Almost all solutions in the tests we ran for this chapter had only a single FSL. When the relative cost of parts to labor (as defined later in this chapter) is around four, the best strategy falls somewhere between the two extremes. We show that an FSL strategy implemented in this fashion can save 10% to 30% over a strategy that stocks spare parts only in technician repair kits.

Section 4.2 reviews the relevant literature for choosing the parts in the repair kit, the number of technicians to hire, and the number of FSLs to maintain. Section 4.3 contains the spare parts supply chain with FSL recourse design problem formulation. Section 4.4 describes the proposed algorithm structure to find a good solution to the design problem. Section 4.5 presents results from computational experiments and Section 4.6 concludes.

4.2 Related Literature

The major contribution of this chapter is a framework for deciding how to design a spare parts supply chain for traveling repair technicians with access to supplemental inventory at locations distributed throughout a single large service region. The design decision includes determining the number of technicians, the number of FSLs, and how to allocate part types between the technicians' repair kits and the FSLs. Since this supply chain structure has not been studied outside of this thesis to the best of our knowledge, there is no literature on this problem, or prior results to report. We report, however, previous related research conducted at the individual problem level for the three components, repair kit, technicians, and FSLs, that make up a complete design of a spare parts supply chain with FSL recourse.

4.2.1 Repair kit literature

The repair kit decision problem where technicians do not have access to any supplemental inventory has been studied by numerous authors in the past. The introduction of supplemental inventory at FSLs into the problem model was made in Chapter 2.

The seminal repair kit paper of Smith et al. [66] develops a model that minimizes an unconstrained objective function that includes both holding costs and penalty costs for a single-job problem with no more than one part type used on a given job. For the same problem setting, Graves [42] proposes a model that minimizes holding cost subject to satisfying a specified job fill rate and thus avoids the determination of a penalty cost parameter. The model in Mamer and Smith [56] also minimizes an unconstrained total expected yearly cost function for a single-job repair kit problem but relaxes the assumption of only one part type failure.

Heeremans and Gelders [45] introduces a model that minimizes holding cost for a multiple-job problem but constrains tour fill probability rather than job fill rate as perceived by the customer. The model in Teunter [71] extends that of Heeremans and Gelders [45] by indeed constraining job fill rate. Bijvank et al. [13] derives a closed-form expression for expected job fill rate in a general multiple-job setting where one or more units of a part type may fail and inventory from the kit is not set aside for a job that cannot be completed. Naturally, all of the cited papers propose algorithms to solve the repair kit problems they have formulated.

Chapter 2 modifies the repair kit problem framework as it is used in the papers cited above to represent a spare parts supply chain with FSL recourse. This earlier chapter also proposes algorithms to find good inventory solutions for the repair kit and FSL in this previously-unstudied problem setting.

4.2.2 Number of technicians literature

Determining the number of technicians to complete repairs is a type of fleet sizing problem, in a sense. Intricately tied with the size of the fleet is the routing of the vehicles themselves. With shorter routes, fewer vehicles are required to satisfy customer demand and a lower-cost fleet can be operated.

However, the first attempts to study the fleet sizing problem did not incorporate the routing aspect of the problem into the model. Kirby [50] used analytical and statistical methods to find the optimal trade-off between owned and hired wagons for a small railway system that minimized total expected cost per day. Wyatt [75] extends the model in Kirby [50] to include a variable cost element. Maskell [58] used a simulation approach to determine the optimal size of a fleet composed of a single type of vehicle to complete local deliveries at minimum cost. Gould [41] extends further [50] and [75] by allowing for a non-homogeneous fleet and models the problem as a linear program.

Later modeling efforts began taking the routing aspect into account when choosing the size of the fleet. Ball et al. [7] examines the fleet sizing problem where demand along certain directed arcs must be covered by a set of cycles and the decision maker has the option of contracting out service to a common carrier. The problem is to determine the optimal fleet size and the appropriate vehicle routes to cover the directed arcs with demand while not exceeding maximum route time restrictions. Etezadi and Beasley [31] develop a mixed-integer programming formulation of a vehicle fleet composition problem that can be used to determine both the types of vehicles to be used and the number of each type. Golden et al. [39] looks at the fleet size and mix problem, which is a generalization of [7] in which the assumption that all vehicles have the same cost and capacity is relaxed. Gheysens et al. [36] evaluates various techniques for solving the fleet size and mix vehicle routing problem (FSMVRP) and proposes a new heuristic based on a lower bound procedure. Ferland and Michelon [35] models a vehicle scheduling problem with multiple vehicle types using integer programming where the objective function can be to minimize investment costs and hence reduce fleet size to a minimum. Desrochers and Verhoog [27] presents a new savings heuristic for the FSMVRP based on successive route fusion that is easy to implement. Salhi and Rand [65] gives a literature review on early papers that study fleet composition without accounting for routing and also proposes an efficient heuristic for determining the composition of a vehicle fleet that makes use of a perturbation procedure to improve utilization of vehicles. More recently, Hoff et al. [46] provides a extensive survey of the FSMVRP.

There are only a few papers about fleet sizing for traveling technicians; the problem has not received much attention from the operations research community. Most papers on the topic include the routing aspect at the same time as fleet sizing but we limit ourselves to only the latter in this chapter. Cortes et al. [26] details a simulation procedure to determine the optimal fleet size for technicians that repair failures of copy machines for Xerox Chile, which is almost exactly what we want to do for a part of the decision problem in this chapter. The authors allow for significant variability in the number of calls per day and in the travel and service times for their model, which they take from historical data on repair operations. In the end, the paper generates a curve for the relationship between fleet size and repair performance to help inform the decision maker during the staffing process. Our model on the other hand assumes only the smallest of variability in the time to complete a repair job, *i.e.*, there are two possible outcomes, the likelihoods of which are also determined by the decision maker through the part types stocked in the repair kit. We represent the situation of choosing the number of technicians in a more simplified fashion and thus are able to generate a single recommendation for the number of technicians to hire.

4.2.3 Number of FSLs literature

To the best of our knowledge there are no papers about deciding how many FSLs to build for a fleet of technicians. There are, of course, many papers on a related problem known by many names including simple plant location problem, capacitated or uncapacitated warehouse location problem, among other similar titles. The literature on this problem has been reviewed by Krarup and Pruzan [53], Aikens [3], and more recently by Klose and Drexl [51].

In the plant location problem, the decision maker must choose a subset of locations at which to open facilities from a larger set of potential locations to satisfy customer demand at minimal cost. The cost of a solution is usually the sum of fixed and variable operating costs for each location that is opened (the latter of which are dependent on total customer demand assigned) and the transportation costs of shipping goods from the locations to the customers. There are varieties of this problem in which the amount of demand that can be satisfied by a single location is capacitated (see Baumol and Wolfe [10], Beasley [11], Akinc and Khumawala [4], Sa [64], and Jacobsen [47]) and those in which amount of demand is uncapacitated (see Efroymson and Ray [30], Kuehn and Hamburger [54], Khumawala [49], Manne [57], and Feldman et al. [34]). The authors use many different methods to solve the problems, including a Lagrangean relaxation of a mixed-integer programming formulation in Beasley [11], branch-and-bound methods in Akinc and Khumawala [4], Efroymson and Ray [30], Sa [64], and Khumawala [49], and yet other heuristics in Jacobsen [47], Kuehn and Hamburger [54], and Feldman et al. [34]. A review of heuristic and exact procedures for solving the capacitated version of the problem was conducted in Sridharan [69].

What sets our model apart from these plant location problems is the nature of the demand at the customer sites as well as the effect of travel on the objective function. In our model, the demand for parts from the FSLs is not deterministic and depends on the part types stocked in the repair kit, which is a choice made by the decision maker. In addition, the travel time from customers to the plant locations (FSLs in our case) does not appear directly in the objective function as we model the problem. Rather, it indirectly affects the objective function by influencing the number of parts that must be stocked in the repair kit to meet the required service level, which is directly tied to the total inventory costs.

4.3 Model of Decision Problem

The service provider is tasked with deciding the complete design of a spare parts supply chain with FSL recourse for a metropolition area, which is the region of interest for this problem. Based on the number of machines with service-level agreements installed within the region and their failure rates, there will be a certain distribution of demand for repair jobs that fluctuates from day to day. The service provider must meet this demand in a satisfactory fashion by providing enough repair capacity and deploying it appropriately. In order to simplify this problem for a first analysis, we ignore that daily variability and assume the service provider aims to have sufficient capacity to serve a target average number of jobs per period, represented by M. A supply chain designed to handle this load will be able to sustain the required repair rate in practice when the number of jobs per day is in actuality a random variable with a distribution where M is a possible outcome, and represents a high percentile of demand, *e.g.*, 90th percentile.

4.3.1 Decision Variables

The decision maker must choose all three aspects of the supply chain design:

- how many technicians to employ, r,

- how many FSLs to operate, e, and
- which part types to stock in the technicians' repair kits, x_i for $i = 1, 2, \ldots, n$.

The allocation of the M jobs to the technicians working on any given day is not a decision the service provider has to make because it is optimal to allocate the jobs uniformly, as defined in Definition 1 and illustrated in Figure 25.

Definition 1. *A job allocation is uniform if the difference between the numbers of jobs assigned to any two technicians is no greater than one.*

For this analysis, we suppose that all jobs are identical to one another and assume that every technician has the same technical competencies and repair capabilities, which gives us a homogeneous set of customers as defined in Definition 2. (Note that we do not consider technician routing in this chapter.)

Definition 2. *A set of customers is homogeneous if the times to complete repairs at all customers, represented by T_j for customer j, are identically distributed random variables.*

In this setting, it is optimal to allocate the jobs uniformly among the technicians, as described in Theorem 1.

Theorem 1. *Consider the multiple-job repair kit problem with FSL recourse where the set of customers is homogeneous. In this setting, a uniform job allocation has the highest overall expected fill rate.*

The proof of Theorem 1 can be found in Appendix A on page 120. We detail how to calculate the expected fill rate in Section 4.3.3.

In this chapter, we assume that each technician is assigned the same number of jobs, m, which we claim best represents an equitable allocation of jobs to technicians that one might expect when the set of customers is homogeneous. The value of m is

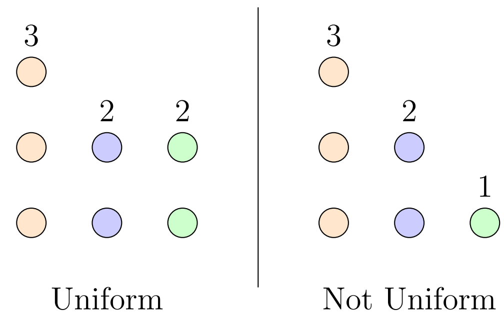

Figure 25: Job allocations

chosen such that roughly M jobs are assigned in total per period, *i.e.*, rm is roughly equal to M. Roughly equal means, for the purposes of this chapter, that rm is the multiple of r closest to M. For example, if $r = 5$ and $M = 32$, $m = 6$. The value of rm does not necessarily need to be less than M. If $r = 5$ and $M = 34$, then we set $m = 7$. Although m is dependent on r, a decision variable, we do not use any subscript notation for sake of brevity.

4.3.2 Objective

The objective of the decision problem is to minimize the total costs of the supply chain per job, represented by C. Total costs include three elements:

1. inventory costs for parts in the repair kits,

2. inventory costs for parts at the FSLs, and

3. labor costs for the employed technicians.

The inventory cost for a part in the repair kit is the annual holding cost of one unit of the part, c_i, times the number of units of the part type in a repair kit, k_i, times the number of technicians, r, a decision variable. Similarly, the inventory cost for a part at the FSL is the annual holding cost of one unit times the number of units at the FSL, f_i, times the number of FSLs, e, also a decision variable. The labor costs are the product of the cost per technician, s, times the number of technicians, r.

Enough units of each part type are kept in the repair kit so that the probability of stocking out is no greater than a parameter ρ, which we set at 0.01 in this analysis. A single repair kit must be used to complete m jobs in a period between replenishments without stocking out of any part type. Each job consumes one spare part from the kit and the parts required from job to job are independent of one another.

The total number of part type i used over the course of m jobs, represented by Y_i, is a binomial random variable with success probability p_i, *i.e.*, the likelihood that part type i is malfunctioning and needs replaced. Hence, for each i, we determine the number k_i to stock in the repair kit such that

$$P(Y_i \leq k_i) = \sum_{j=0}^{k_i} \frac{m!}{j!(m-j)!} p_i^j (1-p_i)^{m-j} \geq 1 - \rho,$$

where $\{Y_i \leq k_i\}$ is the event that part type i does not stock out. The FSL quantity for each part can be found similarly with the number of customer jobs per FSL substituted for the number of jobs per repair kit, m. Assuming that the FSLs are equitably distributed and the technicians will always go to the nearest FSL, the number of customer jobs per FSL is equal to the total number of jobs, rm, divided by the number of FSLs, e.

Once the kit and FSL quantities are calculated, the objective function parameters are completely specified. Note, however, that the objective function parameters depend then on the values of the decision variables for the number of technicians and number of FSLs:

1. Inventory costs for parts in the repair kits: $r \sum_{i=1}^{n} c_i k_i x_i$, where k_i is kit quantity, a function of the number of jobs per technician

2. Inventory costs for parts at the FSLs: $e \sum_{i=1}^{n} c_i f_i (1 - x_i)$, where f_i is the FSL quantity, a function of the total number of jobs and number of FSLs

3. Labor costs for the employed technicians: sr

The objective function is the sum of these three costs divided by the total number of jobs,

$$C(r, e, x_1, \ldots, x_n) = \frac{1}{rm}\left[r\sum_{i=1}^{n} c_i k_i x_i + e\sum_{i=1}^{n} c_i f_i (1 - x_i) + sr\right].$$

4.3.3 Constraints

In practice, the service provider would like to meet the service-level agreements as closely as possible because there is a penalty cost that must be paid whenever a violation occurs. One approach would have been to include this penalty cost in the objective function and minimize the total expected cost per job without any constraints. The downside of this approach is that it introduces the parameter of penalty cost which may be hard to estimate or is subject to negotiation. The approach that we take is to model adherence to service-level agreements as a constraint in the math programming formulation.

All jobs that are assigned in a service period have a common deadline by which they must be completed, the end of the service period, represented by d. The expected fill rate, ϕ, is the number of jobs that are completed by d in expectation divided by the total number of jobs, rm. Letting $\mathcal{J}$ be the set of jobs and $\mathcal{T}_j$ be the time at which job j is completed we have

$$\phi = \frac{1}{rm}\sum_{j \in \mathcal{J}} P(\mathcal{T}_j \leq d).$$

Job j is started after all m_j jobs ahead of it in the to-do list of the technician to which it was assigned are completed. The time to complete a job, T_j, is a random variable and hence job j is completed at

$$\mathcal{T}_j = \sum_{k=1}^{m_j} T_k + T_j.$$

With each technician assigned m jobs as described in Section 4.3.1, the expected fill rate for all technicians is the same as the expected fill rate for a single technician.

The expected fill rate for a single technician (with a slight abuse of notation: let j represent the the fact that a customer job is in the jth position in a technician's to-do list) is

$$\frac{1}{m}\sum_{j=1}^{m} P(\mathcal{T}_j \leq d) = \frac{1}{m}\sum_{j=1}^{m} P(\sum_{k=1}^{j-1} T_k + T_j \leq d).$$

There are two possible outcomes for each T_j, which are independent and identically distributed. The first outcome is that the required spare part is contained in the repair kit, in which case $T_j = \alpha$, a given parameter value. The probability that the required spare part is contained in the repair kit is $\sum_{i=1}^{n} p_i x_i$. The second outcome is that the technician must retrieve the needed spare part from the FSL, in which case $T_j = \beta$, which is strictly greater than α. How much greater β is than α depends on how far away the nearest FSL is, which is a function of the number of FSLs maintained by the service provider.

The more FSLs there are the less extra time is needed; hence, the time penalty for an FSL trip, $\beta - \alpha$, is nonincreasing in the number of FSLs. Beyond being nonincreasing, there is nothing more that we will claim about this time penalty. In fact, we show through a simple example (which can be found in Appendix B) cases where the decreases in time penalty are linear, super-linear, and sub-linear. There is an initial time penalty, $\beta_0 - \alpha$, that is fixed and exogeneous, which is the time penalty when only one FSL is operated. Also we assume that when the number of FSLs equals the number of technicians, the time penalty is zero because at that point there are no cost savings to stocking any part types in the FSL; the naïve solution is to keep everything in repair kits and hence there will never be any extra trips. Between these two points of the extra-time-number-of-FSLs graph we fit three functions, a line, a concave parabola centered at the top left point, and convex parabola centered at the bottom right point, as illustrated in Figure 26. These three relationships help us to test more possibilities of how the function might actually change in practice.

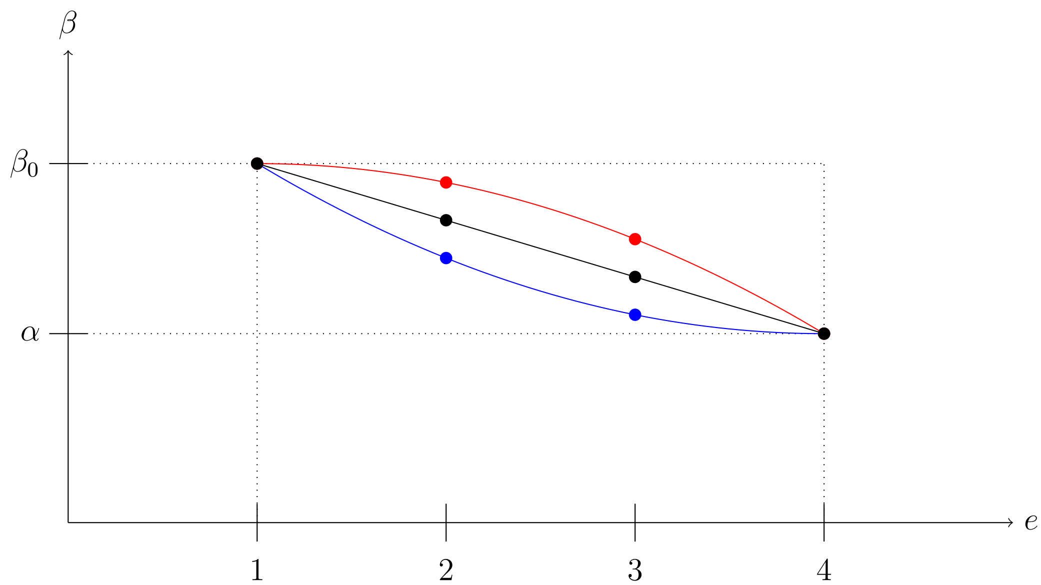

Figure 26: Time for FSL job, β, as a function of the number of FSLs, e, when four technicians, *i.e.*, $r = 4$

With a minimum expected fill rate of γ, the single constraint in the proposed model is

$$\phi(r, e, x_1, \ldots, x_n) = \frac{1}{m} \sum_{j=1}^{m} P(\mathcal{T}_j \leq d) \geq \gamma.$$

The complete optimization formulation follows.

$$\begin{aligned}
\min \quad & C(r, e, x_1, \ldots, x_n) = \tfrac{1}{rm}[\textstyle\sum_{i=1}^{n} c_i k_i x_i + e \sum_{i=1}^{n} c_i f_i (1 - x_i) + sr] \\
\text{s.t.} \quad & \phi(r, e, x_1, \ldots, x_n) = \tfrac{1}{m} \textstyle\sum_{i=1}^{m} P(\mathcal{T}_i \leq d) \geq \gamma \\
& |rm - M| \leq |r'm - M| \, \forall \, r' \neq r, r' \in \mathbb{Z}^+ \\
& |rm - M| \leq |rm' - M| \, \forall \, m' \neq m, m' \in \mathbb{Z}^+ \\
& r \in \mathbb{Z}^+, e \in \mathbb{Z}^+, x_i \in \{0, 1\}, i = 1, 2, \ldots, n
\end{aligned}$$

4.4 Solution Algorithms

The mathematical program developed to represent the service provider's decision problem is an $\mathcal{NP}$-hard optimization problem. The objective function is nonlinear (and nonsmooth), in the decision variables as is the expected fill rate constraint.

Should there exist a polynomial-time algorithm to find the optimal solution with lowest cost then that algorithm could solve a 0/1 nonlinear, nonsmooth knapsack problem in polynomial time, given that such a problem is embedded in the math program. The 0/1 knapsack problem with a linear constraint is an $\mathcal{NP}$-hard optimization problem.

Due to complexity of the problem, we propose heuristic algorithms to find a solution to the math program. The proposed heuristics to find the best repair kit are motivated by the small set of feasible values for the number of technicians and number of FSLs and the fact that choosing a repair kit allocation given these two values is identical to the decision problem from Chapter 2. First, let us show how small the feasible set of technicians and FSLs is.

The number of technicians, r, must be at least one and is technically not bounded from above, but it would be unreasonable and clearly suboptimal to employ more technicians than there would be jobs in a day, *i.e.*, the target number of jobs, M. (We ignore for this chapter any variability in jobs per day and assume that all techncians work every business day.) The number of FSLs, e, must be at least one (zero would not be an FSL strategy) and is also not technically bounded from above. However, it would certainly be unreasonable and suboptimal to have more FSLs than total customers in a region, a number much larger than M. At that point, it would be better to keep an inventory at every customer location, but it would be even better to keep no inventory at the customer sites and give each technician a full repair kit. The part required for the repair would be available with certainty. Thus, a reasonable upper bound on the number of FSLs is the number of technicans, itself a decision variable. A mathematical representation of these constraints follows.

$$1 \leq r \leq M, 1 \leq e \leq r \Rightarrow 1 \leq e \leq r \leq M$$

The total number of feasible combinations (r, e) is $(1 + M)M/2$, which is polynomial in the target number of jobs M.

Given a technician and FSL combination from this small feasible set, the decision problem reduces to that of finding a kit stocking allocation for the appropriate α and β, which are implied by the number of FSLs, and m, implied by the number of technicians (and the target number of jobs). In Chapter 2, we propose numerous heuristics to solve this inventory problem, and we reuse the following for the heuristic in this chapter.

- Subtractive Greedy Algorithm (SGA)
- Initial Ratio Threshold Algorithm (IRTA)
- Kit Probability Search plus Knapsack (KPSK)
- Probability Cost Threshold Algorithm (PCTA)

Recall that we propose three methods to interpolate the FSL job time, convex, concave, and linear. In total, there are twelve variations of the heuristic to solve for a complete solution of number of techncians, number of FSLs, and part types in the repair kit. See Algorithm 9 for pseudocode of the complete algorithm, which exhaustively searches over all feasible combinations of number of technicians and number of FSLs, finds the repair kit for each, and returns the solution with the lowest total cost.

Algorithm 9 Complete Algorithm

```
Best solution (r*, e*, x*_1, ..., x*_n)
for Number of technicians r' ∈ {1, 2, ..., M} do
  for Number of FSLs e' ∈ {1, ..., r'} do
    Solve for parts in repair kit (x'_1, ..., x'_n) given r' and e'
    if (r', e', x'_1, ..., x'_n) better than (r*, e*, x*_1, ..., x*_n) then
      (r*, e*, x*_1, ..., x*_n) ← (r', e', x'_1, ..., x'_n)
    end if
  end for
end for
return Best solution (r*, e*, x*_1, ..., x*_n)
```

The step of solving for parts in the repair kit is executed for each of the twelve variations. The complexity of this heuristic is $(1 + M)M/2$ times the complexity of the heuristic used to solve for the parts in the repair kit, which is at best polynomial in M. The value of M in practice is unlikely to exceed three digits for a region under the auspices of a service provider. The natures of the solutions generated by this heuristic structure as well as the cost saved by using an FSL to stock supplemental inventory are shown in the next section.

4.5 Computational Results

There were two major questions relevant to the spare parts supply chain with FSL recourse design decision that we hoped to address with a thorough computational testing investigation.

1. What does a good solution look like?
2. How much cheaper can repairs be completed with the flexibility offered by FSL inventory recourse?

To answer these questions for a single problem instance, defined by the number of customers and a population of spare parts under scrutiny, we run all twelve algorithm variations plus a no-FSL algorithm to approximate the inventory policy when FSLs are not available for the technicians. We suspect that the answers to these questions will change as the parameters of problem instance, *e.g.*, time for kit job or technician labor cost, change and thus we generated a suite of test instances based on a full factorial experimental design for the factor settings seen in Table 13. The technician cost, s, was set at a value of $100,000$ for all test runs.

4.5.1 What does a good solution look like?

There are three components of a solution to the decision problem in this chapter, which makes it difficult to visually represent a solution in two dimensions. However, it is possible to succinctly describe the nature of a solution with a short description, *e.g.*, few technicians with close-to-full kits. In general, we observed that the solutions

Table 13: Problem characteristics and corresponding levels used in computational testing

Factor	Notation	Levels
Sum of part unit costs/Technician cost	$\sum_{i=1}^{n} c_i/s$	$\{2, 4, 6, 8, 10\}$
Relative time delay for trip to FSL	β_0/α	$\{2.5, 4\}$
Maximum jobs per technician	$\lfloor d/\alpha \rfloor$	$\{4, 8\}$
Number of part types	n	$\{100, 2000\}$
Target number of total jobs	M	$\{20, 60\}$
Length of day	d	$\{7, 9\}$
Minimum fill rate	γ	$\{0.9, 0.95\}$

fell on a spectrum between two extremes: fewer but busy technicians with close-to-full kits and more but less busy technicians with empty kits. Almost all solutions in the tests we ran for this chapter had only a single FSL.

The most important parameter, which determines where on this spectrum a solution falls independent of any other factor, is the relative cost of parts to labor (listed in the first row of Table 13). For most values of this factor, the solution is at one extreme of the spectrum, either few technicians with close-to-full kits or many technicians with empty kits. When the ratio of parts cost to labor cost takes the value 4, the solution may fall somewhere in between the two extreme cases. See Figures 27, 28, and 29 for the fraction of part types stocked in the repair kit as a function of the cost ratio under the three delay function forms. See Figures 30, 31, and 32 for the utilization of the technicians, *i.e.*, the number of jobs each is assigned divided by the maximum number of jobs a technician could complete in a day, as a function of the cost ratio under the three delay function forms.

We can estimate the relative cost of parts compared to labor for our industry partner based on information provided about the costs of their parts. The total unit costs of the spare parts stocked in their supply chain was about \$1.2 million at the time we collaborated with them. If we assume that it costs 30% of the unit cost to hold a part for a year and that it costs $\$100,000$ to employ a technician, the ratio of interest is 3.6. The levels of the parts to labor cost ratio chosen for testing in this

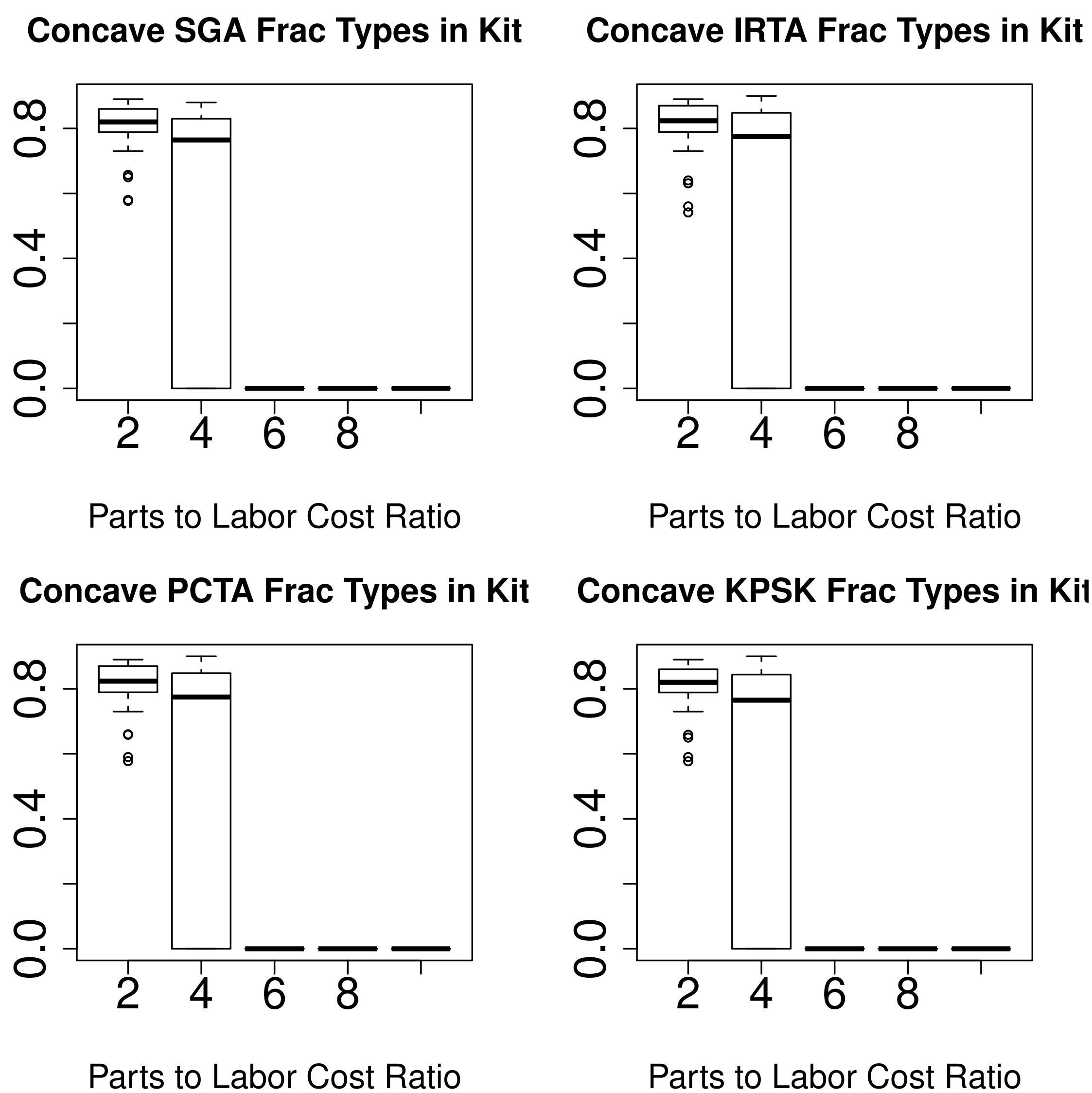

Figure 27: Boxplots of the fractions of part types stocked in the repair kit for all test runs and inventory algorithms under concave delay function form

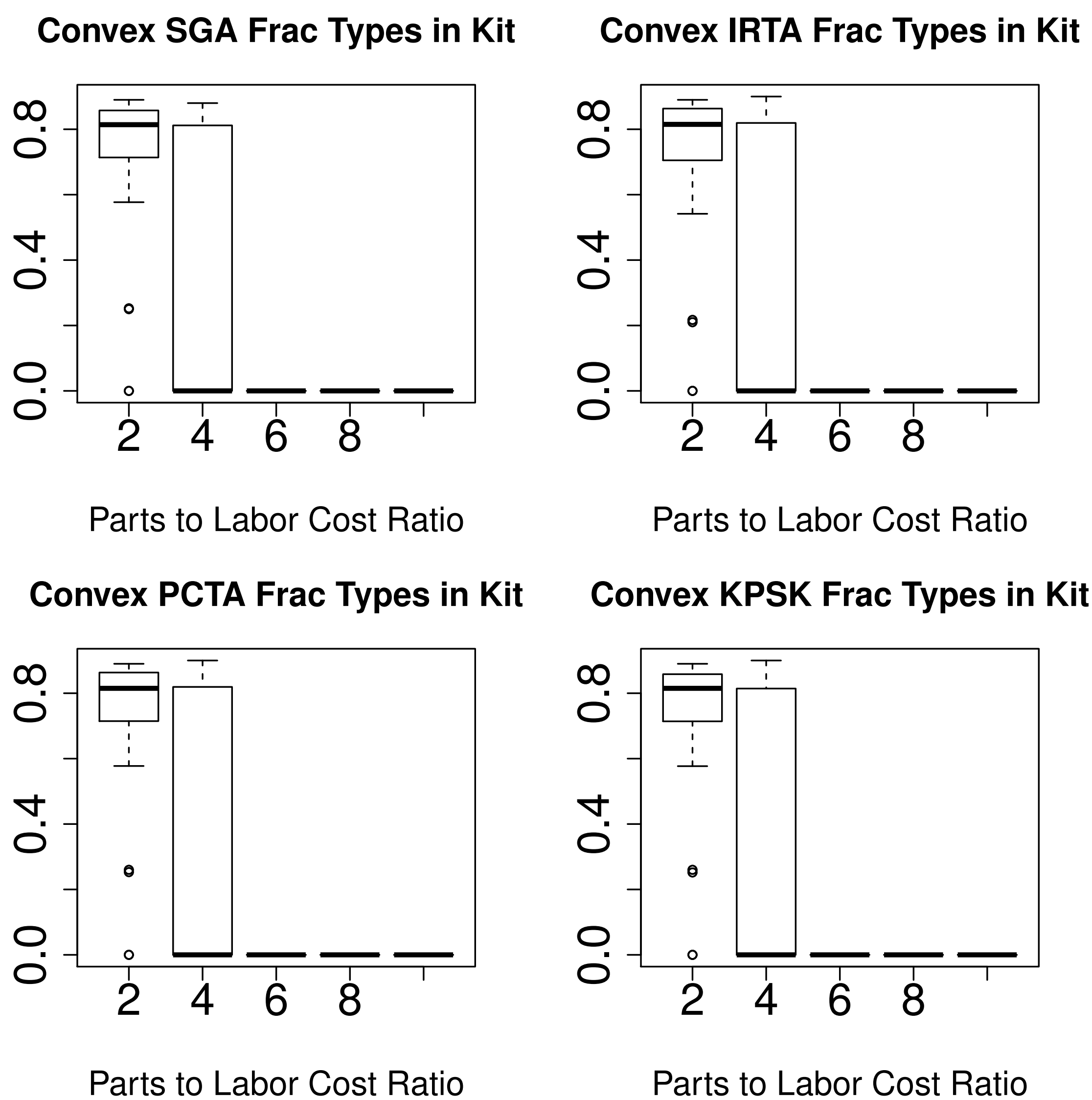

Figure 28: Boxplots of the fractions of part types stocked in the repair kit for all test runs and inventory algorithms under convex delay function form

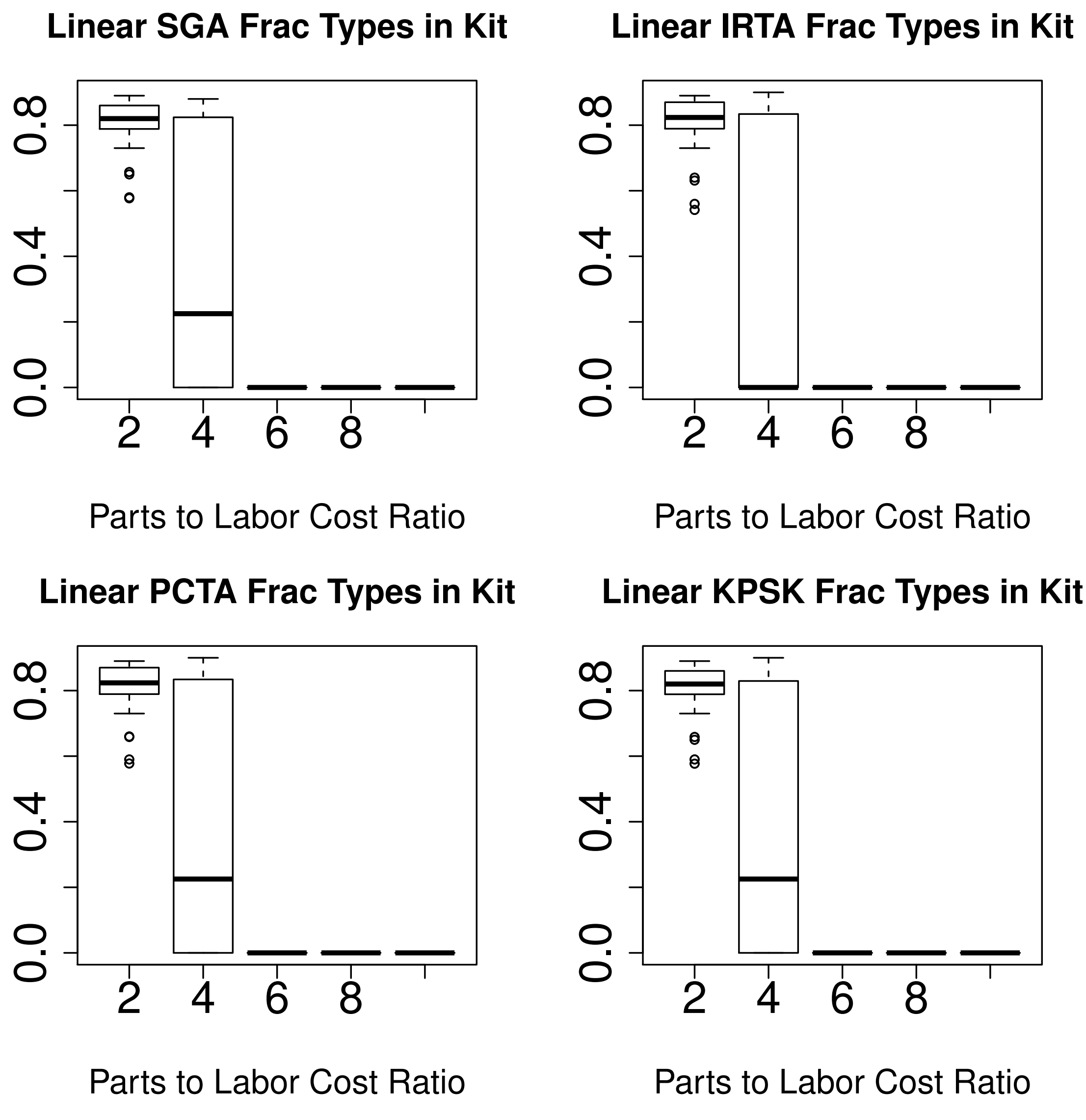

Figure 29: Boxplots of the fractions of part types stocked in the repair kit for all test runs and inventory algorithms under linear delay function form

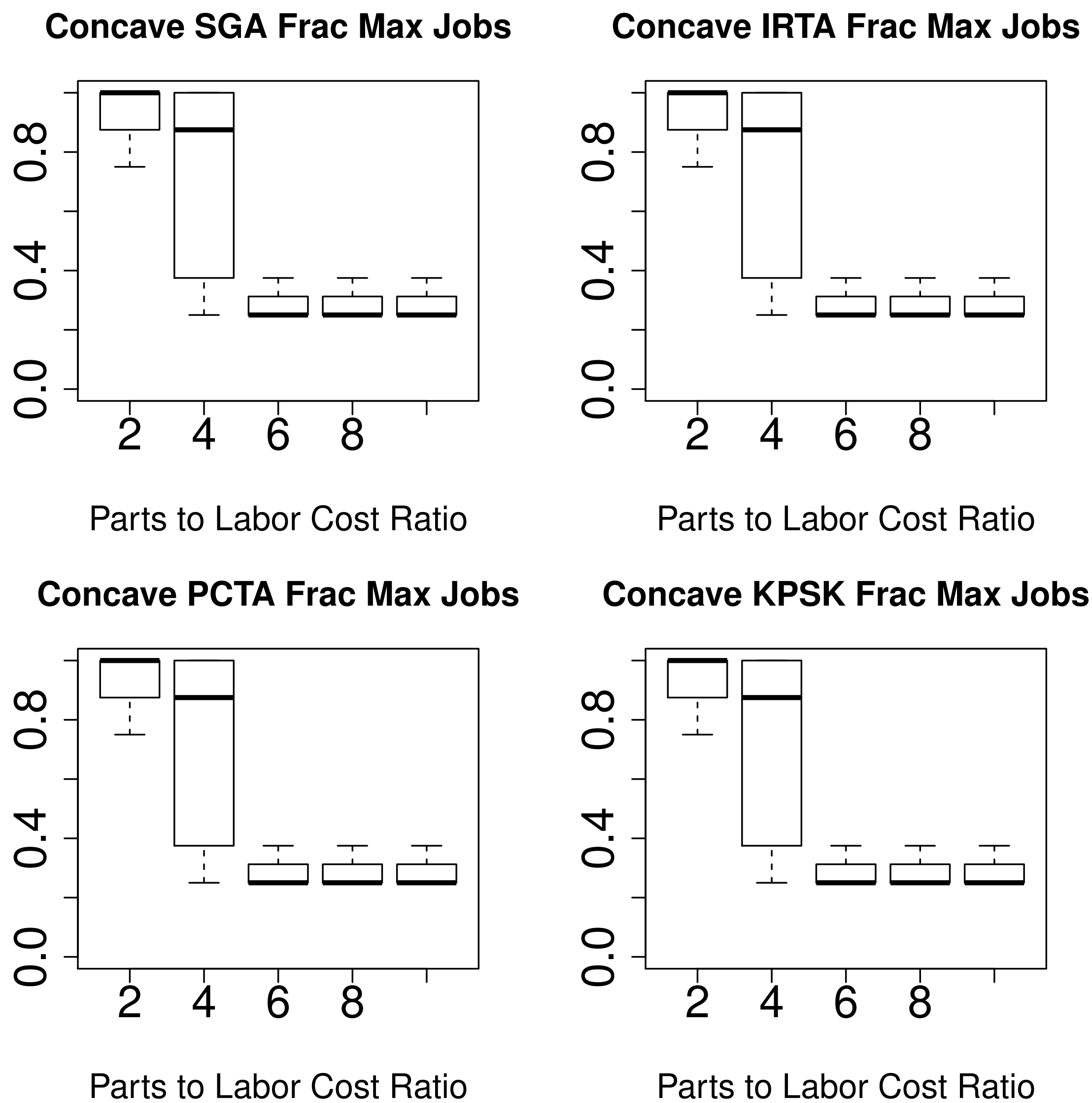

Figure 30: Boxplots of technician utilization for all test runs and inventory algorithms under concave delay function form

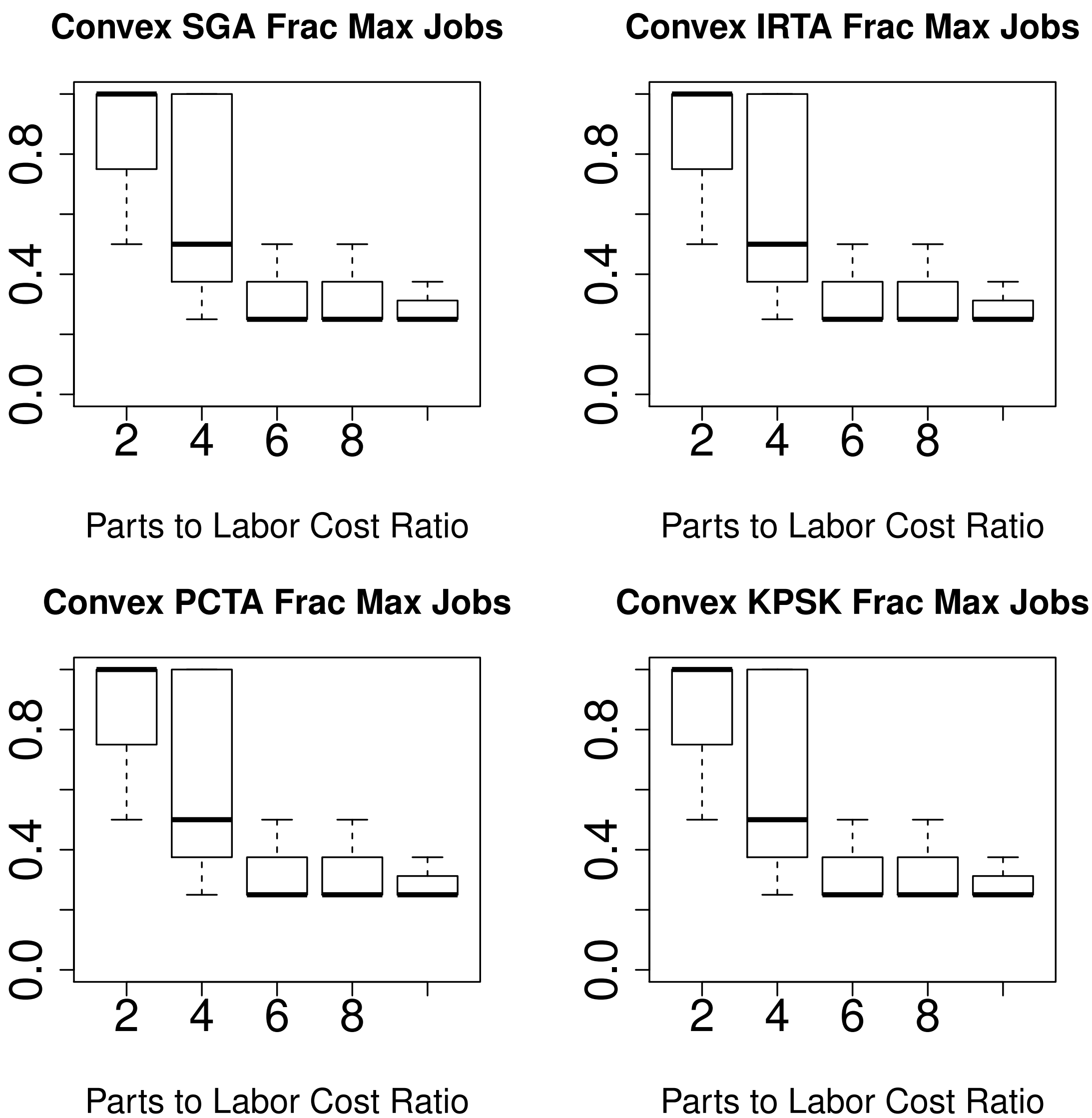

Figure 31: Boxplots of technician utilization for all test runs and inventory algorithms under convex delay function form

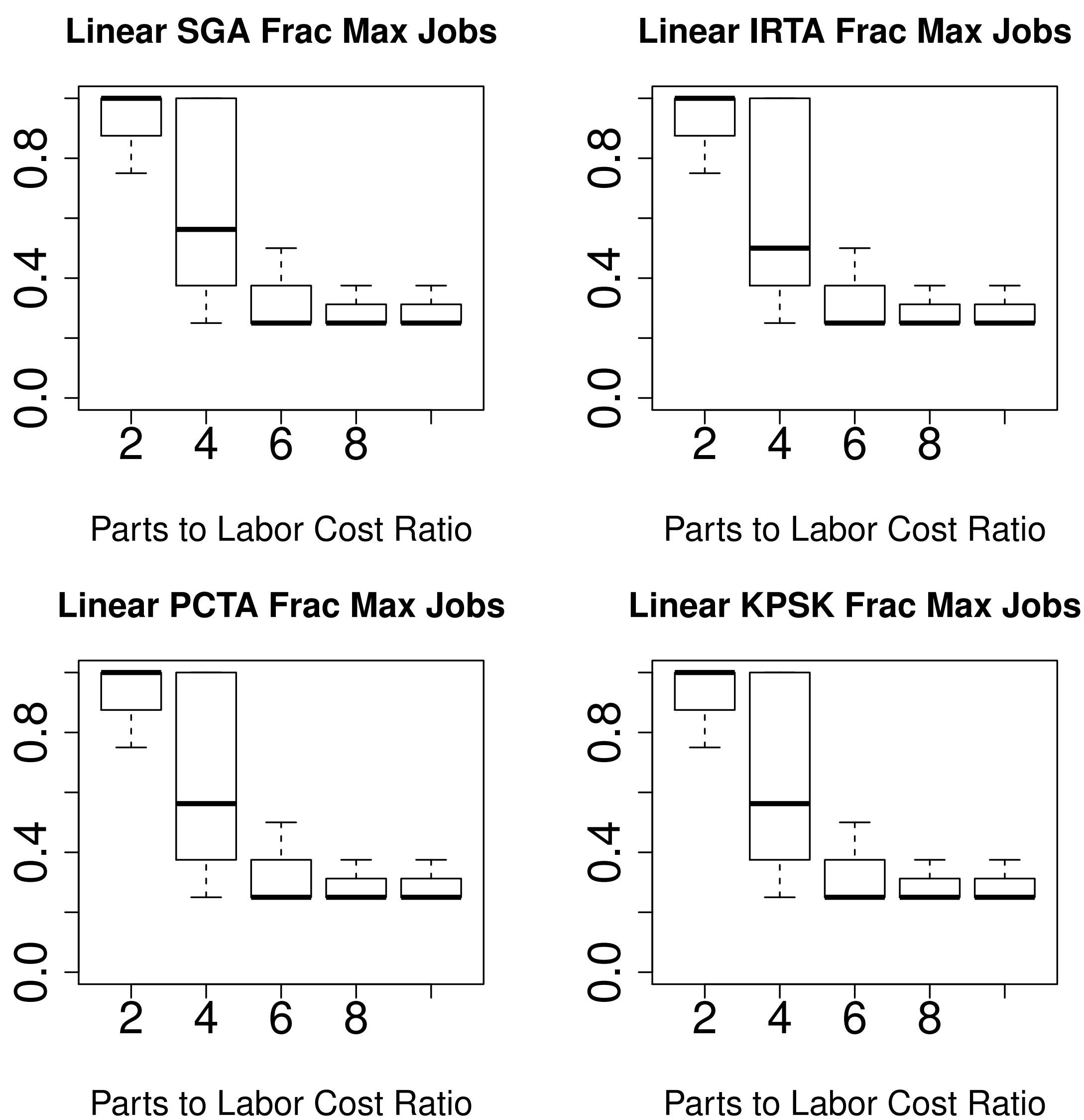

Figure 32: Boxplots of technician utilization for all test runs and inventory algorithms under linear delay function form

chapter were all of an order of magnitude comparable to this rough estimate.

It should also be noted that the proposed configuration under high values of parts to labor cost ratio, a single FSL and many technicians with empty repair kits, is likely not a reasonable solution to implement in practice. Technicians would not stand for carrying no inventory of their own as they travel from job site to job site, especially if they are transitioning from a spare parts supply chain without any FSL recourse where they were required to carry many part types in their repair kits.

The proposed configuration suggests that an alternate tag-team strategy with some technicians responsible for diagnosing required part types and others responsible for bringing those part types and executing the repairs might be beneficial. Such a strategy would of course only be possible if the repairs are actually cut-and-dry part replacements as we assume them to be in this thesis.

4.5.2 How much cost do FSLs save?

In order to determine how much cost the FSLs can save, we first need to specify how the spare parts supply chain without FSL recourse would be configured. The configuration of this no-FSL supply chain consists of the number of technicians and the part types to stock in the repair kit. Without access to recourse inventory at the FSL, jobs that require parts not stocked in the repair kit are considered service failures whereas all other jobs are successes. In this case, the expected fill rate is exactly equal to the probability of a kit job,

$$P(\text{Kit}) = \sum_{i=1}^{n} p_i x_i.$$

The inventory problem is to minimize the total cost of parts in the kit

$$\sum_{i=1}^{n} c_i(k_i - f_i)x_i,$$

while reaching a kit probability at least as large as the minimum required fill rate. Note that neither the objective function nor the constraint function depend on the number of technicians, which means the repair kit will be exactly the same for all possible number of technicians.

We do not need to consider how many technicians to employ because it is clearly optimal to employ the minimum number of fully-utilized technicians such that all M jobs can be completed in a day. We assume that successes, *i.e.*, kit jobs, take α time and failures, *i.e.*, what would be FSL jobs were such a strategy implemented, take no more than α time. Thus, it is never optimal to assign a technician fewer than $\lfloor d/\alpha \rfloor$ jobs in a day, and the optimal (minimal) number of technicians can be calculated using this productivity rate, independent of which parts are stocked in the repair kit.

We compared the four inventory algorithm variations against the no-FSL strategy under all three delay function forms and found that the FSL saves money only when the ratio of parts cost to labor cost is greater than 4, under which circumstances the best solution is to have many technicians with empty kits and low utilization. When this ratio is 6, the FSL saves 10% more than half the time; for a ratio value of 8, the analogous savings is around 25%, and for a value of 10, it is 30%. See Figures 33, 34, 35 for illustrations of these savings. For ratios of 2 and 4, it can be 10% more expensive to the use the FSL under our assumptions that parts must be stocked at the FSL if they are not held in the repair kit and that technicians must go to the FSL when it is revealed that a job requires a part stocked there.

The inventory pooling at an FSL is quite valuable when the cost of spare parts is much larger than the labor cost of technicians, *e.g.*, cost ratio of 10. In this situation, it is cheaper to employ many technicians whose time will be spent mostly in transit between customer sites and the FSL rather than executing repairs. When the cost of spare parts is closer to the labor cost of technicians, it is more expensive to use an FSL, but only because we assume it is required to keep non-kit parts at the FSL and complete every job in order without skipping any FSL jobs. Were it not for these last two assumptions, it would not be more expensive to use an FSL; it would merely be

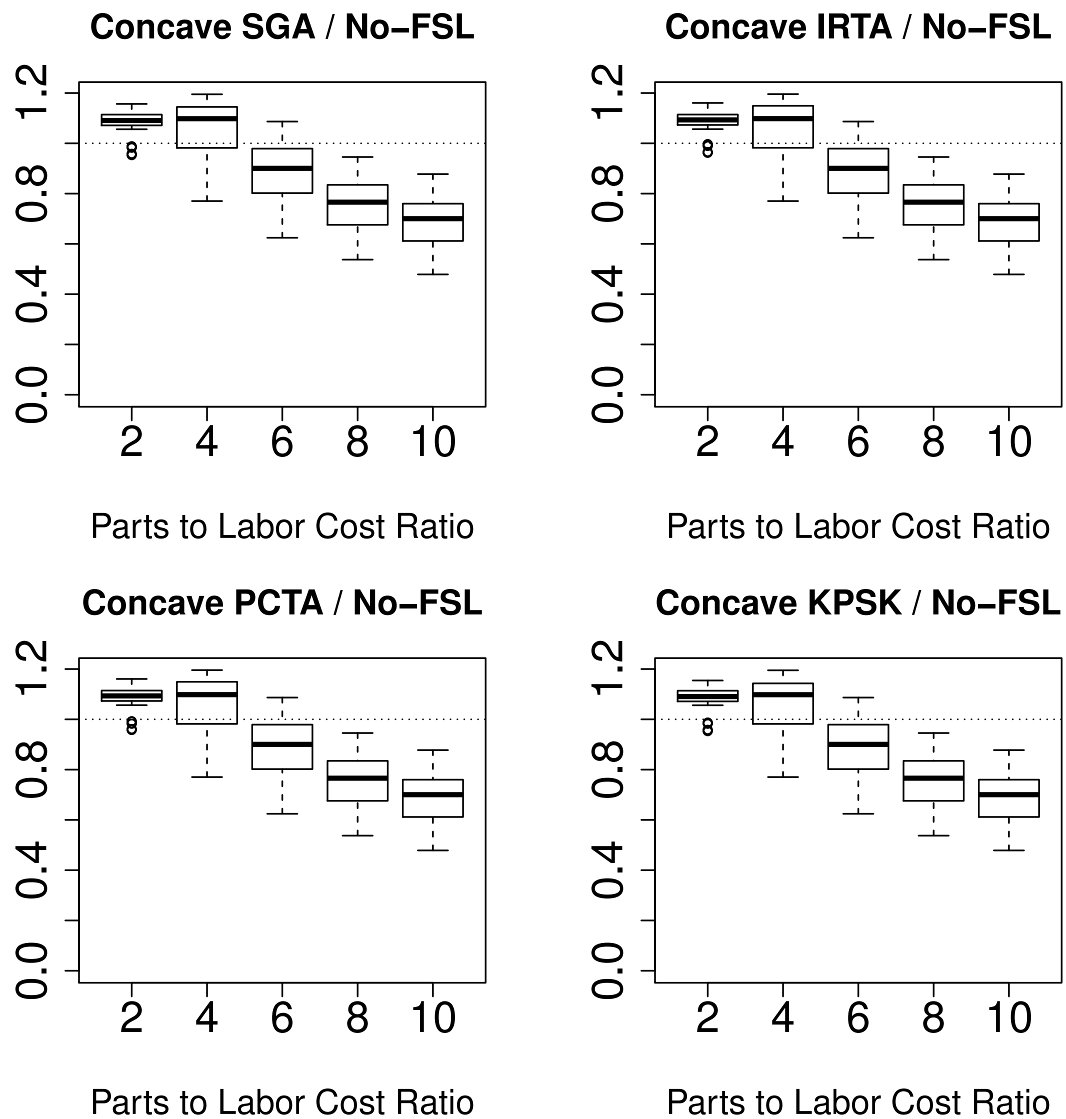

Figure 33: Boxplots of ratios of solution costs with and without FSL inventory for all test runs and inventory algorithms under concave delay function form

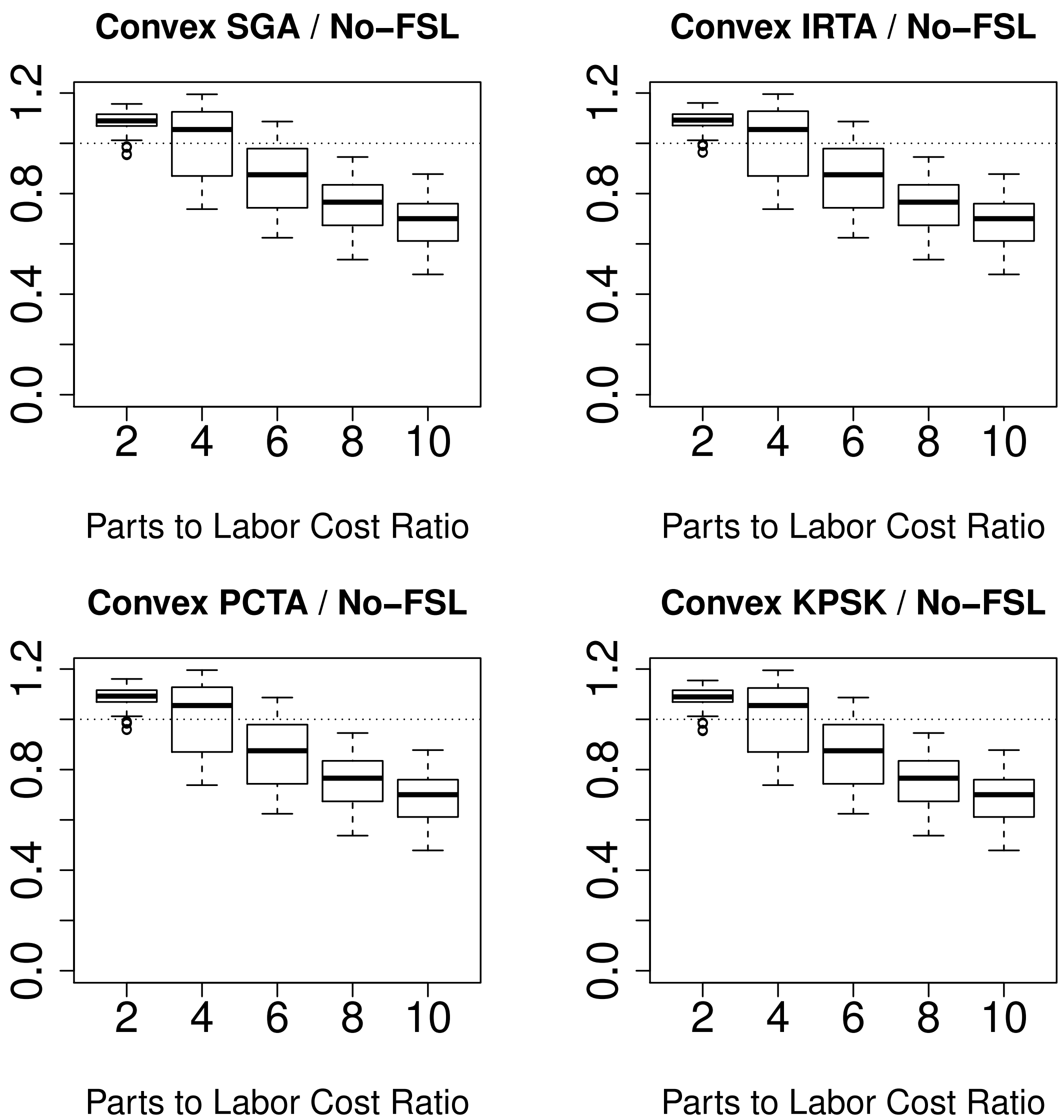

Figure 34: Boxplots of ratios of solution costs with and without FSL inventory for all test runs and inventory algorithms under convex delay function form

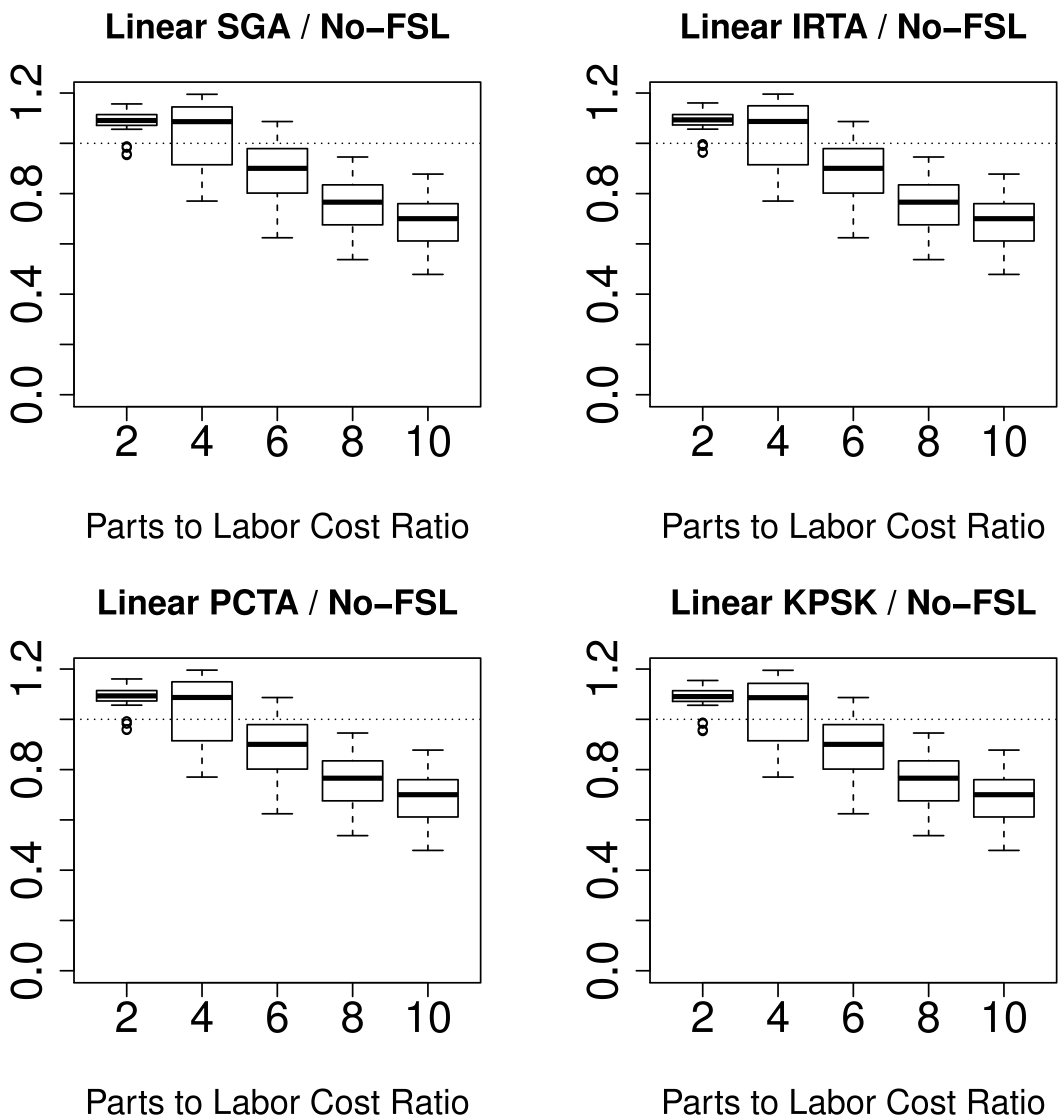

Figure 35: Boxplots of ratios of solution costs with and without FSL inventory for all test runs and inventory algorithms under linear delay function form

no cheaper to do so. A different model that allows for part types to be stocked at neither location and for technicians to skip jobs once it is revealed they require parts from the FSL would show that using an FSL is at worst no better than not using an FSL.

It is important to note that we assume that space is not a constraint, *i.e.*, it is always possible for a technician to carry a complete repair kit stocked with all part types to the job sites. In reality, this may not be the case, especially for supply chains where the number of part types is on the order of thousands or more. In those circumstances, it might be impossible to operate at a satisfactory level of customer service without using FSLs to stock recourse inventory, rendering the cost savings of an FSL strategy a moot point.

4.6 Concluding Remarks

In this chapter, we created a model of the decision problem faced by a repair service provider interested in designing a spare parts supply chain with FSL recourse from scratch. The decision maker must choose the number of technicians to employ, the number of FSLs to operate, and which parts to stock in the repair kit and which in the FSL. We showed that it is optimal to allocate the customer jobs uniformly to the technicians when the routing of the technicians to customer job sites is ignored, but that the decision problem is still difficult despite this restriction. We proposed a straightforward algorithm structure that takes advantage of the inventory algorithms from Chapter 2 to find good solutions to the overall design problem. Through computational testing over a suite of test instances with important parameters varied across reasonable ranges, we demonstrated that the best solutions have either fewer but busy technicians with close-to-full kits or more but less busy technicians with empty kits. When inventory is four times more costly than labor, the solutions fall somewhere between these two extremes. Although the latter extreme may be practically infeasible, it does offer savings of 10% to 30% over an arrangement with no FSL inventory recourse (in which we assume the space available in the repair kit is not a constraint).

CHAPTER V

CONCLUSIONS AND FUTURE WORK

Forward stocking location recourse adds flexibility for inventory storage to a spare parts supply chain. In theory such optionality should allow for solutions at least as good as those available without FSL recourse possible. This thesis investigates how to manage a spare parts supply chain with FSL recourse as well as just how much better a service provider can meet its contractual repair obligations with that approach.

5.1 Conclusions

Chapter 2 develops both a mathematical model for the customer service measure by which the repair performance is assessed and a mathematical programming formulation to represent the decision faced by the service provider of how to stock its supply chain to meet its minimal required level of service. This part of the thesis also proposes six algorithms to make the inventory decision and a lower-bounding procedure helpful in evaluating the quality of the algorithm solutions. We show that two of the algorithms (SGA and PCTA) are not necessarily optimal and that one of those two (SGA) can be arbitrarily bad in the worst case. However, these two algorithms and two others (IRTA and KPSK) produce inventory solutions with costs around ten percent higher than the lower bound on average. In addition, we show through computational testing that the FSL flexibility saves from two to ten percent total inventory cost in most cases.

Chapter 3 generalizes the customer service measure model of Chapter 2 so that the job completion time outcomes for the customer repairs depend on the sequence in which the technician visits the customer sites. More travel time (which varies directly with distance) due to a poorly chosen route leads to more inventory needing to be carried in the repair kits to meet the contractual service level. The expected fill rate measure for a route in a single customer instance is easily constructed by modifying the corresponding measure from Chapter 2 but accurately representing the expected fill rate requires sampling a number of customer instances to account for the daily variation in customer site locations. Six routing algorithms are proposed in this part and each is combined with the four best inventory algorithms from Chapter 3 to determine a repair kit solution. The computational test results show that random routing (perhaps representative of the case where customers arrive dynamically during the course of the service period) leads to significantly higher inventory costs (twenty percent and greater in many cases) but that more complicated routing algorithms are not much better, if at all, than simple greedy routing heuristics.

Chapter 4 broadens the scope of the original decision problem from Chapter 2 to include the number of technicians and the number of FSLs employed in the supply chain. It is shown that when customer repair jobs must be allocated across multiple technicians it is optimal to do so uniformly. Even with the restriction to this job allocation policy it is difficult to solve the problem efficiently to optimality. A straightforward algorithm structure is proposed that takes advantage of the good inventory algorithms from Chapter 2. Test results show that the best spare parts supply chain with FSL recourse has either fewer but busy technicians with close-to-full kits or more but less busy technicians with empty kits, the former when inventory is cheaper relative to labor and the latter otherwise. When inventory is around four times more expensive than labor, the best supply chain configuration falls somewhere between the two extremes. The latter structure may be practically infeasible, but it offers savings of 10% to 30% over a no-FSL strategy.

5.2 Future Work

Regrettably missing from this thesis is a proof that the expected fill rate for a single period and customer instance is monotone in the number of part types stocked in the repair kit, which is equivalent to the probability of a kit job outcome, *i.e.*, no trip to the FSL to retrieve a part. Many of the inventory algorithms proposed herein rely implicitly on this fact which seems intuitive yet whose proof was not so forthcoming. It would better round out the work contained in this thesis if we had been able to show that this relationship holds with certainty. (It is obvious that the other characteristic of a spare parts supply chain, the inventory cost, is monotone in the number of part types stocked in the repair kit.)

Also missing from this thesis (specifically Chapter 4) is a more detailed investigation into the solution structure and cost savings due to the FSL for problem instances where the ratio of parts to labor cost is close to four, the value of that factor around which the solutions shift from fewer but busy technicians with close-to-full kits to more but less busy technicians with empty kits. A closer look at this factor level could be of particular value for practitioners because their cost structures put them close to the transitional inventory ratio of four, if our industry collaborator is representative of companies in its field.

One major direction for new research into spare parts supply chains with FSL recourse in the future is to reframe the problem assumptions so that part types do not necessarily need to be stocked at either the repair kit or the FSL. This restrictive assumption makes the decision problem less complex to model and solve but perhaps does not represent the reality of the situation as closely as possible. Of course, the trade-off between realism and tractability is always present when creating an academic model of a real-world system or process and it is up to the researcher to find the best balance between the two opposing objectives. In defense of the assumption made in this thesis, *i.e.*, that parts types must be stocked at one of the two locations, we argue that it is not unreasonable for the decision maker to pre-select part types that are slowest-moving for exclusion from the model and include only those part types would realistically be stocked at one location or the other. Nevertheless, it is interesting to speculate what might be the case when we allow for part types to go unstocked. Such a problem setting would require modification of the expected fill rate calculation to account for the new third possible outcome of a job, *i.e.*, part type not stocked at either location, as well as new solution algorithms.

Throughout this thesis we have assumed that space is a not a constraint on the number or type of parts that can be carried in the repair kit but this may not necessarily hold in practice for all service providers that utilize this specific type of supply chain structure. Putting an upper limit on the number or volume of parts stocked in the repair kit would be another major direction for new research in the area of spare parts supply chains with FSL recourse. The value of this limit and the delay time for an FSL trip would likely have a significant effect on the feasibility of meeting service requirements in any model that would be developed for this problem setting.

A third assumption that could stand to be relaxed in further investigations is that of a common deadline for all repair jobs. In practice, some of the service agreements require the completion of a repair within a time frame of a few hours, a case that is not particularly well captured under the model in this thesis. Moreover, this thesis considers all customers to have the same priority for repairs when in reality the priority may be dictated by the stringency of the service agreement entered into, which is not necessarily the same. Generalizing the model to allow for different customer classes would be an interesting future foray as well.

APPENDIX A

PROOF THAT UNIFORM JOB ALLOCATION IS OPTIMAL

Lemma 1 is required for the included proof of Theorem 1.

Lemma 1. *Consider the multiple-job repair kit problem with FSL recourse where the set of customers is homogeneous. The probability that job j is completed is no less than the probability that job $j+1$ is completed.*

$$P(\mathcal{T}_j \leq d) \geq P(\mathcal{T}_{j+1} \leq d).$$

Proof. Consider all sample outcomes of kit or FSL jobs for the first j customers. They look something like $\alpha, \alpha, \beta, \ldots, \alpha$ or $\alpha, \beta, \alpha, \ldots, \beta$. Each sample outcome has a positive probability of being realized. The probability of the sample outcome $\alpha, \alpha, \beta, \ldots, \alpha$ is $p \cdot p \cdot (1-p) \cdots p$ where p is the probability of a kit job since all job times are independently distributed.

Now consider all sample outcomes for the first $j+1$ customers. There are exactly twice as many distinct sequences (of length one greater) than there are for the first j customers due to the fact that the $j+1$th customer could take one of two possible outcomes. We can put the sample outcomes for customer j into a one-to-two matching with the sample outcomes for customer $j+1$. For example, we can match $\alpha, \alpha, \beta, \ldots, \alpha$ to $\alpha, \alpha, \beta, \ldots, \alpha, \alpha$ and $\alpha, \alpha, \beta, \ldots, \alpha, \beta$. The probabilities of the latter two outcomes are $p \cdot p \cdot (1-p) \cdots p \cdot p$ and $p \cdot p \cdot (1-p) \cdots p \cdot (1-p)$, respectively, which sum up to $p \cdot p \cdot (1-p) \cdots p(p + (1-p)) = p \cdot p \cdot (1-p) \cdots p$, the probabilty of the sample outcome for customer j to which they are both matched.

Now consider the set C_{j+1} of sample outcomes such that they represent a successful completion of job $j+1$ by the deadline, that is, the αs and βs of the outcome add up to no more than d. Analogously define C_j as the successful outcomes for customer j. The probabilities of all the sample outcomes in C_{j+1} sum to give the probability that job $j+1$ is completed successfully, $P(\mathcal{T}_{j+1} \leq d)$, and likewise for C_j. It is sufficient to show that sum of probabilities for outcomes in C_j is no less than the sum of probabilities for outcomes in C_{j+1}.

Recall our one-to-two matching from the previous paragraph. There are three cases we need to examine: both sample outcomes are elements of C_{j+1}, neither is, and one is but the other is not. In the first case, the outcome for customer j that is matched to the ones for customer $j+1$ is a member of the set C_j and adds the same probability to the total sum of probabilities for outcomes in C_j that its two matched elements add to the total sum of probabilities for outcomes in C_{j+1} (WASH). In the second case, neither outcome has a sum of αs and βs less than d but their matched outcome with one fewer α or β at the end may have such a sum (C_j might have a greater probability sum). In the third case, only one of the pair is contributing its probability to the total sum for C_{j+1} while its match is contributing its probability to the sum for C_j and its match has a probability equal to the sum of the pair's probabilities (C_j definitely has a greater probability sum).

With each of these three cases covered, we can conclude that the sum of probabilities for all outcomes in C_j is no less than the sum of probabilities for all outcomes in C_{j+1} and hence the proof is complete. □

Theorem 1. *Consider the multiple-job repair kit problem with FSL recourse where the set of customers is homogeneous. In this setting, a uniform job allocation has the highest overall expected fill rate.*

Proof. Consider any non-uniform job allocation. We can find two technicians, A and B, such that technician B has at least two more customers assigned to his queue than technician A does. Let b represent the number of jobs in the non-uniform allocation for technician B and similarly a for A, so that $b \geq a + 2$.

The last customer in technician B's queue, call it customer ℓ, has probability $P(\mathcal{T}_b \leq d)$ of being completed under the non-uniform job allocation. If for an alternative allocation we move customer ℓ from B's queue to A's queue then its probability of being completed becomes $P(\mathcal{T}_{a+1} \leq d)$ which we know is no less than $P(\mathcal{T}_b \leq d)$ because $a + 1 < b$.

We have managed to not decrease customer ℓ's probability of being completed without changing the completion probabilties for any other customers. (Remember the completion probabilities only depend on the number of customers ahead in the queue, which did not change for any other customers besides ℓ in this move.)

If this new allocation is still non-uniform we can repeat this step without decreasing (and while likely increasing) the expected fill rate until we reach a uniform allocation. Thus we have shown that for any non-uniform allocation there exists a uniform allocation with an expected fill rate no lower. □

APPENDIX B

EXAMPLE TO SHOW NATURE OF DELAY AS FUNCTION OF NUMBER OF FSLS

Consider a rectangular region with four customers as laid out in Figure 36. We assume the FSLs are positioned to minimize the average distance from a customer to the nearest FSL. With one FSL the average distance is $\sqrt{h^2 + w^2}/2$ as illustrated in Figure 37. With two FSLs the average distance is $\min\{h, w\}/2$ as illustrated in Figures 38 and 39. The two FSLs can be placed anywhere on the short sides of the rectangle to achieve the optimal average distance. With three FSLs the average distance is $\min\{h, w\}/4$ as illustrated in Figures 41 and 40. Two of the FSLs would be located at customer sites that were adjacent along a short side of the rectangle. The third could be located anywhere on the short side between the other two customer sites to achieve the same objective value of average distance.

We now consider three examples to show the relationship between the delay time, which is twice the average distance from the FSL, and the number of FSLs can be

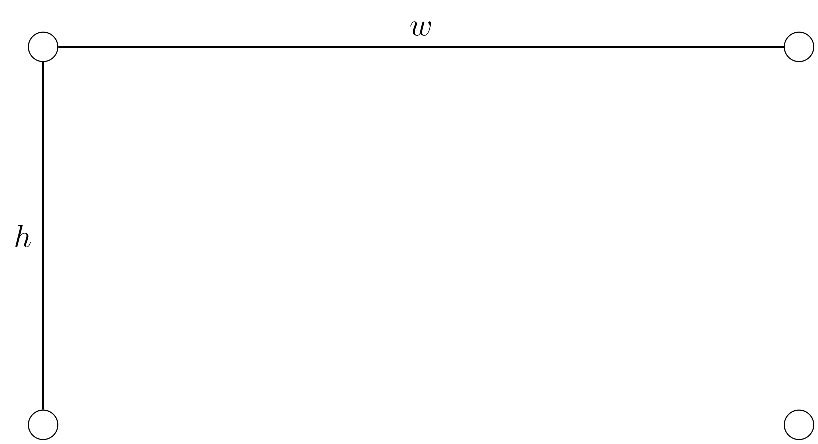

Figure 36: Rectangular region

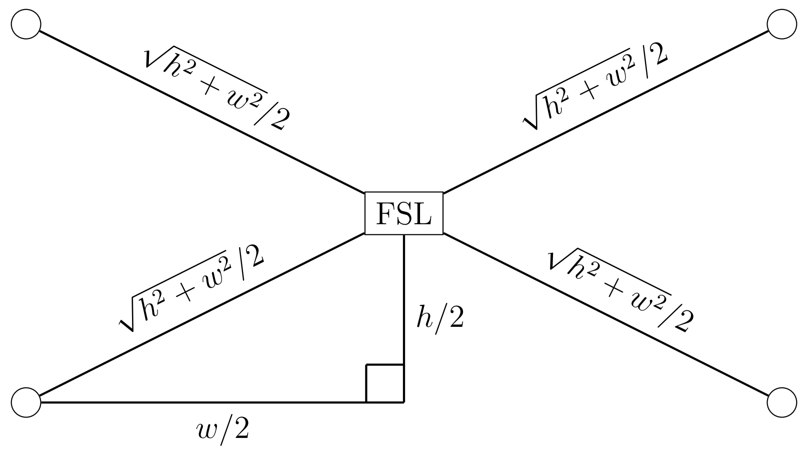

Figure 37: Rectangular region with one FSL

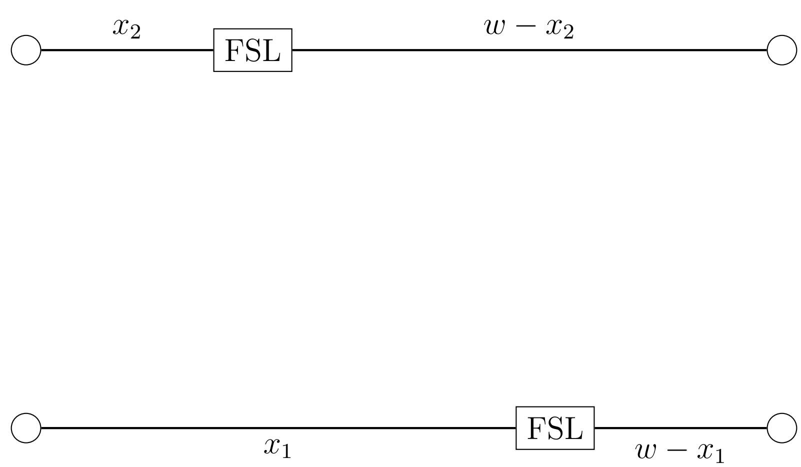

Figure 38: Rectangular region with two FSLs, $h > w$

Figure 39: Rectangular region with two FSLs, $h < w$

Figure 40: Rectangular region with three FSLs, $h < w$

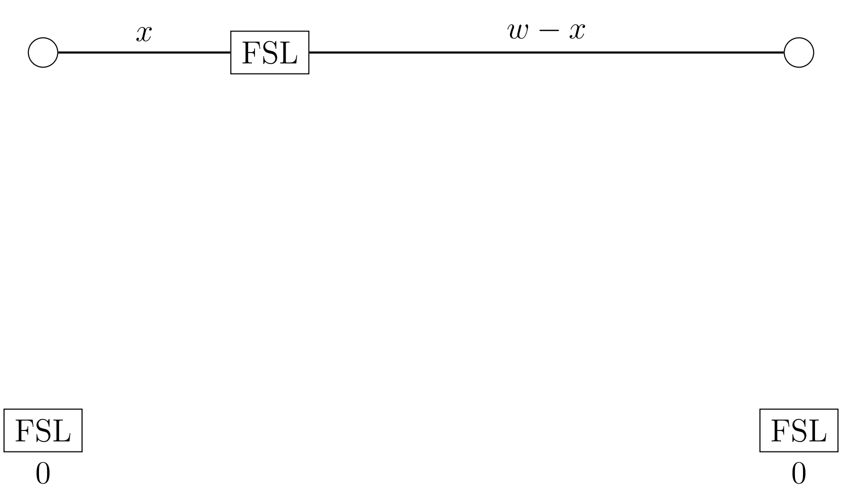

Figure 41: Rectangular region with three FSLs, $h > w$

Table 14: Parameters for three types of functional relationship between delay and number of FSLs

	Sub-Linear	Linear	Super-Linear
h	$3\sqrt{5}/10$	1	$3\sqrt{2}/4$
w	$3\sqrt{5}/5$	$\sqrt{5}/2$	$3\sqrt{2}/4$

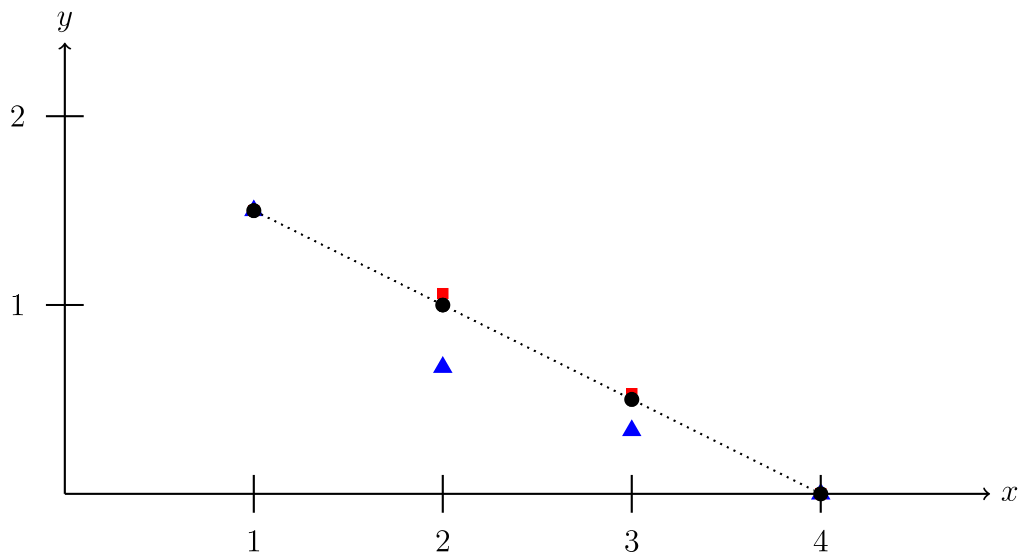

Figure 42: Delay as a function of the number of FSLs

linear, sub-linear, or super-linear. Table 14 summarizes the parameters for the three cases and Figure 42 illustrates the function values.

REFERENCES

[1] "Biography - Matt Ridley." http://www.rationaloptimist.com/biography. Accessed: 2014-11-20.

[2] "Self-sufficiency is another word for poverty - Matt Ridley." http://www.rationaloptimist.com/blog/self-sufficiency-is-another-word-for-poverty.aspx. Accessed: 2014-11-26.

[3] Aikens, C. H., "Facility location models for distribution planning," *European journal of operational research*, vol. 22, no. 3, pp. 263–279, 1985.

[4] Akinc, U. and Khumawala, B. M., "An efficient branch and bound algorithm for the capacitated warehouse location problem," *Management Science*, vol. 23, no. 6, pp. 585–594, 1977.

[5] Archer, A., Levin, A., and Williamson, D. P., "A faster, better approximation algorithm for the minimum latency problem," *SIAM Journal on Computing*, vol. 37, no. 5, pp. 1472–1498, 2008.

[6] Balinski, M., "Notes–on a selection problem," *Management Science*, vol. 17, no. 3, pp. 230–231, 1970.

[7] Ball, M. O., Golden, B., Assad, A., and Bodin, L., "Planning for truck fleet size in the presence of a common-carrier option," *Decision Sciences*, vol. 14, no. 1, pp. 103–120, 1983.

[8] Bard, J. F., Huang, L., Dror, M., and Jaillet, P., "A branch and cut algorithm for the vrp with satellite facilities," *IIE transactions*, vol. 30, no. 9, pp. 821–834, 1998.

[9] Bard, J. F., Huang, L., Jaillet, P., and Dror, M., "A decomposition approach to the inventory routing problem with satellite facilities," *Transportation science*, vol. 32, no. 2, pp. 189–203, 1998.

[10] Baumol, W. J. and Wolfe, P., "A warehouse-location problem," *Operations Research*, vol. 6, no. 2, pp. 252–263, 1958.

[11] Beasley, J. E., "An algorithm for solving large capacitated warehouse location problems," *European Journal of Operational Research*, vol. 33, no. 3, pp. 314–325, 1988.

[12] Bianco, L., Mingozzi, A., and Ricciardelli, S., "The traveling salesman problem with cumulative costs," *Networks*, vol. 23, no. 2, pp. 81–91, 1993.

[13] Bijvank, M., Koole, G., and Vis, I. F., "Optimising a general repair kit problem with a service constraint," *European Journal of Operational Research*, vol. 204, no. 1, pp. 76–85, 2010.

[14] Binart, S., Dejax, P., Gendreau, M., and Semet, F., "A 2-stage method for a field service routing problem with stochastic travel and service times," *Computers & Operations Research*, vol. 65, pp. 64–75, 2016.

[15] Blum, A., Chalasani, P., Coppersmith, D., Pulleyblank, B., Raghavan, P., and Sudan, M., "The minimum latency problem," in *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pp. 163–171, ACM, 1994.

[16] Bostel, N., Dejax, P., Guez, P., and Tricoire, F., "Multiperiod planning and routing on a rolling horizon for field force optimization logistics," in *The vehicle routing problem: latest advances and new challenges*, pp. 503–525, Springer, 2008.

[17] Campbell, A., Clarke, L., Kleywegt, A., and Savelsbergh, M., "The inventory routing problem," in *Fleet management and logistics*, pp. 95–113, Springer, 1998.

[18] Campbell, A. M., Gendreau, M., and Thomas, B. W., "The orienteering problem with stochastic travel and service times," *Annals of Operations Research*, vol. 186, no. 1, pp. 61–81, 2011.

[19] Campbell, A. M. and Savelsbergh, M. W., "A decomposition approach for the inventory-routing problem," *Transportation science*, vol. 38, no. 4, pp. 488–502, 2004.

[20] Castillo-Salazar, J. A., Landa-Silva, D., and Qu, R., "Workforce scheduling and routing problems: literature survey and computational study," *Annals of Operations Research*, pp. 1–29, 2014.

[21] Chao, I.-M., Golden, B. L., and Wasil, E. A., "A fast and effective heuristic for the orienteering problem," *European Journal of Operational Research*, vol. 88, no. 3, pp. 475–489, 1996.

[22] Chao, I.-M., Golden, B. L., and Wasil, E. A., "The team orienteering problem," *European journal of operational research*, vol. 88, no. 3, pp. 464–474, 1996.

[23] Chen, X., Thomas, B. W., and Hewitt, M., "The technician routing problem with experience-based service times," *Omega*, 2015.

[24] Chien, T. W., Balakrishnan, A., and Wong, R. T., "An integrated inventory allocation and vehicle routing problem," *Transportation Science*, vol. 23, no. 2, pp. 67–76, 1989.

[25] Coelho, L. C., Cordeau, J.-F., and Laporte, G., "Thirty years of inventory routing," *Transportation Science*, vol. 48, no. 1, pp. 1–19, 2013.

[26] Cortés, C. E., Gendreau, M., Leng, D., and Weintraub, A., "A simulation-based approach for fleet design in a technician dispatch problem with stochastic demand," *Journal of the Operational Research Society*, vol. 62, no. 8, pp. 1510–1523, 2011.

[27] Desrochers, M. and Verhoog, T., "A new heuristic for the fleet size and mix vehicle routing problem," *Computers & Operations Research*, vol. 18, no. 3, pp. 263–274, 1991.

[28] Dror, M., Ball, M., and Golden, B., "A computational comparison of algorithms for the inventory routing problem," *Annals of Operations Research*, vol. 4, no. 1, pp. 1–23, 1985.

[29] Dutot, P.-F., Laugier, A., and Bustos, A.-M., "Technicians and interventions scheduling for telecommunications," *France Telecom R&D*, 2006.

[30] Efroymson, M. and Ray, T., "A branch-bound algorithm for plant location," *Operations Research*, vol. 14, no. 3, pp. 361–368, 1966.

[31] Etezadi, T. and Beasley, J., "Vehicle fleet composition," *Journal of the Operational Research Society*, pp. 87–91, 1983.

[32] Evers, L., Glorie, K., Van Der Ster, S., Barros, A. I., and Monsuur, H., "A two-stage approach to the orienteering problem with stochastic weights," *Computers & Operations Research*, vol. 43, pp. 248–260, 2014.

[33] Federgruen, A. and Zipkin, P., "A combined vehicle routing and inventory allocation problem," *Operations Research*, vol. 32, no. 5, pp. 1019–1037, 1984.

[34] Feldman, E., Lehrer, F., and Ray, T., "Warehouse location under continuous economies of scale," *Management Science*, vol. 12, no. 9, pp. 670–684, 1966.

[35] Ferland, J. A. and Michelon, P., "The vehicle scheduling problem with multiple vehicle types," *Journal of the Operational Research Society*, pp. 577–583, 1988.

[36] Gheysens, F., Golden, B., and Assad, A., "A comparison of techniques for solving the fleet size and mix vehicle routing problem," *Operations-Research-Spektrum*, vol. 6, no. 4, pp. 207–216, 1984.

[37] Goemans, M. and Kleinberg, J., "An improved approximation ratio for the minimum latency problem," *Mathematical Programming*, vol. 82, no. 1-2, pp. 111–124, 1998.

[38] Golden, B., Assad, A., and Dahl, R., "Analysis of a large scale vehicle routing problem with an inventory component," *Large scale systems*, vol. 7, no. 2-3, pp. 181–190, 1984.

[39] Golden, B., Assad, A., Levy, L., and Gheysens, F., "The fleet size and mix vehicle routing problem," *Computers & Operations Research*, vol. 11, no. 1, pp. 49–66, 1984.

[40] Golden, B. L., Levy, L., and Vohra, R., "The orienteering problem," *Naval research logistics*, vol. 34, no. 3, pp. 307–318, 1987.

[41] Gould, J., "The size and composition of a road transport fleet," *OR*, pp. 81–92, 1969.

[42] Graves, S. C., "Note–a multiple-item inventory model with a job completion criterion," *Management Science*, vol. 28, no. 11, pp. 1334–1337, 1982.

[43] Gupta, A., Krishnaswamy, R., Nagarajan, V., and Ravi, R., "Approximation algorithms for stochastic orienteering," in *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pp. 1522–1538, SIAM, 2012.

[44] Hansen, P., Mladenović, N., Brimberg, J., and Pérez, J. A. M., "Variable neighborhood search," in *Handbook of Metaheuristics*, pp. 61–86, Springer, 2010.

[45] Heeremans, D. and Gelders, L., "Multiple period repair kit problem with a job completion criterion: A case study," *European journal of operational research*, vol. 81, no. 2, pp. 239–248, 1995.

[46] Hoff, A., Andersson, H., Christiansen, M., Hasle, G., and Løkketangen, A., "Industrial aspects and literature survey: Fleet composition and routing," *Computers & Operations Research*, vol. 37, no. 12, pp. 2041–2061, 2010.

[47] Jacobsen, S. K., "Heuristics for the capacitated plant location model," *European Journal of Operational Research*, vol. 12, no. 3, pp. 253–261, 1983.

[48] Jaillet, P., Bard, J. F., Huang, L., and Dror, M., "Delivery cost approximations for inventory routing problems in a rolling horizon framework," *Transportation Science*, vol. 36, no. 3, pp. 292–300, 2002.

[49] Khumawala, B. M., "An efficient branch and bound algorithm for the warehouse location problem," *Management Science*, vol. 18, no. 12, pp. B–718, 1972.

[50] Kirby, D., "Is your fleet the right size?," *Journal of the Operational Research Society*, vol. 10, no. 4, pp. 252–252, 1959.

[51] Klose, A. and Drexl, A., "Facility location models for distribution system design," *European Journal of Operational Research*, vol. 162, no. 1, pp. 4–29, 2005.

[52] Kovacs, A. A., Parragh, S. N., Doerner, K. F., and Hartl, R. F., "Adaptive large neighborhood search for service technician routing and scheduling problems," *Journal of scheduling*, vol. 15, no. 5, pp. 579–600, 2012.

[53] Krarup, J. and Pruzan, P. M., "The simple plant location problem: survey and synthesis," *European Journal of Operational Research*, vol. 12, no. 1, pp. 36–81, 1983.

[54] Kuehn, A. A. and Hamburger, M. J., "A heuristic program for locating warehouses," *Management science*, vol. 9, no. 4, pp. 643–666, 1963.

[55] Lucena, A., "Time-dependent traveling salesman problem–the deliveryman case," *Networks*, vol. 20, no. 6, pp. 753–763, 1990.

[56] Mamer, J. W. and Smith, S. A., "Optimizing field repair kits based on job completion rate," *Management Science*, vol. 28, no. 11, pp. 1328–1333, 1982.

[57] Manne, A. S., "Plant location under economies-of-scale-decentralization and computation," *Management Science*, vol. 11, no. 2, pp. 213–235, 1964.

[58] Maskell, T., "The optimal size of vehicle fleet for local deliveries. msc report," *Management Engineering Section, Imperial College, London 0.05*, 1966.

[59] Méndez-Díaz, I., Zabala, P., and Lucena, A., "A new formulation for the traveling deliveryman problem," *Discrete Applied Mathematics*, vol. 156, no. 17, pp. 3223–3237, 2008.

[60] Minieka, E., "The delivery man problem on a tree network," *Annals of Operations Research*, vol. 18, no. 1, pp. 261–266, 1989.

[61] Mladenović, N. and Hansen, P., "Variable neighborhood search," *Computers & Operations Research*, vol. 24, no. 11, pp. 1097–1100, 1997.

[62] Mladenović, N., Urošević, D., and Hanafi, S., "Variable neighborhood search for the travelling deliveryman problem," *4OR*, vol. 11, no. 1, pp. 57–73, 2013.

[63] Pillac, V., Gueret, C., and Medaglia, A. L., "A parallel matheuristic for the technician routing and scheduling problem," *Optimization Letters*, vol. 7, no. 7, pp. 1525–1535, 2013.

[64] Sa, G., "Branch-and-bound and approximate solutions to the capacitated plant-location problem," *Operations Research*, vol. 17, no. 6, pp. 1005–1016, 1969.

[65] Salhi, S. and Rand, G. K., "Incorporating vehicle routing into the vehicle fleet composition problem," *European Journal of Operational Research*, vol. 66, no. 3, pp. 313–330, 1993.

[66] Smith, S. A., Chambers, J. C., and Shlifer, E., "Note–optimal inventories based on job completion rate for repairs requiring multiple items," *Management Science*, vol. 26, no. 8, pp. 849–854, 1980.

[67] Song, J.-S., "On the order fill rate in a multi-item, base-stock inventory system," *Operations Research*, vol. 46, no. 6, pp. 831–845, 1998.

[68] Song, J.-S. and Yao, D. D., "Performance analysis and optimization of assemble-to-order systems with random lead times," *Operations Research*, vol. 50, no. 5, pp. 889–903, 2002.

[69] Sridharan, R., "The capacitated plant location problem," *European Journal of Operational Research*, vol. 87, no. 2, pp. 203–213, 1995.

[70] Tang, H. and Miller-Hooks, E., "Algorithms for a stochastic selective travelling salesperson problem," *Journal of the Operational Research Society*, vol. 56, no. 4, pp. 439–452, 2005.

[71] Teunter, R. H., "The multiple-job repair kit problem," *European journal of operational research*, vol. 175, no. 2, pp. 1103–1116, 2006.

[72] Tricoire, F., Bostel, N., Dejax, P., and Guez, P., "Exact and hybrid methods for the multiperiod field service routing problem," *Central European Journal of Operations Research*, vol. 21, no. 2, pp. 359–377, 2013.

[73] Tsiligirides, T., "Heuristic methods applied to orienteering," *Journal of the Operational Research Society*, pp. 797–809, 1984.

[74] Vansteenwegen, P., Souffriau, W., and Van Oudheusden, D., "The orienteering problem: A survey," *European Journal of Operational Research*, vol. 209, no. 1, pp. 1–10, 2011.

[75] Wyatt, J., "Optimal fleet size," *Journal of the Operational Research Society*, vol. 12, no. 3, pp. 186–187, 1961.

VITA

Brian Joseph Kues was born on March 25, 1988, to Lawrence and Susan Kues in Cincinnati, Ohio, USA, and raised in the nearby suburb of Loveland with his two younger brothers, Derek and Matthew. In August 2002, he entered Saint Xavier High School, a Jesuit secondary school in Cincinnati, on a one-year scholarship awarded for his top score on the standardized entrance test. He graduated from St. Xavier in May 2006 and at the commencement ceremony was presented a subject award for being the best student in mathematics. Later that year, Brian matriculated to The Ohio State University in Columbus, Ohio, USA, where he was the only student to graduate *summa cum laude* with a Bachelor of Science degree in Industrial and Systems Engineering at the June 2010 commencement. While at Ohio State, Brian was a member of the varsity men's volleyball program and a letterwinner on four conference championship teams. In August 2010, Brian enrolled in the H. Milton Stewart School of Industrial & Systems Engineering at the Georgia Institute of Technology in Atlanta, Georgia, USA, as a Ph.D. student.