Roma Tre University
Ph.D. in Computer Science and Automation

Stochastic optimization for airport inventory management

Annalisa Cesaro

Stochastic optimization for airport inventory management

A thesis presented by
Annalisa Cesaro
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Automation

Roma Tre University
Dept. of Computer Science and Automation

30 March 2010

Committee:
Prof. Dario Pacciarelli

Reviewers:
Prof. A. Agnetis
Prof. G.J.J.A.N. Van Houtum

To my family

Acknowledgments

Con la conclusione di questa tesi, si chiude un periodo intenso della mia vita, in cui le esperienze vissute sono state molteplici e variegate.

Il percorso dottorale è un percorso ricco di libere opportunità.
Ringrazio chi mi ha permesso di iniziare questo percorso, il mio tutor Dario Pacciarelli, che è sempre stato un costruttivo punto di riferimento in tutti questi anni e che mi ha dato tanto.

Ringrazio Gianluca, per il suo amore e sostegno sempre.
Ringrazio i miei 4 angeli custodi: i miei genitori ed i genitori di Gianluca, senza di loro non sarei potuta venire serenamente a lavorare all'università, sapendo di avere affidato mio figlio in mani premurose e sicure.
Ringrazio anche te, Vittorio: anche se così piccolo, doni gioia, grinta e forza.
Vi amo tesori miei.

Ringrazio gli altri dottorandi del laboratorio e del dipartimento, sia chi ancora qui e sia chi lontano, con cui abbiamo condiviso esperienze e momenti di soddisfazione o sconforto.

Ringrazio tutte le persone che ho incontrato nelle varie scuole o conferenze, perchè ognuna sapeva comunicare la propria passione nel raggiungere un traguardo.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1. Airport inventory management

The ever increasing air traffic demand of passengers and cargos all over the world is currently limited by the capacity of the airports, which are expected to become a serious bottleneck for air traffic in a near future [1]. Facing such growth requires significant investments for developing existing airports and/or constructing new ones. Specifically, there is an increasing need for safety equipments in order to grant airport safety, as well as for supporting the correct execution of airport operations. In this scenario airports face every day the challenging task of maintaining high standards of safety at a sustainable cost. In such a context maintenance plays an important role. Preventive maintenance is scheduled in advance and anticipates the realization of equipment failures. Corrective maintenance is carried out upon a failure happens in a system. Typically, when a failure of some equipment takes place, failed components must be promptly replenished with new spare parts, since safety standards are not compatible with long repairing times. As observed by several authors, see e.g. by [43], the logistics of spare parts differs from those of other materials in several ways. Equipments may have remarkable costs, long repairing times and sporadic failures. The latter are difficult to forecast and cause relevant financial effects, due to the economical and legal implications of a lack of safety of airport operations. These characteristics are particularly stringent in the airport context, where therefore maintenance deserves substantial attention. Maintenance concerns aftermarket service, which importance today is in general high. Deloitte [26] discusses of service revolution,

because the high combined revenues due to aftermarket service in Aerospace, Defence, Automotive and High Tech. Specifically, it reports that in Aerospace and Defence on average service revenues account for 47% of the total business and that profitability is much higher that in the primary product business. It follows clearly that improving aftermarket processes and resource allocation is crucial in aerospace business. The sporadic nature of the failure process for a single equipment translates in most cases into very low demand for spare parts. For an item there are typically less than ten working equipments with MTBF equal to six years or more. Therefore, for economical reasons, airports are usually grouped on a regional base and served by a single warehouse. For example, the Italian territory is divided in 17 service regions serving a total of 38 civil airports [30]. Each spare part warehouse manages the aggregated demand of all the airports encompassed in its region. It follows that the aggregated demand rate for some items may be low or high, depending on the number of working equipments in the area. The spare parts supply chain may typically involve at least three actors: airport authorities, logistics companies and equipment suppliers. The latter are responsible for supplying new components and/or repaired items, which usually require long replenishment times. Intermediate logistic companies are in charge of replenishing spare parts in the short term, by granting minimum levels of operational availability (i.e., the fraction of time during which all working equipments are operative) regulated by contracts with airport authorities. Clearly the best design of resupply networks and the optimal allocation of inventories within these networks is of unquestionable importance to the economical maintenance of equipment. The types of decisions that must be made relating to service parts can roughly be divided into three planning categories: strategic planning, tactical planning and operational planning. Strategic planning is an on-going activity that has two primary functions. First, the definition of customer requirements, e.g. the range of service parts needed and timeliness of these needs, today and for the next several years. Second, the definition of the resources to meet customer requirements, e.g. the operating environment, the information systems, the supply chain partners. From the perspective of service parts, tactical planning establishes what inventories will be required to meet operational objectives at some future time, given the design and the operational characteristics of an existent resupply system infrastructure. Finally, the third category of planning concern operational decisions, which are based on real-time execution problems, are formulated over short planning horizons and contain more details on current operating limitations.

1.2. Research motivation

This thesis is devoted to constructing mathematical models that can be effectively used to carry out the tactical planning goal of determining systems stock levels. Our case study is motivated by the practical needs of an Italian logistics company supporting the activity of 38 civil airports spread over the Italian territory. The company handles 17 warehouses and manages the overall processes of purchasing, holding, ensuring that the overall reliability of safety equipments is always within contractual limits. The aim of the company is therefore to grant the prescribed quality of service at minimum cost. To this aim, a two echelon inventory policy without lateral transshipments is currently adopted and the level of stock and geographical allocation of spare parts is obtained with the VARIMETRIC algorithm of [74]. This policy is depicted in figure 1(left).

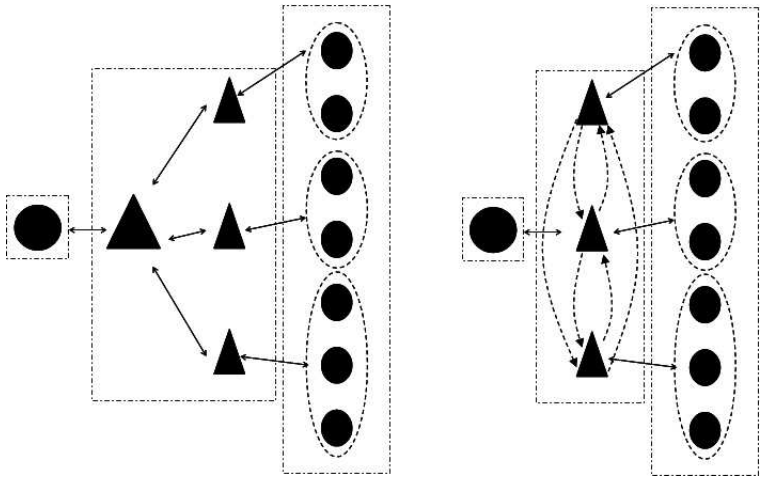

Figure 1.1: Logistics chain system: two echelon (left) and single echelon (right)

Effective supply chain management is currently recognized as a key determinant of competitiveness and success in manufacturing and services, because the implementation of supply chain management has significant impact on cost service and quality. Numerous strategies for achieving these targets have been proposed. One such strategy allows movements of stock between locations at the same echelon level or across different levels. These stock movements are

termed lateral transshipments, or simply transshipments. As a demand occurs under the implementation of transshipment strategy, there will be three possible activities: the demand is met from the stock on hand or it is met via transshipment from another location or it is backordered. In other words, firstly, if a location's on hand inventory level is greater than the demand size, than the demand is met. Secondly, if the on-hand inventory is positive but less than the demand size, than it is used to partially satisfy the demand and the remaining demand is met either via transshipment or is backordered. Thirdly, if the on-hand inventory level is zero, the demand is met via transshipment or is backordered under the assumption of no lost sales. Therefore, transshipment policy can improve stock availability, customer service level. without increasing stock level which may induce higher inventory relevant cost. In other words, transshipments enable the sharing of stock among locations, they facilitate each location as a secondary, random supply source for the remainder. If transshipment is not limited to one direction, the locations which meet their demand via transshipment form a pool. Thus, the locations' replenishment can be coordinated and even combined in order to avoid excessive inventory costs. The improvements in information technology coupled with the substantial reduction in the cost of processing, storing and analyzing data have made sharing of inventories more attractive. Furthermore, logistics companies, such as UPS, have made the rapid movement of parts from one place to another possible and more affordable. The underlying question that must be addressed is: which is the impact of lateral resupply on inventory levels and operations. A number of authors have addressed this issue. Several simulation studies have demonstrated the effects of lateral resupply in multi-echelon systems [32], [33], [73]. While the environments that were analyzed by these authors did differ, their results showed that in a wide variety of circumstances, lateral resupply among locations is a very effective way to improve customer service and lower inventory investments. Other authors have presented and tested many analytic models that explicitly consider the possibility of supplying location through lateral transshipments. For example, see Archibald [3], Lee [59], Alfredsson and Verrijdt [2], Axsater [4], Taragas [79], [81], Taragas and Cohen [80], Sherbrooke [73]. These models differ in many ways. Some are stationary, continuous time models, while other are periodic review. Essentially, though, all these analytic models are tactical planning models. They are either economic models that suggest what quantities of material to buy or they are models used to determine the probabilities for various events occurring. We say that these models are tactical models because they do not consider the possibility of using all state-of-the-world information when representing the operational environment.

For example FIFO inventory allocation rules are often assumed as the basis for shipping parts from a central location (depot) to field stocking location. Real time execution systems would take more information into account. Moreover, as observed, e.g., by [54], single echelon models with complete pooling might be more effective for reducing both reaction time to stockouts and inventory levels. In Figure 1(right) a single echelon policy is depicted. At the time of writing, lateral communication is only used by the company in emergency situations, when couriers or overnight carriers rapidly transfer parts to demand locations. However, lateral transshipment is not explicitly included in the model when deciding the spares allocation among the 17 warehouses. Therefore, the company managers are interested in evaluating the potential benefits deriving from the adoption of a single echelon with a complete pooling policy. To this aim, however, there is a need of effective models for assessing the performance of a single echelon replenishment policy. This is not an easy task for large instances even for steady state analysis, as the size of the instance increases, as in case of high rate demand. And finally there is the need of efficient and effective models for allocating optimally the stock in such a context. In fact, in models where lateral transshipment is taken into account, the resulting optimization problems have non-linear constraints (on service levels) and objective and integer decision variables (like base stock levels). Especially for problem with large numbers of locations optimization is rough. In such models only explicit enumeration [50] has been proposed for solving the Spares Allocation Problem exactly or many heuristics algorithm have been evaluated in [93], [59], [54], [4] [70].

1.3. Research objectives

Based on the needs summarized above, the following research objectives are the subject of this dissertation. We focus on a single echelon one-for-one ordering policy with complete pooling, with a deterministic rule for lateral transshipments.

- The main objective is to formalize mathematically the Spares Allocation Problem (SAP) and understand its mathematical structure for building an exact algorithm for optimally allocating the spares. In fact, in literature to the best of our knowledge no exact algorithm has been proposed for allocating optimally the spares in a continuous review setting rather than a total enumerative algorithm. By exploiting the above algorithm it is interesting

  - Making insight in the SAP and underline which factors influence inventories in such a context.
  - Evaluating fast and accurate heuristics for SAP.

- Efficient and accurate models for assessing the performance of a single echelon replenishment policy are needed especially for large numbers of locations. A drawback of the policy of interest is the state dependent nature of re-forwardings in the systems, it's therefore interesting.
  - understanding the properties of the Markov chain associated to the chosen policy.
  - exploring, despite its state dependent nature, the possibility of expressing the state probabilities of the associated Markov chain model exactly in product form
  - developing fast and accurate approximate models for evaluating the performance and costs in the system, since computing the state probabilities is not practical as the number of states in the Markov chain increases.

The achievement of the first objective clearly requires a strong connection with the resolution of the second objective. In fact, the development of an exact algorithm for allocating the spares may require in contexts with a large number of warehouses and high rates approximate models for assessing the performance and evaluating the costs.

1.4. Research contribution

This thesis presents an innovative contribution to the combined resolution of the research objectives of Section 1.3. Next, we briefly introduce the main achievements. We focus on a single echelon one-for-one ordering policy with complete pooling, with a deterministic rule for lateral transshipments. This policy may be modeled through a Markov chain. In fact, a Markov process allows us to model the uncertainty in many real-world systems that evolve dynamically in time. The basic concepts of a Markov process are those of a state and of a state transition. In specific applications the modeling art is to find an adequate state description such that the associated stochastic process indeed has the Markovian property that the knowledge of the present state is sufficient to predict the future stochastic behavior of the process. A Markov chain is a random sequence in which the dependency of the successive events goes back

only one unit in time. In other words, the future probabilistic behavior of the process depends only on the present state of the process and is not influenced by its past history. The main interest is the long-run behavior of the Markov chain, i.e. long-run averages are usually required in the analysis of practical applications, to this aim it' s necessary defining the equilibrium distribution, if any, and computing this distribution, e.g. $\{\pi_j, j \in I\}$. I is the state space of the stochastic process. Hence, the equilibrium state probabilities π_j may be determined up to a multiplicative constant by the equilibrium equations

$$\pi_j = \sum_{k \in I} \pi_{kkj}$$

where $j \in I$ and $p_i j$ are the one step state probabilities, when the Markov chain is assumed time-homogeneous. The multiplicative constant is determined by the normalizing equation

$$\sum_{j \in I} \pi_h = 1$$

It's known that in case of finite state Markov chain in general there are two methods to solve the Markov chain equations: direct and iterative methods, such as the Gauss-Jordan method and the Gauss-Seidel method respectively. What one usually does to solve numerically the infinite set of equilibrium equations is to approximate the infinite-state Markov model by a truncated model with finitely many states so that the probability mass of the deleted states is very small. Indeed, for a finite-state truncation with a sufficiently large number of states, the difference between the two models will be negligible from a computational point of view. However, such a truncation often leads to a finite but very large system of linear equations whose numerical solution will be quite time-consuming, although an arsenal of good methods is available to solve the equilibrium equations of a finite Markov chain. Moreover, it is somewhat disconcerting that we need a brute-force approximation to solve the infinite-state model numerically. Fortunately, many applications allow for a much simpler and more satisfactory approach to solving the infinite set of state equations. Under rather general conditions the state probabilities exhibit a geometric tail behavior that can be exploited to reduce the infinite system of state equations to a finite set of linear equations. The geometric tail approach results in a finite system of linear equations whose size is usually much smaller than the size of the finite system obtained from a brute-force truncation. Hence, it's clear that computing the state probabilities is not practical as the number of states in the Markov chain increases In fact, the above methods suffer from computer

memory problems and for long computation time.
Hence, we have explored the possibility to express the state probabilities of the associated Markov chain model in product form, thus reducing the computational effort. We have tested it numerically, but as expected the state dependent nature of re-forwarding in the systems does not allow to express these state probabilities in product form.
Therefore, we adapt four approximation techniques to our model and evaluate their performance in terms of computational effort, memory requirement and error with respect to the exact value (objective 2). Three techniques approximate state probabilities with others that can be expressed in product form, so that the Markov chain can be decomposed. Specifically, we adapt a method by Alfredsson and Verrijdt, the Equivalent Random Traffic method and the Interrupted Poisson Process method. The fourth technique is based on the multi-dimensional scaling down approach, which studies an equivalent reduced Markov chain rather than decomposing the original one. The first three methods are based on decomposition approach. State probabilities are approximated with others that can be expressed in product form, so that the Markov chain can be decomposed and operational availability can be easily computed. Specifically, our first method, referred to in the following as the AV method, is a slight modification of a method by Alfredsson and Verrijdt [2]. The second and third method is based on ideas successfully used in the field of telecommunications, and specifically in the Equivalent Random Traffic method (ERT method) [45] and the Interrupted Poisson Process method (IPP method) [56],[61]. With the IPP and ERT methods part of the traffic may be lost when no server is available. In our adaptation of these methods we include the presence of an external supplier to avoid lost requests. The fourth method is based on the multi-dimensional scaling down approach, which studies an equivalent reduced Markov chain rather than decomposing the original one. A scaling down approach is used by Axsater [6] to study a two-echelon policy. We adapt this method to study the single echelon policy with complete pooling in a Markovian framework without decomposing the original chain. Computational experiments, carried on practical data from an airport equipment maintenance context show the accuracy degree and the computational effort required by each approximate method.
A major contribution of this thesis to solving efficiently and timely the Spares Allocation Problem consists in a branch and bound algorithm (objective 1). In literature an exact efficient method to solve SAP in a continuous review setting seems to lack. As noted in [50] it is very likely that no polynomial time optimization algorithm exists for our type of problem. The problem under con-

sideration in this dissertation could also be considered as a nonlinear knapsack problem. For a general description of knapsack problems, see e.g.[46]. Kellerer et al. [46] prove that even the simplest type of knapsack problem belongs to the class of NP-hard problems. This is generally considered as strong theoretical evidence that no polynomial time algorithm exists for computing optimal solutions and thus a good reason for looking for efficient exact enumerative techniques or to apply approximation algorithms. Therefore, after modeling the stock allocation problem as a non convex integer program, we exploit the special structure of the problem to design an efficient branch and bound procedure. Our bounds are obtained by solving a reduced problem with convex objective function, solvable at optimality very efficiently. Computational experiments, carried on practical data from an airport equipment maintenance context show that this method efficiently solves at optimality many practical instances.
Different cost scenario are evaluated for understanding which factors influence inventories and when the proposed procedure is more efficient.
Moreover a simple and fast heuristic is computed by distributing spare parts among warehouses with positive demand and by giving preference to warehouses with larger demand (objective 1). In fact, simulation experiments carried out in [15] show that avoiding concentration of spares in few warehouses is an effective allocation policy. The accuracy of such an heuristic is evaluated by comparing it with the branch and bound allocations.

1.5. Outline of the thesis

This section gives a short introduction to each chapter.

Chapter 2 provides an overview of the state-of-knowledge in spare parts provisioning. In a first part, some relevant contributions to spare parts inventory control are described. In a second part, we focus on transshipment problems in Supply Chain Systems. Finally we give an overview of some relevant analytical and computational methods, used in the analysis of stock allocation problems where base stock policy is assumed. In fact, because of this assumption we can separate evaluation of a given policy and optimization, the latter giving a value to the decision variables of interest. Therefore, specific methods must be used for evaluation and optimization respectively. Note that the stochastic nature of the problem of interest is taken into account in the analysis of a given policy.

Chapter 3 introduces the main notation used in this dissertation and the two main mathematical models involved in the SAP analysis and solution. The first is a Markov chain model useful for evaluating the costs and waiting times in the system, while the second is an optimization model for stock allocation. This chapter also contains a short description of an optimization model used to prove that a product form of the state probabilities of the associated Markov chain model does not exist.

Chapters 4 describes four models for fast approximation of cost and performance.

In Chapter 5 heuristic and exact allocation algorithms are described.

The main results obtained in this thesis are summarized in Chapter 6. Further research is also addressed.

In Appendix 6.2 we give a resume of the Markov chain theory.

In Appendix 6.2 we describe shortly phase-type distribution and its evolutions.

In Appendix 6.2 we describe some optimization techniques applied in this dissertation.

Chapter 2

An overview on spare parts provisioning

2.1. Spare parts inventory control

Taxonomy of service parts inventory systems

Different elements affects the amounts of inventory found in various portions of a service parts inventory system. In general there are numerous reasons for choosing to stock inventory of an item type within a system, often at multiple locations. The underlying echelon or network resupply structure will have a substantial impact on the amount of inventory needed. There are clearly many possible structures. However, for each one, there is usually a well defined resupply plan. Some systems have many echelon, some have fewer. While the basic structure may be similar for several systems, the actual operating environment can vary dramatically. Different item characteristics may create operational differences between systems, whether the systems are of the same type or not. In fact, each service parts resupply system is designed to accommodate the items found in it. The systems, and the items within them, can have varying characteristics.

- Systems differ in the number of items that are managed. In some environments there are just a few hundred or a few thousand items, up to hundreds of thousands of items.
- The demand rate among items can vary substantially. Demand rates of items also differ dramatically by location within a resupply system, as

well as between different resupply systems.

- The unit shortage, holding and transportation costs differ dramatically among the items as well. We also note that also transportation costs can be a substantial component of operating a service parts resupply system. In some instances the total cost of moving material can amount to hundreds of million of dollars per year. The size and weight of each item along with their demand rates obviously determine the volume, weight and quantity of material that must be transported. But the mode of transport selected to move this material is an important factor in determining the annual transportation costs.
- The procurement, transportation and other components of lead times associated with each item determine the amount of inventory carried in the system in two ways. There is pipeline stock that exists because of the time it takes to receive orders after they are placed, that is, the resupply time. Based on Little's Law, this time results in an average number of units in the resupply system. Thus the choices of suppliers, transportation modes and inventory policies, all affect the average resupply lead time and hence the average pipeline stock. Furthermore lead times are not always constant. Another factor which influences the resupply time is the inventory policy followed by the supplying location. When orders are shipped immediately because stock is on hand at the supplier, then resupply lead times are one value. If the supplier does not have stock available to ship, then the resupply action is delayed for some amount of time. This uncertainty in lead times is an important factor when setting the stock levels. We note that the average and uncertain length of resupply lead time also affects the second type of stock that is required: the safety stock. There will be inherent variability in the demand processes for each item. It's common to assume in such a context that demand over replenishment lead times is governed by a random process. The degree of difference in this variation of demand can be substantial. Uncertain demand over uncertain replenishment lead times yields a requirement for safety stock. In many real world situations, safety stock is the predominant component of total stock for most items.
- Some service parts are consumable and some are repairable after they fail.

There are many other characteristics associated with items that are of importance when setting the inventory levels. These include the physical character-

istics of the items (volume, weight and shape), the special temperature and humidity storage requirements, the possibility of items becoming obsolete and the substitutability of one item type for another. In this dissertation we do not consider these other factors. Many types of inventory policies are found in practice. These range from policies that are location specific to echelon-stock-based inventory policies in which total system stock and system performance across all items at all location are considered. The inventory position at a location for an item is equal to its on-hand plus on-order minus backordered inventory. Reorder points are usually expressed in terms of inventory position. When following an (s,S) policy a location places an order when its inventory position falls to s or below and an order is placed to raise the inventory position to S. Echelon stock in a resupply network refers to the inventory position at that location plus all the inventory found in the resupply system for successor (downstream) locations in the resupply network. In some environments, inventory levels are monitored continuously while in others they are monitored only periodically. Policy implementation obviously depends on wether reviews are continuous or periodic. One important class of policies are called base stock, order-up-to or (s-1,s) policies. When employing these policies in a continuous review environment, an order is placed every time a demand arises. The quantity ordered equals the quantity demanded. In periodic review situations, an order is placed in a period to raise the inventory position to some specified level. In both cases, some target inventory level, based on either echelon or installation inventory position, is used to trigger an order. Thus, whenever the inventory position is below s when a review occurs, an order is placed immediately to raise the inventory position for the location to s.

An overview of the literature

Multi-echelon models

In 1968, Sherbrooke [72] published a landmark paper in which he described a mathematical model for the management of repairable items called METRIC (Multi Echelon Technique for Recoverable Item Control). Since that time many extensions and modifications to that model were proposed. The exact distribution of the number of units in the resupply system at each location in a two echelon depot base system is too computationally burdensome to be of practical use, refer to [63] for its computation. Hence the METRIC model is based on an approximation to this distribution that is easy to compute, and therefore has been widely used in many applications. The METRIC approach substan-

tially means that we evaluate the average delay for stocking point orders due to shortages at the depot. This average delay is added to the stocking location transportation times to get exact average lead times for each location. When evaluating the costs at the locations, these averages are, as an approximation, used instead of real stochastic lead times. Hence a METRIC type approximation is quite simple. It means essentially that a multi-echelon system is decomposed into a number of single echelon systems. Substantially under the METRIC approximation the number of backorders of each local warehouse is assumed to be Poisson distributed [72] and the improved two-moment approximations for that [75, 39]. Although the errors may be substantial in some cases, the approach is also often reasonable in practical applications.

Nahamias [65] and Axsater [5], [8] review the literatue. See Graves [39], Axsater [4] and Sherbrooke [73] for enhancements and applications.
In particular, METRIC was extended to represent more complicated environments in which there are both repairable assemblies and subassemblies. This model was originally developed by Muckstadt [62], who included a hierarchical or indented-parts structure (MOD-METRIC).

In METRIC-type models, ample repair capacity is assumed. A complete stream of research is devoted to the situation with limited repair capacity. When limited repair resources are available, it pays off to set certain repair priorities. For an overview of work that studies limited repair capacity, see Sleptchenko [78].

Cohen et al. [22] have considered the problem of determining stocking policies for low usage items in multi-echelon inventory systems. The problem of determining the stocking quantities for the various parts so as to yield an optimal trade off between holding costs and transportation cost is made worse due to innovations and competitive pressures resulting in complex echelon structures, high priority for service, low demand probabilities, etc. Their paper develops a formula to find the target stocking levels which minimizes the total cost of the system subject to the satisfaction of the service level constraint. When the number of stocking points or stocking levels becomes high, the possible number of stocking policies will also be high, and they use a branch and bound algorithm to obtain the solution, merging all stocking points at all levels to obtain the cost of the full structure. This is followed by branching the structure starting from the toplevel and finding an optimal stocking policy at each level. In a separate but related paper, Cohen et al. [23] discuss the situation where

the requirement for rapid response is concurrent with a need for low levels of inventory. In this situation, both the common competitiveness requirements facing organizations today and the allocation decision for service support becomes crucial. This paper fixes attention on a given product or product family and defines a multi-echelon inventory system based on level by level decomposition using the single location problem as their basic building block. Such rapid response implies regional and local suppliers for final products and spare parts for repair. They establish two types of demand: customers (emergency) demand and normal. Part stocking is assumed to follow a (s, S) policy for which a cost function is formulated.

Most of the previous study is focused on dealing with problems in a two echelon supply chain network, where it includes a single source supplier-warehouse at the higher level and multiple (two or more than two) stocking locations at the lower level. The assumption for simple problem structure are necessary for the reason of computational tractability in the process of finding the optimal solution. Especially, the earlier study addressed relatively simple model with two stock points and/or one single period, thus limiting their practical application. To alleviate the loss of realism, the recent researchers have attempted heuristic approximation and/or simulation approaches in their analysis for the supply chain system with increased members, e.g. [71], [66], [19].
For relevant research specifically devoted to lateral transshipments (inventory pooling) refer to Section 2.2.

Number of items

Most of inventory related research deals with single item problems in which only one item at a time is considered. Such problems are typical when we use an item approach, where inventory levels for each individual item are set independently. An alternative approach, denoted as the system approach by Sherbrooke [74], considered all items in the system when making inventory levels decisions and may lead to large reductions in inventory costs in comparison to an item approach.

Archibald et al. [3] considered a two location, multi item, multi period, periodic review inventory system subject to a storage space limitation for all items. The demand is assumed to follow Poisson distribution and transshipments are possible during a period in response to stockouts.

Wong et al. [94] investigated a two location, multi item continuous review system for repairable items wit one for one replenishment. The optimization problem is to determine stocking policies for all items minimizing the total system cost subject to a target level for the average waiting time for an arbitrary request for a ready-for-use part at each of the two locations. In their model, the decisions with respect to different items are coupled because of the multi-item service measure that is used. The solution procedure requires a long computation time to solve rather large problems.

To overcome that limitation Wong et al [93] developed a simple and efficient solution procedure to obtain close-to-optimal solutions for the multi-item problem with lateral transshipments. The model is further extended to the case with multiple (and not limited to two) locations. Further, they also analyze the magnitude of the savings obtained by using the multi-item approach and lateral transshipments. An efficient heuristic algorithm may be found also in [70].

Performance criteria

The commonly used performance measures are the cost and the service levels. The relevant costs are for short stockout costs, holding cost, transportation cost and ordering cost. There are two relevant criteria for the performance of the system: we do not want to order too frequently, because of scale economies, nor do we want to carry to much inventory. Typically these are translated into more precise criteria focusing on long-run averages over time.

In fact, in a system in general for example after developing the steady state probabilities for the number of units that are in the resupply system (both via transshipments or via emergency shipments) at a random point in time when the demand process has an assumed behavior, it is possible calculate different measures of system performance.

The first performance measure we want to underline, the fill rate, is the most commonly used measure in practice and is defined as follows. Given a stock level s, the fill rate, F(s), is the expected fraction of demands that can be satisfied immediately from on-hand stock. As is intuitively clear, as s increases the fill rate will increase.

In spare part literature refer to Thonemann et al. [85] and to Vliegen and

Van Houtum [89] for fill rate usage as service level measure.

A second performance measure is called the ready rate corresponding to stock level s. The ready rate measures the probability that an item observed at a random point in time has no backorders, that is its net inventory is non negative. We denote the ready rate by R(s). Either there are backorders or there are no backorders at a random point in time.

Silver [76] make use of this performance measure in spare part context.

Observe that when computing either a fill rate or ready rate we are not concerned with the duration of backorders when they occur. Thus for example a fill rate of 95% implies that on average 95 of every 100 units that are ordered have that request satisfied immediately. But we are not measuring how long it takes satisfy the other 5% of the units requested. Thus is not always clear that a firm which maintains a high fill rate is truly satisfying its customer needs.

A third performance criterion measures the expected number of backorders outstanding at a random point in time and is denoted by B(s). It is a response-time focused measure.
Observe that B(s) is equal to the demand rate times the average "waiting time" of a demand. As noted in Muckstadt [63] this is a consequence of Little's law, $L = \lambda W$, where B(s) is L, λ the demand rate and W the average waiting time. We could also compute the conditional value of W, given that backorders exist.

Sheerbrooke [73] considered multi-item, continuous review policies in a spare part setting and showed that maximizing the equipment availability is approximately equivalent to minimizing the sum of expected backorders, suggesting the use of total expected backorders as service measure.

We next shortly describe how these performance measure can be computed.
We assume that backorders are allowed.
Let us now denote the random variable representing the number of units that are in resupply as X. Therefore $P\{X = x\}$ is the probability of having x units in resupply. $P\{X = x\} = p(x|\lambda\tau)$, where λ is the demand rate and τ is the average resupply time.

The ready rate is the probability that there are no backorders existing at a random point in time. This is the probability that the number of units in re-

supply is s or less. $R(s) = \sum_{x=0}^{s} p(x|\lambda\tau)$.

The computation of the fill rate is more difficult, but it is obtained from the steady state probabilities, $p(x|\lambda\tau)$.
Suppose a customer order is received. There will be one unit of the order satisfied if there are $s-1$ or fewer units in resupply. A second unit will be sent to the customer if the order is for two or more units and there are $s-2$ or fewer units in resupply.
Hence the expected number of units filled per customer order is given by

$$\begin{aligned} F_1(s) \quad &= \textstyle\sum_{x-1} p(x|\lambda\tau) + (1-u_1) \sum_{x\leq s-2} p(x|\lambda\tau) + \\ &+ (1-u_1-u_2) \textstyle\sum_{x\leq s-3} p(x|\lambda\tau) + (1 - \sum_{j\leq s-1} u_j) p(0|\lambda\tau) \end{aligned} \tag{2.1}$$

u_j measures the probability that a customer order is exactly for j units.
For example when the demand process is Poisson $F_1(s) = F(s) = R(s) - p(s|\lambda\tau)$ and $F(s) < R(s)$.
In case of compound Poisson demand $\lambda F_1(s)$ measures the expected number of units that can be shipped on time per day, when λ is the expected daily rate at which customers place orders.
$\lambda\overline{u}$ measures the expected number of units demanded per day, where $\overline{u}$ is the expected number of units demanded per order. Thus $\frac{F_1(s)}{\overline{u}}$ measures the fraction of the units ordered that are sent to customers on time. Here is the fill rate $F(s)$. Next we see that the expected number of units in backorder status in steady state is

$$B(s) = \sum_{x>s} (x-s) p(x|\lambda\tau)$$

Properties of the performance measures

Some properties of the performance measures that will be important in the analysis of stock allocation problems are described in what follows. Such properties are shown for the general representations of the service measure given above and may be commonly found also in specific service measures used in practice.

We begin by analyzing the fill rate measure. Let us assume for simplicity that the demand process is a simple Poisson process with rate λ. Furthermore, assume that resupply times for each order are independent and identically distributed with mean τ. From the Palm's Theorem, the probability that x units

are in the resupply system in steady state is given by

$$p(x|\lambda\tau) = e^{-\lambda\tau}\frac{(\lambda\tau)^x}{x!}$$

In fact, Palm Theorem states that if demand follows a Poisson process with mean λ and the replenishment lead time is independently and identically distributed according to an arbitrary distribution with mean τ, then the steady-state probability distribution for the number of items in the replenishment pipeline follows a Poisson distribution with mean $\lambda\tau$. Since the demand process is a simple Poisson process, the fill rate, given a stock level s is given by

$$F(s) = 1 - \sum_{x \geq s} p(x|\lambda\tau) = \sum_{x<s} p(x|\lambda\tau)$$

Perhaps the optimization goal might be to choose stock levels so that the average fill rate is maximized given some target investment level in inventory. This type of optimization problem would be easy to solve if F(s) were a discretely concave function. Unfortunately it is not. We have

$$\Delta F(s) = F(s+1) - F(s)$$

and

$$\Delta^2 F(s) = (s+1) - (s)$$

Hence with a demand process being a Poisson process $\Delta^2 F(s) > 0$ when $\lambda\tau > s+1$ and F(s) is not concave in that region. Hence F(s) is discretely concave only when $s \geq \lambda\tau - 1$ when $\lambda\tau$ is an integer. Hence typically in practical cases s may be constrained to assume values that are greater or equal to $\lceil\lambda\tau\rceil$ to assure that the fill rate is concave over the feasible region.

The backorder function B(s) has very desirable mathematical properties, however.

$$B(s) = \sum_{x>s}(x-s)p(x|\lambda\tau)$$

For B(s) to be discretely convex and strictly decreasing requires

$$\Delta B(s) = B(s+1) - B(s) < 0$$

and

$$\Delta^2 B(s) = (s+1) - (s) > 0$$

We have

$$\begin{aligned} \Delta B(s) &= \sum_{x \geq s+1}(x-(s+1))p(x|\lambda\tau) - \sum_{x \geq s+1}(x-s)p(x|\lambda\tau) \\ &= -\sum_{x \geq s+1} p(x|\lambda\tau) < 0 \end{aligned} \tag{2.2}$$

and

$$\begin{aligned} \Delta^2 B(s) &= -\sum_{x \geq s+2} p(x|\lambda\tau) + \sum_{x \geq s+1} p(x|\lambda\tau) \\ &= p(s+1|\lambda\tau) > 0 \end{aligned} \tag{2.3}$$

and hence B(s) is a strictly (discretely) convex function of s for all $s \geq 0$.

Space, capacity and time

Space, capacity and time constraints are three factors that can affect significantly the system performance, either costs or service level. Not many works has been done in the areas of transshipment problem accounting for these factors.

Wong [93, 94] investigated multi item spare parts system, minimizing total costs for inventory, holding, lateral transshipments and emergency shipments subject to a target level for the average waiting time per demanded part at each location. Recent similar studies may be found in [70].

Van Houtum and Zijm [87] classified inventory systems as two categories: service model and cost model. In a service model the objective is to minimize the total system costs subject to a set of service level constraints, such as space, time and capacity constraints. In a cost model, however, the service constraints are replaced with shortage penalty costs. Although in general the cost models are analytically more tractable, they have a serious limitation in that the penalty costs are generally hard to estimate. Archibald et al. [3] analyzed a multi-period, periodic review model of a two locations inventory system with limited storage space.

These kind of optimization problem with space, capacity and waiting time constraints is appropriate to be analyzed by Lagrange relaxation [94].

Background: analysis of one-for-one and order up policies

In what follows we focus on a one-for-one (s-1,s) policy in the continuous review case. This policy is appealing intuitively in our context, where item costs are usually high. It is nonetheless important that they are optimal in many circumstances.

It is possible to show the optimality of the (s-1,s) policy for managing a single item by considering a single location and a serial system both when inventory levels are monitored periodically or continuously.

Classical proofs are based on dynamic programming. In their seminal paper, Clark and Scarf [21] proved the optimality of base stock policies for incapacitated, periodic review, finite horizon, serial systems using dynamic programming approach.

A different approach to prove the same result was introduced by Federgruen and Zipkin [34] to prove the optimality of echelon base stock policies in the infinite horizon case. The arguments are based on a lower bound on cost.

A third approach for establishing the forms of optimal policies in inventory systems is the single item single customer approach introduced by Muharremoglu and Tsitsiklis [64]. They proved that state dependent echelon base-stock policies are optimal for incapacitated multi-echelon serial systems for both the finite and the infinite horizon models when lead times and demands are Markov modulated.

In Muckstadt [63] their approach is presented and discussed. Under the hypothesis of compound Poisson demand, continuous review, constant lead times and a serial system of locations, the key idea in this innovative proof is to decompose the system in a collection of countably infinite subsystems, each having a single stock unit and a single customer demanding one single part.

Let us now consider the continuous time divided into periods of different length: the length of a period is the time between the arrival of two consecutive customer orders.

Let us now consider each unit of demand as an individual customer. Suppose at the beginning of period 1 there are v_0 customers waiting to have their

demands satisfied. Let us now index these customers as $1, 2, 3, \ldots, v_0$ in any order. All subsequent customers are indexed $v_0 + 1, v_0 + 2, \ldots$ in the order of the period of their arrivals, arbitrarily breaking ties among customers that arrive in the same period.

Next, define the concept of the distance of a customer at the beginning of any period. Every customer who has been served is a at distance 0. Every customer who has arrived, placed in the actual order, but who has not yet received inventory, is at distance 1. All customers arriving in future periods are said to be at distance $2, 3, \ldots$, corresponding to the sequence in which they will arrive. Distances are assigned to customers that arrive in the same period in the same order as their indices. This ensures that customers with higher indices are always at higher distances.

Next, define the concept of a location of a unit. At location 0 there are the units already used to satisfy a customer order. At location 1 there is the stock on hand in the last warehouse in the serial system. Then there are as many artificial locations as the maximum possible lead time between the last stage in the serial system and the stage which precedes it. Then there is another physical location representing the stage which precedes the last one and so on up to have a location representing the supplier. For short there are as many physical location as many stages the serial system has, up to the supplier, and as many artificial location as the sum of the maximum possible lead time between two consecutive stages for every consecutive stages couple in the serial system. If the unit has not been ordered from the supplier it is in the location with the greatest index, which represents the supplier.
At the beginning of period 1, an index is assigned to all units in a serial manner, starting with units at location 1, then at location2, and so on. Arbitrarily assign an order to units present at the same location. Assume a countably infinite number of units available at the supplier.

The state of the system at the beginning of period n is a vector with an element which stores the realization of the demand in the period n and a countable infinite number of couples. In the j-th couple the first element stores the distance of customer j at the beginning of period n and the second stores the location of unit j at the beginning of period n.

Define a release action as an order placement for a unit, which enter in the distribution system of the supplier/intermediate warehouses in the serial sys-

tem, otherwise an hold action is realized.
Let a policy be a vector of release/hold actions for each unit in the location representing the supplier and those representing the intermediate warehouses in the serial system.
Let a committed policy be such that it ensures that the only customer that the j-th unit can satisfy is customer j's demand and that the only unit that customer j can receive is the j-th unit. Let a monotone policy be such that it always releases the units with the smallest indices from the supplier and the intermediate locations in the serial system.

In each period the demand is observed. At the beginning of period n each stage places an order to the former one, which release the number of units requested to the successive location, i.e. these units enter in its distribution process. By observing the number of units in the artificial locations, we know how many units were ordered a known number of periods before.

The demand D_n is realized, these new customer arrive and are at distance 1. All customer at distance $2, 3, \ldots, 2 + D_n - 1$ all arrive at distance 1. All customers at distance $2 + D_n, 3 + D_n, \ldots$ at the beginning of the period move D_n steps toward distance 1.

Units on hand at the first warehouse and waiting customers are matched to the extent possible.

Then, h monetary units are charged per unit of inventory remaining on hand and b monetary units are charged per waiting customer at distance 1. We assume $b > h$, thus ensuring that if the inventory position is negative in some period, then the optimal policy will be to increase the inventory position to some non negative level.

The outline of the proof is as follows.
Each pair "j-th unit - j-th customer" represents the j-th subsystem.
The cost for the overall system is the sum of the expected costs for the subsystems because of the linear cost structure. In fact, every monotone and committed policy for the entire system corresponds to a set of a monotone and committed policies for the subsystems and any set of monotone and committed policies for the subsystems yields a feasible policy for the system.
When the individual subsystems are managed independently and optimally, the resulting policy for the entire system is optimal.

The optimal policy for a subsystem should be such that if it is optimal to release a unit from a stage or physical location when the corresponding customer is at distance y, then it would also be optimal to release the unit from that stage if the customer were any closer. Consequently, an appropriately defined so called "critical distance" policy is optimal for every subsystem. There is a critical distance corresponding to each stage. When this policy is used for every subsystem, the resulting policy is a state dependent echelon base stock policy for the entire system.

Observe that the subsystems are operationally independent in the sense that each subsystem can be managed independently without being affected by the policies used to manage the other subsystems. Those parts of the state vector that pertain to unit j and customer j are a sufficient state descriptor for the j-th subsystem. A subtle point to be noted is that the subsystems, though operationally independent, are stochastically dependent through the demand process.

Theorem 1 *For any starting state x_1 in period 1, the optimal expected cost in periods $1, 2, \ldots$ for an entire system S equals the optimal expected costs in periods $1, 2, \ldots$ for the group of subsystems. Furthermore, when each subsystem w is managed independently and optimally the resulting policy is optimal for the entire system.*

The proof is given in [63]. Next we show the existence of an optimal policy with a very special structure for every subsystem, a so-called a "critical distance" policy.

Theorem 2 *If it is uniquely optimal for subsystem w to release unit w (if it is at the supplier or at any physical location) in period n when the system is in the Markovian state s_n and customer w is at a distance $y + 1$, then it is optimal to release it if the customer were any closer.*

The proof is given in [63] and is by contradiction: it is suboptimal for a subsystem to hold unit w if customer w is at distance $y+1$ while it is suboptimal for a subsystem to release unit w if customer w were at a distance y. Therefore, a "critical distance" policy y for a certain stage in period n and Markovian state s_n for every subsystem is the maximum distance in which it is optimal to release: it is optimal to release unit w if and only if customer w is at distance of y or closer. When the critical distance policy is used in period n for every subsystem, the resulting policy for the original entire system is an order-up-to policy.

Theorem 3 *The optimal policy for the system is to release as many units as necessary to raise the inventory position in each physical location to its critical distance minus 1 in period n when in Markovian state s_n.*

The proof is given in [63].

2.2. Transshipment problems in Supply Chain Systems

Effective supply chain management has become an important management paradigm. Basically, it is an effective an systematic approach of managing the entire flow of information, material and services in fulfilling a customer demand. In this dissertation we are mainly focused on material flow management in the supply chain system. At present many quantitative models have been proposed to provide decision support for the management of materials in supply chain [83]. Moreover, since the network of entities that constitute the entire supply chain is typically too complex to analyze and optimize globally, it is often desirable to focus on smaller parts of the system. One such part that is attracting growing attention is the local distribution network, consisting of multiple stocking locations, which are supplied by one or more sources.
The overall performance of the distribution network, whether evaluated in economic terms or in terms of customer service, can be substantially improved if the stocking locations collaborate in the occurrence of unexpectedly high demand, which may result in shortages in one or more locations. Collaboration usually takes the form of lateral inventory transshipment from a stock point with a surplus of on-hand inventory to another location that faces a stockout. Since the cost of transshipment in practice is generally lower than both the shortage cost and the cost of an emergency delivery from the designated warehouse and the transshipment time is shorter than the regular replenishment lead time, lateral transshipment simultaneously reduces the total system cost and increases the fill rates at the locations. A group of stocking locations that share their inventory in this manner is to form a pooling group, since they effectively share their stock to reduce the risk of shortages and provide better service at lower cost.

Common assumptions

As pointed by Chiou [20] there are several basic assumptions that are commonly seen in the literature of transshipment such as the behaviors of demand occurrence, transshipment time, repair time and transshipping priority rule.

The behaviors of demand occurrence are usually characterized by the time between demands and the distribution of demand size. The time between demands is commonly assumed to follow an Exponential or Gamma distribution. However, the distributions of demand size per each demand occurrence depend on the characteristics of the investigated industry. For example, it was taken as Weibull distribution for spare parts which have slow moving, expensive and lumpy demand pattern [55]. Needham and Evers [66] assume the normal distribution for military spare parts. A drawback of using the normal distribution is that it is less appropriate for low volume items [77], however, it does not place any restriction on the values of the mean and variance. Besides, Wong et al. [94] assumed the demand occurs according to the Poisson process with constant rate for repairable parts in equipment-intensive industries such as airlines, nuclear power plants and manufacturing plants using complex machines. Furthermore in a large amount of transshipment literature the behaviors of demand are alternatively characterized by assuming what distribution the average demand per time period follows [66, 81, 94].

In the majority of the literature transshipment time is assumed to be negligible. Kukreja and Schmidt [55] assumed that a part can be transshipped between any two locations within a working day. This transshipment time is assumed to be negligible. At present only some papers account for the non-negligible transshipment time. In any case transshipment times are assumed to be shorter than emergency supply. Lateral transshipment are faster and cheaper than emergency supplies. Otherwise it makes no sense to pool the item inventories. Wong et al [93, 94] addressed the analysis of a multi item, continuous review model of a multi location inventory system of repairable spare parts with lateral transshipment and waiting time constraints, in which lateral and emergency shipments occur in response to stockouts. He considered non negligible transshipment times. For the case of transshipment for spare parts, the repair time is usually assumed exponentially distributed. This assumption is probably not very realistic. However Axsater [4] and Alfredsson and Verrijdt [2] showed that the service performance of the system is insensitive to the choice of the lead time distribution.
Wong et al.[91] showed that delayed lateral transshipments can improve the system performance, i.e. if a location having no backorders receives a repaired part and at the same time at least one location in the pooling group has backorders, than it is reasonable to send the repaired part to the location with backorders.

One common transshipping priority rule for fulfilling the demands is that a location receiving an order first satisfied its own backorder, if one exists and then uses the remaining units to satisfy backorders at other locations in a way that minimizing transshipping costs. The requested backorders are to be fulfilled according to FIFO policy. A significant amount of literature in transshipment assumed that complete pooling policy is to be applied. A unit demand is backordered if it cannot be satisfied via transshipment, in other words when there are no units in the system. In case last parts cannot be shared, one may introduce threshold parameters, having a situation of partial pooling,and agree that a stocking point does not supplies a part by lateral transshipment if the physical stock of the requested item is at or below the threshold level. A rule has to be added for how the values of the threshold parameters are chosen, e.g. [7].

Preventive and Emergency transshipments

There are two classes of transshipment. Lee [60] proposed that lateral transshipment can be divided into two categories: emergency lateral transshipment (ELT) and preventive lateral transshipment (PLT).
ELT is an emergency redistribution from a stocking point with ample stock to a location that has reached stock out. However, PLT reduces risk by redistributing stock between retailers thus anticipating stockout before the realization of costumer demands. In short, ELT responds to stockouts, while PLT reduces the risk of future stockout.

Lee [59] presented a model that allows ELT between local warehouses that are part of a group. If a local warehouse cannot satisfy costumer demands with its on-hand stock, ELT is used to fill the demands from a warehouse in the same group that has enough stock on-hand. If ELT is impossible due to group-wide stockout, the unmet demand will be backordered. Lee [59] derived expressions that approximate the fractions of demand that can be satisfied by stock on-hand, ELT and backordering, and in doing so proved that applying lateral transshipment reduces total costs.

Axsater [4] analyzed a system similar to that of Lee but with the modification of assuming that warehouses within each group are not identical. Axsater derived steady state probability by assuming exponentially distributed replenishment time. Analytical results were compared with simulation results to

show that in case of non identical warehouses the proposed model gives better results. Rather than describing all the approaches for incorporating lateral resupply into models, e.g. Archibald [3], Lee [59], Alfredsson and Verrijdt [2], Axsater [4], Taragas [81], Taragas and Cohen [80], Sherbrooke [73], we want to focus just on two relevant models: Lee's model [59] and Axsater one [4], respectively. Both are approximations.
In the first model Lee constructs probability distributions for key random variables, but also constructs an economic model that can be used to set stock levels in a two-echelon system. This model is a METRIC like (Multi Echelon Technique for Recoverable Item Control) model, refer to the landmark paper of Sherbrooke [72]. A METRIC model is based on an approximation to the distribution of the number of units in the resupply system at each location in a two-echelon depot base system. The second model is a queueing like model based on the assumption that the underlying system is governed by a continuos-time Markov process. In such a model Axsater focuses on computing probabilities of system performance resulting from given stock levels. The models pertain to repairable items. Lee conducted experiments showing that the approximations are accurate when service levels are high. Axsater developed the alternative model, which substantially differs from the previous in a couple of ways. The most important difference is as follows. In Lee's model the location demand rate is implicitly independent of wether or not there is stock on hand at the location.The other difference between the models arises because Axsater represents the entire operating environment as a continuous-time Markov process. Axsater model is more accurate.

Policies

Transshipment policies are incorporated with traditional inventory control policies, which are classified based on two fundamental questions: when to replenish and how much to order. We focus here just on (s-1,s) policies, because we assume this policy in this dissertation. In fact, continuous one-for-one stock replenishment is a commonly used inventory control policy for a system in cooperation with transshipment. It means whenever any stock is withdrawn, a replenishment order is released. This control policy is especially suitable for slow-moving and expensive items. The first to deal with continuous (s-1,s) policy in multi echelon systems with transshipment were Dada [24] and Lee [59]. One can refer to the following research for more in depth description [59], [4], [73], [2], [38], [54], [92, 95], [57].

Lee [59] developed a method for determining the minimum cost inventory position for a system that allows transshipments between identical locations and finds approximations to measures of system performance including the expected number of backorders and transshipments. Derived expressions approximate the fractions of demands that can be satisfied by stock on-hand, ELT and backordering, and in doing so proved that applying lateral transshipment reduces total cost.

Both Axsater [4] and Sherbrooke [73] proposed similar approximations for systems that allow transshipments between non identical locations. Axsater analyzed a system similar to that of Lee, but with the modification of assuming that warehouses within each group are not identical. Steady state probability is derived by assuming exponentially distributed replenishment time. Analytical results were compared with simulation results to show that, in case of non identical warehouses, gives accurate results.

Recently, Grahovac and Chakravarty [38] formulated and solved the proposed model based on (s-1,s) policy. They reached some counter-intuitive conclusion that is saving is not always accompanied by a reduction in the overall inventory in the supply chain. These opposing trends suggest that new extra incentives are needed to enforce the transshipment arrangement.

In addition Kukreja et al.[54] developed a heuristic to determine replenishment and transshipment policies for a system with non identical locations under the objective of minimizing cost.

Wong et al. [94] extended the single item model of [92] to a model of multiple items. They analyze a two location, multi item, continuous review system for repairable items with one-for-one stock replenishment and determine stock levels for all items minimizing the total cost subject to a target level for average waiting time. These models are appropriate specifically for slow moving, expensive and repairable items.

In [93] Wong et al. extended their study to more than two locations.
Finally Kranenburg [50] studied the multi-item spare parts inventory models in which the different features of commonality, service differentiation, lateral transshipment, two-echelon structure, and two transportation modes are incorporated. These results has been collected in [51], [52], [53].

Performance criteria

The implementation of supply chain management has significant impact on cost, service level and quality. Lateral transshipments represent one way in which logistics managers can reduce inventories while simultaneously maintaining customer service levels. Therefore, the commonly used performance measures to evaluate the effectiveness of transshipment are the cost and the service levels. The relevant costs considered in the transshipment model are similar to those of inventory research. Stockout costs, holding cost, transportation cost and ordering cost. Hence in short there are two relevant criteria for the performance of the system: we do not want to order too frequently, because of scale economies, nor do we want to carry to much inventory. Typically these are translated into more precise criteria focusing on long-run averages over time.

2.3. Methodology

In this section we describe some techniques that can be used in the analysis of the Spares Allocation Problem. Recall that we assume that a single echelon one-for-one (base stock) ordering policy with complete pooling, with a deterministic rule for lateral transshipments is used. Because of this assumption we can separate evaluation and optimization. Evaluation constitutes the analysis of a given policy. Evaluation techniques can be exact or approximate. We discuss examples of both. Optimization concerns the process of finding optimal values for the decision variables. The type of problems we are looking at have integer-valued decision variables and non-linear constraints.

Evaluation

We focus on continuous review setting. In all problems assumptions must be made on the behaviors of demand occurrence, which are usually characterized by the time between demands and the distribution of demand size. As in [94] we assume the demand occur according to Poisson process with constant rate. The repair time is assumed exponentially distributed. The assumption is probably not very realistic, however Axsater [4] and Alfredsson and Verrijdt [2] showed that the service performance of the system is insensitive to the choice of the lead time distribution. With these assumptions the equilibrium probabilities for the number of items in stock and in replenishment may be found easily by modeling

the operational environment as a Markov chain, which is a random sequence in which the dependency of the successive events goes back only one unit in time. Since Poisson arrivals see time averages, the equilibrium probabilities may be used for determining service levels. The equilibrium distribution of a Markov chain may be computed exactly, by solving the linear system describing the long-run behavior of the Markov chain, or approximately. Examples of the latter that we will use in Chapter 4 are:

- First-order approximation, where the effective demand to each warehouse is described by the first moment of its distribution [4].
- Second-order approximation, which characterizes the effective demand just by its mean and variance [45].
- Third-order approximation, which takes into account the first three ordinary moments in computations [56], [61].

Depending on the model, extensions are possible to situations where the demand does not follow a Poisson process.
The power of Markovian modeling is high in inventory and many other applications. A general description of Markov chain theory is given in appendix 6.2. In [55], a lumpy demand per parts is considered. Kukreja and Schmidt derived analytical results for the mean and variance of the lead time demand at various locations in an inventory system with transshipments, and then used simulation methodology for determining inventory control policies for such a system.

Axsater [7] took into account a compound Poisson demand.

Queueing models are closely related to Markov processes. The results obtained in the analysis of queueing models may be useful as well. Typically, if the demands are considered as arrivals, and the items in replenishment are considered to be in service in the queueing model, the analogy is straightforward. For example Axsater [4] and therefore Alfredsson and Verrijdt [2] and Kutanoglu [57] uses the Erlang loss model. In this model, arrivals (demands) occur according to a Poisson process. The service time (replenishment time) can follow any distribution. Queueing models for spares inventory and repair capacity have been proposed in [40], [97], and [92]. Recently, also closed queueing network models have been proposed by Kranenburg and Van Houtum [50] In such networks,they let one of the stations represent the demand process, and another station represent the replenishment process. The closed queueing network representation has the following advantages.

- In the replenishment process other service disciplines can be modeled and analyzed as well.

- For the determination of the steady-state probabilities of inventory in system it's not necessary to assume Poisson demand.

However, at this moment, in case of transshipments only the case with equal demand rates can be analyzed. Its counterpart is a Markov process (the one described above) proposed in [91], where there are no restrictions on the base stock levels, and also, asymmetric demand can be dealt with. However, exact evaluation of this Markov process is only possible for a limited number of local warehouses, since the state space grows exponentially in the number of local warehouses.

In two echelon context the METRIC approximation is applied. The exact distribution of the number of units in the resupply system at each location in a two echelon depot base system is too computationally burdensome to be of practical use, refer to [63] for its computation. Hence the METRIC model is based on an approximation to this distribution that is easy to compute, and therefore has been widely used in many applications. Substantially under the METRIC approximation the number of backorders of each local warehouse is assumed to be Poisson distributed [72] and the improved two-moment approximations for that [75, 39].

Due to the complexities involved in the analytical modeling and solution of this kind of supply chain problems, some researchers have attempted heuristic approximations and / or simulation approaches, in effort to maintain some degree of realism in their analysis.

Needham and Evers [66] investigated the interaction of relevant costs and transshipment policies via simulation study and presented a method for determining a threshold value at which the benefits of transhipments outweigh their costs. They found that the cost of a stockout is the primary determinant in the transshipment decision, with lower stockout cost levels generally decreasing the likelihood of transshipment usage.

Ozdemir et al. [69] analyzed a capacitated transshipment problem. They modeled it as a network flow problem embedded in a stochastic optimization problem.They solved it by applying an Infinitesimal Perturbation Analysis (IPA),

which is a simulation based optimization technique.

Recently, Van Utterbeeck et al. [86] studied the effects of resupply flexibility on the design of service parts supply systems by applying simulation optimization. They investigated both the effect of the number of echelons and the degree of resupply flexibility, considering single echelon and two echelon systems with either no resupply flexibility, lateral transshipments only or both lateral transshipments and emergency deliveries. They showed that increased resupply flexibility enables increasingly important cost savings.

An interesting new stream of research is on systems under truly decentralized management. There, not only is the policy implemented locally, but also the policy variables are chosen by local managers, according to their own local incentives. See Cachon [13] for an interesting review.The cost allocation problem in spare parts inventory pooling is analyzed through the use of game theoretic models in [96].

Optimization

Setting stock levels for items managed using an (s-1,s) policy will depend on the objectives and constraints that are stipulated. For example, we could choose to minimize the average number of outstanding backorders across n item types subject to a constraint on investment in inventory. We could also select stock levels that minimize investment cost subject to an average fill rate constraint. Other optimization models could be formulated as well for complex resupply networks.
Specifically in this dissertation we focus on a stock allocation problem with transshipments. Typically, the latter problems for spare parts allocation with lateral transshipments, emergency shipments and base stock policy may be structured as integer programming problems with a non linear objective function and non linear constraints.
Therefore enumerative methods, such as branch and bound may solve them.

Branch and bound, proposed by Land and Doig [58] for linear programming problems, consists of a systematic enumeration of all candidate solutions, where large subsets of fruitless candidates are discarded en masse, by using upper and lower estimated bounds of the quantity being optimized. However such methods may be time consuming for real life problem instances.

A branch and bound procedure has been applied in spare part context by

Cohen [22].

Therefore, to find a feasible solution for the optimization problem heuristic algorithms may be used. Fast and easy to implement methods are the greedy methods, also known as marginal analysis methods. Greedy methods iteratively select at each step the single spare part allocation which is the most promising in objective function minimization until the feasibility check is satisfied.

For example several heuristic procedures can be found in the literature for allocating spares to warehouses in a single echelon context with complete pooling [4], [55]. In particular, for the problem of spares allocation Wong et al. [93] has shown that the greedy-type heuristic has the best performance. Also the landmark METRIC approach proposed by Sherbrooke [74] in a two echelon context solves the optimization problem through a marginal analysis algorithm.

However we want here examine solution methods that may be employed in general problems.

One solution approach that may be used is to construct a Lagrangian relaxation of a particular optimization problem, which finds a lower bound to the optimization problem. We begin by solving the resulting relaxed problem for a given set of Lagrange multiplier values. We then adjust these multiplier values and re-solve the relaxed problem. We continue in this manner until a stopping criterion of some sort is satisfied. An important observation on Lagrange relaxation is due to Everett [31]. In fact, Everett's theorem gives a relationship between the solution of a Lagrange relaxation and an optimization problem where only one constraint $g(x)$ is relaxed, when x is a vector, and both the objective and the constraint are convex functions. For short we have.

$$\begin{array}{ll} \min & f(x) \\ s.t.: & \\ & g(x) \leq b \\ & x \in S \end{array} \tag{2.4}$$

where S is a set of vectors that constraints the choice of an optimal solution. Its Lagrangian relaxation is

$$min_{x \in S}[f(x) + \theta(g(x) - b)] \tag{2.5}$$

for a given scalar $\theta \geq 0$.
is the Lagrangian multiplier associated with the constraint $g(x) \leq b$.

Theorem 4 *Suppose to have an optimization problem of the form of Problem 2.4 where only one constraint $g(x)$ is relaxed, when x is a vector, and both the objective and the constraint are convex functions.*
Suppose Problem 2.5 is its Lagrangian relaxation.
Suppose finally that $x^0(\theta)$ is an optimal solution to Problem 2.5 with the Lagrangian multiplier set to θ. Let $b^{'} = g(x^0(\theta))$.
Then $x^0(\theta)$ also solves

$$\begin{array}{ll} \min & f(x) \\ s.t.: & \\ & g(x) \leq b^{'} \end{array} \tag{2.6}$$

The proof is given in Muckstadt [63]. Thus by varying the value of θ, we can find optimal solutions to problems of the form of Problem 2.6. If $b^{'} = b$ for some choice of the Lagrange multiplier θ, then we have also solved Problem 2.4. For short for every choice of θ there exists a corresponding value of $b^{'}$ In general by investigating an appropriate range of values for the Lagrangian multipliers we can provide good lower bounds, if not necessarily optimal solutions to the optimization problem we want to solve. Besides a lower bound, a feasible solution is desidered as well. Having obtained the lower bound, often a feasible solution can be obtained by making judicious observations on it. For this kind of approximate methods Fisher [35] proposes the name of Lagrangian heuristic. A strength of Lagrangian relaxation is that it provides us with a lowerbound on the optimal cost, which can be used to evaluate the accuracy of some heuristics. It generally requires more computational effort than the greedy algorithm and the quality of the solution obtained by a Lagrangian heuristic is not necessarily better than a solution obtained by the greedy method. Optimization problems with space, capacity and time constraints are appropriate to be analyzed by Lagrangian relaxation [91].

Chapter 3

Spares allocation problem: an exact evaluation

With the multi-dimensional Markovian approach [91], the behavior of the inventory system is modeled with a Markov chain. Studying the Markov chain allows to compute structural properties as well as the state probabilities of the chain, which allows to evaluate the performance of the inventory system. This model becomes impractical for large instances, due to the extremely large number of states of the Markov chain.

3.1. The model

The model addressed in this chapter is a single item, single echelon, N-locations, continuous review, one-for-one replenishment policy inventory system, which allows for lateral transshipments with complete pooling, emergency transshipments from an external supplier and no negligible transfer times. A deterministic *closest neighbor rule* is used for lateral transshipment.

Main notation

In order to formally define the problem, let us introduce the following notation.

Let $A = \{1, 2, \ldots, a\}$ be a set of operational sites (e.g., airports) where working equipments are located.
We assume that operational sites are grouped on a regional basis, with a ware-

house of spare parts for each regions. Let $W = \{1, 2, \ldots, w\}$ be the set of regional warehouses.

Let s_i be the number of spare parts to allocate to each warehouse $i \in W$, $S = \sum_{i \in W} s_i$ be the total stock level and $s = (s_1, ..., s_w)$ be an allocation of spares to warehouses, i.e., the vector of decision variables.

We also denote with $MTTR$ the mean time to return, i.e. the average replenishment time of the external supplier, with $MTBF$ the mean time between failures, with OS the order and ship time, with $MCMT$ the mean corrective maintenance time and with OA the operational availability of all the a sites.

Let μ be the service rate of a server at any warehouse h. We assume that $\frac{1}{\mu}$ is given by the sum of the repair time (MTTR) and the time needed for ordering the ready for use part, sending the failed item and shipping the ready for use one (OS). Let T_{hi} be the transfer time for a spare from warehouse h to warehouse i and $T_s(j, h)$ be the substitution time, i.e., the time needed to transfer a spare part to the site $j \in A$ from the warehouse $h \in W$ and to physically replace the failed item, and T_{0i} be the mean emergency replenishment time from the external supplier to warehouse i, taking into account also the time needed to issue an order and the transfer time.

Let λ_{jh} be the rate of failure processes from site j to warehouse h and $\lambda_h = \sum_{j \in A} \lambda_{jh}$ be the arrival rate at warehouse h.

Given an allocation s, the network blocking probability is the probability that a failure occurs at some site and no warehouse can satisfy the spare demand. We show in equation (5) that the network blocking probability only depends on the total stock level S rather than on the particular allocation s, and denote it as $P_B(S)$.

Given an allocation s, let $\pi_{hi}(s)$ be the probability of the event: there are no spares in warehouse $h \in W$ and $i \in W$ is the closest warehouse with available spares (i.e., every warehouse l such that $T_{hl} < T_{hi}$, including the case $l = h$, is in stockout condition).

Let $n = (n_1, \ldots, n_w, n_{w+1})$ be a vector representing the state of the network, in which n_i is the number of outstanding requests at warehouse $i \in W$, and n_{w+1} is the number of outstanding emergency requests to the external supplier.

Let $p(n)$ be the probability of being in state n for the whole warehouses network.

Let p_{hi} be the probability of re-forwarding a spare requests from warehouse h, in stockout condition, towards warehouse i. It is a static probability reflecting the closeness between two warehouses.
Let j be a state for the whole warehouses network. Let $p^h(j)$ be the steady state probability that there are exactly j_h spares available at warehouse h. Specifically denote $p^i(s)$ as the marginal probability of having s_i outstanding orders at warehouse i, i.e., the probability of having no stock available at warehouse i:

$$p^i(s) = \sum_{n:n_i=s_i} p(n)$$

Let c^h be the inventory holding cost for warehouse h, c^t_{ij} be the cost for a lateral transshipment from warehouse j to warehouse i, in stockout condition, and c^e be the emergency transshipment cost.

System processes and assumptions

When a failure occurs for some component at some airport j, a demand for a new spare part is issued to the associated regional warehouse h. If spare parts are locally available, the component is immediately replaced in the airport using the stock on hand at the local warehouse. Then, the failed component is sent to an external supplier, which can either repair or replace the component with a new item, so that warehouse h can restore the local stock level for that specific component, after a replenishment time. If no spare part is locally available, warehouse h forwards the request to the nearest warehouse i with available spares to satisfy the demand through a lateral transshipment. Then, warehouse i will issue a replenishment order to the external supplier to restore its stock level. If no spare is available in any warehouse the demand must be satisfied directly by the external supplier through an emergency transshipment, i.e., by using the first repaired/new component available at the supplier. In such a case we say that the warehouses network is blocked, since the failed equipment will not be working at airport j until after the substitution. Since the replenishment time from the supplier to a warehouse can range up to several months for expensive components, in order to guarantee the high operational availability required by contract with airport authorities, the blocking probability must be kept at a very low level.

In our model we use the Poisson distribution for the demand process, which is a typical assumption for modeling low demand processes [82]. It is worthwhile to mention that the MTBF of an equipment depends on exogenous agents, such as the damp, the temperature and other operational conditions. Therefore, in our model we use specific values for each airport.
The replenishment time of the external supplier is a random variable, exponentially distributed, with known mean value $T_{0i} = \frac{1}{\mu}$ for $i \in W$. By contract its mean value is the same for all warehouses and it is equal to the sum of the mean time to return (MTTR) and the order and ship time (OS).
The capacity of the supplier repair shop is assumed to be infinite. It follows that also the number of replenishments from the external supplier follows the Poissonian distribution.

These common assumptions make possible to use the Markovian analysis for modeling the multi-dimensional inventory system.
However, as shown by [54], the results obtained with these assumptions holds under less restrictive hypothesis on the replenishment distribution.
We may describe the system dynamics with a suitable queueing network.
Each warehouse acts as a single queueing system without buffer, in which the number of servers equals the number of spares in the warehouse. The arrival process (i.e. demands for spare parts) is stochastic. The service time of each request equals the time needed to repair/replenish a spare part. Therefore, the number of busy servers corresponds to the number of outstanding orders of spare parts.
In this way each warehouse is a G/M/s_i/0/∞ system.

The operational availability OA is defined as in [74]:

$$OA = \frac{MTBF}{MTBF + MCMT} \tag{3.1}$$

This is the performance measure established by contract between the logistic company and the airport authorities.
MCMT is the average time occurring from the failure of an item to its physical substitution. This is the substitution time if the spare is available at the regional warehouse. If no spares are locally available, the request is forwarded to the closest warehouse with available spares and MCMT increases by the deterministic transfer time between the two warehouses. When no warehouse has spares available, MCMT equals the substitution time plus the replenishment

time from the external supplier.

The MCMT can be therefore computed as follows:

$$\begin{array}{rl} MCMT = & \sum_{h\in W}\sum_{j\in A}\lambda_{jh}T_s(j,h)+ \\ & \sum_{h\in W}(\lambda_h\sum_{i\in W}\pi_{hi}(s)T_{ih})+ \\ & (\sum_{h\in W}\lambda_h)P_B(S)T_{0h} \end{array} \tag{3.2}$$

We observe that the first term $\sum_{h\in W}\sum_{j=1}^{A}\lambda_{jh}T_s(j,h)$ of Equation (3.2) only depends on the failure process and on the distance between the sites and their respective regional warehouses. In other words, it does not depend on the specific spare parts management policy being used. Moreover, this quantity is typically small with respect to the other terms of Equation (3.2), therefore, for sake of simplicity we assume it negligible in our model and omit its computation in the rest of this chapter.

As for the quantity $\pi_{hi}(s)$, we assume that a strict deterministic nearest chosen neighbor rule is adopted for sourcing a lateral transshipment, as in Kukreja [54]. Differently form [54], we use equation (4.2) to compute this value, which by assuming independence directly follows from the definition of $\pi_{hi}(s)$ and therefore may be used as an approximation for it.

$$\pi_{hi}(s) = (1-p^i(s))\prod_{l:T_{hl}<T_{hi}}(p^l(s)). \tag{3.3}$$

3.2. Multi-dimensional Markovian approach

We model the system under study with a queueing network with blocking, and study it by using a Markov chain model, with a very similar approach to that of Wong et al. [91]. The main difference is that we explicitly include the external supplier in the Markov chain while in [91] a failure of a part occurring when all warehouses are in stockout condition is lost.

Theoretical considerations on Markov chains will be given in Appendix 6.2, where some background material is given. In the Markov chain, a state $n = (n_1,\ldots,n_w,n_{w+1})$ is a vector, in which n_i is the number of outstanding requests at warehouse $i \in W$, and n_{w+1} is the number of outstanding emergency transshipments issued from all warehouses to the external supplier. Note that the overall number of outstanding requests is $\sum_{i=1,\ldots,w+1} n_i$. In case of blocked

network, if $n_{w+1} \geq 1$ the first repaired item returned by the external supplier is used for replacing a failed item at some operative site.

There are direct transitions among states just in case of a single arrival event (i.e., a request for a spare at some warehouse) or a single departure event (i.e., the replenishment of a repaired item by the external supplier). Let e^i be a vector with $w+1$ elements, all equal to 0 but the element in position i that is equal to 1, and let $\psi(h,i)$ be equal to 1 if i is the warehouse closest to h with spares available, and be equal to 0 otherwise. More precisely, $\psi(h,i) = 1$ if $n_i < s_i$ and $n_l = s_l$ for each $l \in W$ such that $T_{lh} < T_{ih}$, included $l = h$. With this notation, $n + e_i$ is the state of the Markov chain representing an arrival at the i-th warehouse (with $n_i < s_i$), due either to a failure in the i-th service region or to a re-forwarded request from some other warehouse h in stockout conditions for which $\psi(h,i) = 1$. Similarly, $n - e_i$ is the state with a departure from the i-th warehouse (if $n_i > 0$). For the external supplier, $n + e_{w+1}$ represents a new emergency request (if $n_i = s_i$ for each $i \in W$) and $n - e_{w+1}$ represents the fulfillment of an emergency request (if $n_{w+1} > 0$). The transition rate $q(n,m)$ from state n towards state $m = n \pm e_i$ and $n \pm e_{w+1}$ is as follows.

- $q(n, n+e_i) = \lambda_i + \sum_{h \in W - \{i\}} \psi(h,i) \cdot \lambda_h$, for $i \in W$ and $n_i = 0, 1, \ldots, s_i - 1$;
- $q(n, n + e_{w+1}) = \sum_{i \in W} \lambda_i$, if $n_i = s_i \ \forall i \in W$;
- $q(n, n - e_i)_i \cdot \mu$, for $i \in W$ and $n_i > 0$ and $n_{w+1} = 0$;
- $q(n, n - e_{w+1}) = \sum_{i=1}^{w+1} n_i \cdot \mu$, for $n_i = s_i \ \forall i \in W$ and $n_{w+1} \geq 1$.

Figure 4.1(left) shows an example of a Markov chain for two warehouses, the first having two spares and the second having three available spares. Theorem 5 shows that the blocking probability P_B of the Markov chain can be easily computed. Let $S = \sum_{i \in W} s_i$ be the total stock level in the network, and let $\rho = \frac{\sum_{i \in W} \lambda_i}{\mu}$.

Theorem 5 *Given a set W of warehouses, with total stock level S, in which the service process is exponentially distributed with average rate μ for each server and the demand flow to warehouse $i \in W$ is Poissonian with average rate λ_i, the blocking probability of all warehouses is $P_B(S) = 1 - \sum_{k=0}^{S-1} \frac{\rho^k}{k!} e^{-\rho}$.*

Proof. Let us consider a cut in the Markov chain grouping all the states n such that $\sum_{h=1}^{w+1} n_h = k$ (in Figure 4.1(left) is highlighted the case for $k = 2$).

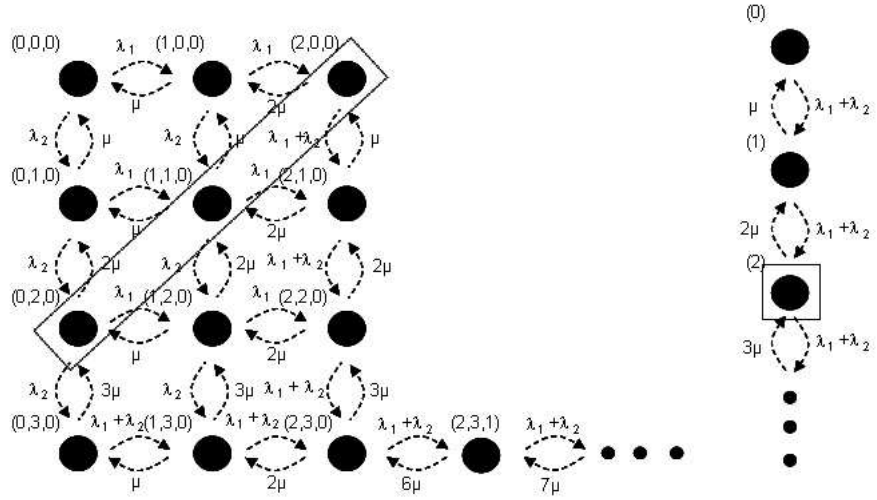

Figure 3.1: A Markov chain (left) and the aggregated birth death model (right)

For each value $k = 0, 1, \ldots, \infty$ the state aggregation property described in [67, 47] applies, and the states contained in each cut can be replaced with an aggregated state k. The demand rate for each aggregated state $k = 1, \ldots, \infty$ is $\sum_{i \in W} \lambda_i$ and the service rate is $k\mu$, as in figure 4.1(b). The network is therefore equivalent to a virtual single warehouse with combined stock level $S = \sum_{i \in W} s_i$, demand rate $\sum_{i \in W} \lambda_i$ and service rate for each server μ. The overall stockout probability $P_B(S)$ is the probability that the total number of requests is greater or equal to the total number of spares available, i.e., $P_B(S) = \sum_{k=S}^{\infty} p_k = 1 - \sum_{k=0}^{S-1} p_k$, where p_k is the probability of state k in a queue $M/M/S$, i.e. $p_k = \frac{\rho^k}{k!} \cdot e^{-\rho}$. The blocking probability of all warehouses is therefore

$$P_B(S) = 1 - \sum_{k=0}^{S-1} \frac{\rho^k}{k!} e^{-\rho}. \tag{3.4}$$

□

Unfortunately, this result does not allow to compute the OA of the system, since to this aim the marginal blocking probability of each warehouse is necessary. However, steady state probabilities can be computed for each state in the Markov chain by solving a linear system. To this aim, Theorem 5 can be

used to reduce the infinite state space Markov chain to an equivalent one with a finite number of states. Specifically, all the states in which all warehouses are in stockout condition can be replaced with a single state, with probability P_B and with suitable modified departure transition rates.

Let n^B be the state such that $n_i^B = s_i$ for each $i \in W$ and $n_{w+1}^B = 0$. Let $p_{n^B} = \frac{\rho^{n^B}}{n^B!} \cdot e^{-\rho}$ be the probability of state n^B and let $q(n^B, n^B - e_i) = s_i\mu$ the departure transition rates from state n^B to state $n^B - e_i$. Let us now replace all states such that $n_i^B = s_i$ for each $i \in W$ and $n_{w+1}^B \geq 0$ with a single state $\hat{n}^B$. To achieve the equivalence with the original Markov chain it is sufficient to set the departure transition rates from state $\hat{n}^B$ to state $\hat{n}^B - e_i$ equal to $q(\hat{n}^B, \hat{n}^B - e_i) = s_i\mu F$ for each $i \in W$, where the factor F is equal to

$$F = \frac{p(n^B)}{P_B(S)} = \frac{\frac{\rho^{n^B}}{n^B!} \cdot e^{-\rho}}{1 - \sum_{k=0}^{S-1} \frac{\rho^k}{k!} e^{-\rho}}$$

For instance, Figure 3.2(left) shows a Markov chain with infinite number of states and Figure 3.2(right) shows its equivalent Markov chain with a finite number of states. In general, the number of states in the finite state space Markov chain is equal to

$$\prod_{i \in W} (s_i + 1). \tag{3.5}$$

This number can be exceedingly large as the number of warehouses and spares increases. Therefore, there is a need for approximate methods to compute the OA of large networks.

Remark

The above Markovian model is very similar to that of Wong et al. [94]. The main difference is that we assume that the emergency shipments enter a queue at the external supplier, together with the replenishment orders and therefore we explicitly include the external supplier in the Markov chain, while in [94] a failure of a part occurring when all warehouses are in stockout condition is lost. In our case the resulting overall blocking probability is strictly greater and therefore more conservative. The last fact is important in our applicative context, where the operational availability requirements are strict. The following holds. Let 0 be a $w + 1$ dimensional vector representing the state of the network, in which 0 is the number of outstanding requests at any warehouse $i \in W$, and 0 is the number of outstanding emergency requests to the external

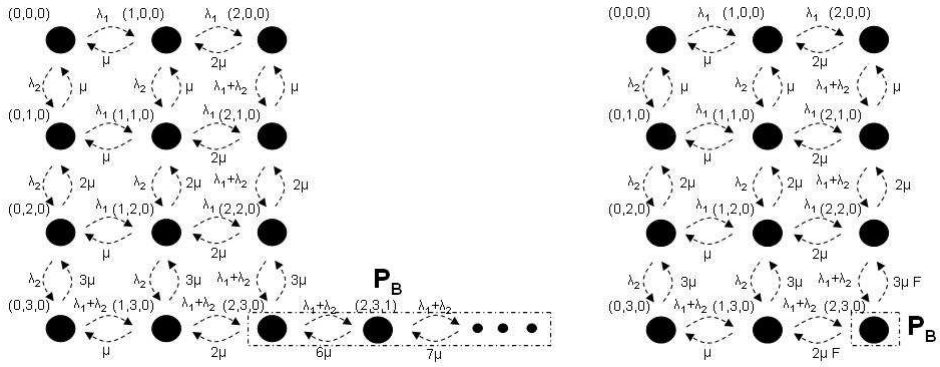

Figure 3.2: A Markov chain with infinite (left) and finite (right) number of states.

supplier. Let $p(0)$ and $\hat{p}(0)$ be the probability of being in state 0 for the whole warehouses network respectively in our model and in the model of Wong et al. [94]. Finally, given an allocation s with total stock level S, denote as $P_B(S)$ and $\hat{P}_B(S)$ the network blocking probability, which is the probability that a failure occurs at some site and no warehouse can satisfy the spare demand, respectively in our model and in the model of Wong et al. [94].

$$p(0) = \frac{1}{\sum_{k=0}^{\infty} \frac{\rho^k}{k!}} = e^{-\rho} < \hat{p}(0) = \frac{1}{1+\frac{\rho}{1!}+\frac{\rho^2}{2!}+\frac{\rho^3}{3!}+...+\frac{\rho^S}{S!}}$$

$$\frac{\rho^k}{k!}e^{-\rho} < \frac{\rho^k}{k!}\frac{1}{1+\frac{\rho}{1!}+\frac{\rho^2}{2!}+\frac{\rho^3}{3!}+...+\frac{\rho^S}{S!}}$$

$$P_B(S) = 1 - \sum_{k=0}^{S-1} \frac{\rho^k}{k!}e^{-\rho} > \hat{P}_B(S) = 1 - \sum_{k=0}^{S-1} \frac{\rho^k}{k!}\frac{1}{1+\frac{\rho}{1!}+\frac{\rho^2}{2!}+\frac{\rho^3}{3!}+...+\frac{\rho^S}{S!}} \tag{3.6}$$

3.3. General methods for the computation of the state probabilities of a Markov chain

In what follows we will give a sketch of how computationally or analytically burdensome can be computing the state probabilities of a Markov chain. We

will describe in brief some general, common techniques useful to state probabilities computation, pointing our attention on methods for finite state Markov chains and infinite state Markov chains.

Methods for a finite state Markov chain

In general there are two methods to solve the Markov chain equations.

- Direct methods.
- Iterative methods.

To discuss these methods, let us assume that the states of the Markov chain are numbered as $1, \ldots, N$.

Direct methods

A convenient direct method is a Gaussian elimination method such as the Gauss-Jordan method. This method in general eliminates the first variable present in the first equation from all equations below the first equation, and then eliminate a second variable present in the second linear equation from all equations below and so on up to the last variable and last equation. This will put the system into triangular form. Then, using back-substitution, each unknown can be solved for. This reliable method is recommended as long as the dimension N of the system of linear equations does not exceed the order of thousands. The computational effort of Gaussian elimination method is proportional to N^3. Reliable and ready to use methods for Gaussian elimination are widely available. A Gaussian elimination method requires that the whole coefficient matrix is stored, since this matrix must be updated at each step of the algorithm. This explains why a Gaussian elimination method suffers from computer memory problems when N gets large.

Iterative method of successive overrelaxation

Iterative methods have to be used when the size of the system of linear equations gets large. In specific applications an iterative method can usually avoid computer memory problems by exploiting the sparse structure of the application. An iterative method does not update the matrix of coefficient each time. In applications these coefficients are usually composed from a few constants. Then only these constants have to be stored in memory when using an iterative

method. In addition to the advantage that the coefficient matrix need not be stored, an iterative method is easy to program for specific applications.
The iterative method of successive overrelaxation is a suitable method for solving the linear equations of large Markov chains. The well known Gauss Siedel method is a case of the method of successive overrelaxation The iterative methods generate a sequence of vectors converging towards a solution of the equilibrium equations. The normalization is done at the end of the calculations. The methods starts with an initial approximation vector and the procedure is generally continued until the changes made by an iteration are below some tolerance. The Gauss siedel method is convergent in all practical cases but the convergence speed may be very high. The ordering of the states may also have a considerable effect on the convergence speed of the successive overrelaxation algorithm. In general Tijms [84] suggests to order the states such that the upper diagonal part of the matrix of coefficients is as sparse as possible. In specific applications the transition structure of the Markov chain often suggests an appropriate ordering of the states. There are then others methods, such as the Krylov iteration method and the recursive method, which are strongly dependent of the structure of the system of linear equations to be solved and are typically a matter of experimentation.

Methods for an infinite state Markov chain

It is shown that brute-force truncation is not necessary to get a finite system of linear equations when the state space I and the state probabilities exhibit a geometric tail behavior in the infinite-state model. For this situation, which naturally arises in many applications, an elegant computational method for the state probabilities can be given. Markov chains with a multidimensional state space are prevalent in stochastic networks and in such applications it often happens that the equilibrium probabilities are known up to a multiplicative constant. If the number of states is too large for a direct computation of the multiplicative constant, the Metropolis-Hastings algorithm may be used to obtain the equilibrium probabilities.

Geometric tail approach for an infinite state space

Many applications of Markov chains involve an infinite state space. What one usually does to solve numerically the infinite set of equilibrium equations is to approximate the infinite-state Markov model by a truncated model with finitely many states so that the probability mass of the deleted states is very

small. Indeed, for a finite-state truncation with a sufficiently large number of states, the difference between the two models will be negligible from a computational point of view. However, such a truncation often leads to a finite but very large system of linear equations whose numerical solution will be quite time-consuming, although an arsenal of good methods is available to solve the equilibrium equations of a finite Markov chain. Moreover, it is somewhat disconcerting that we need a brute-force approximation to solve the infinite-state model numerically. Fortunately, many applications allow for a much simpler and more satisfactory approach to solving the infinite set of state equations. Under rather general conditions the state probabilities exhibit a geometric tail behavior that can be exploited to reduce the infinite system of state equations to a finite set of linear equations. The geometric tail approach results in a finite system of linear equations whose size is usually much smaller than the size of the finite system obtained from a brute-force truncation. As an example, consider a Markov chain whose state space is one dimensional and is given by

$$I = \{0, 1, \ldots\}$$

Let us assume that the equilibrium probabilities $p(n)$, $n \in I$, exhibit the geometric tail behavior $\mathrm{p(n)} \sim \gamma\eta^n$ as $n \to \infty$ for some constant $\gamma > 0$ and $0 < \eta < 1$. Below conditions under which 3.3 holds will be discussed. First we demonstrate how the geometric tail behavior can be exploited to reduce the infinite system of state equations to a finite system of linear equations. It will be seen below that the decay factor η in 3.3 can usually be computed beforehand by solving a non-linear equation in a single variable. Solving a non-linear equation in a single variable is standard fare in numerical analysis. In most applications it is not possible to compute the constant γ beforehand. Fortunately, we do not need the constant γ in our approach. In fact, for sufficiently large integer M,

$$p(n) \sim p(M)\eta^{n-M}$$

, with $n \geq M$. Replacing $p(n)$ by $p(M)\eta^{n-M}$ for $n \geq M$ in equilibrium equations for the steady state behavior of the Markov chain leads to a finite set of M linear equations for the remaining M state probabilities and 1 linear equation for the normalization condition. How large an M should be chosen has to be determined experimentally and depends, of course, on the required accuracy in the calculated values of the equilibrium probabilities. However, Tijms [84] assert that empirical investigations show that in specific applications remarkably small values of M are already good enough for practical purposes.

Conditions for the geometric tail behavior

A useful but technical condition fro the existence of a geometric tail approximation can be given in terms of the generating function $\sum_{n=0}^{\infty} p(n)z^n$ of the equilibrium probabilities $p(n)$, details may be found in Tijms [84]. Such a condition seems not to have a probabilistic interpretation. Next we give a probabilistic condition from the existence of a geometric tail behavior in the Markov chain. This condition is in terms of the one step transition probabilities of the Markov chain. It is as follows:

- There is an integer $r \geq 0$ such that the one step transition probabilities between state i and state j depends on i and j only through $j - i$ when $i \geq r$ and $j \geq 1$.
- There is a n integer $s \geq 1$, such that the one step transition probabilities between states i and j are equal to 0 for $j > i + s$ and $i \geq 0$.
- Letting α_{j-i} denote the one step transition probability between states i and j for $i \geq r$ and $1 \leq j \leq i + s$, the constant α_k satisfy $\alpha_s > 0$ and $\sum_{k=-\infty}^{s} k\alpha_k < 0$.

Under this condition the equilibrium equations for $p(j)$ for $j \geq r + s$ are homogeneous linear difference equations with constant coefficients. A method to solve such a linear difference equation is the method of particular solutions. Substituting a solution of the form $p(n) = w^n$ in the equilibrium equations for $p(n)$ with $n \geq r + s$, we find the so-called characteristic equation, which has s roots.

Metropolis - Hastings algorithm

The Metropolis-Hasting algorithm is a method for constructing a Markov chain with a given limiting probability distribution. In the context of stochastic networks, the Markov chains have generally multidimensional state space. However the number of possible states is soon very large so direct calculation of the normalization constant is not piratically feasible. This raises the following question. Suppose that N positive numbers each of the form $\overline{p}(n_i)$ for $i = 1, \ldots, N$, where n_i are specific states, have finite sum $S = \sum_{i=1^N} \overline{p}(n_i)$. How do we construct a Markov chain whose equilibrium probabilities are given by $\frac{\overline{p}(n_i)}{S}$ for $i = 1, \ldots, N$? To answer the question we need the concept of a reversible Markov chain. Let $\{X_n\}$ be a Markov chain with a finite state space and one step transition probabilities t_{ij}. It is assumed that $\{X_n\}$ has

no two disjoint closed sets. Then the Markov chain has a unique equilibrium distribution $\{p(j)\}$. Assume now that a non null vector (g_j), $j \in I$ exists such that

$$g_j t_{jk} = g_k t_{kj} \tag{3.7}$$

and $j, k \in I$. Then, for some constant $c \neq 0$,

$$g_j = cp(j). \tag{3.8}$$

The proof is simple. Fix $j \in I$ and sum both sides of 3.7 over k. This gives

$$g_j = \sum_{k \in I} g_k p_{kj}$$

, with $j \in I$. This gives

$$g_j = \sum_{k \in I} g_k t_{kj} \tag{3.9}$$

with $j \in I$. These equations are exactly the equilibrium equations of the Markov chain $\{X_n\}$. Therefore

$$p(j)t_{jk} = p(k)t_{kj}$$

with $j, k \in I$.
A Markov chain having this property is called a reversible Markov chain.
An excellent description of such Markov chains may be found in [47] and [67].
The property states that the long run average number of transitions from state j to state k per time unit is equal to the long run average number of transitions from state k to state j far all $j, k \in I$.

Let us return to the problem of constructing a Markov chain with equilibrium probabilities $\{p(j) = \frac{\overline{p}(j)}{S}, j = 1, \ldots, N\}$ when N positive numbers $\overline{p}(n_i)$ are given and have finite sum S.
To do so, choose any Markov matrix $M = (m_{ij})$, with $i, j = 1, \ldots, N$ with positive elements m_{ij}. Next construct a Markov chain $\{X_n\}$ with state space $I = \{1, \ldots, N\}$ and one step transition probabilities

$$t_{ij} = \begin{array}{l} m_{ij}\alpha_{ij}, j \neq i \\ \\ m_{ii}\alpha_{ii} + \sum_{k=1}^{N} m_{ik}(1 - \alpha_{ik}), j = i \end{array} \tag{3.10}$$

where $_{ij}$ are appropriately chosen numbers between 0 and 1 with $\alpha_{ii} = 1$ for $i = 1, \ldots, N$.

The state transition of the Markov chain are governed by the following rule: if the current state of the Markov chain $\{X_n\}$ is i, then a candidate state k is generated according to the probability distribution $\{m_{ij}, j = 1, \ldots, N\}$. The next state of the Markov chain $\{X_n\}$ is chosen equal to the candidate state k with probability α_{ik} and is chosen equal to the current state i with probability $1 - \alpha_{ik}$. By an appropriate choice of the $_{ij}$, we have

$$\overline{p}(j)t_{jk} = \overline{p}(k)t_{kj} \tag{3.11}$$

with $j, k = 1, \ldots, N$, implying that the Markov chain has the equilibrium distribution

$$p(j) = \frac{\overline{p}(j)}{\sum_{k=1}^{N} \overline{p}(k)} \tag{3.12}$$

for $j = 1, \ldots, N$. The expression 3.11 holds for the choice

$$\alpha_{ij} = \min(\frac{\overline{p}(j)m_{ji}}{\overline{p}(i)m_{ij}}, 1) \tag{3.13}$$

for $i, j = 1, \ldots, N$ Summarizing, the Metropolis-Hastings algorithm generate a sequence of successive states of a Markov chain $\{X_n\}$, whose equilibrium distribution is given by 3.12.

3.4. Computational experience

In this section we describe our computational experience on 60 practical instances from the Italian airport maintenance context plus other 990 randomly generated instances. The experiments are carried out by varying the mean demand to each warehouse, the number of warehouses and the stock levels. The departure transition rate μ of each server is fixed equal to $\mu = \frac{1}{3 \text{ months}}$ for all instances. This is the value used by the managers in the practical application.

All the experiments are executed on a PC equipped with a processor Intel Core2 Duo CPU (3 GHz), 3.25 GB Ram and Windows operating system.

Random instances are classified according to the average arrival rates from all airports $\bar{\lambda} = \sum_{i \in W} \lambda_i$ and to the stock levels s_i, $i \in W$. Instances with $\bar{\lambda} = 0.001$ are in the *low* demand class, instances with $\bar{\lambda} = 0.01$ are in the *medium* demand class, and instances with $\bar{\lambda} = 0.1$ are in the *high* demand class. As for the stocking policy, for each demand class we generate 17 *distributed* instances, in which $s_i = 1 \ \forall i \in W$ and $w = 1, 2, \ldots, 17$. Other 169

centralized instances are obtained by storing all spares in a single warehouse $j \in W$ that acts as a central depot, i.e., $s_i = 0$ for $i \in W - \{j\}$ and $s_j = S$ for $S = 1, 2, \ldots, 10$ and for $w = 1, 2, \ldots, 17$ (excluding the case $S = w = 1$, which is included in the previous group). Finally, 144 *hybrid* instances are obtained by setting $s_i = 1$ for $i \in W - \{j\}$ and $s_j \in \{2, 3, \ldots, 10\}$, for $w = 2, \ldots, 17$. For each instance, the subset of warehouses, as well as the central depot j in the latter two groups of instances, is chosen at random among the 17 regional warehouses of the practical application. Overall, there are 17+169+144 random instances for each demand class, for a total of 990 random instances, besides the 60 practical instances.

In practical data experiments, the exact computation with the Markov chain model fails in finding a solution in 18 out of 60 practical instances, due to memory limits. As for the random instances, the exact computation fails in about 30% of the cases. Table 4.1 shows some practical instances for which the Markov chain approach fails in finding the exact OA value. For each item we show the cost in euro, the replenishment time (months) and the MTBF (hours) of each item. The last three columns report the number of warehouses where at least one item is installed, the total number of installed items and the total number of allocated spares.

As far as the multi-dimensional scaling down approach is concerned, the computer used in our experiments can easily manage Markov chains with 2^{14} states.

The smallest instances that cannot be solved with the Markov chain model, in terms of the quantity $\prod_{i \in W}(s_i + 1)$, have 12 warehouses and 19 allocated spares for the practical instances, and 12 warehouses and 13 allocated spares for the random instances. However, we observe that also the arrival rate Λ affects the computation time.

The time and memory effort needed to solve the Markov chain increase rapidly with the number of warehouses and installed items. In figure 4.5 we show the computation time required to solve the Markov chain model for the 17 distributed instances and for the three levels of demand $\bar{\lambda}$. Instances with low demand require significantly lower computation time. Moreover, the Markov chain model fails in computing the exact solution in one case for $\bar{\lambda} = 0.001$, in three cases for $\bar{\lambda} = 0.01$ and in four cases for $\bar{\lambda} = 0.1$. Figure 4.6 shows the memory effort required to solve the same instances with the Markov chain model.

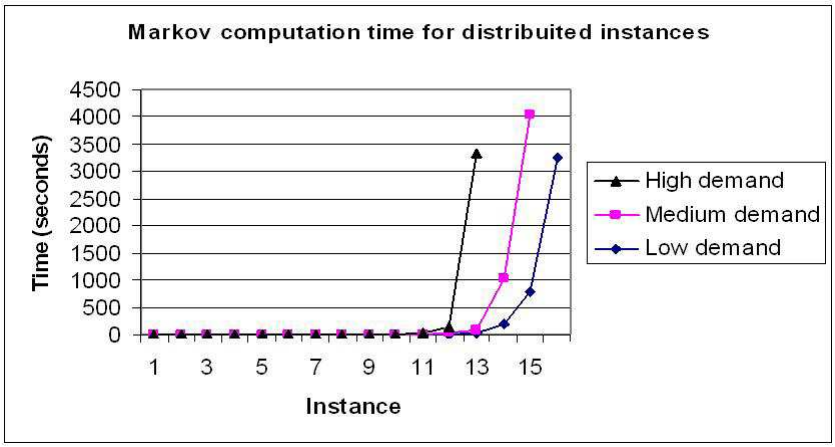

Figure 3.3: Computation time for the Markov chain model and distributed instances.

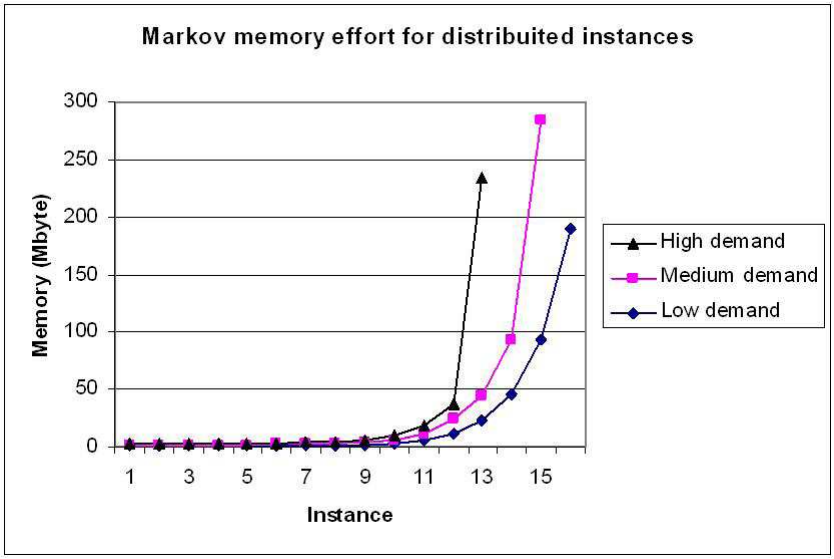

Figure 3.4: Memory effort for the Markov chain model and distributed instances.

Item	Cost (euro)	Repair time (months)	MTBF (hours)	num. warehouses	installed items	$\sum_{i \in W} s_i$
1	$\simeq 7000$	3	76000	12	25	25
2	$\simeq 6000$	3	12000	12	19	29
3	$\simeq 1800$	3	45000	14	20	20
4	$\simeq 3000$	3	109000	14	25	17
5	$\simeq 3000$	3	26000	12	20	24

Table 3.1: Two non solvable instances.

3.5. Markov chain structure: a remark

In this section, we present shortly an optimization model used to prove that a product form of the state probabilities of the Markov chain model describing the behavior of the system does not exist. To this aim we show two numerical examples. By using a suitable optimization model we have shown that the Markov chain cannot be decomposed exactly in product form. In fact, the best product form approximation returns a positive accuracy error, which implies that an exact product form does not exist.

After a theoretical insight in such a task, by using classical Markov chain theory [84] and advanced one [67], [16], it is clear that a known product form for such a network had not been found.

A question arises about the decomposability of the equilibrium probabilities in product form.

To this aim note that we refer to the Markov chain model with finite state space, introduced in Section 4.2, e.g. for a pictorial example refer to Figure 3.2(right).

Let N be the number of warehouses in the maintenance network. Let us use for spare parts management a single echelon policy with complete pooling and with the presence of the external supplier managing the unfilled requests. Let n be a w dimensional vector representing a possible state, so that each element n_i of n, with $i = 1, \ldots, w$, records the number of outstanding requests at each warehouse i.
Let us fix one state n, by solving the linear system describing the equilibrium

behavior of the system, we may find the steady state probability of state n, denote it as $p(n)$. Denote as $\hat{p}_i(n_i)$ a value, which should represent the probability that warehouse i is in state n_i, without care about the state of the other warehouses. Finally, denote as $\hat{p}(n)$ the product of the w $\hat{p}_i(n_i)$ values, when n is fixed.

We ask the following: is it possible to find for each warehouse i and any possible state n_i some numbers, let us call them $\hat{p}_i(n_i)$, so that for any possible state the following holds: $\hat{p}(n) = p(n)$?

Otherwise we look for those $\hat{p}_i(n_i)$ values either minimizing the sum of the square errors between $\hat{p}(n)$ and $p(n)$ or minimizing the maximum square error among the corresponding $\hat{p}(n)$ and $p(n)$.
To this aim we have implemented a MatLab routine for computing both the marginal state probabilities, which minimize the maximum square error with respect to the joint ones (min-max program), refer to the model 3.14, and the marginal state probabilities, which minimize the sum of the square errors with respect to the joint ones (non linear program), refer to the model 3.15.

Denote M as the number of states in the state space and order the states so that each state is matched with a single integer number between 1 and M. Denote as $s = (s_i), i = 1, \ldots, N$ a given stock allocation. Recall $\hat{p}(n) = \hat{p}_1(n_1)\hat{p}_2(n_2)\ldots\hat{p}_w(n_w)$.
Minimization of the maximum square error between $\hat{p}(n)$ and $p(n)$:

$$\begin{array}{lll} \min_{p_i(n_i)} & \max & (p(n) - \hat{p}(n))^2 \\ s.t.: & & \\ & & \sum_{n_i=0}^{s_i} \hat{p}_i(n_i) = 1 \\ & & \\ & & 0 < \hat{p}_i(n_i) < 1 \end{array} \tag{3.14}$$

Minimization of the sum of the square errors between $\hat{p}(n)$ and $p(n)$:

$$\begin{array}{ll} \min_{p_i(n_i)} & \sum_{n=1}^{M} (p(n) - \hat{p}(n))^2 \\ s.t.: & \\ & \sum_{n_i=0}^{s_i} \hat{p}_i(n_i) = 1 \\ & \\ & 0 < \hat{p}_i(n_i) < 1 \end{array} \tag{3.15}$$

In both cases constraints exists for computing the values $\hat{p}_i(n_i)$ representing probabilities: they must be non negative and less than 1 and such that the

joint state probabilities computed through their product sum to 1.

For what concerns the constrained non linear programming the MatLab engine implements methods for large scale optimization and medium scale optimization. The large scale algorithm is a subspace trust region method and is based on the interior-reflective Newton method. While, the medium scale optimization is solved through a sequential quadratic programming (SQP) method. In this method, the function solves a quadratic programming (QP) subproblem at each iteration. An estimate of the Hessian of the Lagrangian is updated at each iteration.

For what concerns the min-max program the MatLab engine implements a sequential quadratic programming (SQP) method, too.

In both cases tests have been carried on simple instances have shown that the best product form approximation returns a positive accuracy error, which implies that an exact product form does not exist.

The characteristic data for two tests are shown in Table 3.5. Let as refer

Parameter name	Unit	Values
Warehouses with positive demand		4,5
Average MTBF	hours	11000, 60000
Average MTTR+OS	hours	2160, 2160
Allocation	vector	[1,1,1,1],[1,1,1,1,1]

Table 3.2: Parameter values for the computational experiment

to the two tests respectively as test A and test B. By solving the models 3.14 and 3.15 we have computed the following accuracy errors.

test	Minimum error on the sum of square errors	Minimum maximum square error
test A	0.0017	$3.2\ 10^{-5}$
test B	$6.1\ 10^{-7}$	$1.8\ 10^{-7}$

Table 3.3: Numerical results

3.6. The optimization model

In this section we define formally the optimization problem we aim to solve. In our model, a logistic company aims to compute the stock level s_i of each warehouse $i \in W$ such that a minimum level of service is granted at the operational sites and the overall cost is minimum. Costs are related to inventory holding, transshipments and emergency shipments.

Given an allocation s of spares to warehouses, the model used to compute the level of service is a single item, single echelon, w-locations, continuous review, one-for-one replenishment policy inventory system, whit lateral and emergency shipments, complete pooling and non-negligible transshipment times.

The Spares Allocation Problem is the problem of finding an allocation s which minimizes the overall cost for inventory holding, lateral and emergency shipments, subject to a constraint on the minimum operational availability of the system.
The contractual service level to grant is the operational availability OA of all operational sites for each item, computed as in [74], and expressed in equation 3.1.

We assume the Poisson distribution for the demand process, which is a typical assumption for modeling low demand processes [82]. We also use location dependent MTBF values. The replenishment time of the external supplier is a random variable, exponentially distributed, with known mean value $T_{0j} = MTTR+OS$, which is the same for any warehouse j. The capacity of the supplier repair shop is assumed to be infinite. It follows that also the number of replenishment from the external supplier follows the Poissonian distribution. These common assumptions make possible to use the Markovian analysis for modeling the multi-dimensional inventory system. However, as shown by [54], the results obtained with these assumptions holds under less restrictive hypothesis on the replenishment distribution.

Finally, the following assumptions are made for optimization purposes.

1. Lateral transshipment is always more convenient than emergency shipment, i.e., the time and cost needed for a transshipment from warehouse i to warehouse j is always smaller than the time and cost required for an

emergency shipment from warehouse j:

$$\max_{i,j \in W} \{T_{ij}\} < T_{0j} \tag{3.16}$$

$$\max_{i,j \in W} \{c^t_{ij}\} < c^e \tag{3.17}$$

2. The cost for a lateral transshipment from warehouse i to warehouse j increases linearly with the transfer time T_{ij}, i.e.,

$$c^t_{ij} = \alpha T_{ij} \tag{3.18}$$

Lateral transshipments are made only when a location experiences a demand with no on-hand stock.

Therefore, our optimization model is as follows.
Let L be the minimum operational availability level to be achieved by a feasible allocation. It is easy to check that this quantity corresponds to allowing a maximum waiting time $\frac{(1-L)MTBF}{L}$ to substitute failed items. Then, the Spares Allocation Problem P_0 can be formulated as the following integer program with non-convex objective function:

Problem P_0:

$$\begin{array}{ll} \min & \sum_{i=1}^{w} c^h s_i + \lambda_i \sum_{j \in W} \pi_{ij}(s) c^t_{ji} + \lambda_i P_B(S) c^e \\ s.t.: & \\ & \sum_{i=1}^{w} [\lambda_i \sum_{j \in W} \pi_{ij}(s) T_{ji} + \lambda_i P_B(S) T_{0i}] \leq \frac{(1-L)MTBF}{L} \end{array} \tag{3.19}$$

Let $f_1(S) = \sum_{i=1}^{w} c^h s_i$ be the total inventory holding cost, the cost for lateral transshipments be defined by $f_2(s) = \left\{ \sum_{i=1}^{w} \lambda_i \sum_{j \in W} \pi_{ij}(s) c^t_{ji} \right\}$ and let $f_3(S) = \sum_{i=1}^{w} \lambda_i P_B(S) c^e$ be the cost for the emergency shipments. Similarly, for the waiting times we let $t_2(s) = \left\{ \sum_{i=1}^{w} \lambda_i \sum_{j \in W} \pi_{ij}(s) T_{ji} \right\}$ be the waiting times due to lateral transshipments and $t_3(S) = \sum_{i=1}^{w} \lambda_i P_B(S) T_{0i}$ be the emergency waiting times.

The above model is with integer-valued decision variables and non-linear objective and constraints.
No polynomial time optimization algorithm exists for our type of problems.

The problems under consideration in this dissertation could also be considered as a complex type of knapsack problems: nonlinear knapsack problems with multiple constraints, where more than one copy of each item can be selected. For a general description of knapsack problems, see e.g. Kellerer et al. [46]. Kellerer et al. [46] prove that even the simplest type of knapsack problem belongs to the class of NP-hard problems.

Therefore, no other optimization procedures than enumerative methods exists for solving it optimally.
Branch and bound is such a technique, which should be more efficient with respect a total enumeration algorithm.
Otherwise, approximation algorithms may be applied for looking for feasible and hopefully good solutions.

Chapter 4

Lateral transshipment: approximate performance models

The model addressed in this chapter is a single item, single echelon, N-locations, continuous review, one-for-one replenishment policy inventory system, which allows for lateral transshipments with complete pooling, emergency transshipments from an external supplier and no negligible transfer times. A deterministic *closest neighbor rule* is used for lateral transshipment.

Specifically, in this chapter we compare approximation techniques for computing the operational availability of a practical corrective maintenance system in charge of the maintenance of 38 Italian Airports for varying the stock level and the demand.

A drawback of this policy is the state dependent nature of re-forwarding in the systems, which does not allow to express the state probabilities of the associated Markov chain model in product form. Therefore, computing the state probabilities is not practical as the number of states in the Markov chain increases.

We adapt four approximation techniques to our model and evaluate their performance in terms of computational effort, memory requirement and error with respect to the exact value. Three techniques approximate state probabilities with others that can be expressed in product form, so that the Markov chain can be decomposed. Specifically, we adapt a method by Alfredsson and Verrijdt, the

Equivalent Random Traffic (ERT) method and the Interrupted Pisson Process (IPP) method. The fourth technique is based on the multi-dimensional scaling down approach, which studies an equivalent reduced Markov chain rather than decomposing the original one.

4.1. Literature review

The literature on spare part logistics with lateral transshipments is strictly related to the more general context of inventory management. Most contributions focus on the analysis of different inventory management models. Several authors [2, 38] demonstrated the benefits of inventory sharing flexibility provided by complete pooling policies. An extensive overview of the research concerning transshipment modeling in supply chain systems is given by [20]. Kennedy et al. review the different modeling issues in spare part management [48]. Here we limit ourselves to present the foremost works on techniques for assessing the performance of single-echelon systems, applicable in continuous review policies with complete pooling.

As observed, e.g. in [92, 48], at least two main streams of research can be distinguished for approaching the modeling tasks, namely the multi-dimensional Markovian approach and the decomposition approach. A third stream of research that can be cited is based on simulation.

With the multi-dimensional Markovian approach [91], the behavior of the inventory system is modeled with a Markov chain. Studying the Markov chain allows to compute structural properties as well as the state probabilities of the chain, which allows to evaluate the performance of the inventory system. This model becomes impractical for large instances, due to the extremely large number of states of the Markov chain. In order to overcome this drawback, one possibility is to study an equivalent Markov chain with a smaller number of states, even if this approach did not receive much attention in the literature. Axsater [6] suggests and evaluates a similar technique in a two echelon context with continuous review policy and lumpy demand at each warehouse. With the *scaling down* approach of Axsater, a high-demand system is approximated by a low-demand system. The real customer demand is scaled down such that the ratio between standard deviation and mean value is preserved. According to the author, the scaling down technique is quite effective to speed up the analysis of single queueing systems. However, to the best of our knowledge, little work has been done to assess the effectiveness of this technique in the context of multi-dimensional Markov chains.

A second stream of research is based on an approximate decomposition approach [2]. This approach consists of estimating the state probabilities of the Markov chain, rather than computing their exact values, by studying each queueing system independently from each other, so that the system performance can be easily computed. The basic idea is to adjust the demand flow, so that the lateral transshipments are taken into account. Within this stream of research, Alfredsson and Verrijdt [2] use an iterative method proposed by Axsater [4] in a two echelon context to compute fraction of demand satisfied by different sources by assuming exponentially distributed replenishment times and Poissonian demand. The AV method allows each warehouse to share inventory with every other warehouse, so that all warehouses act as a single big pool. Specifically, lateral transshipment is used when no spare is locally available and the request is directed toward a randomly chosen closest neighbor warehouse with spares available. An external supplier manages the requests that cannot be filled by other local warehouses or by the central warehouse. The demand and the stock level may differ from one facility to another. The numerical experiments show that the approximate results are very close to the simulation results for low demand rate, while the error may increase remarkably with the demand rate.

The Interrupted Poisson Process (IPP) [56, 61] and the Equivalent Random Traffic (ERT) [45] methods are decomposition approaches used in particular in the design of telecommunication networks to assess the blocking probability of a network [44]. With these methods there are no external entities, such as the external supplier, and requests arriving, when all service centers are busy, are lost. Multiple re-forwardings are possible, by letting requests jumping among the service centers more than one time. With both methods a lateral transshipment from a warehouse (i.e., a re-forwarded request) is viewed as an overflow from the demand arriving at the warehouse, and therefore with variance larger than the mean value. The *peakedness* of a distribution is the ratio between variance and mean value, which is equal to one for the Poisson process and it is greater than one when dealing with the overflow process at a queue with Poissonian demand and exponential service time.

There are significant differences among the AV, ERT and IPP methods. Besides the presence/absence of the external supplier, a further difference is that AV method models the effective demand at each warehouse as a Poissonian flow, which is therefore described by the first moment of its distribution. The ERT method characterizes the effective demand just by its mean and variance assuming a peakedness greater than one, while the IPP method models the effective demand at each warehouse as hyperexponential, and takes into account

its first three ordinary moments in computations.

As observed, e.g. by [18, 17], reliable models of service systems with overflow should take the peakedness into account. In order to explore the potential benefits of including peakedness in the model, we compare two methods, the first assuming Poissonian demand and the second assuming demand with peakedness greater than one.

A third, quite different, stream of research is based on simulation (see, e.g. [82, 55]). In this case the inventory system is modeled as a discrete event system, whose evolution allows to evaluate the original system behavior. Simulation models allows to easily incorporate all relevant practical details of the system, but particular care is necessary to guarantee the statistical relevance of the results achieved. On the other hand, they may require very long computational times in low demand contexts in order to achieve reliable results. Rare event techniques, such as the importance sampling policy [41], can be used in such cases to reduce the simulation times, but the technique can still remain very time consuming [14]. Therefore, such models are not further explored in this chapter.

4.2. Multi-dimensional Markovian approach

We model the system under study with a queueing network with blocking, and study it by using a Markov chain model, with a very similar approach to that of Wong et al. [91]. The main difference is that we explicitly include the external supplier in the Markov chain while in [91] a failure of a part occurring when all warehouses are in stockout condition is lost.

In the Markov chain, a state $n = (n_1, \ldots, n_w, n_{w+1})$ is a vector, in which n_i is the number of outstanding requests at warehouse $i \in W$, and n_{w+1} is the number of outstanding emergency transshipments issued from all warehouses to the external supplier. Note that the overall number of outstanding requests is $\sum_{i=1,\ldots,w+1} n_i$. In case of blocked network, if $n_{w+1} \geq 1$ the first repaired item returned by the external supplier is used for replacing a failed item at some operative site.

There are direct transitions among states just in case of a single arrival event (i.e., a request for a spare at some warehouse) or a single departure event (i.e., the replenishment of a repaired item by the external supplier). Let e^i be a vector with $w+1$ elements, all equal to 0 but the element in position i that is equal to 1, and let $\psi(h, i)$ be equal to 1 if i is the warehouse closest to h with spares available, and be equal to 0 otherwise. More precisely, $\psi(h, i) = 1$

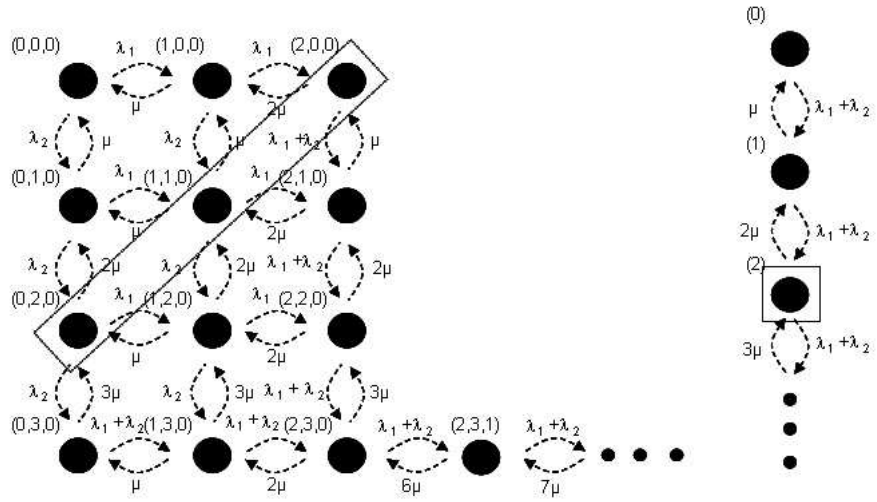

Figure 4.1: A Markov chain (left) and the aggregated birth death model (right)

if $n_i < s_i$ and $n_l = s_l$ for each $l \in W$ such that $T_{lh} < T_{ih}$, included $l = h$. With this notation, $n + e_i$ is the state of the Markov chain representing an arrival at the i-th warehouse (with $n_i < s_i$), due either to a failure in the i-th service region or to a re-forwarded request from some other warehouse h in stockout conditions for which $\psi(h, i) = 1$. Similarly, $n - e_i$ is the state with a departure from the i-th warehouse (if $n_i > 0$). For the external supplier, $n + e_{w+1}$ represents a new emergency request (if $n_i = s_i$ for each $i \in W$) and $n - e_{w+1}$ represents the fulfillment of an emergency request (if $n_{w+1} > 0$). The transition rate $q(n, m)$ from state n towards state $m = n \pm e_i$ and $n \pm e_{w+1}$ is as follows.

- $q(n, n+e_i) = \lambda_i + \sum_{h \in W - \{i\}} \psi(h, i) \cdot \lambda_h$, for $i \in W$ and $n_i = 0, 1, \ldots, s_i - 1$;
- $q(n, n + e_{w+1}) = \sum_{i \in W} \lambda_i$, if $n_i = s_i \ \forall i \in W$;
- $q(n, n - e_i) = n_i \cdot \mu$, for $i \in W$ and $n_i > 0$ and $n_{w+1} = 0$;
- $q(n, n - e_{w+1}) = \sum_{i=1}^{w+1} n_i \cdot \mu$, for $n_i = s_i \ \forall i \in W$ and $n_{w+1} \geq 1$.

Figure 4.1(left) shows an example of a Markov chain for two warehouses, the first having two spares and the second having three available spares. Theorem

5 shows that the blocking probability P_B of the Markov chain can be easily computed. Let $S = \sum_{i \in W} s_i$ be the total stock level in the network, and let $\rho = \frac{\sum_{i \in W} \lambda_i}{\mu}$.

4.3. Approximate performance computation

In this section we describe four methods for estimating the OA for the single echelon model with complete pooling. The first three methods are based on decomposition. w independent single-dimensional queueing systems approximate the multi-dimensional original one. The decomposition approach is an exact solution method when the steady state probabilities can be expressed in product form. Unfortunately, the Markov chain model studied in this chapter cannot be expressed in product form. However, as described in [9], product form networks provide the basis for many approximate algorithms to solve more general non product form ones. In this section we describe three decomposition methods. The first adapts the AV method of Alfredsson and Verrijdt [2] to the case of deterministic re-forwarding [54]. The AV method models the demand at each queueing system with a Poissonian independent distribution with adjusted demand rate. The second method assumes non-Poissonian independent demand distributions with adjusted demand rate at each queueing system. It is based on IPP [56, 61] and ERT [45] methods. The third method is IPP itself, which assumes as effective demand process a simple renewal process, the Interrupted Poisson Process, with adjusted demand rate at each queueing system. According to Iversen [44], IPP and ERT methods are particularly suitable to model the peakedness of overflow processes.

The fourth method is based on the scaling down concept [6]. We apply this concept to the multi-dimensional single echelon with complete pooling context.

Decomposition approach

The basic ideas of the decomposition methods studied in this chapter consists of computing the fraction of the demand λ_i at warehouse i that is satisfied by one of three different sources. The first fraction β_i (the *local fill rate*) is directly satisfied by the stock available at warehouse i, the second fraction α_i (the *transshipment fraction*) is satisfied through lateral transshipments from the other warehouses, the third fraction is satisfied by the external supplier through emergency shipments and it is equal to the joint *blocking probability* $P_B(S)$ of all warehouses, computed as in Theorem 5. Figure 4.2 shows a pictorial

representation of the three flows of spare parts which satisfy the three demand fractions at the first service region of Figure 1.1(right).

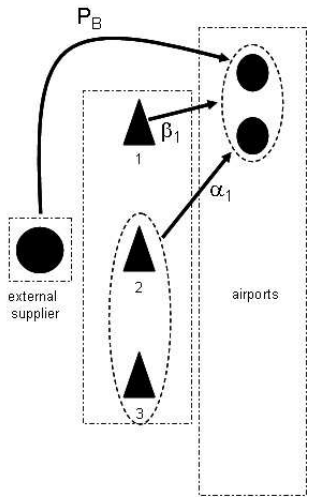

Figure 4.2: The three fractions α_1, β_1 and $P_B(S)$ of demand at warehouse 1.

With a decomposition method, each warehouse is studied separately. To this aim, let $p^i(s)$ denote the probability of having s_i outstanding orders at warehouse i, i.e., the probability of having no stock available at warehouse i. The local fill rate β_i and the transshipment fraction α_i are therefore:

$$\begin{aligned} \beta_i &= (1 - p^i(s)) \\ \alpha_i &= 1 - \beta_i - P_B(S) \end{aligned} \tag{4.1}$$

The value $p^i(s)$ for each warehouse can be computed only if the effective arrival rates $\lambda_i^{'}$ at each warehouse $i \in W$ are known. To compute the latter values, we let o_i be the overflow from all other warehouses that is re-forwarded to warehouse i. The closest neighbor sourcing rule for lateral transshipment is taken into account by the probability $\pi_{hi}(s)$, defined in Section 3.1 as probability of having stock available at warehouse i and having no stock available at every warehouse l such that $T_{hl} < T_{hi}$ (including the case $l = h$):

$$\pi_{hi}(s) = \beta_i \cdot \prod_{l:T_{hl}<T_{hi}} (1 - \beta_l). \tag{4.2}$$

Equation 4.2 approximates $\pi_{hi}(s)$, as explained in section 3.1. If values $\pi_{hi}(s)$ are known, then the values o_i and $\lambda_i^{'}$ can be computed as follows:

$$\begin{aligned} o_i &= \sum_{h=1,h\neq i}^{w} \pi_{hi}(s) \cdot \lambda_h \\ \lambda_i^{'} &= \lambda_i + o_i. \end{aligned} \tag{4.3}$$

This expressions are similar to those used in AV method by Alfredsson and Verrijdt [2], with the difference that we use the closest neighbor sourcing rule for lateral transshipments instead of the random sourcing rule used in [2]. Similar modification is applied by Kukreja [54].

In order to compute the above quantities, we still need the values $p^i(s)$, $\forall i \in W$. The two decomposition methods analyzed in this chapter differs in the approach used to estimate these values.

AV method

With the AV method [2], the demand flow at each warehouse is considered Poissonian. Therefore, the steady state probabilities of having j outstanding requests at warehouse i are the same that in a Markovian queueing system with s_i servers and zero buffer. In such a case $p^i(s)$ may be computed as follows.

$$\begin{aligned} p^i(0) &= \frac{1}{\sum_{j=0}^{s_i} \frac{(\lambda_i^{'})^j}{\mu^j \cdot j!}} \\ p^i(s) &= \frac{\frac{(\lambda_i^{'})^{s_i}}{\mu^{s_i}}}{s_i!} \cdot p^i(0) \end{aligned} \tag{4.4}$$

In order to compute α_i and β_i in steady state an iterative procedure is followed. The iterative procedure starts with $\beta_i = 1 - P_B(S)$ and $\alpha_i = 0$, which implies that o_i is initially zero for all $i \in W$. Then, at each iteration, quantities $\pi_{hi}(s)$ are computed with equation 4.2 while quantities o_i and $\lambda_i^{'}$ are computed with equations 4.3 and then used to update steady-state probabilities $p^i(s)$. In the next iteration the values of β_i, α_i are recalculated and used to update the other quantities. This procedure is repeated until the β_i, α_i and o_i values do not change anymore. These values converge after a few iterations (usually less then 30 in our computational experiments), as experienced also by Axsater [4], Alfredsson and Verrijdt [2] and Kutanoglu [57].

Modified ERT method

With the ERT method [2], demand at each warehouse is only characterized by its mean value and its variance. The basic idea is that the peaked demand at warehouse i can be viewed as the overflow of another queue $M/M/k_i/0/\infty$ with k_i servers, Poissonian demand and exponential service process. Therefore, this method models warehouse i with a Markovian queueing system with a number of servers $k_i + s_i$, Poissonian demand flow with average A_i and zero buffer. The first k_i servers act as a primary queue, whose peaked overflow is sent to a secondary queue with s_i servers. Figure 4.3 is a pictorial representation of the ERT basic idea. The quantities k_i and $A_i$must be determined in order to model the desired peaked effective demand at the secondary queue, with average $\lambda_i^{'}$ and variance $v_i^{'}$. In such a case, the values $p^i(s)$ of Equation (4.4)

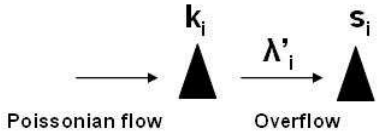

Figure 4.3: Equivalent system

are computed as the ratio between the overflow of the queue with $k_i + s_i$ servers and Poissonian demand A_i and the effective demand λ_i':

$$\frac{A_i \cdot E_{k_i+s_i}(A_i)}{\lambda_i^{'}} \tag{4.5}$$

In order to compute the values $p^i(s)$ we therefore need to compute A_i and k_i values. To this aim, we compute the mean of the effective demand as in equations 4.3 and the variance of the effective demand, $v_i^{'}$, as follows. Let v_i be the variance of the regular flows for warehouse i and Z be the peakedness factor.

$$v_i^{'} = v_i + \sum_{j=1,j\neq i}^{N} \pi_{ji}(s) \cdot v_j \tag{4.6}$$

Therefore we may solve the following equations 4.7 w.r.t. k_i and A_i:

$$\begin{aligned} \lambda_i^{'} &= A_i \cdot E_{k_i}(A_i) \\ \frac{v_i^{'}}{\lambda_i^{'}} &= Z = 1 - \lambda_i^{'} + \frac{A_i}{k_i + 1 - A_i + \lambda_i^{'}} \end{aligned} \tag{4.7}$$

This non-linear system has a unique solution [44], and we compute it by using the non linear equations methods of [11].

Differently from [2], we compute the values α_i and β_i in steady state with an iterative procedure, similar to the one used in the IPP method [56, 61]. The iterative procedure starts with $\beta_i = 1 - P_B(S)$ and $\alpha_i = 0$, which implies that o_i are initially zero. Then in each iteration $\lambda_i^{'}$ and $v_i^{'}$ are computed, as in equations 4.3 and 4.6, and used to compute steady-state blocking probabilities $p^i(s)$. The values of β_i, α_i are then recalculated, used as input to next iteration, and the whole procedure is repeated until the β_i, α_i and o_i do not change anymore. These values converge after a few iterations.

IPP method

With the IPP model, the demand at each warehouse, including transshipments, can be adequately characterized by a simple renewal process. The inter arrival time distribution of an IPP is hyperexponential, commonly used in the literature to model high-variability arrival processes [90]. Such a distribution has 4 parameters, in what follows we will denote them as $a_1, a_2, \gamma_1, \gamma_2$. Its mean is denoted as δ^i. In appendix 6.2 theoretical details of such a process are given, in this section we present just the relevant formulas for the computation of marginal service measures. Under the IPP hypothesis, the steady state probabilities of having s_i outstanding requests at warehouse i are the same that in a queueing system with s_i servers and zero buffer, where the blocking probability is computed through a generalized Erlang loss function. In such a case, assuming to have already estimated $a_1^i, a_2^i, \gamma_1^i, \gamma_2^i$, $p^i(s)$ may be computed as follows.

$$\begin{aligned}
&\phi(z) = \tfrac{a_1^i \gamma_1^i}{z+\gamma_1^i} + \tfrac{a_2^i \gamma_2^i}{z+\gamma_2^i} \\
&C_j(\xi) = \textstyle\prod_{k=1}^{j} \tfrac{\phi(k\mu+\xi)}{1-\phi(k\mu+\xi)} \\
&C_0 = 1 \\
&C_{-1} = 1 \\
&p^i(s) = \tfrac{1}{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!} \frac{1}{C_j(0)}}
\end{aligned} \tag{4.8}$$

In order to compute values $p_{s_i}^i$, we therefore need to compute $a_1^i, a_2^i, \gamma_1^i, \gamma_2^i$. By characterizing $a_1^i, a_2^i, \gamma_1^i, \gamma_2^i$ for each queueing system i, substantially the

effective demand process to each warehouse is estimated.
To this aim, we compute the first three moments of the effective demand as follows.
Let m_1^i, m_2^i, m_3^i be the first three moments of the random variable for the transshipments of the effective flows for warehouse $i$and let r_1^i, r_2^i, r_3^i be the first three moments of the regular demand flow. We have:

$$\begin{aligned} m_1^i &= \frac{r_1^i}{\mu} p^i(s) \\ m_2^i &= m_1^i + \frac{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!} \frac{1}{C_{j-1}(\mu)}}{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!} \frac{1}{C_j(\mu)}} m_1^i \\ m_3^i &= 3m_2^i - 2m_1^i + 2(m_2^i - m_1^i) \frac{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!} \frac{1}{C_{j-1}(2\mu)}}{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!} \frac{1}{C_j(2\mu)}} \end{aligned} \tag{4.9}$$

In the above expressions the $C_j(\xi)$ are computed as in A.36, but this time the $\phi(z)$ are always referred to the Laplace transform of the inter arrival times of the regular demand and not of the effective one. We use this modification for taking into account the presence of the external supplier.
Therefore, under independence assumption the first three moments of the effective demand to each warehouse are computed by applying the multinomial theorem, thus aggregating transshipped and regular flows to each warehouse i.
Let us denote the m_k^{ij} the k-th moment of the flow from i to j, we compute it as

$$m_k^i \pi_{ij}(s)$$

Such an expression depends on the presence of the external supplier, i.e. transshipments cannot be again re-forwarded. Substantially after using the multinomial theorem the mean effective demand is computed as in equation 4.3.

Finally the above first three moments of the effective flows are used for estimating $a_1^i, a_2^i, \gamma_1^i, \gamma_2^i$ for each warehouse i. In fact the characteristics of the effective demand flow, which is assumed to be an Interrupted Poisson Process, are completely specified by the Laplace transform of the inter arrival distribution [56], therefore the parametric moments may be matched with the numerics ones, just evaluated through the multinomial theorem.

We directly solve the non linear matching equations, which leads us to achieve more reliable results with respect to the approximate computation proposed in

[56]. To this aim we use a trust region Newton method for unconstrained nonlinear equations and an interior affine scaling down approach for constrained optimization problems [11]. In appendix 6.2 details on such a technique will be given.

In order to compute α_i and β_i in steady state an iterative procedure is followed. The iterative procedure starts with $\beta_i = 1 - P_B(S)$ and $\alpha_i = 0$, which implies that o_i is initially zero for all $i \in W$. Then, at each iteration, quantities $\pi_{hi}(s)$ are computed with equation 4.2 while quantities $a_1^i, a_2^i, \gamma_1^i, \gamma_2^i$ for each warehouse i are computed through moment matching and then used to update steady-state probabilities $p^i(s)$. In the next iteration the values of β_i, α_i are recalculated and used to update the other quantities. This procedure is repeated until the β_i, α_i and o_i values do not change anymore.

A sketch of the iterative algorithm is described in Figure 4.4.

Scaling down approach

With the scaling down approach, a system is approximated by scaling the demand, the replenishment time and the stock level of each warehouse using a *scale factor* K. The main purpose is to reduce the stock level at each warehouse $i \in W$ to a new value $\hat{s}_i$, in order to achieve a Markov chain with an affordable number of states, whose probabilities can be efficiently computed.

The intuition behind this method is that the performance levels of the original inventory system depend more on the ratios between demand, replenishment time and stock level than on their absolute values. Their relative sizes may not linearly influence the OA approximation goodness. There are two critical issues in the method. The first issue is the choice of the K, the second one is the rounding of the scaled stock levels, which clearly must be integer values. Axsater [6] chooses the scale factor K by keeping the same standard deviation-to-mean ratio in the scaled system as in the original one, while the rounding problem is not addressed since all parameters are assumed to be multiples of K. In our procedure we relax the latter assumption and choose a scale factor such that the scaled Markov chain can be solved efficiently.

Our procedure is as follows. Let MAX be the maximum number of states of the Markov chain that can be efficiently managed. In view of equation 3.5, the number of states that must be taken into account in the scaled Markov chain is equal to $\prod_{i \in W}(\hat{s}_i + 1)$, therefore we set the scale factor K as the minimum integer value such that:

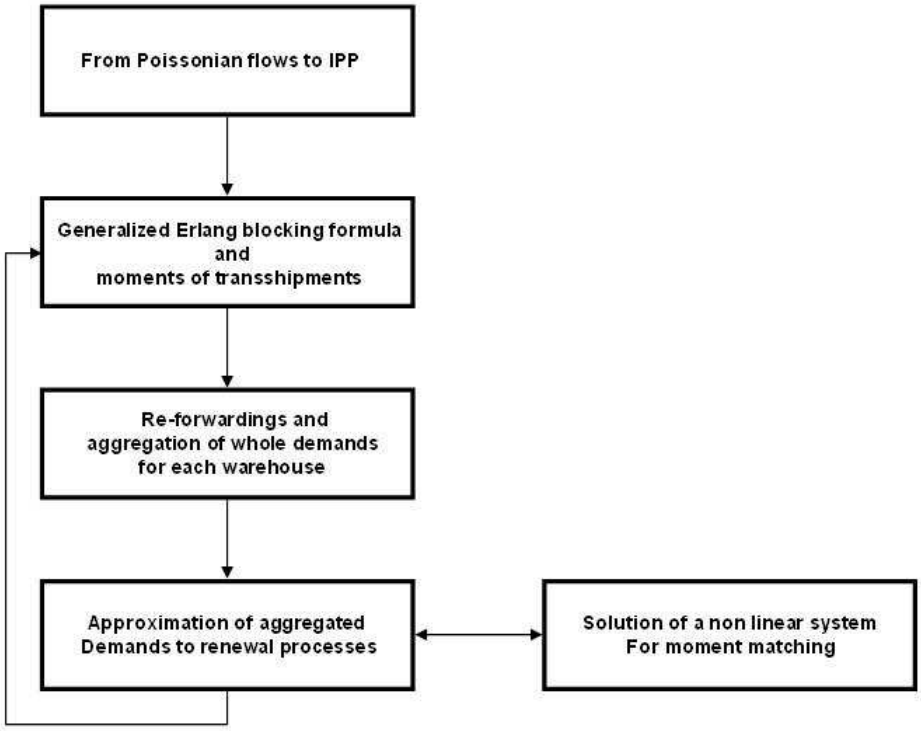

Figure 4.4: Iterative procedure for estimating steady state IPP parameters

$$\prod_{i \in W} (\hat{s}_i + 1) \leq MAX. \tag{4.10}$$

Specifically, K is obtained iteratively as follows. Starting from $K = 2$, we set the overall stock level in the reduced system equal to

$$\hat{S} = \lfloor \frac{\sum_{i \in W} s_i}{K} \rfloor + 1 \tag{4.11}$$

and then allocate a provisional number of spares $\lfloor \frac{s_i}{S} \cdot \hat{S} \rfloor$ to each warehouse $i \in W$, where S is the overall number of spares in the system. The remaining

number of spares $\hat{S}-\sum_{i\in W}\lfloor\frac{s_i}{S}\cdot\hat{S}\rfloor$ (smaller than w) is allocated by ordering the warehouses for decreasing value of $\frac{s_i\cdot\hat{S}}{S}-\lfloor\frac{s_i}{S}\cdot\hat{S}\rfloor$ and allocating an additional spare to the first warehouses until all spares are allocated, thus obtaining the values $\hat{s}_i$ associated to the given value of K. In case of tie, the spare is allocated with priority to the warehouse with higher demand. Then, inequality (4.10) is checked. If $\prod_{i\in W}(\hat{s}_i+1)\leq MAX$ holds, we set K and $\hat{s}_i$. Otherwise, we increase K and repeat the procedure until inequality (4.10) holds. Finally, the scaled $\hat{\lambda}_i$ and $\hat{\mu}$ are fixed as $\hat{\lambda}_i=\frac{\lambda_i}{K}$ and $\hat{\mu}=\frac{\mu}{K}$, respectively.

4.4. Numerical study

In this section we describe our computational experience. We compare the exact results in terms of OA values, computed by directly solving the Markov chain model, with the results obtained with the approximate techniques described in section 4.3. Besides the percentage error in terms of OA values between the approximate OA value and the Markov chain one, we report on the computation time and on the memory required by the different models. All the experiments are executed on a PC equipped with a processor Intel Core2 Duo CPU (3 GHz), 3.25 GB Ram and Windows operating system.

The set of instances used for our computational study is composed by 60 practical instances from the Italian airport maintenance context plus other 990 randomly generated instances. The experiments are carried out by varying the mean demand to each warehouse, the number of warehouses and the stock levels. The departure transition rate μ of each server is fixed equal to $\mu=\frac{1}{3\text{ months}}$ for all instances. This is the value used by the managers in the practical application.

Random instances are classified according to the average arrival rates from all airports $\bar{\lambda}=\sum_{i\in W}\lambda_i$ and to the stock levels s_i, $i\in W$. Instances with $\bar{\lambda}=0.001$ are in the *low* demand class, instances with $\bar{\lambda}=0.01$ are in the *medium* demand class, and instances with $\bar{\lambda}=0.1$ are in the *high* demand class. As for the stocking policy, for each demand class we generate 17 *distributed* instances, in which $s_i=1\ \forall i\in W$ and $w=1,2,\ldots,17$. Other 169 *centralized* instances are obtained by storing all spares in a single warehouse $j\in W$ that acts as a central depot, i.e., $s_i=0$ for $i\in W-\{j\}$ and $s_j=S$ for $S=1,2,\ldots,10$ and for $w=1,2,\ldots,17$ (excluding the case $S=w=1$, which is included in the previous group). Finally, 144 *hybrid* instances are obtained by setting $s_i=1$ for $i\in W-\{j\}$ and $s_j\in\{2,3,\ldots,10\}$, for $w=2,\ldots,17$. For each instance, the subset of warehouses, as well as the central depot j in the latter

two groups of instances, is chosen at random among the 17 regional warehouses of the practical application. Overall, there are 17+169+144 random instances for each demand class, for a total of 990 random instances, besides the 60 practical instances.

As far as the multi-dimensional scaling down approach is concerned, the computer used in our experiments can easily manage Markov chains with 2^{14} states. Using a value $MAX = 2^{14}$ in equation (4.10) results in a scale factor $K = 2$ for all the 1050 instances. Hence, we denote with $K2$ this case. In subsection 4.4 we evaluate the effect of the scale factor on the overall performance of the multi-dimensional scaling down approach.

Moreover, in what follows we do not show explicitly the results concerning the IPP method, because we found approximately the same results as the modified ERT method. IPP method is substantially more difficult than ERT method and more time consuming, for these reason only the modified ERT method will be taken into account for analyzing the influence on performance estimation of the investigated peaky nature of the flows in our system.

Time and memory effort

In practical data experiments, the exact computation with the Markov chain model fails in finding a solution in 18 out of 60 practical instances, due to memory limits. As for the random instances, the exact computation fails in about 30% of the cases. Table 4.1 shows some practical instances for which the Markov chain approach fails in finding the exact OA value. For each item we show the cost in euro, the replenishment time (months) and the MTBF (hours) of each item. The last three columns report the number of warehouses with at least one spare, the total number of installed items and the total number of allocated spares.

The smallest instances that cannot be solved with the Markov chain model, in terms of the quantity $\prod_{i \in W}(s_i + 1)$, have 12 warehouses and 19 allocated spares for the practical instances, and 12 warehouses and 13 allocated spares for the random instances. However, we observe that also the arrival rate $\bar{\lambda}$ affects the computation time.

The time and memory effort needed to solve the Markov chain increase rapidly with the number of warehouses and installed items. In figure 4.5 we show the computation time required to solve the Markov chain model for the 17 distributed instances and for the three levels of demand $\bar{\lambda}$. Instances with low demand require significantly lower computation time. Moreover, the Markov chain model fails in computing the exact solution in one case for $\bar{\lambda} = 0.001$,

Item	Cost (euro)	Repair time (months)	MTBF (hours)	num. warehouses	installed items	$\sum_{i \in W} s_i$
1	$\simeq 7000$	3	76000	12	25	25
2	$\simeq 6000$	3	12000	12	19	29
3	$\simeq 1800$	3	45000	14	20	20
4	$\simeq 3000$	3	109000	14	25	17
5	$\simeq 3000$	3	26000	12	20	24

Table 4.1: Five non solvable instances.

in three cases for $\bar{\lambda} = 0.01$ and in four cases for $\bar{\lambda} = 0.1$. Figure 4.6 shows the memory effort required to solve the same instances with the Markov chain model.

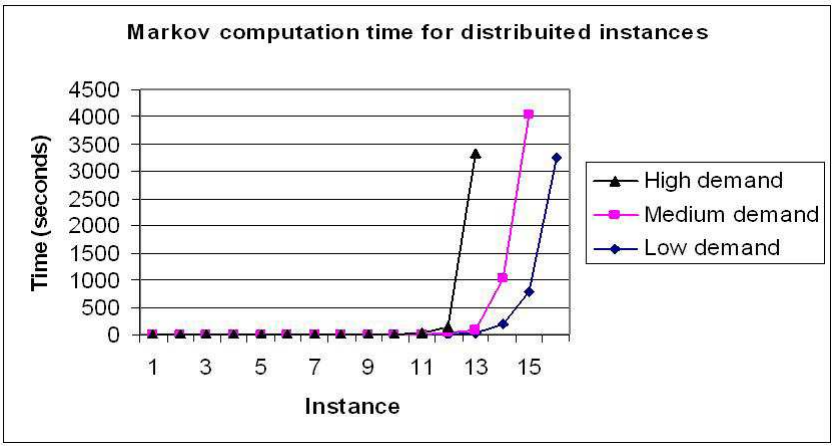

Figure 4.5: Computation time for the Markov chain model and distributed instances.

In figures 4.7 and 4.8 we compare the computation time and memory effort required by the three approximate models to solve the 60 practical instances. The instances are ordered for increasing number of warehouses and, if equality holds, for increasing number of spares. It can be observed that, even if all the three methods are quite efficient in computing a solution, the multi-dimensional scaling down approach is very fast with all instances. The maximum time

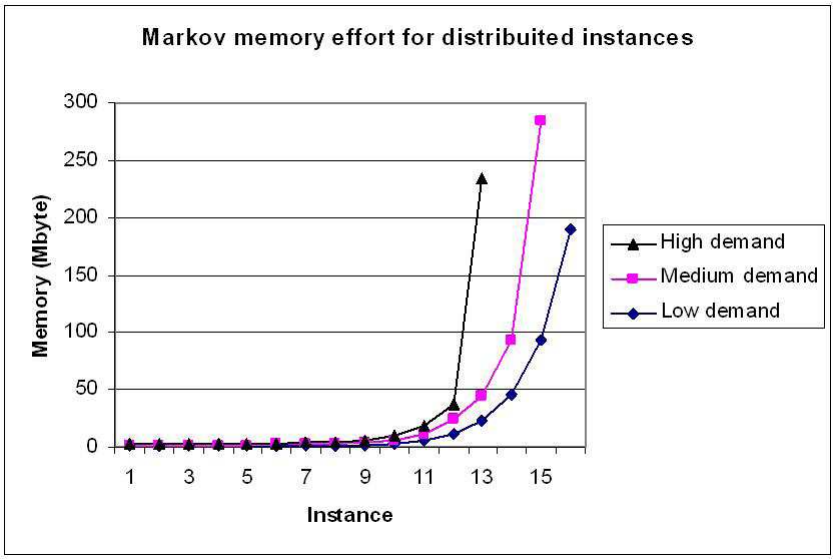

Figure 4.6: Memory effort for the Markov chain model and distributed instances.

required to solve a practical instance is 0.11 seconds. As expected, the ERT method is slightly more time consuming than the AV method, due to the need to solve a non-linear system instead of using a closed form expression as in the AV method. Similar behavior can be observed for the memory effort, shown in Figure 4.8.

In figure 4.9 we compare the computation time and memory effort required by the three approximate models to solve the 990 random instances.

In the figure we report the average computation time (respectively, the memory effort) required to solve all the instances with the same demand $\bar{\lambda}$, the same number of warehouses w and the same stock level S. The computation times of the three approximate methods increase with $\bar{\lambda}$ and w. With the two decomposition methods AV and ERT, the computation time decreases with S, differently from the Markov chain approach. This is due to the overflow reduction caused by an higher S, which results in a reduced number of iterations required by the two methods to achieve convergence. As for the memory effort, we observe that with all the three methods the memory occupation is

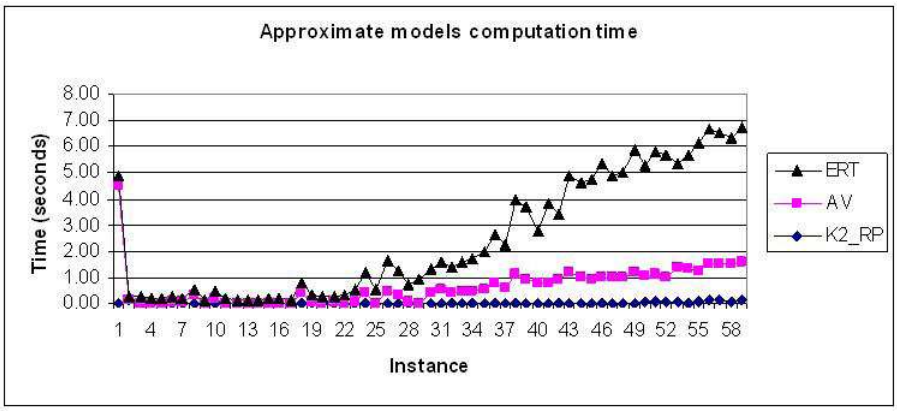

Figure 4.7: Computation time for the approximate models and practical instances.

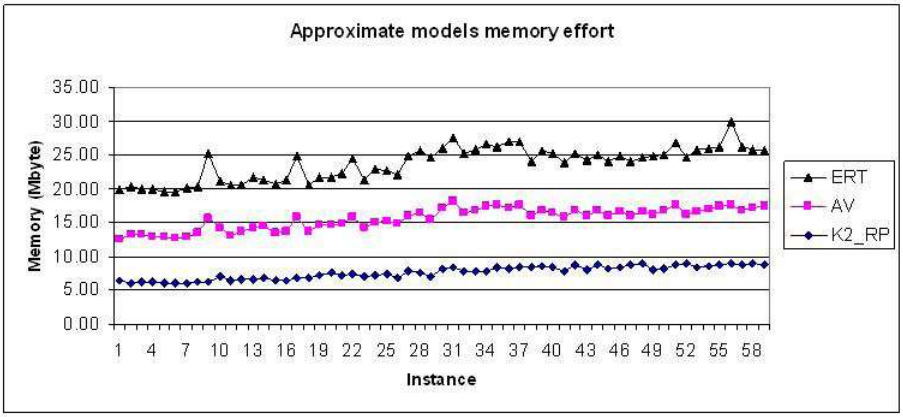

Figure 4.8: Memory effort for the approximate models and practical instances.

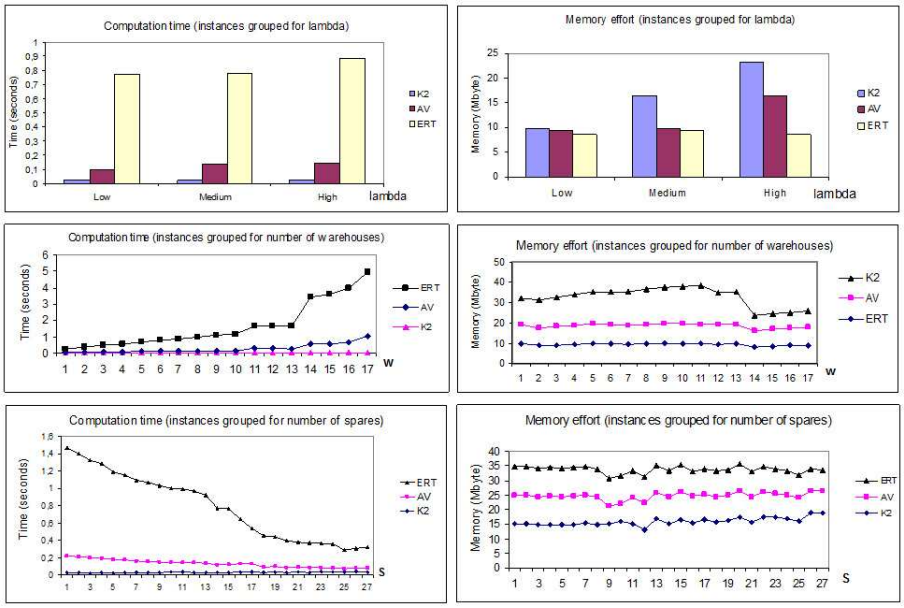

Figure 4.9: Computation time (left) and memory effort (right) for the approximate models and random instances.

not an issue. For the AV and ERT methods, the memory required to solve every instance only slightly increases with $\bar{\lambda}$ and w, while is almost constant with S. The multi-dimensional scaling down approach is more sensitive to the increase of $\bar{\lambda}$, w and S, but the memory required is always very limited in our experiments, and can be controlled by increasing the scale factor K.

Accuracy analysis

We now analyze the percentage error in OA evaluation for the three approximate methods with respect to the Markov chain solution. Figure 4.10 shows the percentage error achieved for the 42 practical instances for which the exact OA value can be computed with the Markov chain model. The 42 instances are ordered for increasing value of their exact Operational Availability. The two decomposition techniques ERT and AV provides the same values in practice, since their percentage difference is always smaller than 10^{-6}. This is mainly

due to the fact the estimated peakedness factor is almost 1 for all instances, and therefore modeling the arrival process as a peaked process, as in the ERT method, does not provide benefits with respect to approximating it with a Poisson process, as in the AV method. Therefore, in Figure 4.10 only one curve is shown for the two decomposition methods. The scaling down method clearly outperforms the decomposition techniques for small OA values ($OA < 0.997$), while the percentage error is similar for larger OA values. Besides the better performance shown in figure, in our experiments the scaling down method provides OA values smaller than the exact ones in more than 80% of the experiments while the decomposition methods find OA values always larger than the exact ones. The scaling down method is therefore more conservative than the decomposition methods, and this is an important feature when the method has to be used within an optimization procedure for spares allocation.

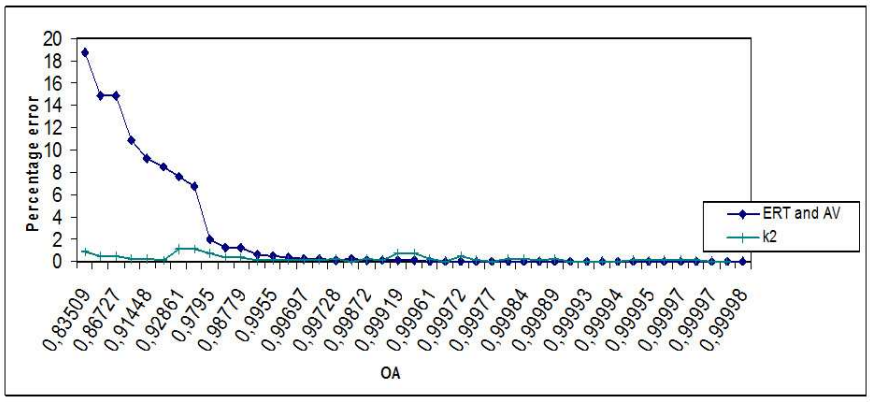

Figure 4.10: Percentage error for practical instances

Figure 4.11 shows the percentage error achieved for the 695 random instances for which the exact OA value can be computed with the Markov chain model. Instances with similar exact OA values are grouped together and the average error is shown in figure for the three models. Also for the random instances, the scaling down method clearly outperforms the decomposition techniques for small OA values while it behaves similarly for larger OA values.

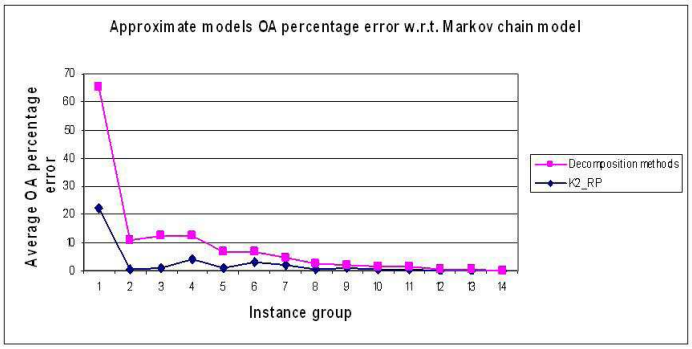

Figure 4.11: Percentage error for random instances

Influence of the scale factor in the multi-dimensional scaling down approach

In this section we show the influence of the scale factor K in the multidimensional scaling down method on the evaluation accuracy. In Figure 4.12 we report the OA value computed with the multidimensional scaling down method for four random instances and for K varying from 1 to 10, the value for $K = 1$ corresponding to the exact Markov chain computation. The four instances mainly differ each other for the number of spares S. It is interesting to notice that the approximate values are very similar to the exact one when K is sufficiently smaller than S, while the estimation deteriorates for $K \geq S$. In fact, the critical points of this method are the computation of the scaled number of spares $\hat{S}$ in Equation (4.11) and the allocation of these spares, which make the scaled model quite different from the original model as K approaches S.

Figure 4.12: 4 sample instances: OA varying for different scale factor values

4.5. Conclusions

In this chapter, we have presented three approximate evaluation techniques for estimating the operational availability of a maintenance supply chain with single echelon inventory and complete pooling. An external supplier manages

un-satisfied spare part requests when no spare is available in the warehouses of the network. From our computational experiments, carried out on practical and randomly generated instances, it turns out that the two decomposition approaches AV and ERT generate quite the same OA values. Therefore, taking into account the peaky nature of overflows, as in ERT, does not improve the accuracy of the solutions and is more time consuming with respect to modeling overflows with Poissonian flows, as in the AV method. However, the scaling down approach clearly outperforms the two decomposition approaches in terms of both accuracy and computation time.

Future developments of this research include the incorporation of these fast approximation methods in an optimization framework, to optimize the amount of spares and their allocation in order to grant the minimum levels of operational availability required by airport authorities at the minimum cost. To this aim the scaling down approach is preferable to the decomposition methods since the approximate OA values are usually smaller than the exact values computed by solving the Markov chain, and therefore more conservative.

Chapter 5

Spares allocation problem: optimization algorithms

This chapter deals with spare optimization in multi-location inventory systems with single item and repairable spare parts. Lateral and emergency shipments occur in response to stockouts. A continuous review base stock policy is assumed for the inventory control of the spare parts. The objective is to minimize the total costs for inventory holding, lateral transshipments and emergency shipments subject to a target level of operational availability of the whole system. For a given allocation, the computation of costs and the feasibility check requires solving a queueing network with blocking, which can be studied using a Markov chain modeling approach.
We model the stock allocation problem as a non convex integer program. We exploit the special structure of the problem to design an efficient branch and bound procedure. Our bounds are obtained by solving a reduced problem with convex objective function, solvable at optimality very efficiently. Computational experiments, carried on practical data from an airport equipment maintenance context show that this method "efficiently solves at optimality many practical instances."

5.1. Introduction

Single echelon inventory systems are experiencing an increasing interest in practice, in particular for the management of expensive spare parts. In such a context, the supply chain involves at least three actors: equipment users, logistics

companies and equipment suppliers. The users need spare parts to carry on their business without interruptions. Intermediate logistic companies are in charge of replenishing spare parts in the short term, by granting the contractual service level to the users at minimum cost. The suppliers are responsible for supplying new components and/or repaired items to the logistic companies.

As observed by several authors, see e.g. by [43], the logistics of spare parts differs from those of other materials in several ways. Equipments may have remarkable costs, long repairing times and sporadic failures. The latter are difficult to forecast and may cause relevant financial effects, due to the economical implications of a lack of equipment at the operational sites. In such cases, continuous review policies are typically adopted to reduce both reaction time to stockouts and inventory levels [2, 38]. Several heuristic procedures can be found in the literature for allocating spares to warehouses in a single echelon context with complete pooling [4, 54, 93].

This chapter addresses the problem of spare parts allocation in a single echelon inventory system with complete pooling characterized by expensive spares, long repairing time and strict requirements of operational availability (i.e., the fraction of time during which all operational sites are working). Our work is motivated by a practical problem faced by a large Italian logistics company. The company handles 17 warehouses supporting the daily activity of 38 airports spread over the Italian territory. Stock levels are currently determined with the VARIMETRIC algorithm of Sherbrooke (Sherbrooke 2004), based on a stiff hierarchic structure. However, in operation lateral transshipments take place between stocking points whenever there is an emergency requirement for parts, using couriers and overnight carriers to rapidly move parts. The company is therefore interested in determining the potential savings deriving from explicit inclusion of lateral transshipments in the model. To this aim, we propose and evaluate a new branch and bound procedure for stock level definition and spare parts allocation. The procedure exploits the particular cost structure of the maintenance supply chain under study and is very effective in this context. However, the method is general and we discuss the algorithm performance in a more general context.

This chapter is organized as follows. Section 5.2 describes the single echelon one-for-one ordering model with complete pooling and the spare parts allocation problem is formulated as a non-convex integer program. Section 5.3 studies the mathematical structure of the optimization problem. In Section 4.2 the Markov chain model used for computing transshipment costs and times. Heuristic and exact allocation algorithms are described in Section 5.4. Computational experiments are presented in Section 5.5, based on practical data

from the airport maintenance context. Some conclusions follow in Section 5.6.

5.2. The problem

In our model, a logistic company aims to compute the stock level s_i of each warehouse $i \in W$ such that a minimum level of service is granted at the operational sites and the overall cost is minimum. Costs are related to inventory holding, transshipments and emergency shipments.

Given an allocation s of spares to warehouses, the model used to compute the level of service is a single item, single echelon, w-locations, continuous review, one-for-one replenishment policy inventory system, whit lateral and emergency shipments, complete pooling and non-negligible transshipment times.

The Spares Allocation Problem is the problem of finding an allocation s which minimizes the overall cost for inventory holding, lateral and emergency shipments, subject to a constraint on the minimum operational availability of the system.

The contractual service level to grant is the operational availability OA of all operational sites for each item, computed as in chapter 3

We assume the Poisson distribution for the demand process, which is a typical assumption for modeling low demand processes [82]. We also use location dependent MTBF values. The replenishment time of the external supplier is a random variable, exponentially distributed, with known mean value $T_{0j} = MTTR+OS$, which is the same for any warehouse j. The capacity of the supplier repair shop is assumed to be infinite. It follows that also the number of replenishment from the external supplier follows the Poissonian distribution. These common assumptions make possible to use the Markovian analysis for modeling the multi-dimensional inventory system. Finally, we make the following assumptions.

1. Lateral transshipment is always more convenient than emergency shipment, i.e., the time and cost needed for a transshipment from warehouse i to warehouse j is always smaller than the time and cost required for an emergency shipment from warehouse j:

$$\max_{i,j \in W} \{T_{ij}\} < T_{0j} \tag{5.1}$$

$$\max_{i,j \in W} \{c^t_{ij}\} < c^e \tag{5.2}$$

2. The cost for a lateral transshipment from warehouse i to warehouse j increases linearly with the transfer time T_{ij}, i.e.,

$$c_{ij}^t = \alpha T_{ij} \tag{5.3}$$

Let L be the minimum operational availability level to be achieved by a feasible allocation. It is easy to check that this quantity corresponds to allowing a maximum waiting time $\frac{(1-L)MTBF}{L}$ to substitute failed items. Then, the Spares Allocation Problem P_0 can be formulated as the following integer program with non-convex objective function:

Problem P_0:

$$\begin{array}{ll} \min & \sum_{i=1}^{w} c^h s_i + \lambda_i \sum_{j \in W} \pi_{ij}(s) c_{ji}^t + \lambda_i P_B(S) c^e \\ s.t.: & \\ & \sum_{i=1}^{w} [\lambda_i \sum_{j \in W} \pi_{ij}(s) T_{ji} + \lambda_i P_B(S) T_{0i}] \leq \frac{(1-L)MTBF}{L} \end{array} \tag{5.4}$$

Let $f_1(S) = \sum_{i=1}^{w} c^h s_i$ be the total inventory holding cost, the cost for lateral transshipments is defined by $f_2(s) = \left\{ \sum_{i=1}^{w} \lambda_i \sum_{j \in W} \pi_{ij}(s) c_{ji}^t \right\}$ and let $f_3(S) = \sum_{i=1}^{w} \lambda_i P_B(S) c^e$ be the cost for the emergency shipments. Similarly, for the waiting times we let $t_2(s) = \left\{ \sum_{i=1}^{w} \lambda_i \sum_{j \in W} \pi_{ij}(s) T_{ji} \right\}$ be the waiting times due to lateral transshipments and $t_3(S) = \sum_{i=1}^{w} \lambda_i P_B(S) T_{0i}$ be the emergency waiting times.

5.3. Problem structure

In this section, we exploit the special structure of problem P_0. Let us first analyze the three functions $f_1(S)$, $f_2(s)$, $f_3(S)$. $f_1(S)$ is clearly linear and increasing with the total stock level S. Kranenburg and Van Houtum [49] proved that $f_3(S)$ is convex and decreasing with S. Finally, $f_2(s)$ is non-convex. The latter property can be easily shown by observing that $f_2(s)$ equals zero when $s = 0$ or when $s_i \to \infty$ for all $i = 1, \dots w$, since there are no transshipments in these two cases. On the other hand, $f_2(s) > 0$ otherwise, thus implying that $f_2(s)$ is non-convex. Similarly, it can be easily proved that $t_3(S)$ is convex [49] and $t_2(s)$ is non-convex. We next show that the quantities $t_2(s) + t_3(S)$ and $f_2(s) + f_3(S)$ are decreasing. To this aim, let us consider an allocation s and a warehouse $i \in W$. Denote with $\hat{s}$ the allocation such that

$\hat{s}_i = s_i + 1$ and $\hat{s}_h = s_h$ for all $h \in W$, $h \neq i$. Let also denote $S = \sum_{i=1}^{w} s_i$ and $\hat{S} = \sum_{i=1}^{w} \hat{s}_i = S + 1$.

Let us first observe that when passing from s to $\hat{s}$ the number of spares at each warehouse does not decrease and therefore the probability of an outstanding request at each warehouse cannot increase. Specifically, the following properties must hold.

- $P_B(s) > P_B(\hat{s})$.
- As for warehouse i, the probability that the aggregated arrival rate λ_i (without transshipments) is satisfied by the local stock increases when passing from s_i to $s_i + 1$. Thus, the probability of reforwarding the demand decreases, i.e. $\pi_{ij}(\hat{s}) < \pi_{ij}(s)$. It follows that $\sum_{j \in W} \pi_{ij}(s) T_{ji} + P_B(s) T_{0i} > \sum_{j \in W} \pi_{ij}(\hat{s}) T_{ji} + P_B(\hat{s}) T_{0i}$.
- For what concerns warehouse $h \neq i$, in view of rule (4.2) and assumption (5.1), the transshipment probability remains the same since λ_h and s_h are the same with s and $\hat{s}$. However, the probability of reforwarding the request to warehouse i increases and the probability of reforwarding the request to a most far warehouse (or to the external supplier) decreases, i.e., $\sum_{j \in W} \pi_{hj}(s) T_{jh} + P_B(s) T_{0h} > \sum_{j \in W} \pi_{hj}(\hat{s}) T_{jh} + P_B(\hat{s}) T_{0h}$.

In conclusion,

$$\begin{aligned} & t_2(s) + t_3(S) - t_2(\hat{s}) - t_3(\hat{S}) = \\ = & \sum_{i=1}^{w} \lambda_i \Big[\sum_{j \in W} \left(\pi_{ij}(s) - \pi_{ij}(\hat{s}) \right) T_{ji} + \\ & + \left(P_B(S) - P_B(\hat{s}) \right) T_{0i} > 0. \end{aligned}$$

Using assumption (5.2), a similar discussion for the costs leads to $f_2(s) + f_3(S) > f_2(\hat{s}) + f_3(\hat{S})$.

Given an upper bound UB on the optimum of problem P_0, an upper bound MAX on the total stock level S of an optimal solution can be efficiently computed by considering only the terms $f_1(S)$ and $f_3(S)$ of the objective function.

$$MAX = \min \{ S : f_1(S) + f_3(S) \geq UB \} \tag{5.5}$$

This value is quite close to the optimal stock level S^* when the transshipment cost $f_2(s^*)$ is small with respect to $f_1(S^*) + f_3(S^*)$, as in our practical application. Similarly, a lower bound MIN on S^* can be efficiently computed by

considering only the term $t_3(S)$, decreasing with S, in the constraint of the problem.

$$MIN = \min\left\{S : t_3(s) \leq \frac{(1-L)MTBF}{L}\right\} \tag{5.6}$$

These bounds can be used to refine the formulation of Problem P_0, thus leading to the new formulation P_1.

Problem P_1:

$$\begin{array}{ll} \min & f_1(S) + f_2(s) + f_3(S) \\ s.t.: & \\ & t_2(s) + t_3(S) \leq \frac{(1-L)MTBF}{L} \\ & \\ & MIN \leq S \leq MAX \end{array} \tag{5.7}$$

We next introduce a Lagrangian relaxation $P_2(\gamma)$ of problem P_1 by relaxing the waiting time constraint. We use the notation γ to denote the Lagrangian multiplier.

Problem $P_2(\gamma)$:

$$\begin{array}{ll} \min & f_1(S) + f_2(s) + f_3(S) + \gamma\left(t_2(s) + t_3(S) - \frac{(1-L)MTBF}{L}\right) \\ s.t.: & \\ & MIN \leq S \leq MAX \end{array} \tag{5.8}$$

It is well known that, for varying γ, $P_2(\gamma)$ is a concave, piecewise linear function. Calling breakpoint the values of $P_2(\gamma)$ in which the slope of $P_2(\gamma)$ changes, there is an optimal solution γ^* for the Lagrangian dual $\max\{P_2(\gamma) : \gamma \geq 0\}$ which is a breakpoint. If we let $\overline{s}$ be an optimal allocation for $P_2(\overline{\gamma})$, and $\overline{\gamma}$ is not a breakpoint, then the slope of $P_2(\gamma)$ in $\overline{\gamma}$ is [35]:

$$t_2(\overline{s}) + t_3(\overline{S}) - \frac{(1-L)MTBF}{L} \tag{5.9}$$

Theorem 6 *If $\overline{\gamma}$ is not a breakpoint there is a single optimal stock level for $P_2(\overline{\gamma})$.*

Proof. By contradiction, let us suppose that in $\overline{\gamma}$ there are two optimal allocations s and $\overline{s}$ with different stock levels S and $\overline{S}$, respectively. Therefore:

$$\begin{aligned} & f_1(S) + f_2(s) + f_3(S) + \overline{\gamma}\left(t_2(s) + t_3(S) - \frac{(1-L)MTBF}{L}\right) = \\ & = f_(\overline{S}) + f_2(\overline{s}) + f_3(\overline{S}) + \overline{\gamma}\left(t_2(\overline{s}) + t_3(\overline{S}) - \frac{(1-L)MTBF}{L}\right). \end{aligned}$$

¿From equation (5.9) it follows that the constraint violation is the same for s and $\overline{s}$, i.e., $t_2(s) + t_3(S) = t_2(\overline{s}) + t_3(\overline{S})$. It follows from the proportionality assumption (5.3) that also $f_2(s) + f_3(S) = f_2(\overline{s}) + f_3(\overline{S})$ must hold. Hence, we obtain $f_1(S) = f_1(\overline{S})$, i.e., $c^h S = c^h \overline{S}$. This implies the thesis $S = \overline{S}$. □

Theorem 7 *If $\overline{\gamma}$ is a breakpoint and the slope of $P_2(\overline{\gamma})$ decreases from $t_2(s^1) + t_3(S^1) - \frac{(1-L)MTBF}{L}$ to $t_2(s^2) + t_3(S^2) - \frac{(1-L)MTBF}{L}$, then $S^2 > S^1$.*

Proof. At the breakpoint $\overline{\gamma}$ there are at least the two optimal solutions s^1 and s^2 for problem $P_2(\overline{\gamma})$, i.e.,

$$f_1(S^1) + f_2(s^1) + f_3(S^1) + \overline{\gamma}\left(t_2(s^1) + t_3(S^1) - \frac{(1-L)MTBF}{L}\right) =$$
$$= f_1(S^2) + f_2(s^2) + f_3(S^2) + \overline{\gamma}\left(t_2(s^2) + t_3(S^2) - \frac{(1-L)MTBF}{L}\right).$$

Since the slope of $P_2(\gamma)$ decreases, then $t_2(s^1) + t_3(S^1) > t_2(s^2) + t_3(S^2)$ and, from the proportionality assumption (5.3), also $f_2(s^1) + f_3(S^1) > f_2(s^2) + f_3(S^2)$ must hold. Hence, it follows that $f_1(S^1) < f_1(S^2)$, which implies the thesis $S^1 < S^2$. □

Theorem 8 *If the breakpoint γ^* is an optimal solution of the Lagrangian dual $\max\{P_2(\gamma) : \gamma \geq 0\}$ and the slope of $P_2(\gamma^*)$ decreases from $t_2(s^1) + t_3(S^1) - \frac{(1-L)MTBF}{L} \geq 0$ to $t_2(s^2) + t_3(S^2) - \frac{(1-L)MTBF}{L} \leq 0$, then:*

1. *s^2 is feasible for problem P_1;*
2. *either s^2 is optimal for P_1 or S^2 is greater than the optimal stock level for P_1.*

Proof. The feasibility of s^2 directly follows from $t_2(s^2) + t_3(S^2) - \frac{(1-L)MTBF}{L} \leq 0$. If s^2 is not optimal, let s^* be an optimal allocation and S^* be the corresponding stock level. From the optimality of S^* it follows that:

$$f_1(S^*) + f_2(s^*) + f_3(S^*) < f_1(S^2) + f_2(s^2) + f_3(S^2) \tag{5.10}$$

On the other hand at γ^* the objective function of the Lagrangian relaxation computed in s^* must be greater or equal than in s^2, i.e.,

$$f_1(S^*) + f_2(s^*) + f_3(S^*) + \gamma^*\left(t_2(s^*) + t_3(S^*) - \frac{(1-L)MTBF}{L}\right) \geq$$
$$= f_1(S^2) + f_2(s^2) + f_3(S^2) + \gamma^*\left(t_2(s^2) + t_3(S^2) - \frac{(1-L)MTBF}{L}\right).$$

Therefore, $t_2(s^*) + t_3(S^*) > t_2(s^2) + t_3(S^2)$ must hold. From assumption (5.3) it must hold also $f_2(s^*) + f_3(S^*) > f_2(s^2) + f_3(S^2)$. Therefore, from inequality (5.10), $f_1(S^*) < f_1(S^2)$, i.e., $S^* < S^2$. □

Despite the nice properties of $P_2(\gamma^*)$ shown in Theorem 8, the computation of $P_2(\gamma^*)$ requires the computation of quantity $f_2(s^*) + \gamma^* t_2(s^*)$, which is computationally expensive. Moreover, we observe that in practical applications it often occurs $f_2(s^*) << f_1(S^*) + f_3(S^*)$, as in our case study. In order to efficiently compute a lower bound to $P_2(\gamma^*)$, let us introduce problem $P_3(\gamma)$:

Problem $P_3(\gamma)$:

$$\begin{array}{ll} \min & f_1(S) + x + f_3(S) + \gamma\left(y + t_3(S) - \frac{(1-L)MTBF}{L}\right) \\ s.t.: & \\ & MIN \leq S \leq MAX \\ & x \leq f_2(s) \\ & y \leq t_2(s) \end{array} \tag{5.11}$$

Suitable values for x and y can be computed by exploiting the properties that $f_3(S)$ and $f_2(s) + f_3(S)$ are decreasing with S. Given any feasible allocation s and the corresponding $S = \sum_{i=1}^{w} s_i$, the following must hold:

$$\begin{array}{l} f_3(S) \leq f_3(MIN) \\ t_3(S) \leq t_3(MIN) \\ f_2(s) + f_3(S) \geq \min_{\overline{s}:\sum_{i=1}^{w}\overline{s}_i = MAX}\{f_2(\overline{s})\} + f_3(MAX) \\ t_2(s) + t_3(S) \geq \min_{\overline{s}:\sum_{i=1}^{w}\overline{s}_i = MAX}\{t_2(\overline{s})\} + t_3(MAX) \end{array} \tag{5.12}$$

Therefore, the values $x = \min_{\overline{s}:\sum_{i=1}^{w}\overline{s}_i = MAX}\{f_2(\overline{s})\} + f_3(MAX) - f_3(MIN)$ and $y = \min_{\overline{s}:\sum_{i=1}^{w}\overline{s}_i = MAX}\{t_2(\overline{s})\} + t_3(MAX) - t_3(MIN)$ guarantee that constraints $x \leq f_2(s)$ and $y \leq t_2(s)$ are satisfied by any allocation s such that $MIN \leq \sum_{i=1}^{w} s_i \leq MAX$. In what follows, we fix x and y to these values and omit the two latter constraints from the formulation of Problem $P_3(\gamma)$, which can be written as follows:

Problem $P_3(\gamma, MIN, MAX)$:

$$\begin{array}{ll} \min & f_1(S) + \min_{\overline{s}:\sum_{i=1}^{w}\overline{s}_i = MAX}\{f_2(\overline{s})\} + f_3(MAX) - f_3(MIN) + f_3(S) + \\ & \gamma\left(\min_{\overline{s}:\sum_{i=1}^{w}\overline{s}_i = MAX}\{t_2(\overline{s})\} + t_3(MAX) - t_3(MIN) + t_3(S) - \frac{(1-L)MTBF}{L}\right) \\ s.t.: & \\ & MIN \leq S \leq MAX \end{array} \tag{5.13}$$

Since $f_1(S)$ is linearly increasing while $f_3(S)$ and $t_3(S)$ are convex and decreasing with S, the objective function of Problem $P_3(\gamma, MIN, MAX)$ is convex for any given $\gamma \geq 0$. Therefore, given the values x and y, the optimal S can be efficiently computed by using a binary search approach in the interval $[MIN, MAX]$. We compute x as in [91]. The next theorem shows that computing y is not necessary in order to compute the maximum of $P_3(\gamma, MIN, MAX)$.

Theorem 9 *The value $\gamma = 0$ maximizes $P_3(\gamma, MIN, MAX)$.*

Proof. To prove the theorem it is sufficient to prove that quantity

$$\min_{\overline{s}:\sum_{i=1}^{w}\overline{s}_i=MAX} \{t_2(\overline{s})\} + t_3(MAX) - t_3(MIN) + t_3(S) - \frac{(1-L)MTBF}{L}$$

is always non positive for $MIN \leq S \leq MAX$. This property follows by observing that $t_3(S) \leq t_3(MIN)$ and that a feasible solution exists for $S = MAX$, i.e.,

$$\min_{\overline{s}:\sum_{i=1}^{w}\overline{s}_i=MAX} \{t_2(\overline{s})\} + t_3(MAX) \leq \frac{(1-L)MTBF}{L}$$

□

5.4. Solution procedure

In this section, a branch-and-bound algorithm for finding an optimal allocation of spares to warehouses is described. At each node of the enumeration tree the lower bound is computed by solving $P_3(0, MIN, MAX)$, where MIN and MAX are computed according to equations (5.6) and (5.5) at the root node and then updated by the branching rule. The heuristic algorithm described in subsection 5.4 provides an initial upper bound UB, then updated whenever a new feasible solution is found, and a sketch of the branch and bound procedure is presented in subsection 5.4.

Upper-bound computation

A simple upper bound to Problem P_0 is computed by distributing spare parts among warehouses with positive demand and by giving preference to warehouses with larger demand. In fact, simulation experiments carried out in [14]

Procedure ISA
set $S = 0$ and $s_i = 0, \quad i = 1, \ldots, w$.
set $rhs = \frac{(1-L)MTBF}{L}$ and $k = 1$.
repeat
repeat
set $S = S + 1$, $s_k = s_k + 1$ and $k = k + 1$;
if $(k = |\overline{W}|$ then set $k = 1$;
until $\sum_{i=1}^{w}[\lambda_i \sum_{j \in W} \hat{\pi}_{ij}(s)T_{ji} + \lambda_i P_B(S)T_{0i}] \leq rhs$
if $\sum_{i=1}^{w}[\lambda_i \sum_{j \in W} \pi_{ij}(s)T_{ji} + \lambda_i P_B(S)T_{0i}] > \frac{(1-L)MTBF}{L}$
then $rhs = rhs - \sum_{i=1}^{w}\left[\lambda_i \sum_{j \in W} (\pi_{ij}(s) - \hat{\pi}_{ij}(s))\, T_{ji}\right]$
until $\sum_{i=1}^{w}[\lambda_i \sum_{j \in W} \pi_{ij}(s)T_{ji} + \lambda_i P_B(S)T_{0i}] \leq \frac{(1-L)MTBF}{L}$
return s.

Figure 5.1: Pseudocode of the heuristic for Initial Spares Allocation

show that avoiding concentration of spares in few warehouses is an effective allocation policy. The heuristic procedure ISA (Initial Spares Allocation), sketched in Figure 5.4, finds an allocation s, feasible for P_0, by greedily allocating one spare at a time to warehouses in the set $\overline{W} = \{i \in W : \lambda_i > 0\}$. Without loss of generality we assume that the warehouses are numbered for decreasing value of λ_i, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{|\overline{W}|}$. ISA terminates when the quantity $\sum_{i=1}^{w}[\lambda_i \sum_{j \in W} \pi_{ij}(s)T_{ji} + \lambda_i P_B(S)T_{0i}]$ becomes smaller than $\frac{(1-L)MTBF}{L}$. In order to speed up the computation of state probabilities $\pi_{ij}(s)$ at each step of the procedure, the heuristic computes approximated values $\hat{\pi}_{ij}(s)$ with the fast multi-dimensional scaling down method described in [15]. When

$$\sum_{i=1}^{w}[\lambda_i \sum_{j \in W} \hat{\pi}_{ij}(s)T_{ji} + \lambda_i P_B(S)T_{0i}] \leq \frac{(1-L)MTBF}{L} \tag{5.14}$$

the feasibility of allocation s is checked by solving the associated Markov chain exactly. In case of feasible solution, Procedure ISA stops and returns the feasible allocation s, otherwise the constraint (5.14) is strengthened by replacing

the right-hand side $\frac{(1-L)MTBF}{L}$ with the smaller value

$$\frac{(1-L)MTBF}{L} - \sum_{i=1}^{w} \left[\lambda_i \sum_{j \in W} (\pi_{ij}(s) - \hat{\pi}_{ij}(s)) T_{ji} \right]$$

Procedure ISA then continues allocating one spare at a time and checking feasibility with the multi-dimensional scaling down method until a new apparently feasible solution is found.

Branch and bound algorithm

Our BB (Branch and Bound) procedure maintains a queue Q of intervals $[MIN, MAX]$ for the stock level S, each corresponding to an instance of P_1.

Procedure ISA provides an initial solution $Bestsol$, from which the first upper bound UB on the optimum is derived. At the root node, Q is initialized with one open problem in which MIN and MAX are computed according to (5.6) and (5.5).

At each iteration of the BB procedure an open problem is removed from Q according to First In First Out rule and an optimal solution S^* to $P_3(0, MIN, MAX)$ is computed. If the lower bound $P_3(0, MIN, MAX) \geq UB$ the problem is closed. Otherwise, an allocation $s^* = \text{argmin}\{t_2(s) : \sum_{i=1}^{w} s_i = S^*\}$ is computed as in [91].

If $t_2(s^*) + t_3(S^*) \leq \frac{(1-L)MTBF}{L}$, then s^* is feasible for P_0 and, in view of assumption (5.3), it is also an optimal allocation for the restricted version of P_1 in which $S = S^*$. In this case, two new open problems are added to Q with $MIN \leq S \leq S^* - 1$ and $S^* + 1 \leq S \leq MAX$, respectively, and the upper bound UB is updated if $f_1(S^*) + f_2(s^*) + f_3(S^*) < UB$.

If $t_2(s^*) + t_3(S^*) > \frac{(1-L)MTBF}{L}$, then s^* is infeasible for P_1. Thus, for all values $MIN \leq S \leq S^*$ there is no feasible solution to P_0 and only the open problem with $S^* + 1 \leq S \leq MAX$ is added to Q. The procedure terminates when Q is empty and the current allocation $Bestsol$ is an optimal solution to P_0.

5.5. Case study from the corrective airport maintenance context

In this section, we report on our computational experiments with the algorithms for spares allocation presented in section 5.4 applied to solve the practi-

Algorithm BranchAndBound
Find an allocation $BestSol = ISA$
set $UB = f_1(BestSol) + f_2(BestSol) + f_3(BestSol)$
$MIN = \min\left\{S : t_3(s) \leq \frac{(1-L)MTBF}{L}\right\}$
$MAX = \min\{S : f_1(S) + f_3(S) \geq UB\}$
push $[MIN, MAX]$ in queue Q
while $Q \neq \emptyset$ do
pop $[x, y]$ from Q
if $P_3(0, x, y) < UB$ then
let S^* be the optimal solution to $P_3(0, x, y)$
set $s^* = \text{argmin}\{t_2(s) : \sum_{i=1}^{w} s_i = S^*\}$
if $t_2(s^*) + t_3(S^*) \leq \frac{(1-L)MTBF}{L}$ then
if $f_1(S^*) + f_2(s^*) + f_3(S^*) < UB$ then
set $BestSol = s^*$ and $UB = f_1(S^*) + f_2(s^*) + f_3(S^*)$
set $y = \min\{y; \min\{S : f_1(S) + f_3(S) \geq UB\}\}$
end if
push $[x, S^* - 1]$ in Q
push $[S^* + 1, y]$ in Q
else
push $[S^* + 1, y]$ in Q
end if
end if
end while
Return UB and $Bestsol$

Figure 5.2: Pseudocode of the BB algorithm

cal problem from the airport maintenance context described in the introduction of the chapter. The case study originates from the practical needs of an Italian logistics company supporting the activity of 38 civil airports spread over the Italian territory. The company manages the overall processes of purchasing, holding and replacing failed items, ensuring that the overall reliability of safety equipments is always within contractual limits. The aim of the company is therefore to grant the prescribed quality of service at minimum cost. While the company currently follows a two echelon policy for spare part management, the company managers are interested in evaluating the potential benefits deriving from the adoption of a single echelon policy, which is generally acknowledged to achieve better performance in similar contexts. To this aim, the algorithms have been tested on a set of twelve instances from our case study, each based on the warehouse locations and demand rates of a particular item. However, in order to test the algorithms on a wider context than the real situation, we generated several scenarios by varying holding, transshipment and emergency costs of each item. For the replenishment time of an item we use the exponential distribution with average equal to three months for all items and scenarios while for the transshipment time and cost we use a deterministic value proportional to the distance between warehouses. Each pair holding-emergency cost defines a scenario for each of the twelve items. We consider 21 scenarios by choosing the cost of an item from the interval $[200, 1200]$ and fixing the emergency cost equal to 7000. Three additional scenarios are defined in order to analyze the influence of the emergency cost on the algorithms performance. In this second set of scenarios the item cost is fixed equal to 300 for each item while the emergency cost varies from 7000 to 200000. In total we obtain 288 instances. In Table 5.1 we summarize the values of the main parameters used in the computational experiment.

Parameter name	Unit	Values
Warehouses with positive demand		2,3,4,5,6,7,8,9
Number of installed items		3,5,8,9,10,11,16,18
Average MTBF	hours	16000, 17000, 26000, 38000, 61000, 79000, 81000, 94000, 101000, 132000, 191000, 200000
Holding cost	euros	$200, 250, 300, \ldots, 1000, 1150, 1200$
Emergency cost	euros	$7000, 50000, 100000, 200000$
Min-Max average transshipment lead time	hours	$[5, 37.5]$
Emergency lead time	hours	2160

Table 5.1: Parameter values for the computational experiment

In tables 5.2, 5.3 and 5.4 we show the results obtained for the 12 instances and the 24 scenarios. Table 5.2 shows the solutions and the computation time (in seconds) of ISA and BB and the relative error of ISA, computed as $\frac{ISA-BB}{BB}$. Each row in table shows the average results obtained for an item by varying the costs over the 24 scenarios. We observe that the BB algorithm is able to find the optimal solution within less than 100 seconds of computation for eight of the twelve instances, while the optimum is found within approximately 30 minutes for other three instances. ISA always finds a feasible solution within less than one second. The ISA value turns out to be the optimal solution for 50 out of 288 instances and the average error over the 24 scenarios varies in the range $[0.04, 0.31]$. These experiments show that ISA provides good solutions within short computation time, even if it is worth using the exact algorithm to find better solutions.

Item	# wareh.	BB cost values			ISA	$\frac{ISA-BB}{BB}$	Computation time	
		holding	trans.	emer.	value		ISA	BB
1	2	1635.00	107.98	16.54	2210.32	0.22	0.13	0.82
2	3	1245.00	369.07	26.07	2257.27	0.31	0.15	0.94
3	4	1740.00	452.22	8.38	2300.32	0.10	0.14	1.40
4	4	2260.00	528.79	43.12	3501.11	0.18	0.16	2.02
5	4	1485.00	580.49	23.25	2789.49	0.27	0.17	1.39
6	5	1350.00	670.44	38.49	2213.43	0.07	0.14	1.79
7	6	1920.00	886.66	72.86	3378.51	0.14	0.16	31.69
8	6	1448.33	924.44	62.50	2757.65	0.19	0.14	70.03
9	7	1513.33	1114.43	22.83	3102.12	0.26	0.13	1808.64
10	7	1530.00	1105.76	28.53	3111.92	0.26	0.13	1936.89
11	8	2112.00	1349.72	12.96	3596,77	0.05	0.15	1080.56
12	9	2014.00	1304.32	34.13	3436,47	0.04	0.13	4652.70

Table 5.2: Performance of ISA and BB algorithms for the 12 items

In Table 5.3 we analize the performance of ISA and BB for varying the holding cost of the items. Each row in table reports the average results over the 12 items for one of the first 21 scenarios. We also show the three components of the optimum cost, i.e., holding, transshipment and emergency cost. It can be observed that the transshipment cost is often comparable with the holding cost, and therefore it cannot be neglected in the solution of the problem. For a holding cost lower than 350 the number of spares allocated by ISA is smaller than the optimal value, while for holding costs higher than 350 ISA always allocates a number of spares optimal or strictly larger than the optimal solution. In fact, we observe that the number of spares allocated by ISA does not depend on the spare holding cost and therefore the number of spares allocated by ISA is always the same for all scenarios. A consequence of this behavior is that the

gap between ISA and the optimum is influenced by the holding cost. When the holding cost of an item increases from 200 to 350 the error decreases from 15% down to 8%. For larger holding costs the error increases regularly until the maximum of 32%, achieved for a holding cost equal to 1200. Smaller errors are attained when the number of spares allocated by ISA is approximately the optimal one and the error only depends on the warehouses to which they are allocated.

Hold.	Emerg.	BB cost values			ISA	$\frac{ISA-BB}{BB}$	# spares	
cost	cost	holding	trans.	emerg.	value		BB	ISA
200	7000	1050.00	225.57	2.40	1538.80	0.20	5.2	3.75
250	7000	1229.17	302.33	2.71	1726.30	0.13	4.9	3.75
300	7000	1250	511.14	3.38	1913.80	0.10	4.1	3.75
350	7000	1341.67	608.98	12.51	2101.30	0.09	3.8	3.75
400	7000	1433.33	702.55	13.23	2288.80	0.08	3.5	3.75
450	7000	1537.50	770.17	15.87	2476.30	0.08	3.4	3.75
500	7000	1541.67	914.79	29.99	2663.80	0.09	3.0	3.75
550	7000	1604.17	985.84	45.09	2851.30	0.09	2.9	3.75
600	7000	1700.00	1030.90	49.58	3038.80	0.10	2.8	3.75
650	7000	1733.33	1125.55	56.92	3226.30	0.12	2.6	3.75
700	7000	1866.67	1125.55	56.92	3413.80	0.13	2.6	3.75
750	7000	1937.50	1182.59	62.38	3601.30	0.14	2.5	3.75
800	7000	2066.67	1182.59	62.38	3788.80	0.15	2.5	3.75
850	7000	2195.83	1182.59	62.38	3976.30	0.17	2.5	3.75
900	7000	2325.00	1182.59	62.38	4163.80	0.18	2.5	3.75
1000	7000	2454.17	1182.59	62.38	4351.30	0.19	2.5	3.75
1050	7000	2583.33	1182.59	62.38	4538.80	0.20	2.5	3.75
1100	7000	2712.50	1182.59	62.38	4726.30	0.21	2.5	3.75
1150	7000	2841.67	1182.59	62.38	4913.80	0.21	2.5	3.75
1200	7000	2875	1182.59	62.38	5101.30	0.22	2.5	3.75

Table 5.3: Performance of ISA and BB for different holding costs

In Table 5.3 we analize the performance of ISA and BB for varying the item emergency costs. Each row in table shows the average results over the 12 instances for one of the 4 scenarios. Similarly to the previous scenario, the number of spares allocated by ISA is the same for all scenarios since this value does not depend on the spare emergency cost. The gap between ISA and the optimum depends therefore on the emergency cost, even if the emergency cost has a smaller influence on the error of ISA with respect to the holding cost.

As a concluding observation, computational experiments show that the overall behavior of ISA is acceptable as an initial solution for subsequent optimization. In general, the performance of ISA depends on the specific holding and emergency costs being considered and, therefore, it may be quite erratic. BB algorithm seems to be more promising, since it finds the proven optimum within acceptable computation time for all tested instances.

Hold. cost	Emerg. cost	BB cost values			ISA value	$\frac{ISA-BB}{BB}$	# spares	
		holding	trans.	emerg.			BB	ISA
300	7000	1250	511.14	3.38	1913.80	0.10	4.17	3.75
300	50000	1250	511.14	24.12	1964.42	0.11	4.17	3.75
300	100000	1325	442.93	38.76	2023.27	0.12	4.42	3.75
300	200000	1400	406.72	15.40	2140.99	0.17	4.67	3.75

Table 5.4: Performance of ISA and BB for different emergency costs

5.6. Conclusions

In this chapter we propose and evaluate a solution methodology for optimizing inventory stock allocation of repairable spare parts in a single echelon, w-locations system, where lateral and emergency shipments occur in response to stockouts. We model our problem as a non-convex integer program and develop a new heuristic and a new branch and bound algorithm for allocating the spare parts optimally. Both algorithms are evaluated by using practical data from the Italian airport corrective maintenance context. Computational experiments demonstrate that the branch and bound technique is able to optimally solve almost all tested instances within reasonable computation time. The heuristic algorithm finds sub-optimal solutions within very limited computation time, thus being a promising approach for finding feasible solutions for difficult instances.

Future research should focus on the development of faster exact methods and effective metaheuristics for the solution of large and difficult instances, as well as on the application of the ideas proposed in this chapter to manage the maintenance of different critical infrastructures, such as medical equipments in hospitals or communication or energy distribution networks and so on.

Chapter 6

Conclusions

6.1. Summary of main achievements

Effective supply chain management is currently recognized as a key determinant of competitiveness and success in manufacturing and services, because the implementation of supply chain management has significant impact on cost, service and quality. Numerous strategies for achieving these targets have been proposed.

The improvements in information technology coupled with the substantial reduction in the cost of processing, storing and analyzing data have made new strategies more attractive. On such strategy allows movements of stock between locations at the same echelon level via lateral transshipment.

Despite the above technology improvements, the implementation of such transshipment strategy requires still great efficiency especially in real life problems, because it suffers from computer memory limits and long computation times when the number of warehouses gets large, or when the number of parallel items to ba analyzed following an item approach gets large, too.In fact, a drawback of the policy of interest is the state dependent nature of the re-forwardings in the systems.

Therefore an effective tactical planning requires joint contribution from various disciplines in order to be implemented efficiently, such as engineering, mathematics, economics and computer science. New solution methods have to

be explored in order to effectively implementing new management strategies.

This thesis uses operations research techniques in order to study a single echelon, one-for-one ordering policy with complete pooling, with a deterministic rule for lateral transshipments.
Specifically we propose new evaluation and optimization methods thus handling real life problems within a reasonable amount of computation time. In fact, we test all the proposed methods on the practical case study motivated by the practical needs of an Italian logistics, supporting the activity of 38 civil airports spread over the Italian territory. The company handles 17 warehouses and manages the overall process of purchasing, holding, ensuring that the overall reliability of safety equipments is always within contractual limits. The aim of the company is therefore to grant the prescribed quality of service at minimum cost.

The items to be managed in such a context are typically expensive ones and with low demand, but we clearly recognize that there are many different types of service parts and that they perform many different functions. Therefore, in such a context also parts with a lower ratio between holding and transshipment costs may be encountered and managed. Thus with all the uncertainties that exist, a tactical plan should be created that will provide the flexibility needed to meet a wide range of scenarios, pointing the attention on the characteristics of the majority of items. Common techniques models the management policy with a Markov chain approach, thus evaluating such a policy given a spare parts allocation. The optimal stock allocation problem is formulated as an integer program with non linear objective function and non linear constraints. Therefore total enumeration methods or approximation algorithms can be employed for optimally solve it.

In this thesis efficient evaluation and optimization techniques are proposed. Accurate models have been developed for assessing the performance of a single echelon replenishment policy, and then evaluated on the basis of practical data with an increasing numbers of locations. The properties of the Markov chain associated to the chosen policy have been analyzed, with particular reference to the possibility of expressing the state probabilities in product form.

Using a suitable optimization model we have shown that the Markov chain cannot be decomposed exactly in product form. In fact, the best product form approximation returns a positive accuracy error, which implies that an exact

product form does not exist.

Hence, we have adapted four approximation techniques to our model and evaluate their performance in terms of computational effort, memory requirement and error with respect to the exact value. Three techniques approximate state probabilities with others that can be expressed in product form, so that the Markov chain can be decomposed. Specifically, we adapt a method by Alfredsson and Verrijdt, the Equivalent Random Traffic (ERT) method and the Interrupted Pisson Process (IPP) method. The last two techniques have been proposed for exploring the influence of peakedness in approximation models with respect to the accuracy of performance estimation due to the state dependent nature of the re-forwardings in the system.
The fourth technique is based on the multi-dimensional scaling down approach, which studies an equivalent reduced Markov chain rather than decomposing the original one. Concerning the IPP method, we found approximately the same results as the modified ERT method. IPP method is substantially more difficult than ERT method and more time consuming, for these reason only the modified ERT method has been taken into account for analyzing the influence on performance estimation of the investigated peaky nature of the flows in our system. Even if all the three methods are quite efficient in computing a solution, the multi-dimensional scaling down approach is the most efficient with all instances. The maximum time required to solve a practical instance is 0.11 seconds. As expected, the ERT method is slightly more time consuming than the AV method, due to the need to solve a non-linear system instead of using a closed form expression as in the AV method. Similar behavior can be observed for the memory effort.
With the two decomposition methods AV and ERT, the computation time decreases with S, differently from the Markov chain approach. This is due to the overflow reduction caused by an higher S, which results in a reduced number of iterations required by the two methods to achieve convergence. As for the memory effort, we observe that with all the three methods the memory occupation is not an issue. For the AV and ERT methods, the memory required to solve every instance only slightly increases with $\bar{\lambda}$ and w, while is almost constant with S. The multi-dimensional scaling down approach is more sensitive to the increase of $\bar{\lambda}$, w and S, but the memory required is always very limited in our experiments, and can be controlled by increasing the scale factor K.
By analyzing the percentage error in OA evaluation, it turns out that the two decomposition techniques ERT and AV provides very similar values, since their percentage difference is always smaller than 10^{-6}. This is mainly due to the

fact the estimated peakedness factor is almost 1 for all instances, and therefore modeling the arrival process as a peaked process, as in the ERT method, does not provide benefits with respect to approximating it with a Poisson process, as in the AV method.
The scaling down method clearly outperforms the decomposition techniques for small OA values ($OA < 0.997$), while the percentage error is similar for larger OA values. Besides the better performance shown in figure, in our experiments the scaling down method provides OA values smaller than the exact ones in more than 80% of the experiments while the decomposition methods find OA values always larger than the exact ones. The scaling down method is therefore more conservative than the decomposition methods, and this is an important feature when the method has to be used within an optimization procedure for spares allocation.
We have therefore analyzed the influence of the scale factor K in the multidimensional scaling down method on the evaluation accuracy. The approximate values are very similar to the exact one when K is sufficiently smaller than S, while the estimation deteriorates for $K \geq S$. In fact, the critical points of this method are the computation of the scaled number of spares $\hat{S}$ in Equation (4.11) and the allocation of these spares, which make the scaled model quite different from the original model as K approaches S.

The formulation and solution of the Spares Allocation Problem (SAP) is one of the main achievements of this thesis. The mathematical structure of the problem has been investigated to build an efficient exact algorithm for optimally allocating the spares. Two assumption on the cost structure of the problem leads to prove properties of the cost function that in turns allow to design a new efficient branch and bound procedure. The lower bound is obtained by solving a reduced problem with convex objective function, solvable at optimally very efficiently. A new fast heuristic algorithm is also developed to find a feasible allocation within small computation time.

Both algorithms are evaluated by using practical and realistic data from the Italian airport corrective maintenance context. Computational experiments demonstrate that the branch and bound technique is able to optimally solve almost all tested instances within reasonable computation time. The heuristic algorithm finds quite good solutions within very limited computation time, thus being a promising approach for finding feasible solutions to difficult instances.

Moreover we have analyzed several cost structure scenarios and we have observed that the transshipment cost is often comparable with the holding cost and therefore it cannot be neglected in the solution of the problem.

For small holding cost values the number of spares allocated by the heuristic algorithm is often smaller than the optimal value, while for high holding costs it always allocate a number of spares optimal or strictly larger than the optimal solution. As a consequence of this behavior the performance of the proposed heuristic algorithm depends on the holding cost. and therefore the number of spares allocated by ISA is always the same for all scenarios. A consequence of this behavior is that the gap between ISA and the optimum is influenced by the holding cost. When the holding cost of an item increases from 200 to 350 the error decreases from 15% down to 8%. For larger holding costs the error increases regularly until the maximum of 32%, achieved for a holding cost equal to 1200. Smaller errors are attained when the number of spares allocated by ISA is approximately the optimal one and the error only depends on the warehouses to which they are allocated.

Similarly when varying only the emergency costs, the number of spares allocated by ISA is the same for all scenarios since this value does not depend on the spare emergency cost. The gap between ISA and the optimum depends therefore on the emergency cost, even if the emergency cost has a smaller influence on the error of ISA with respect to the holding cost.

As a concluding observation, computational experiments show that the overall behavior of ISA is acceptable as an initial solution for subsequent optimization. In general, the performance of ISA depends on the specific holding and emergency costs being considered and, therefore, it may be quite erratic. BB algorithm seems to be more promising, since it finds the proven optimum within acceptable computation time for all tested instances.

6.2. Direction for future research

The methods proposed in this thesis could be implemented in the Italian corrective maintenance context in order to effectively support the logistics company in its tactical planning. Furthermore, this tactical planning instruments should be able to manage inventory decisions in a large network consisting of different regional groups. Moreover the ideas proposed in this chapter could be applied

to manage the maintenance of different critical infrastructures, such as medical equipments in hospitals or communication or energy distribution networks and so on.

To achieve this goals, a number of issues remain that need further development.

Future research should be focused on the development of faster exact methods and effective metaheuristics for the solution of large and difficult instances.

Anyhow, extensive tests should be made on the branch and bound actually proposed for solving SAP. In fact, is is essential in developing and understanding of how optimal policy responds to the key cost and service factors in the system. This insight is essential in determining subsequent directions for model and algorithm development. Subsequent research in fact should investigate the ideas suggested by the above exploration of the optimal stock allocation, in order to use them to develop an efficient algorithm for solving the stocking problem discussed in this thesis. The latter algorithm should be compared in terms of computational effort and accuracy with the heuristic algorithms suggested in literature as the best ones.

Further research should address also not only methodological studies, but also modeling tasks. In fact in this thesis we attempt to increase the understanding of the properties, characteristics and methodologies of the transshipment problem in single echelon context, but there still exist rich research opportunities either for considering more complex systems with more echelons, items and locations or for integrating different planning needs, e.g. strategic and tactical planning or tactical and operational planning. Some extensions are pointed out as follows.

The discussed model could be for example enriched with other operational details, e.g. different transportation modes, maintaining substantially the same basic mathematical structure, but making the problem larger and more difficult, therefore the knowledge of SAP and the development of evaluation techniques for difficult instances, may be helpful in applying the same techniques in quite different contexts.

Especially the approximate evaluation techniques could be adapted to more flexible management policies, e.g. transshipments thresholds, different transshipment policies, multiple echelon contexts, thus evaluating several alternative tactical policies. In fact when dealing with more than two locations for example some alternative transshipment policies and priority rules may be taken into account.

Finally tactical planning could be realized together operational planning, e.g. integrating the inventory and transportation problem.

Appendices

Markov chain theory

In this appendix we want just recall the basic notions about Markov chain, especially pointing our attention on theoretical consideration around equilibrium probabilities and equilibrium distribution.

Stochastic processes

A stochastic process is a model of a quantity that evolves over time, just like a dynamical system, but influenced by random factors. In the study of inventory systems stochastic processes are used to model demands and supplies that are affected by unpredictable events, as well as key variables such as the inventory position.

Formally a stochastic process is a collection of random variables $\mathbf{X} = \{X(t) : t \geq 0\}$, where $X(t)$ denotes the quantity of interest at time t. Time may be discrete, where t ranges over the non negative integers, or continuous. Each $X(t)$ itself may be discrete or continuous or a vector of such variables. To define a stochastic process $\mathbf{X}$ means to specify the joint range and probability distributions of all these random variables. Information is revealed gradually over time, this notion is called a filtration.

A sample path of $\mathbf{X}$ describes a realization of all random variables $X(t)$. Thus a sample path is a definite function of time.

In principle, the ranges of the $X(t)$ may be different, but in most practical cases the $X(t)$ share a common range, denoted S, called the state space of $\mathbf{X}$. An element of S is a state. $X(t)$ itself is referred to as state variable.

Classification

The time parameter t can be continuous (real) or discrete (integer). Depending on this choice, we say $\mathbf{X}$ is a continuous-time process or discrete-time process. The

specification of time is a basic modeling choice. In some cases there is a natural, predetermined time scale; observations and decisions may take place at regular scheduled points in time or they may occur continuously. Often the choice is more of convenience: one type may be easier to analyze than the other.

Regardless of how we model time, the random variables $X(t)$ themselves may be discrete or continuous. In such cases, respectively, $\mathbf{X}$ is called a discrete state or a continuous state process.
There are four possible combinations for time and space characteristics.

A continuous state process $\mathbf{X}$ can model the quantity (demand or inventory) of an infinitely divisible product. A discrete state process in contrast models a discrete item.

In continuous time a jump process is one which sample paths are all step functions, i.e. constant except for certain distinct time points. Both the time points and the jump sizes may be random. A jump process then has discontinuous sample paths. A common assumption is for simple discontinuities: all sample paths are right continuous. That is, there is a jump at time u, this jump is included in $X(u)$, but not in $X(t)$ with $t < u$. Also each sample path is assumed to have left limits, that is the limit $\lim\{X(t) : t \to u, t < u\}$ exists for all u.
A process with non decreasing sample paths is called an accumulation process. Usually demand is modeled as such a process ($D(t)$ is the cumulative demand up to time t).
A jump-accumulation process with $X(0) = 0$, where every jump size is precisely 1 is a counting process, which, then, is a continuous time, discrete state process: each sample path is a non decreasing step function, starting at 0, with unit steps. A counting process models events occurring randomly over time; $X(t)$ counts the number of events during time interval $(0, t]$.

An important distinction, both in continuous and in discrete time, is whether or not a process changes predictably over time. If so, the process is nonstationary, otherwise it is called stationary.
More precisely a process $\mathbf{X}$ is stationary if a shift in the time axis leaves its probability law unchanged. That is, the $X(t)$ all have the same marginal distribution, and for any fixed $u > 0$, the pair $[X(t), X(t+u)]$ has the same distribution for all t. Likewise, the distribution of any group of three or more variables remains constant over time. In particular a stationary process has $E[X(t)] = E[X(0)]$, with $t \geq 0$. Of course or any give sample path the quantity $X(t)$ typically changes with t. Stationarity is a property of distributions not sample paths.

Here is a related definition: for any fixed $u > 0$, let $X_u(t) = X(t + u) - X(t)$ and $\mathbf{X}_u$ defined as $\{X_u(t) : t \geq 0\}$. $X_u(t)$ is called increment of $\mathbf{X}$ over the interval

$(t, t+u]$.
The process $\mathbf{X}$ has stationary increments if $\mathbf{X}_u$ is stationary for all $u > 0$.
This is the natural concept of time invariance for an accumulation process. It implies that $E[X(t)] = E[X(1)]t$, $t \geq 0$.
If the cumulative demand process has stationary increments, then demand during any week has the same distribution.

Among stationary processes $\mathbf{X}$, some have an important property, ergodicity. Its major consequence is worth mentioning: let us consider any sample path and calculate the long run frequency distribution of $X(t)$ on that path. For example, if $X(t)$ is integer valued, we are interesting in the computation of of the long run fraction of time $X(t)$ spends on each of the integers. If $\mathbf{X}$ is ergodic, then this frequency distribution is identical to the probability distribution of $X(0)$.

To appreciate this property, let us now consider a stationary $\mathbf{X}$ that is not ergodic: time is discrete and the state space is $S = \{0, 1\}$. For $X(0)$, each value has probability $\frac{1}{2}$. Then for all $t > 0$, $X(t) = X(0)$. Thus, each sample path is either $\{0, 0, 0, \ldots\}$ or $\{1, 1, 1, \ldots\}$. Therefore, the proportion of 0's is never $\frac{1}{2}$, which is $Pr\{X(0) = 0\}$. Here of course the $X(t)$ are not independent.

Turning now on a different property, a process $\mathbf{X}$ has independent increments if its increments over disjoint time intervals are independent. This implies that $X(t)$ and $X_u(t)$ are independent, for all t and u. That is the current value provides no information about future increments. This property is especially relevant for an accumulation process, e.g. the demand. In such a case the demands during different weeks are independent random variables.

Markov Processes

The notion of what is nowadays called a Markov process was devised by the Russian mathematician A.A. Markov when, at the beginning of the twentieth century, he investigated the alternation of vowels and consonants in Pushkin's pome Onegin. He developed a probability model in which the outcomes of successive trials are allowed to be dependent on each other such that each trial depends only on its immediate predecessor. This model, being the simplest generalization of the probability model of independent trials, appeared to give an excellent description of the alternation of vowels and consonants and enabled Markov to calculate very accurate estimate of the frequency at which consonants occur in Pushkin's poem. A Markov process allows us to model the uncertainty in many real world systems that evolve dynamically in time. The basic concept of a Markov process are those of state and state transition. The state constitutes a full description of the system at any point in time. In a

deterministic system, the state summarizes all the available information relevant to predicting the future evolution of the system. Markov introduced a similar concept in stochastic systems.

Let us now consider time t. We observe the current value $X(t)$ and also we have recorded the realizations of past values, i.e. $X(s)$ for $0 \leq s < t$. We are interested in some future value, $X(t+u)$ for $u > 0$. Now we canot predict $X(t+u)$ perfectly, but the information we have may tell us something on it. The question is, "how much of that information is really salient?". If $\mathbf{X}$ is a Markov process, then $X(t)$ itself embodies all the relevant information, the other $X(s)$ add no further information. It is enough to condition on the present, the future is independent of the past. For this reason $X(t)$ is called the state variable. Therefore the following holds.

$$\begin{aligned} &[X(t+u)|\{X(s) : 0 \leq s \leq t\}] \\ &[X(t+u)|X(t)] \end{aligned} \tag{A.1}$$

The first one conditions on all the past realizations X(s), while the second conditions only on the current state. For a Markov process these variables have the same distribution, for all t and u.

The dynamics of a Markov process can be expressed in relatively simple terms. For instance in the discrete time case, it sufficies to specify the conditional distributions of $[X(t+u)|X(t)]$ for all t. The initial conditions give the distribution of $X(0)$.

A Markov process X is time homogeneous if the conditional distribution of $[X(t+u)|X(t)]$ remains constant over t, for all fixed u. That is, the rules governing changes from present to future, i.e. the dynamics of the process, remain constant over time.

The simplest kinds of Markov processes are called Markov chains.

Discrete time Markov chains

A Markov chain is a discrete-time, discrete-state,time-homogeneous Markov process. Thus the state space S is countable.Therefore, a discrete time Markov chain is a stochastic process which is a random sequence, in which the dependency of the successive events goes back only one unit at a time. In other words, the future probabilistic behavior of the process only depends on the present state of the process and is not influenced by its past history. This is the above described Markovian property. As time passes, $\mathbf{X}$ jumps from state to state within S, such jumps are called transitions. The one step transition probabilities are the numbers

$$t_{ij} = Pr\{X(t+1) = j|X(t) = i\}$$

with $i, j \in S$.

Because $\mathbf{X}$ is time homogeneous, these do not depend on t. By the Markov property the t_{ij}, together with the knowledge of the probability distribution of the initial state $X(0)$, fully specify the dynamics of $\mathbf{X}$. By definition, the t_{ij} satisfy

$$\begin{aligned} t_{ij} &> 0 \\ \textstyle\sum_{j \in S} t_{ij} &= 1 \end{aligned} \tag{A.2}$$

In applications of Markov chains the art is.

- To choose the state variables such that the Markovian property holds.
- To determine the one step transition probabilities t_{ij}.

Once this modeling step is done, the rest is simply a matter of applying the theory.

Let us collect these probabilities in a matrix $P = (t_{ij})$. Each row i represents the current state and the columns j correspond to the possible subsequent states. If the state space S is infinite, then P is an infinite matrix. P is called the transition probability matrix of $\mathbf{X}$. Let $\mathbf{e}$ denote a column vector of ones, the conditions above can be written.

$$\begin{aligned} P &\geq 0 \\ P\mathbf{e} &= \mathbf{e} \end{aligned} \tag{A.3}$$

We now show that the one step transition probabilities determine the probability of going from state i to state j in the next n steps.
The n-step transition probabilities are defined by

$$t_{ij}^{(n)} = Pr\{X(n) = j | X(0) = i\}$$

for $i, j \in S$ for any n. Obviously $t_{ij}^{(1)} = t_{ij}$.

Theorem 10 *For all* $n, m = 0, 1, 2, \ldots$

$$t_{ij}^{(n+m)} = \sum_{k \in S} t_{ik}^{(n)} t_{kj}^{(m)} \tag{A.4}$$

for $i, j \in S$.

Proof. A formal proof is as follows. By conditioning on the state of the Markov chain at time $t = n$, we find.

$$\begin{aligned} Pr\{X(n+m) = j | X(0) = i\} &= \textstyle\sum_{k \in S} Pr\{X(n+m) = j | X(0) = i, X(n) = k\} Pr\{X(n) = k | X(0) = i\} \\ &= \textstyle\sum_{k \in S} Pr\{X(n+m) = j | X(n) = k\} Pr\{X(n) = k | X(0) = i\} \\ &= \textstyle\sum_{k \in S} Pr\{X(m) = j | X(0) = k\} Pr\{X(n) = k | X(0) = i\} \end{aligned} \tag{A.5}$$

The second equality uses the Markovian property and the last equality uses the assumption of time homogeneity. □

The theorem states that the probability of going from i to j in $n+m$ steps is obtained by summing the probabilities of the mutually exclusive events of going first from state i to some state k in n steps and then going from state k to state j in m steps.
Hence the n step transition probabilities can be recursively computed from the one step transition probabilities t_{ij}. In fact, the $t_{ij}^{(n)}$ are the elements of the n-fold matrix product P^n.
To aid modeling and intuition, it is useful to represent S and P by a directed graph called state transition diagram. There is a node for each state in i in S. There is an arc from node i to node j when $t_{ij} > 0$. Thus the arcs indicates which transitions can occur.
When S is infinite we cannot draw the whole graph. Still there may be a pattern or structure in P that can be indicated graphically.

A useful Markov chain model is the model with one or more absorbing states. A state is absorbing if the process cannot leave this state once it entered this state, hence

$$t_{ii} = 1.$$

Let us now consider a Markov chain $\{X(n)\}$ for which the state space S is finite and there is some state r such that for each state $i \in S$ there is an integer n such that $t_{ir}^{(n)} > 0$.
What is the mean return time from state r to itself? Let $\tau = \min\{n \geq 1 | X(n) = r\}$. To calculate $\mu_{rr} = E[\tau | X(0) = r]$ we need the mean visit times μ_{ir} for each state $i \neq r$.
By conditioning on the next state after state r, the following holds

$$\mu_{rr} = 1 + \sum_{j \in S, j \neq r} t_{rj} \mu_{jr}.$$

The μ_{ir} are found by solving a system of linear equations. Let us number the states as $1, 2, \ldots, N$ and let state r be numbered as N.

Theorem 11 *The mean visit times μ_{iN} for $i \neq N$ are the unique solution to the linear equations*

$$\mu_{iN} = 1 + \sum_{j=1}^{N-1} t_{ij} \mu_{jN} \tag{A.6}$$

for $i = 1, \ldots, N-1$.

Proof. Tijms [84] to prove that the linear equations have a unique solution uses the trick of making state N absorbing for a modified Markov chain. Let $\overline{P}$ be the Markov matrix obtained by replacing the Nth row in the matrix P by the vector $(0, 0, \ldots, 1)$. The mean first passage times μ_{jN} for $j = 1, \ldots, N-1$ are not changed by making state N absorbing. Let us now denote by Q the $(N-1)x(N-1)$ submatrix that results by omitting the Nth row and the Nth column in the matrix P. Let the vectors $\mu = (\mu_{1N}, \ldots, \mu_{N-1,N})$ and $\mathbf{e} = (1, \ldots, 1)$. Then we can write A.6 in matrix notation as

$$\mu = \mathbf{e} + Q\mu.$$

Since state N is absorbing for the Markov matrix $\overline{P}$, the following holds $q_{ij}^{(n)} = \overline{t}_{ij}^{(n)}$ with $i, j = 1, \ldots, N-1$, where the $q_{ij}^{(n)}$ and $\overline{t}_{ij}^{(n)}$ are the elements of the n-fold matrix products Q^n and $\overline{P}^n$.
State N can be reached from each starting state $i \neq N$ under the Markov matrix $\overline{P}$. Further state N is absorbing under $\overline{P}$. this implies that

$$\lim_{n\to\infty} \overline{t}^{(}n)_{ij} = 0$$

for all $i, j = 1, \ldots, N-1$. Hence for the above reasoning

$$\lim_{n\to\infty} Q^n = 0.$$

Therefore by a standard result from linear algebra, it now follows that the unique solution

$$\mu = (I - Q)^{-1}\mathbf{e}$$

exists. □

Types of transitions

Many applications of Markov chains involve chains in which some of the states are absorbing and other states are transient.

The first passage time probability $f_{ij}^{(n)}$ be defined by

$$f_{ij}^{(n)} = Pr\{X(n) = j, X(k) \neq jfor1 \leq k \leq n-1 | X(0) = i\}$$

for $i, j \in S$. Next, let us denote as $f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$, which denotes the probability that the process ever makes a transition into state j when the process starts in state i.

A state i is said to be transient if $f_{ii} < 1$ and is said to be recurrent if $f_{ii} = 1$.

Theorem 12 *Suppose that state j is transient. Then, for any state $i \in I$*

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

A proof for this theorem is given in [84]. We can consider these kind of transitions also in terms of subsets of the state space S. We have.

- State j is reachable from state i if, in the state transition diagram there is a path from i to j. States i and j communicate if each is reachable from the other. In terms of P, j is reachable from i if $(t_{ij}^u) > 0$ for some $u > 0$, where t_{ij} are the one step transition probability recorded in the Markov matrix P.
  Clearly we can partition S into subsets S_k, such that all the states within each subset communicate, but states in different subsets do not.
  Also we say that each subset is reachable from another if any state in the second subset is reachable from any state in the first.
- Therefore we say that a subset S_k is transient if some other subset is reachable from it. Otherwise S_k is recurrent. Hence, if $\mathbf{X}$ starts in a transient subset, sooner or later $\mathbf{X}$ must leave it, never to return. Conversely, if $\mathbf{X}$ starts in a recurrent subset, it stays there forever.
- A Markov chain is reducible if there is more than one recurrent subset. Otherwise it is irreducible.
- The long term behavior of a reducible chain thus depends crucially on its initial conditions: starting in one recurrent subset, $\mathbf{X}$ can never reach another one.
- As we will see an irreducible Markov chain has a unique stationary probability vector.

A question that is important to be addressed is whether the n-step probabilities $t_{ij}^{(n)}$ always have a limit as $n \to \infty$. The answer is negative. If we consider a Markov chain with state space $S = \{1, 2\}$ and one step transition probabilities t_{ij} with $t_{12} = t_{21} = 1$ and $t_{11} = t_{22} = 0$. Therefore the n step transition probabilities $t_{ij}^{(n)}$ alternate between 0 and 1 and hence have no limit as n tend to infinity. The reason is the periodicity in this Markov chain example. In what follows we show that the Cesaro limit of the n step transition probabilities always exists. Then by following Tijms treatment we analyze the effect of the initial state on the Cesaro limit of the n step transition probabilities and we describe when this limit does not depend on the initial state. Such elements are necessary for granting the existence of a unique equilibrium distribution for the Markov chain.

Theorem 13 *For all $i, j \in S$, $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)}$ always exists and*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} t_{jj}^{(k)} = \begin{cases} \frac{1}{\mu_{jj}} & \text{if state } j \text{ is recurrent} \\ 0 & \text{if state } j \text{ is transient} \end{cases} \tag{A.7}$$

Here μ_{jj} denote the mean recurrence time from state j to itself.
Also

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}t_{ij}^{(k)}=f_{ij}\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}t_{jj}^{(k)} \tag{A.8}$$

for any $i,j\in S$, where f_{ij} is the probability that the process ever makes a transition into state j when the process starts in state i.

Proof. For a transient state j it is possible to prove that $\lim_{n\to\infty}t_{ij}^{(n)}=0$ for all $i\in S$. By using the result that the Cesaro limit is equal to the ordinary limit whenever the latter limit exists the theorem follows for transient states j.
Let us now consider a recurrent state j, therefore $f_{jj}=1$ holds. The times between successive visits to state j are independent and identically distributed random variables with mean μ_{jj}. In other words, visits of the Markov chain to state j can be seen as renewals. If $N(t)$ is their number during the first t transition epochs, by renewal theory (see Tijms [84] lemma 2.2.2) we have that the long run average number of transitions to state j per time unit is given by

$$\lim_{t\to\infty}\frac{N(t)}{t}=\frac{1}{\mu_{jj}}$$

with probability 1 when the process starts in state j.
If I_k is such that

$$I_k=\begin{cases}1 & \text{if the process visits state } j \text{ at time } k\\ 0 & \text{otherwise}\end{cases} \tag{A.9}$$

The following holds

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}I_k=\frac{1}{\mu_{jj}}$$

with probability 1. The latter limit is bounded by 1. Therefore we have

$$E[I_k|X(0)=j]=Pr\{X(k)=j|X(0)=j\}=t_{jj}^{(k)}$$

and

$$\begin{aligned}\tfrac{1}{\mu_{jj}} &= E[\lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{k=1}^{n}I_k|X(0)=j]\\ &= \lim_{n\to\infty}E[\tfrac{1}{n}\textstyle\sum_{k=1}^{n}I_k|X(0)=j]\\ &= \lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{k=1}^{n}E[I_k|X(0)=j]=\lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{k=1}^{n}t_{jj}^{(k)}.\end{aligned} \tag{A.10}$$

By exploiting the useful relation

$$t_{ij}^{(n)}=\sum_{k=1}^{n}t_{jj}^{(n-k)}f_{ij}^{(k)}$$

, averaging this relation over $n = 1, \ldots, m$, interchanging the order of summation and letting $m \to \infty$, the theorem follows. □

Next, we ask under which condition the influence of the initial state does not affect any more the process as time increases. That is, $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} t_{ij}^{(k)}$ does not depend on the initial state $X(0) = i$ for each $j \in S$.

For a finite state Markov chain having no two disjoint recurrent sets in [84] is proved that $f_{ij} = 1$ for all $i \in S$, when j is a recurrent state. Therefore for the above reasoning for such a Markov chain $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} t_{ij}^{(k)}$ does not depend on the initial state i when j is recurrent. This statement is also true for a transient state j, since then the Cesaro limit is always equal to 0.

For an infinite state Markov chain we have to assume the existence of some state r such that $f_{ir} = 1$ for all $i \in S$ and $\mu_{rr} < \infty$.
Substantially, the Markov chain has a regeneration state r that is ultimately reached from each initial state with probability 1 and the number of steps needed to return from state r to itself has a finite expectation.

Therefore, with the assumptions made we have that both for a finite state and an infinite state Markov chain

$$lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} t_{ij}^{(k)}$$

does not depend on the initial state i for all $j \in S$.

Thanks to these elements it will be possible to give the equilibrium distribution of the Markov chain.

The equilibrium distribution

An explanation of the term equilibrium is as follows: starting the process according to the equilibrium distribution leads to a process that operates in an equilibrium mode. Therefore we have.

$$\begin{aligned} &Pr\{X(0) = j\}, j \in S \\ &Pr\{X(n) = j\}, j \in S \end{aligned} \tag{A.11}$$

The proof is based on induction. Let the m-th state probability be $Pr\{X(m) = j\} = \pi_j$ for some $m \geq 0$.

$$\begin{aligned} Pr\{X(m+1) = j\} \quad &= \textstyle\sum_{k\in S} Pr\{X(m+1) = j | X(m) = k\} Pr\{X(m) = k\} \\ &\textstyle\sum_{k\in S} t_{kj}\pi_k = \pi_j \end{aligned} \tag{A.12}$$

with $j \in S$.

Theorem 14 *Assuming the existence of some state r such that $f_{ir} = 1$ for all $i \in S$ and $\mu_{rr} < \infty$, then the Markov chain has a unique equilibrium distribution $\{\pi_j, j \in S\}$, which is equal to*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} t_{ij}^{(k)} = \pi_j$$

independently of the initial state i.

Moreover if $\{x_j, j \in S\}$ is any solution to the equilibrium equations

$$x_j = \sum_{k\in S} x_k t_{kj}$$

for $j \in S$.
Then for some constant c, $x_j = c\pi_j$ for all $j \in S$.

Therefore the equilibrium probabilities π_j are the unique solution to the equilibrium equations together with the normalizing equation

$$\sum_{j\in S} \pi_j = 1.$$

The dynamics of state probabilities

The above analysis is made in stationary condition. Here we give in brief some detail about the dynamics of state probabilities.
For each $t \geq 0$ it is possible to define the row vector $\pi(t)$, which components are

$$\pi_i(t) = Pr\{X(t) = i\}.$$

We call the $\pi(t)$ probability vectors. Evidently, $\pi(t)$ expresses the distribution of $X(t)$ in the form of a vector. The initial vector $\pi(0)$ is part of the specification of $\mathbf{X}$. For $t > 0$, $\pi(t)$ describes $X(t)$ as viewed from just before time t.
The definitions of the $\pi_i(t)$ and t_{ij} imply

$$\pi_i(t+1) = \sum_{j\in S} t_{ij}\pi_j(t)$$

for $i \in S$. In matrix vector notation, this becomes

$$\begin{aligned} \pi(t+1) &= \pi(t)P \\ \Delta\pi(t) &= \pi(t)[-(I-P)] \end{aligned} \tag{A.13}$$

Here, I is the identity matrix. Thus, the $\pi(t)$ satisfy a discrete time linear system. Thus, while $X(t)$ itself jumps unpredictably from state to state, the probabilities describing this behavior evolve in an orderly, predictable fashion.

Continuous time Markov chains

A continuous time Markov chain $\mathbf{X}$ is a discrete state, continuous time, time homogeneous Markov process. This is identical to the earlier definition of a Markov chain except for the continuous time parameter. Just as in the discrete time case, the Markov property expresses that the conditional distribution of a future state given the present state and past states depends only on the present state and is independent of the past.

Continuous case in summary

Here is one way to specify such a process. Suppose we are given.

- A discrete time Markov chain $\mathbf{X}_D$ with state space S and transition matrix P, where each $t_{ii} = 0$.
- A vector $\Theta = (\theta_i)_S$, where $\theta_i > 0$.

The convention $t_{ii} = 0$ is convenient and natural. This convention ensures that the sojourn time in a state is unambiguously defined. If there are no absorbing states, it is no restriction to make this convention. Let us now construct the process $\mathbf{X} = \{X(t) : t \geq 0\}$ as follows: the state space of $\mathbf{X}$ too is S. The sequence of values is precisely that of the discrete time chain $\mathbf{X}_D$. The time $\mathbf{X}$ spends in state i on each visit there, however, is a random variable, distributed exponentially with parameter θ_i. These times are independent of $\mathbf{X}_D$ and of each other.
Therefore on entering state i $\mathbf{X}$ stays there for an exponential amount of time with mean $\frac{1}{\theta}$. Then a transition occurs to another state, according to the probabilities in the matrix P. The discrete time chain X_D is said to be embedded in $\mathbf{X}$, which is a continuous time Markov chain. The above process is a Markov jump process.
Most of the key properties of $\mathbf{X}$ are inherited from the embedded discrete time chain $\mathbf{X}_D$. Furthermore the definitions of recurrent and transient are the same as in the discrete time case.Even if in general the theory of continuous time Markov chains is much more intricate than the theory of discrete time Markov chains. There are very

difficult technical problems. Therefore in the subsequent analysis we follow Tijms approach, which imposes a regularity condition that is not too strong from a practical point of view, but avoids all technical problems.

Another equivalent way to specify the data and to construct $\mathbf{X}$ is the following. Define the matrix $Q = (q_{ij})$, where

$$q_{ij} = \begin{cases} \theta_i t_{ij}, j \neq i \\ \\ -\theta_i, j = i \end{cases} \tag{A.14}$$

Clearly the fact that $P\mathbf{e} = \mathbf{e}$ implies $Q\mathbf{e} = 0$. Q is called the infinitesimal generator of $\mathbf{X}$. Then, $\mathbf{X}$ operates as follows: consider $\mathbf{X}$ is in state i. Over a small increment of time Δt, the probability of jumping to $j \neq i$ is approximately $q_{ij}\Delta t$ and that of staying at i is approximately $1 - \delta_i \Delta t$. Moreover, these probabilities are independent of how $\mathbf{X}$ arrived at i and how long it has been there.

Thus, as long as $\mathbf{X}$ remains in state i, q_{ij} measures the potential of a jump to j, $j \neq i$. It is called the transition rate from i to j.
The transition rates are not probabilities, however for Δt very small $q_{ij}\Delta t$ can be interpreted as the probability of moving from state i to state j within the next Δt time units when the current state is state i. Let us assume that the rates $\theta_i = \sum_{j \neq i} q_{ij}$ are positive and bounded in $i \in S$. In what follows we will understand the reasoning of such an hypothesis. The generator Q thus fully describes the behavior of $\mathbf{X}$. To formulate $\mathbf{X}$, we can specify Q directly, bypassing P and

The flow rate equation method is useful for obtaining the equilibrium distribution of a continuous time Markov chain.

Let $\nu_i = \sum_{j \neq i} q_{ij}$ be the parameter of the exponentially distributed sojourn time in state i. It is assumed that the rates $\theta_i = \sum_{j \neq i} q_{ij}$ are positive and bounded in $i \in S$. Finally let us define the probability $p_{ij}(t) = Pr\{X(t) = j | X(0) = i\}$ with $i, j \in S$, which are the transient transition probabilities.

To ensure that the limits of the $p_{ij}(t)$ are independent of the initial state i and constitute a probability distribution, we need the following assumption. The process $\{X(t), t \geq 0\}$ has a regeneration state r such that

$$\begin{aligned} &Pr\{\tau_r < \infty | X(0) = i\} = 1, \forall i \in S \\ &E[\tau_r | X(0) = r] < \infty \end{aligned} \tag{A.15}$$

where τ_r is the first epoch beyond epoch 0 at which the process $\{X(t)\}$ makes a transition into state r.

In other words, state r will ultimately be reached with probability 1 from any other state and the mean recurrence time from state r to itself is finite. Under this assumption it can be proved that there is a probability distribution

$$\{t_j, j \in S\}$$

such that

$$\lim_{t\to\infty} p_{ij}(t) = t_j$$

for $j \in S$ independently of the initial state i.

The limiting probability t_j can be interpreted as the probability that an outside observer finds the system in state j when the process has reached the statistical equilibrium and the observer has no knowledge about the past evolution of the process. The notion of statistical equilibrium relates not only to the length of the time the process has been in operation but also to our knowledge of the past evolution of the system. Substantially t_j with probability 1 represents the long run fraction of time the process will be in state j independently of the initial state $X(0) = i$.

Theorem 15 *Suppose the rates $\theta_i = \sum_{j\neq i} q_{ij}$ of the continuous time Markov chain $\{X(t)\}$ are positive and bounded in $i \in S$ and that a regenerative state r exists, such that it will ultimately be reached with probability 1 from any other state and the mean recurrence time from state r to itself is finite. Then the probabilities t_j, $j \in S$ are obtained as follows. x_j-s are the unique solution to the linear equations*

$$\begin{aligned} &\nu_j x_j = \textstyle\sum_{k\neq j} q_{kj} x_k, j \in S \\ &\textstyle\sum_{j\in S} x_j = 1 \end{aligned} \tag{A.16}$$

with $j \in S$. Then for some constant c $x_j = ct_j$ for all $j \in S$.

The above linear equations are called the equilibrium equations or balance equations of the Markov process. Note that there is also a normalizing equation. t_j are called the equilibrium probabilities.

A physical explanation of the equilibrium equations can be given by using the obvious principle that over the long run the average number of transition out of state j per time unit is equal to the average number of transitions into state j per time unit. Recall that the t_j is the long run fraction of time the process is in state j and the leaving rate out of state j is ν_j. Therefore in the equilibrium equation the flow out of state j is balanced with the rate into state j. This principle is the flow rate equation method.

More generally for any set A of states with $A \neq I$ the rate out of the set A is

equal to the rate into the set A in equilibrium.

Moreover we have

Theorem 16 *Suppose the rates $\theta_i = \sum_{j \neq i} q_{ij}$ of the continuous time Markov chain $\{X(t)\}$ are positive and bounded in $i \in S$ and that a regenerative state r exists, such that it will ultimately be reached with probability 1 from any other state and the mean recurrence time from state r to itself is finite. Then*

- *The continuous time Markov chain $\{X(t)\}$ has a unique equilibrium distribution $t_j, j \in S$. Moreover*

$$t_t = \frac{\pi_j / \nu_j}{\sum_{k \in S} \pi_k \ \nu_k}$$

for $j \in S$, where $\{\pi_j\}$ is the equilibrium distribution of the embedded Markov chain $\{X(n)\}$.

- *Let $\{x_j\}$ be any solution to*

$$\nu_j x_j = \sum_{k \neq j} x_k q_{kj}$$

with $j \in S$, where the summation of x_j over j is finite. Then for some constant c $x_j = ct_j$ for all $j \in S$.

Transient state probabilities

The computation of transient solutions for Markov systems is a very important issue that arises in numerous problems in queueing, inventory and reliability. As already defined the transient probabilities of a continuous time Markov chain $\{X(t), t \geq 0\}$ are

$$p_{ij}(t) = Pr\{X(t) = j | X(0) = i\}$$

, with $i, j \in S$ and $t > 0$. A first method is the method of linear differential equations. The following theorem is about the Kolmogoroff's forward differential equations.

Theorem 17 *Suppose the rates $\theta_i = \sum_{j \neq i} q_{ij}$ of the continuous time Markov chain $\{X(t)\}$ are positive and bounded in $i \in S$. Then for any $i \in S$*

$$p_{ij}^{'}(t) = \sum_{k \neq j} q_{kj} p_{ik}(t) - \nu_j p_{ij}(t)$$

for $j \in S$ and $t > 0$.

Proof. A sketch of the proof for the finite state space S. Let us fix $i \in S$ and $t > 0$ and consider an interval $(t, t + \Delta t]$ with Δt very small. The following holds

$$\begin{aligned} p_{ij}(i + \Delta t) &= Pr\{X(t + \Delta t) = j | X(0) = i\} \\ &= \textstyle\sum_{k \in S} Pr\{X(t + \Delta t) = j | X(0) = i, X(t) = k\} Pr\{X(t) = k | X(0) = i\} \\ &= \textstyle\sum_{k \in S} Pr\{X(t + \Delta t) = j | X(t) = k\} p_{ik}(t) \\ &\textstyle\sum_{k \neq j} q_{kj} \Delta t p_{ik}(t) + (1 - \nu_j \Delta t) p_{ij}(t) + o(\Delta t) \end{aligned} \tag{A.17}$$

□

A second method for the computation of the state probabilities is the uniformization method. In this treatment the process leaves state i after an exponentially distributed time with mean $\frac{1}{\nu_i}$ and then jumps to another state j with probability t_{ij}. Letting $X(n)$ denote the state of the process just after the $n-th$ state transition, the discrete time stochastic process $X(n)$ is an embedded Markov chain with one step transition probabilities t_{ij}. The uniformization method transforms the original Markov chain with non identical leaving rates into an equivalent stochastic process in which the transition epochs are generated by a Poisson process at a uniform rate. To this aim, choose a number ν such that $\nu \geq \nu_i$ for $i \in S$. Lat us now define a discrete time Markov chain $\{\overline{X}(n)\}$ which one step transition probabilities are given by

$$\bar{t}_{ij} = \begin{cases} \frac{\nu_i}{\nu} t_{ij}, j \neq i \\ \\ 1 - \frac{\nu_i}{\nu}, j = i \end{cases} \tag{A.18}$$

for $i \in S$. Define a Poisson process $N(t), t \geq 0$ with rate ν such that the process is independent of the discrete time Markov chain $\{\overline{X}(n)\}$ and a continuous time stochastic process $\{\overline{X}(t), t \geq 0\}$, which makes transitions at epochs generated by the above Poisson process with rate ν and the state transitions are governed by the discrete time Markov chain $\{\overline{X}(n)\}$. A heuristic way to see that the two processes are identical is as follows. For any $i, j \in S$ with $j \neq i$

$$\begin{aligned} &Pr\{\overline{X}(t + \Delta t) = j | \overline{X}(t) = i\} \quad = \nu \Delta t \bar{t}_{ij} + o(\Delta t) \\ &\qquad = \nu_i \Delta t t_{ij} + o(\Delta t) = q_{ij} \Delta t + o(\Delta t) \\ &= Pr\{X(t + \Delta t) = j | X(t) = i\} \end{aligned} \tag{A.19}$$

for $\Delta t \to 0$. A formal proof is given in [84].

Phase type distribution and its evolutions

In this appendix we give only some details on markov modulated Poisson processes and Interrupted Poisson processes.

Steep distributions

Steep distributions are also called hypo exponential distributions or generalized Erlang distributions. This generalized distribution function is obtained by convolving k exponential distributions with parameter λ. Here we only consider the case where all k exponential distributions are identical. Then we obtain the following density function which is called the Erlang-k distribution.

$$\begin{aligned} f(t) &= \tfrac{(\lambda t)^{k-1}}{(k-1)!}\lambda e^{-\lambda t}, \lambda > 0, t \geq 0, k = 1, 2, \ldots \\ F(t) &= \textstyle\sum_{j=k}^{\infty} \tfrac{(\lambda t)^j}{j!} e^{-\lambda t} \\ &= 1 - \textstyle\sum_{j=0}^{k-1} \tfrac{(\lambda t)^j}{j!} e^{-\lambda t} \end{aligned} \tag{A.20}$$

Flat distribution

The general distribution function is in this case a weighted sum of exponential distributions (compound distribution)

$$F(t) = \int_0^{\infty} (1 - e^{-\lambda t}) dW(\lambda) \tag{A.21}$$

with $\lambda > 0$ and $t \geq 0$, where the weight function may be discrete or continuous. This distribution class correspond to a parallel combination of the exponential dis-

tributions. The density function is called complete monotone due to the alternating signs.

Hyper-exponential distribution

In this case $W(\lambda)$ is discrete. Suppose we have the following given values $\lambda_1, \lambda_2, \ldots, \lambda_k$ and that $W(\lambda)$ has the positive increases $a_1, a_2, \ldots, a_k$ where $\sum_{i=1}^{k} a_i = 1$. For all the other values $W(\lambda)$ is constant. In this case A.21 becomes:

$$F(t) = 1 - \sum_{i=1}^{k} a_i e^{-\lambda_i t}, t \geq 0 \tag{A.22}$$

Its moments may be found with the aid of the Laplace transform [88]. Through it its mean value is $m_1 = \sum_{i=1}^{k} \frac{a_i}{\lambda_i}$.

If $k = 1$ or all λ_i are equal, we get the exponential distribution.

This class of distributions is called hyper-exponential distributions and can be obtained by combining k exponential distributions in parallel, where the probability of choosing the i-th distribution is given by a_i. The distribution is called flat because its distribution function increases more slowly from 0 to 1 than the exponential distribution does.

In practice it is difficult to estimate more than one or two parameters. The most important case is for $n = 2$ ($a_1 = a, a_2 = 1 - a$)

$$F(t) = 1 - a_1 e^{-\lambda_1 t} - a_2 e^{-\lambda_2 t} \tag{A.23}$$

In such a case the probability density function is:

$$f(t) = a_1 \lambda_1 e^{-\lambda_1 t} + a_2 \lambda_2 e^{-\lambda_2 \cdot t} \tag{A.24}$$

The Laplace-Stieltjes transform is:

$$\phi(s) = \frac{a_1 \lambda_1}{s + \lambda_1} + \frac{a_2 \lambda_2}{s + \lambda_2} \tag{A.25}$$

Cox distributions

By combining the steep and flat distributions we obtain a general class of distributions (phase type distributions), which can be described with exponential phase in both series and parallel. They can be analyzed through the theory of Markov processes.

We have already defined in appendix 6.2 the meaning of absorbing state. Let us recall that an absorbing state i is one having $p_{ii} = 1$, therefore $p_{ij} = 0$ and $j \neq i$. Thus once $\mathbf{X}$ arrives in an absorbing state, it never leaves it. Clearly, every absorbing state forms its own subset and each is positive recurrent. An absorbing Markov chain is one where every state is either transient or absorbing; therefore $\mathbf{X}$ will enter an absorbing state sooner or later.

An absorbing chain is irreducible, then when it has precisely one absorbing state. In this case the long run behavior of the chain is obvious and uninteresting: the absorbing state i has $\pi_i = 1$. Its short term (transient behavior, however, can be interesting.

Discrete phase type distribution

Next, we describe a type of probability distribution defined in terms of an absorbing chain. Let us consider a finite state, absorbing irreducible chain $\mathbf{X}$, with a single absorbing state and in case number the states so that the absorbing state comes last.

We use P to denote not the full transition matrix, but rather only the submatrix corresponding to transient states, leaving out the last row and column. P_{all} denotes the full matrix. The following holds

$$P_{all} = \begin{pmatrix} P & (I-P)\mathbf{e} \\ \mathbf{0} & 1 \end{pmatrix} \tag{A.26}$$

Similarly, let $\pi_{all}(t)$ denote the full probability vector and $\pi(t)$ the subvector corresponding to transient states. Therefore we have $\pi(0), 1 - \pi(0)\mathbf{e}$.

$$P_{all}^t = \begin{pmatrix} P^t & (I-P^t)\mathbf{e} \\ \mathbf{0} & 1 \end{pmatrix} \tag{A.27}$$

and

$$\pi_{all}(t) = \pi_{all}(0)P_{all}^t = (\pi(0)P^t, 1 - \pi(0)P^t\mathbf{e}) \tag{A.28}$$

In particular the π(t) satisfy a linear recursion of the same form as

$$\pi(t+1) = \pi(t)P$$

. Let T be the first time the chain enters the absorbing state, that is the time until absorbtion. Let us observe that for each $t \geq 0$ the event $\{T > t\}$ is equivalent to $\{X(t) is in a transient state\}$. Consequently

$$F^0(t) = \pi(t)\mathbf{e} = \pi P^t \mathbf{e} F(t) = 1 - \pi P^t \mathbf{e} \tag{A.29}$$

The initial state vector is the unique vector which gives information on how to restart the state transitions after being arrived in an absorbing state.

Continuous phase type distribution

The continuous phase type distributions are continuous probability distributions analogous to the discrete phase type one.Just as the latter is defined in terms of a discrete time Markov chain, this new distribution is constructed from a continuous time chain.

Let us suppose $\mathbf{X}$ is a continuous time Markov chain with generator Q_{all} and a specified initial vector μ_{all}. The last state is an absorbing state and we partition the data as follows

$$Q_{all} = \begin{pmatrix} -M & M\mathbf{e} \\ \mathbf{0} & 0 \end{pmatrix} \tag{A.30}$$

and

$$\mu_{all}(t) = (\mu, 1 - \mu\mathbf{e}) \tag{A.31}$$

The vector μ describes the initial probabilities of the transient states. As for M, its diagonal is positive, the off-diagonal entries are non positive entries. Again, let T be the time until absorbtion. As before we have

$$F^0(t) = \mu e^{-Mt}\mathbf{e} F(t) = 1 - \mu e^{-Mt}\mathbf{e} \tag{A.32}$$

That is the subvector $\pi(t)$ solves the system of linear differential equations

$$\pi^{'}(t) = \pi(t)(-M)$$

with $t \geq 0$ and initial conditions $\pi(0) = \mu$.

MMPP

The Markovian Arrival Processes (MAP) models are evolutions of the phase type distributions. The latter may be obtained from MAP distributions.
In the phase type distributions we note the memoryless property in restarting the state transitions after being arrived in an absorbing state. It is no important which is the last transient state before arriving in the absorbing state, the initial state vector is in any case the same.
A MAP has as many initial state vector as is the number of transient states in the chain. Any time the absorbing state is reached the system knows which is the last transient state just before being absorbed.
Therefore a MAP is formally defined by two matrices and not anymore by one submatrix for transient state transitions and one initial state vector. The first matrix has the same role as the matrix P or Q in the discrete and continuous case respectively, the second one trace the last transient state before being absorbed and gives information on how to restart the state transitions.

The MMPP is the so-called Markov Modulated Poisson Process and it is a MAP.

The MMPP-2 is commonly used, it has only two transient states. In particular when the absorbing state is reached an arrival happens and the next transient state will be the same before the absorbtion.

Such a system is equivalent to two Poisson processes modulated by a Markov chain.

IPP

Due to its lack of memory the Poisson process is very easy to apply. In some cases, however, the Poisson process is not sufficient to describe a real arrival process as it has only one parameter. Kuczura [56] proposed a generalisation which has been widely used. The idea of generalisation comes from the overflow problem. Customers arriving at the system will first try to be served by a primary system with limited capacity (n servers). If the primary system is busy, then the arriving customers will be served by the overflow system. Arriving customers are routed to the overflow system only when the primary system is busy. During the busy periods customers arrive at the overflow system according to the Poisson process. During the non-busy periods no calls arrive to the overflow system, i.e. the arrival intensity is zero. Substantially, the Interrupted Poisson Process (IPP) is the easiest MMPP, where one of the two Poisson processes modulated by the Markov chain has a null parameter. In fact, the IPP model is used to model on-off sources. The blocking probability is computed using the generalized Erlang loss function. With IPP model [44], the demand at each warehouse (including transshipments) can be adequately characterized by a simple renewal process. The inter arrival time distribution of an IPP is hyperexponential, commonly used in the literature to model high-variability arrival processes [90].

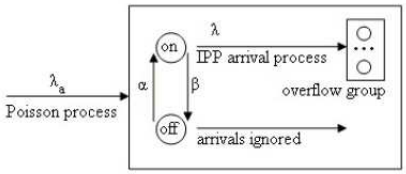

Figure A.1: Sketch of IPP process

As before we have

$$F(t) = 1 - a_1 e^{-\lambda_1 t} - a_2 e^{-\lambda_2 t} \tag{A.33}$$

In such a case the probability density function is:

$$f(t) = a_1 \lambda_1 e^{-\lambda_1 t} + a_2 \lambda_2 e^{-\lambda_2 \cdot t} \tag{A.34}$$

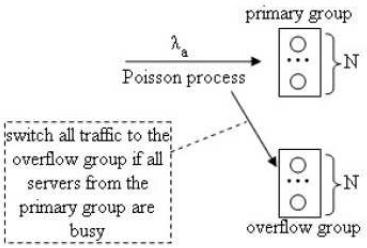

Figure A.2: Sketch of IPP overflow process

The Laplace-Stieltjes transform is:

$$\phi(s) = \frac{a_1\lambda_1}{s+\lambda_1} + \frac{a_2\lambda_2}{s+\lambda_2} \tag{A.35}$$

The generalized Erlang loss function is the following.

$$\begin{aligned}
&\phi(z) = \tfrac{a_1^i\gamma_1^i}{z+\gamma_1^i} + \tfrac{a_2^i\gamma_2^i}{z+\gamma_2^i} \\
&C_j(\xi) = \textstyle\prod_{k=1}^{j} \tfrac{\phi(k\mu_i+\xi)}{1-\phi(k\mu_i+\xi)} \\
&C_0 = 1 \\
&C_{-1} = 1 \\
&p_{s_i}^i = \tfrac{1}{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!}\frac{1}{C_j(0)}}
\end{aligned} \tag{A.36}$$

In order to compute values $p_{s_i}^i$, we therefore need to compute $a_1^i, a_2^i, \gamma_1^i, \gamma_2^i$. By characterizing $a_1^i, a_2^i, \gamma_1^i, \gamma_2^i$ for each queueing system i, substantially the effective demand process to each warehouse is estimated.
To this aim, we compute the first three moments of the effective demand as follows. Let m_1^i, m_2^i, m_3^i be the first three moments of the random variable for the transshipments of the effective flows for warehouse $i$and let r_1^i, r_2^i, r_3^i be the first three moments

of the regular demand flow. We have:

$$
\begin{aligned}
m_1^i &= \tfrac{r_1^i}{\mu_i} p_{s_i}^i \\
m_2^i &= m_1^i + \frac{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!} \frac{1}{C_{j-1}(\mu_i)}}{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!} \frac{1}{C_j(\mu_i)}} m_1^i \\
m_3^i &= 3m_2^i - 2m_1^i + 2(m_2^i - m_1^i) \frac{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!} \frac{1}{C_{j-1}(2\mu_i)}}{\sum_{j=0}^{s_i} \frac{s_i!}{s_i-j!j!} \frac{1}{C_j(2\mu_i)}}
\end{aligned}
\tag{A.37}
$$

In the above expressions the $C_j(\xi)$ are computed as in A.36, but this time the $\phi(z)$ are always referred to the Laplace transform of the inter arrival times of the regular demand and not of the effective one.

Optimization algorithms

Optimization and convexity in brief

The presented concepts are only the main ones useful in the analysis of inventory systems.

Let X be a set, x a n-dimensional vector and $f(x)$ a real valued function defined on X. We say that $x^* \in X$ minimizes f on X or is a global minimizer of f on X, if $f(x^*) \leq f(x)$ with $x \in X$.The value $f(x^*)$ is the minimum of f on X. We write $f(x^*) = \min\{f(x) : x \in X\}$ and $x^* = argmin\{f(x) : x \in X\}$. Similar expression hold for the maximum.

Not every function has a minimizer and a minimum. Every function, however, has an infimum, the greatest lower bound of the range $\{f(x) : x \in X\}$, which could be equal to $-\infty$.

Let us suppose that $X \subseteq I\!R^n$. x^* is a local minimizer and $f(x^*)$ a local minimum of f on X, if $f(x^*) \leq f(x)$ for all $x \in X$ in some neighborhood of x^*.

Let us suppose $X \subseteq I\!R^n$ is an open set and f a real valued function defined on X. Let f be continuously differentiable. Define the gradient of f, denoted $\nabla f(x)$, as the n-dimensional vector of partial derivatives of f evaluated at x; that is

$$\nabla f(x) = [\frac{\partial f(x)}{\partial x_j}]_{j=1}^n$$

In case f is twice continuously differentiable. The Hessian of f, denoted as $\nabla^2 f(x)$, is the $n \times n$ - matrix of second partial derivatives of f evaluated at x:

$$\nabla^2 f(x) = [\frac{\partial^2 f(x)}{\partial x_i \partial x_j}]_{i,j=1}^n$$

This matrix is symmetric.

In the one-dimensional case ($n = 1$), the gradient is just the derivative of f and the hessian is just the second derivative of f.

In general, it is hard to find a global minimizer or even a local one. Optimality condition help us to find them or to recognize them. The first-order optimality condition is as follows: let f be continuously differentiable on the open set X. If x^* is a local minimizer, then

$$\nabla f(x^*) = 0$$

This result suggests a method to find x^*: solve the system of n equations in n unknowns, which nullify the gradient. There is no guarantee, however, that a solution if one exists really solves the problem. To resolve this issue, the concept of convexity is needed.

There are corresponding results for constrained optimization. At first, we focus on the linear case. Let A be a $m \times n$ - matrix, where $m < n$ and let b be a m-dimensional vector, X_0 be an open set in $\mathbb{R}^n$, f be a real valued, continuously differentiable function on X_0. We want to minimize f over X_0, subject to the constraints $Ax = b$ If x^* is a local minimizer of f over X, then there exists an m-dimensional vector μ^*, such that

$$\nabla f(x^*) + A^{'}\mu^* = 0$$

If the rows of A are linearly independent, moreover μ^* is unique.

The components of μ^* are called dual variables or Lagrange multipliers.

Therefore if we have a constraint set of an optimization problem, specified in terms of equality and inequality constraints, we a have a sophisticated collection of optimality conditions, involving the auxiliary variables, that we have defined as Lagrange multipliers. These variables facilitate the characterization of optimal solutions, but also provide valuable sensitivity information, quantifying up to first order the variation of the costs (objective function) caused by variation of problem data. The theory of Lagrange multipliers can be developed in a variety of ways.

- The penalty viewpoint, where we disregard the constraints, while adding to the cost a high penalty for violating them. By then working with "penalized" unconstrained problems and by passing to the limit as the penalty increases.

- The feasible direction viewpoint, which relies on the fact that at a local minimum there can be no cost improvement, when traveling a small distance along a direction that leads to feasible points.

The preceding informal discussion will be made now rigorous. We now consider a problem involving both equality and inequality constraints

$$\begin{aligned} &\min_{x\in X} && f(x) \\ &s.t.: && h_1(x)=0,\ldots,h_m(x)=0 \\ & && g_1(x)\leq 0,\ldots,g_r(x)\leq 0 \end{aligned} \tag{A.38}$$

where f, h_i, g_j are continuously differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$. For any feasible point x the set of active inequality constraints id denoted as $A(x)=\{j|g_j(x)=0\}$. If $j\notin A(x)$, we say that the j-th constraint is inactive at x. We note that if x^* is a local minimum of the inequality constrained problem, then x^* is also a local minimum for a problem identical to the inequality constrained problem except that the inactive constraints at x^* have been discarded. Thus in effect inactive constraints at x^* do not matter: they can be ignored in the statement of the optimality conditions. On the other hand, at a local minimum, active equality constraints can be treated to a large extent as equalities. In particular, if x^* is a local minimum of the inequality constrained problem then x^* is a local minimum for the following corresponding equality constrained problem.

$$\begin{aligned} &\min_{x\in X} && f(x) \\ &s.t.: && h_1(x)=0,\ldots,h_m(x)=0 \\ & && g_j(x)=0,\forall j\in A(x^*) \end{aligned} \tag{A.39}$$

Thus if x^* is regular (i.e. the equality constraints gradients are linearly independent) for the latter problem, there exist Lagrange multipliers $\lambda_1^*,\ldots,\lambda_m^*$ and μ_j^*, $j\in A(x^*)$ such that

$$\nabla f(x^*)+\sum_{i=1}^{m}\lambda_i^*\nabla h_i(x^*)+\sum_{j\in A(x^*)}\mu_j^*\nabla g_j(x^*)=0$$

Assigning zero Lagrange multipliers to the inactive constraints we obtain

$$\nabla f(x^*)+\sum_{i=1}^{m}\lambda_i^*\nabla h_i(x^*)+\sum_{j=1}^{r}\mu_j^* g_j(x^*)=0$$

$$\mu_j^*=0,\forall j\notin A(x^*)$$

More formally we have the so-called Karush-Kuhn-Tucker necessary conditions. Firstly, let us define a feasible vector as regular if the equality constraint gradients $\nabla h_i(x)$ and the active inequality constraint gradients $\nabla g_j(x)$ with $j\in A(x)$ are linearly

independent. Define the Lagrangian function as

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x)$$

Theorem 18 *Let x^* be a local minimum of the problem*

$$\begin{aligned} &\min_{x \in X} f(x) \\ &s.t.: \qquad h_1(x) = 0, \ldots, h_m(x) = 0 \\ &\qquad\quad g_1(x) \leq 0, \ldots, g_r(x) \leq 0 \end{aligned} \tag{A.40}$$

where f, h_i, g_j are continuously differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$, and assume that x^* is regular. Then there exist unique Lagrange multiplier vectors $\lambda^* = (\lambda_i^*, \ldots, \lambda_m^*)$, $\mu^* = (\mu_1^*, \ldots, \mu_r^*)$, such that

$$\nabla L(x^*, \lambda^*.\mu^*) = 0$$

$$\mu_j^* \geq 0, j = 1, \ldots, r$$

$$\mu_j^* = 0, \forall j \notin A(x^*)$$

where $A(x^*)$ is the set of active constraints at x^*. If in addition f, h and g are twice continuously differentiable there holds

$$y^{'} \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) y \geq 0$$

for all $y \in \mathbb{R}^n$ such that

$$\nabla h_i(x^*)^{'} y = 0, \forall i = 1 \ldots, m$$

$$\nabla g_j(x^*)^{'} y = 0, \forall j \in A(x^*)$$

Proof. This proof is present in [12]. Let $v(x) = (h_1(x), \ldots, h_m(x), g_j(x))$ with $j \in A(x^*)$. $v(x)$ may be treated as a set of equality constraints. Here we follow the penalty approach and approximate the original constrained problem by an unconstrained optimization problem that involves a penalty for violation of the constraints. In particular we introduce the objective function

$$F^k(x) = f(x) + \frac{k}{2} \parallel v(x) \parallel^2 + \frac{\alpha}{2} \parallel x - x^* \parallel^2$$

where x^* is the local minimum of the constrained problem and α is some positive scalar. The term $\frac{k}{2} \parallel v(x) \parallel^2$ imposes a penalty for violating the constraints $v(x) = 0$, while the term $\frac{\alpha}{2} \parallel x - x^* \parallel^2$ is introduced for technical proof-related reasons, to ensure that x^* is a strict local minimum of the function $f(x) + \frac{\alpha}{2} \parallel x - x^* \parallel^2$ subject to

$v(x) = 0$. Since x^* is a local minimum, we can select $\epsilon > 0$ such that $f(x^*) \leq f(x)$ for all feasible x in the closed sphere $S = \{x| \parallel x - x^* \parallel \leq \epsilon\}$. Let x^k be an optimal solution of the problem

$$\begin{aligned} \min \quad & F^k(x) \\ s.t.: \quad & x \in S \end{aligned} \tag{A.41}$$

S is compact and therefore an optimal solution for Weierstrass' theorem exists. It is possible to show that the sequence $\{x^k\}$ converges to x^*. We have for all k

$$F^k(x^k) = f(x^k) + \frac{k}{2} \parallel v(x^k) \parallel^2 + \frac{\alpha}{2} \parallel x^k - x^* \parallel^2 \leq F^k(x^*) = f(x^*)$$

and since $f(x^k)$ is bounded over S, we obtain $\lim_{k\to\infty} \parallel v(x^k) \parallel = 0$. Therefore, every limit point $\bar{x}$ of $\{x^k\}$ satisfies $v(\bar{x}) = 0$ Furthermore, by taking the limit as $k \to \infty$, we obtain

$$f(\bar{x}) + \frac{\alpha}{2} \parallel \bar{x} - x^* \parallel^2 \leq f(x^*)$$

Since $\bar{x} \in S$ and $\bar{x}$ is feasible, we have $f(x^*) \leq f(\bar{x})$, which yields $\bar{x} = x^*$. Thus the sequence $\{x^k\}$ converges to x^* and it follows that x^k is an interior point of the closed sphere S for sufficiently large k. Therefore x^k is an unconstrained local minimum of $F^k(x)$ for sufficiently large k. We can work now directly with the corresponding unconstrained necessary optimality conditions, refer to pag. 13 in [12]. From the first order necessary condition, we have for sufficiently large k

$$0 = \nabla F^k(x^k) = \nabla f(x^k) + k\nabla v(x^k)h(x^k) + \alpha(x^k - x^*)$$

Since $\nabla v(x^*)$ has full rank for regularity, the same is true for $\nabla v(x^k)$ if k is sufficiently large. For such k, $\nabla v(x^k)^{'}\nabla v(x^k)$ is invertible and by pre multiplying with $(\nabla v(x^k)^{'}\nabla v(x^k))^{-1}\nabla v(x^k)^{'}$ the previous equation we obtain

$$kv(x^k) = -(\nabla v(x^k)^{'}\nabla v(x^k))^{-1}\nabla v(x^k)^{'}(\nabla f(x^k) + \alpha(x^k - x^*))$$

By taking the limit as $k \to \infty$ and $x^k \to x^*$, we see that $\{kv(x^k)\}$ converges to the vector

$$\delta^* = -(\nabla v(x^*)^{'}\nabla v(x^*))^{-1}\nabla v(x^*)^{'}\nabla f(x^*)$$

By taking the limit as $k \to \infty$ we finally obtain

$$\nabla f(x^*) + \nabla v(x^*)\delta^* = 0$$

proving the first order Lagrange multiplier condition. By using the second order unconstrained optimality condition for the problem we see that the following matrix is positive semi definite for all sufficiently large k and for all $\alpha > 0$:

$$\nabla^2 F^k(x^k) = \nabla^2 f(x^k) + k\nabla v(x^k)\nabla v(x^k)^{'} + k\sum_i v_i(x^k)\nabla^2 v_i(x^k) + \alpha I$$

where I is the identity matrix. Let us fix any y such that $\nabla v(x^*)^{'} y = 0$ and let y_k be the projection of y on the null space of $\nabla v(x^k)^{'}$, that is

$$y^k = y - \nabla v(x^k)(\nabla v(x^k)^{'} \nabla v(x^k))^{-1} \nabla v(x^k) y$$

Since $\nabla v(x^k)^{'} y^k = 0$ and $\nabla^2 F^k(x^k)$ is positive semi definite, we have

$$0 \leq (y^k)^{'} \nabla^2 F^k(x^k) y^k = (y^k)^{'} (\nabla^2 f(x^k) + k \sum_i v_i(x^k) \nabla^2 v_i(x^k)) y^k + \alpha \parallel y^k \parallel^2$$

Since $k v_i(x^k) \rightarrow \delta_i^*$ and from the definition of y^k as projection of y on the null space of $\nabla v(x^k)^{'}$, together with the fact $x^k \rightarrow x^*$ and $\nabla v(x^*)^{'} y = 0$, we have $y^k \rightarrow y$ and obtain

$$0 \leq y^{'} (\nabla^2 f(x^*) + \sum_i \delta_i^* \nabla^2 v_i(x^*)) y + \alpha \parallel y^2$$

Since α can be taken arbitrarily close to 0, we have our result, which is the second order Lagrange multiplier condition. We next must show the assertion $\mu_j^* \geq 0$ for $j \in A(x^*)$. We give a proof of this assertion by using again the penalty approach. Let $h(x) = (h_1(x, \ldots, h_m(x)))$. We introduce the following functions

$$g_j^+(x) = max\{0, g_j(x)\}$$

and the penalized problem

$$\begin{aligned} &\min \mathrm{F}^k(x) \equiv f(x) + \tfrac{k}{2} \parallel h(x) \parallel^2 + \tfrac{k}{2} \textstyle\sum_{j=1}^r (g_j^+(x))^2 + \tfrac{\alpha}{2} \parallel x - x^* \parallel^2 \\ &s.t. : \qquad\qquad x \in S \end{aligned} \tag{A.42}$$

where as before α is a fixed positive scalar, $S = \{x | \parallel x - x^* \parallel \leq \epsilon\}$, and > 0 is such that $f(x^*) \leq f(x)$ for all feasible x with $x \in S$. Note that the function $(g_j^+(x))^2$ is continuously differentiable with gradient $2 g_j^+(x) \nabla g_j(x)$. If x^k minimizes $F^k(x)$ over S, as before it is possible to show that $x^k \rightarrow x^*$ and that the Lagrange multipliers λ_i^* and μ_j^* are given by

$$\lambda_i^* = \lim_{k \rightarrow \infty} k h_i(x^k)$$

$$\mu_j^* = \lim_{k \rightarrow \infty} k g_j^+(x^k)$$

Since $g_j^+(x^k) \geq 0$ we obtain $\mu_j^* \geq 0$ for all j. ∎

One approach for using necessary conditions to solve inequality constrained problems is to consider separately all the possible combinations of constraints being active or inactive.

Convexity

The set $X \subseteq I\!R^n$ is convex if, for every pair of vectors x^1, x^2 in X and every scalar $s \in [0,1]$

$$x = sx^1 + (1-s)x^2 \in X$$

Thus, X is convex if it contains the whole line segment connecting x^1 and x^2.

Let f be a real-valued function defined on the convex set X. The function f is convex if, for every pair of vectors x^1 and x^2 in X and every scalar $s \in [0,1]$, defining $x \equiv sx^1 + (1-s)x^2$ as above

$$f(x) \leq sf(x^1) + (1-s)f(x^2)$$

The expression on the right is an approximation of $f(x)$, obtained by linear interpolation between the values at x^1 and x^2. Therefore, f is convex, when the true value $f(x)$ always lies at or below this approximation. In other words, the graph of f lies below the line segment connecting $[x^1, f(x^1)]$ and $[x^2, f(x^2)]$.

Suppose X is open and convex and the function $f : X \to I\!R$ is continuously differentiable. Then f is convex if and only if for all x and y in X

$$f(x) \geq f(y) + (x-y)\nabla f(y)$$

The right hand side of the above expression is the first order linear approximation of $f(x)$ centered at the point y. Therefore f is convex when the true value $f(x)$ lies at or above the approximation.

Suppose f is twice continuously differentiable. Then f is convex if and only if the Hessian $\nabla^2 f(x)$ is non negative definite for all x.

We care about convexity because it simplifies optimization. In fact, if X is a convex set the following holds.

Suppose f is strictly convex on X. Then, f has at most one, unique local minimizer x^*. Thus if we want to minimize f and we know that f is convex, we need only search for a local minimum, and if f is strictly convex, we know that the optimal point is unique, provided one exists. Convexity also provides necessary and sufficient conditions for local optimality itself. Denote a function f as locally convex if it is convex on some neighborhood of x. If x^* is a local minimizer of f, then f is locally convex at x^*. Let us suppose f continuously differentiable. If f is locally convex at x^* and $\nabla f(x^*) = 0$, then x^* is a local minimizer of f. In the last result, if f is in fact convex everywhere, then x^* is a true global minimizer. The first order necessary condition thud becomes sufficient also.

Convexity is preserved by optimization. Let us suppose that X and Y are convex sets, and $f = f(x, y)$ is convex on $X \times Y$. If $g(x)$ is defined as the minimal value of f for fixed x (i.e. $g(x) = \min\{f(x, y) : y \in Y\}$). Then g is convex on X.

The Lagrangian relaxation method for integer programming

Lagrangian relaxation has grown from a successful but largely theoretical concept to a tool that is the backbone of a number of large scale applications. In the past Fisher wrote an excellent survey of Lagrangian relaxation [35], which in 2004 has been recognized as one of the Ten Most Influential Titles of "Management Science's", and he wrote an "how to do it" exposition, too [36]. This appendix follows closely those studies.

Lagrangian relaxation is based upon the observation that many difficult integer programming problems can be modeled as a relatively easy problem complicated by a set of side constraint. To exploit this observation, we create a Lagrangian problem in which the complicating constraints are replaced with a penalty term in the objective function involving the amount of violation of the constraints and their dual variables. The Lagrangian problem is easy to solve and provide a lower bound (upper bound) for a minimization (maximization) problem on the optimal value of the original problem. It can thus be used in place of a linear programming relaxation to provide bounds in a branch and bound algorithm. The Lagrangian approach offers a number of important advantages over a linear programming relaxation.

Let us formulate the Lagrangian relaxation concept in the following general terms. We take into account a combinatorial optimization problem formulated as an integer program.

$$\begin{aligned} Z = \quad & \min \quad cx \\ & s.t.: \quad Ax = b \\ & \qquad\quad Dx \leq e \\ & \qquad\quad x \geq 0, integral \end{aligned} \tag{A.43}$$

where x is a n-dimensional vector, b an m-dimensional vector and e a r-dimensional vector and all other matrices have conformable dimensions. Let us assume that the following Lagrangian problem is easy to solve relative to problem A.43 and hopefully

in polynomial or pseudo-polynomial time.

$$\begin{aligned} Z_D(u) = \quad & \min \quad cx + u(Ax - b) \\ & s.t.: \quad Dx \leq e \\ & \qquad x \geq 0, integral \end{aligned} \tag{A.44}$$

It is clear that the optimal value of this problem for fixed u at a non negative value is a lower bound on Z.

$$Z_D(u) \leq cx^* + u(Ax^* - b) = Z$$

There are three major questions in designing a Lagrangian relaxation based system.

- Which constraints should be relaxed.
- How to compute good multipliers u.
- How to deduce a good feasible solution to the original problem, given a solution to the relaxed problem.

Fisher remarks that answer to the first question is that the relaxation should make the problem significantly easier but not too easy. For the second question there is a choice between a general purpose procedure called the sub gradient method and smarter methods, which may be better but which are highly problem specific.

Ideally, u should solve the following dual problem.

$$Z_D = \quad \max_u \quad Z_D(u) \tag{A.45}$$

Most schemes for determining u have as their objective finding optimal or near optimal solutions to problem A.45. Such a problem has a number of structural properties that makes it feasible to solve. Let us assume that the set $X = \{x|Dx \leq e, x \geq 0\text{and integral}\}$ of feasible solutions for problem A.44 is finite, so we can represent X as $X = \{x^t, t, 1, \ldots, T\}$. This allow to express problem A.45 as follows.

$$\begin{aligned} Z_D = \quad & \max \quad w \\ & s.t.: \quad w \leq cx^t + u(Ax^t - b), t = 1, \ldots, T \end{aligned} \tag{A.46}$$

The dual of problem A.46 is a linear program with many columns

$$\begin{aligned} Z_D = \quad & \min \quad \textstyle\sum_{t=1}^{T} \lambda_t c x^t \\ & s.t.: \quad \textstyle\sum_{t=1}^{T} \lambda_t A x^t = b \\ & \qquad \textstyle\sum_{t=1}^{T} \lambda_t = 1 \\ & \qquad \lambda_t \geq 0, t = 1, \ldots, T \end{aligned} \tag{A.47}$$

Problem A.47 with λ_t required to be integral is equivalent to problem A.43, although problem A.47 and the linear relaxation of problem A.43 are not in general equivalent problems. Problem A.46 makes it apparent that problem A.44 is the lower envelope of a finite family of linear functions. The function $Z_D(u)$ in problem A.44 has nice properties like continuity and concavity (convexity in case of maximization of the Lagrangian function, due to a minimization original problem). However it differentiable except at points where the Lagrangian problem has multiple optimal solutions. Therefore it is differentiable almost everywhere and it generally is not differentiable at an optimal point. At differentiable points, the derivative of the function $Z_D(u)$ with respect to u is given by $Ax - b$. These observations suggest that it may be fruitful to apply a gradient method for the minimization of $Z_D(u)$ with some adaptation at the points where it is not differentiable. This has been nicely accomplished in a procedure called the sub gradient method. At points where $Z_D(u)$ is not differentiable, the sub gradient method chooses arbitrarily from the set of alternative optimal Lagrangian solutions and uses the vector $Ax - b$ for this solution as though it were the gradient of $Z_D(u)$. Specifically an m-dimensional vector y is called a sub gradient of $Z_D(u)$ at $\bar{u}$ if it satisfies

$$Z_D(u) \leq Z_D(\bar{u}) + y(u - \bar{u})$$

for all u. It is clear that $Z_D(u)$ is sub differentiable everywhere. The vector $(Ax^t - b)$ is a sub gradient at any u for which x^t solves problem A.44. The result is a procedure that determines a sequence of values for u by beginning at an initial point u^0 and applying the formula

$$u^{k+1} = \max\{0, u^k - t_k(b - Ax^k)\}$$

. Recall that u is a non negative vector, therefore the choice of the maximum between 0 and $u^k - t_k(b - Ax^k)$ is due to a projection of $u^k - t_k(b - Ax^k)$ in the convex set defined as follows.

$$M = \{u | u \geq 0, Z_D(u) > -\infty$$

In this formula t_k is a scalar step size and x^k is an optimal solution to problem A.44, the Lagrangian problem with dual variables set to u^k. The non differentiability also requires some variation in the way the step size is normally set in a gradient method. Numerically it is possible to note that if the step size converges to 0 too quickly, then the sub gradient method will converge to a point other than the optimal solution. Therefore the step size in the sub gradient method should converge to 0 but not too quickly. These observations have been confirmed in a result [42] that states that if $k \to \infty$, $t_k \to 0$ and $\sum_{i=1}^{T} t_i \to \infty$, then $Z_D(u^k)$ converges to its optimal value Z_D. Note that these conditions are sufficient but not necessary, in fact, the second condition could be violated but the optimal solution for u could be found same way. A formula for t_k that have proven effective in practice is

$$t_k = \frac{\delta_k(Z_D(u^k) - Z^*)}{\sum_{i=1}^{m}(b_i - \sum_{j=1}^{n} a_{ij}x_j^k)^2}$$

In this formula Z^* is the objective value of the best known feasible solution to problem A.43 and δ_k is a scalar chosen between 0 and 2. Frequently, the sequence δ_k is determined by starting with $\delta_k = 2$ and reducing it by a factor of two, whenever $Z_D(u^k)$ has failed to decrease in a specified number of iterations. The feasible value Z^* initially can be set to 0 and then updated using the solutions that are obtained on those iterations in which the Lagrangian problem solution turns out to be feasible in the original problem A.43. Unless we obtain a u^k for which $Z_D(u^k) = Z^*$, there is no way of proving optimality in the sub gradient method. To resolve this difficulty, the method is usually terminated upon reaching a specified iteration limit. Other procedures that have been used for setting multipliers are called multiplier-adjustment methods. They are heuristics for the dual problem that are developed for a specific application and exploit some special structure of the dual problem in that application. Specifically, in these methods a sequence u^k is generated by the rule $u^{k+1} = u^k + t_k d_k$, where t_k is a positive scalar and d_k is a direction. To determine d_k a finite and usually small set of primitive directions S for which it is easy to evaluate the directional derivative of $Z_D(u)$ is defined. Usually directions in S involve changes in only one or two multipliers. Directions in S are scanned in fixed order and d_k is taken to be either the first direction found along which $Z_D(u)$ increases or the direction of steepest ascent within S. The step size t_k can be chosen either to maximize $Z_D(u^k + t_k d_k)$ or to take us to the first point at which the directional derivatives changes. If S contains no improving direction the algorithm is terminated, which of course can happen prior to finding an optimal solution to problem A.45. With these methods one is usually able to improve on the sub gradient method. However, because its simplicity and robust behavior in a wide variety of applications. it is usually at least the initial choice for setting the multipliers in Lagrangian relaxation.

The Lagrangian relaxation approach can be compared with more traditional linear programming based branch and bound algorithms, by comparing the lower bound obtained by relaxing the integrality requirement on x and solving the resulting linear program. Let $Z_L P$ denote the optimal value of problem A.43 with integrality on x relaxed. Geoffrion [37] stated that $Z_D \geq Z_L P$ for any Lagrangian relaxation. This fact is established by the following sequence of relations between optimization

problems.

$$\begin{array}{lcll} Z_D & = & \max_{u\geq 0} & \{\min_x (cx + u(Ax - b))\} \\ & & s.t.: & Dx \geq e \\ & & & x \geq 0, integral \\ & \geq & \max_{u\geq 0} & \{\min_x (cx + u(Ax - b))\} \\ & & s.t.: & Dx \geq e \\ & & & x \geq 0 \\ (ByLPduality) & = & \max_{u\geq 0} & \max_{v\geq 0} ve - ub \\ & & s.t.: & vD \leq c + uA \\ (ByLPduality) & = & \min_x & cx \\ & & s.t.: & Ax = b \\ & & & Dx \geq e \\ & & & x \geq 0 \\ & = & Z_L P & \end{array} \tag{A.48}$$

Besides showing that $Z_D \geq Z_L P$, the preceding relationships indicate when $Z_D = Z_L P$ and when $Z_D > Z_L P$. The inequality in the sequence of relations connecting Z_D and $Z_L P$ is between the Lagrangian problem and the Lagrangian problem with the integrality constraint relaxed. hence, we can have $Z_D > Z_L P$ only if this inequality holds strictly and hence if the Lagrangian problem is affected by removing the integrality constraint on x. This result shows that we can improve the lower bound by using a Lagrangian relaxation in which the variables are not naturally integral (the continuous and integer solutions can differ). Therefore with careful choice of which constraints to dualize, Lagrangian relaxation can provide results that are significantly superior to LP-based branch and bound. The choice of which constraint to dualize is to some extent an art. Typically several alternative relaxations may be built and evaluated (both empirically and analytically). One way is to begin with an integer programming formulation and select different constraint to dualize. Alternatively, one can begin with some easy to solve model which is close to the problem one wishes to solve and then try to add a set of side constraints to represent those aspects of the

problem of interest which are missing in the simpler model. A Lagrangian relaxation can be obtained by dualizing the side constraints that have been added.

Trust-region and interior affine scaling methods

In section 4.3 for realizing the moment matching between estimated numerical moments and their corresponding analytical expressions, under the hypothesis of hyper-exponential interarrival times, a non linear system of equations have been solved. To this aim an affine scaling trust region approach has been used to solve bound constrained nonlinear systems. The method is a recent iterative one. It combines ideas from the classical trust region Newton method for unconstrained nonlinear equations and an interior affine scaling approach for constrained optimization problems. The method generates feasible iterates and handles the bounds implicity. In subsequent sections the trust region basic theory will be presented and then the specific affine scaling trust region approach will be described. Before presenting the basic theory let us point out that our aim is to consider the problem of the numerical solution of bound-constrained nonlinear systems. It is standard to express these problems as.

$$F(x) = 0, x \in \Omega \tag{A.49}$$

where $F(x) = (F_1(x), F_2(x), \ldots, F_n(x))^T$ and $\Omega = \{x \in R^n | l \leq x \leq u\}$. The vectors $l \in (I\!R \bigcup \infty)^n$ and $u \in (I\!R \bigcup \infty)^n$ are specified lower and upper bounds on the variables such that Ω has a nonempty interior. It's worth noting that a possible approach to solving problem A.49 consists in reformulating it as a bound constrained non linear least squares problem.

$$\min_{x \in \Omega} f(x) = \min_{x \in \Omega} \frac{1}{2} \|F(x)\|_2^2 \tag{A.50}$$

Recently, Bellavia et al. [10] generalized the trust-region strategy for unconstrained systems of nonlinear equations to bound-constrained systems and proposed in [10] a new reliable method for the numerical solution of problem A.49 in its original form.

Trust region methods

Trust region methods were developed at first for unconstrained optimization of smooth functions. You may refer to the excellent books [12] and [68] for a detailed description of this theory. For unconstrained optimization of smooth funtions a powerful collection of algorithms have been developed. We now give a broad description of their main properties and then describe specifically the trust region approach in more detail. All algorithm for unconstrained optimization require the user to supply a starting point, which we usually denote as x_0. The user with knowledge about the application and the data set may be in a good position to choose x_0 to be a reasonable estimate of

the solution. Otherwise, the starting point must be chosen by the algorithm by a systematic approach or in some arbitrary manner.

Beginning at x_0, optimization algorithms generate a sequence of iterates $\{x_k\}_{k=0}^{\infty}$ that terminates when either no more progress can be made or when it seems that a solution point has been approximated with sufficient accuracy. In deciding how to move from one iterate x_k to the next, the algorithms use information about the objective function f at x_k and possibly information from earlier iterates $x_0, x_1, \ldots, x_{k-1}$. They use this information to find a new iterate x_{k+1} with a lower function value than x_k. There exist non monotone algorithms that do not insist on a decrease in f at every step, but even these algorithms require f to be decreased after some prescribed number m of iterations.

There are two fundamental strategies for moving from the current point x_k to a new iterate x_{k+1}: the line search and the trust region.

In the line search strategy, the algorithm chooses a direction p_k and searches along this direction from the current iterates x_k for a new iterate with a lower function value. The distance to move along p_k can be found by approximately solving the following one dimensional minimization problem to find a step length α:

$$\min_{\alpha>0} f(x_k + \alpha p_k) \tag{A.51}$$

By solving A.51 exactly, we would derive the maximum benefit from the direction p_k, but an exact minimization may be expensive and is usually unnecessary. Instead, the line search algorithm generates a limited number of trial step lengths until it finds one that loosely approximates the minimum of problem A.51. At the new point, a new search direction and step length are computed and the process is repeated. The steepest descent direction $-\nabla f_k$ is the most obvious choice for search direction for a line search method. However line search methods may use search directions other than the steepest descent direction. In general any descent direction (i.e. one that makes am angle of strictly less than $\frac{\pi}{2}$) radians with $-\nabla f_k$) is guaranteed to produce a decrease of f, provided that the step length is sufficiently small. We can verify this claim by using Taylor's theorem. We have that.

$$f(x_k + \epsilon p_k) = f(x_k) + \epsilon_k^T \nabla f_k + O(\epsilon^2). \tag{A.52}$$

When p_k is a downhill direction, the angle θ_k between p_k and ∇f_k has $cos\theta_k < 0$, thus

$$p_k^T \nabla f_k = \|p_k\| \|\nabla f_k\| cos\theta_k < 0$$

It follows that

$$f(x_k + \epsilon p_k) < f(x_k)$$

for all positive but sufficiently small values of ϵ. Another important search direction is the Newton direction. This direction is derived from the second order Taylor series approximation to $f(x_k + p)$, which is

$$f(x_k + p) \approx f_k + p^T \nabla f_k + \frac{1}{2} p^T \nabla^2 f_k p = m_k(p)$$

Assuming for the moment that $\nabla^2 f_k$ is positive definite, we obtain the Newton direction by finding the vector p that minimizes $m_k(p)$. By simply setting the derivative of $m_k(p)$ to 0 we obtain the following explicit formula

$$p_k^N = -(\nabla^2 f_k)^{-1} \nabla f_k$$

The Newton direction is reliable when the difference between the true function $f(x_k + p)$ and its quadratic model $m_k(p)$ is not too large. When $\nabla^2 f_k$ is not positive definite, the Newton direction may not even be defined, since $(\nabla^2 f_k)^{-1}$ may not exist. Even when it is defined, it may not satisfy the descent property $\nabla f_k^T p_k^N < 0$, in which case it is unsuitable as a search direction. In these situations, line search methods modify the definition of p_k to make it satisfy the descent condition while retaining the benefit of the second order information.

On the other hand Quasi Newton search directions provide an attractive alternative to Newton's method in that they do not require computation of the Hessian and yet still attain superlinear rate of convergence. In place of the true Hessian $\nabla^2 f_k$ they use an approximation B_k, which is updated after each step to take into account the additional knowledge gained during the step. The updates make use of the fact that changes in the gradient g provide information about the second derivative of f along the search direction.

In the second algorithmic strategy, known as trust region, the information gathered about f is used to construct a model function m_k whose behavior near the current point x_k is similar to that of the actual objective function f. Because the model m_k may not be a good approximation of f when x is far from x_k, we restrict the search for a minimizer of m_k to some region around x_k. In other words, we find the candidate step p by approximately solving the following subproblem.

$$min_p\{m_k(x_k + p)\}, x_k + p \text{ inside the trust region} \tag{A.53}$$

If the candidate solution does not produce a sufficient decrease in f, we conclude that the trust region is too large and we reduce it. Usually the trust region is a ball defined by

$$\|p\| \leq \Delta$$

where the scalar $\Delta > 0$ is called the trust region radius. Elliptical and box shaped trust regions may also be used. The model m_k is usually defined to be a quadratic

function of the form

$$m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p$$

where f_k, ∇f_k and B_k are a scalar, vector, matrix, respectively. As the notation indicates, f_k and ∇f_k are chosen to be the function and gradient values at the point x_k, thus m_k and f are in agreement to the first order at the current iterate x_k. The matrix B_k is either the Hessian $\nabla^2 f_k$ ar some approximation to it.

In a sense, the line search and trust region approaches differ in the order in which they choose the direction and distance of the move to the next iterate. Line search starts by fixing the direction p_k and then identifying an appropriate distance, namely the step length α_k. In trust region, we first choose a maximum distance, i.e. the trust region radius Δ_k and then seek a direction and step that attain the best improvement possible subject to this distance constraint. If this step proves be unsatisfactory, we reduce the distance measure Δ_k and try again. In general, the direction of the step changes whenever the size of the trust region is altered.

The size of the trust region is critical to the effectiveness of each step. If the region is too small, the algorithm misses an opportunity to take a substantial step that will move it much closer to the minimizer of the objective function in the region. If too large, the minimizer of the model may be far from the minimizer of the objective function in the region. In practical algorithms, we choose the size of the region according to the performance of the algorithm during previous iterations.

Let us note that the trust region approach requires us to solve a sequence of subproblems as follows.

$$\begin{array}{ll} min_{p \in Re^n} & m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p \\ s.t.: & \|p\| \leq \Delta_k \end{array} \tag{A.54}$$

in which the objective function and the constraint (which can be written as $p^T p \leq \Delta_k^2$) are both quadratic. If $B_k = \nabla^2 f_k$ we work with a trust region Newton method.

When B_k is positive definite and $\|B_k^{-1} \nabla f_k\| \leq \Delta_k$, the solution of problem A.54 is easy to identify: it is simply the unconstrained minimum $p_k^B = -B_k^{-1} \nabla f_k$ of the quadratic $m_k(p)$. It is the full step. The solution of problem A.54 is not so obvious in other cases, but it can usually be found without too much computational effort. In any case we need only an approximate solution to obtain convergence and good practical behavior. We will point out this fact in the subsequent.

An important observation here is that even if B_k is not positive definite or, more

generally, even if it is not a descent direction, the restricted step p_k improves the objective function, provided $\nabla f_k \neq 0$ and Δ_k sufficiently small. The reason is that $m_k(p_k)$ is smaller than f_k (which is equal to $m_k(0)$) and $f(x_k + p_k)$ is very close to its second order expansion $m_k(p_k)$ when $\|p_k\|$ is small. More specifically, let us show this fact when $B_k = \nabla^2 f_k$, we have for all p with $\|p\| \leq \Delta_k$

$$f(x_k + p) = m_k(p) + o(\Delta_k^2)$$

so that

$$\begin{aligned} f(x_k + p_k) &= m_k(p_k) + o(\Delta_k^2) \\ &= f_k + \min_{\|p\| \leq \Delta_k} \{\nabla f_k^T p + \tfrac{1}{2} p^T \nabla^2 f_k p\} + o(\Delta_k^2) \end{aligned} \tag{A.55}$$

Therefore denoting $\bar{p}_k = -\frac{\nabla f_k}{\|\nabla f_k\|} \Delta_k$ we have

$$\begin{aligned} f(x_k + p_k) &\leq f_k + \nabla f_k^T \bar{p}_k + \tfrac{1}{2} \bar{p}_k^T \nabla^2 f_k \bar{p}_k + o(\Delta_k^2) \\ &= f_k - Delta_k \|\nabla f_k\| + \tfrac{\Delta_k^2}{2\|\nabla f_k\|^2} \nabla f_k^T \nabla^2 f_k \nabla f_k + o(\Delta_k^2) \end{aligned} \tag{A.56}$$

For Δ_k sufficiently small, the negative term $-\Delta_k \|\nabla f_k\|$ dominates the last two terms on the right hand side, showing

$$f(x_k + p_k) < f(x_k)$$

We will state this result more formally in the lemma 20. It can be seen in fact from the preceding relations that a cost improvement is possible even when $\nabla f_k = 0$, provided Δ_k is sufficiently small and f has a direction of negative curvature at x_k, that is, $\nabla^2 f_k$ is not positive semi definite. Thus the preceding procedure will fail to improve the cost only if $\nabla f_k = 0$ and $\nabla^2 f_k$ is positive semi definite, that is x_k satisfies the first and second order necessary conditions. In particular, one can typically make progress even if x_k is a stationary point that is not a local minimum. We are thus motivated to consider a method of the form

$$x_{k+1} = x_k + p_k$$

where p_k is the restricted Newton step corresponding to a suitably chosen scalar Δ_k. Here, for a given x_k, Δ_k should be small enough so that there is cost improvement; one possibility is to start from an initial trial Δ_k and successively reduce Δ_k by a certain factor as many times as necessary until a cost reduction occurs. The choice of the initial trial value for Δ_k is crucial here; if it is chosen too large, a large number of reductions may be necessary before a cost improvement occurs; if it chosen too small, the convergence rate may be too poor. Therefore a reasonable way to adjust the initial trial is as follows.

The strategy for choosing the trust region radius Δ_k at each iteration may be based on the agreement between the model function m_k and the objective function f at previous iterations. Given a step p_k, we define the ratio.

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)} \tag{A.57}$$

the numerator is called the actual reduction and the denominator is the predicted reduction (i.e. the reduction in f predicted by the model function). Let us note that

---

Trust region

Given $\bar{\Delta} > 0$, $\Delta_0 \in (0, \bar{\Delta})$ and $\eta \in [0, \frac{1}{4})$
set $k = 0$.
repeat
  Obtain p_k by (approximately) solving problem A.54;
  Evaluate ρ_k from expression A.57
  if $(\rho_k < \frac{1}{4})$
  then set $\Delta_{k+1} = \frac{1}{4}\Delta_k$;
  else if $(\rho_k > \frac{3}{4})$ and $(\|\rho_k\| = \Delta_k)$
    then set $\Delta_{k+1} = \min(2\Delta_k, \bar{\Delta})$
    else set $\Delta_{k+1} = \Delta_k$
  if $(\rho_k > \eta)$
  then set $x_{k+1} = x_k + p_k$
  else set $x_{k+1} = x_k$
until $x_{k+1} \neq x_k$

Figure A.3: Pseudocode of the algorithm for the trust region approach

since the step p_k is obtained by minimizing the model m_k over a region that includes $p = 0$, the predicted reduction will always be non negative. Hence, if ρ_k is negative, the new objective value $f(x_k + p_k)$ is greater than the current value, so the step must be rejected. On the other hand, if ρ_k is close to 1, there is a good agreement between the model m_k and the function f over this step, so it is safe to expand the trust region for the next iteration. If ρ_k is positive but significantly smaller than 1, we do not alter the trust region, but if it close to 0, we reduce the trust region radius at the next iteration. Here $\bar{\Delta}$ is an overall bound on the step lengths. Let us note that the radius is increased only if $\|p_k\|$ actually reaches the boundary of the trust region.

To turn Algorithm A.5 into a practical algorithm, we need to focus on solving the trust region subproblem A.54. A first step to characterizing exact solutions of it is

given by the following theorem, which shows that the solution p^* of problem A.54 satisfies

$$(B_k + \lambda_k I)p^* = -\nabla f_k \tag{A.58}$$

for some $\lambda \geq 0$. This fact may be seen as a correction of the Hessian matrix (or of its approximation) by a positive semi definite matrix.

Theorem 19 *The vector p^* is a global solution of the trust region problem*

$$\begin{array}{ll} \min_{p \in I\!R^n} m_k(p) & = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p \\ s.t. : & \|p\| \leq \Delta_k \end{array} \tag{A.59}$$

if and only if p^* is feasible and there is a scalar $\lambda \geq 0$ such that the following conditions are satisfied

$$\begin{array}{l} (B_k + \lambda I)p^* = -\nabla f_k \\ \\ \lambda(\Delta_k - \|p^*\|) = 0 \\ \\ (B + \lambda I) \text{ is positive semi definite} \end{array} \tag{A.60}$$

The proof relies on the following technical lemma, which deals with the unconstrained minimizer of quadratics and is particularly interesting in the case where the Hessian is positive semi definite.

Theorem 20 *Let m be the quadratic function defined by*

$$m(p) = g^T p + \frac{1}{2} p^T B p \tag{A.61}$$

where B is any symmetric matrix. Then the following statements are true.

- *m attains a minimum if and only if B is positive semi definite and g is in the range of B. If B is positive semi definite, then every p satisfying $Bp = -g$ is a global minimizer of m.*
- *m has a unique minimizer if and only if B is positive definite.*

Proof.

- The if part. Since g is in the range of B, there is a p with $Bp = -g$. For all $w \in I\!R^n$, we have

$$\begin{array}{ll} m(p+w) & = g^T(p+w) + \frac{1}{2}(p+w)^T B(p+w) \\ \\ & = (g^T p + \frac{1}{2} p^T B p) + g^T w + (Bp)^T w + \frac{1}{2} w^T B w \\ \\ & = m(p) + \frac{1}{2} w^T B w \\ \\ & \geq m(p) \end{array} \tag{A.62}$$

since B is positive semi definite. Hence p is a minimizer of m. For the only if part, let p be a minimizer of m. Since $\nabla m(p) = Bp + g = 0$, we have that g is in the range of B. Also we have $\nabla^2 m(p) = B$ positive semi definite, giving the result.

- For the if part, the same argument as before suffices with the additional point that $w^T Bw > 0$ whenever $w \neq 0$. For the only if part, it is possible to proceed as before to deduce that B is positive semi definite. If B is not positive definite, there is a vector $w \neq 0$ such that $Bw = 0$. Hence from relations A.62, we have $m(p + w) = m(p)$, so the minimizer is not unique, giving a contradiction.

□

Proof theorem 19. Let us assume first that there is $\lambda \geq 0$ such that the conditions A.60 are satisfied. Lemma 20 implies that p^* is a global minimum of the quadratic function

$$\bar{m}_k(p) = \nabla f_k^T p + \frac{1}{2} p^T (B_k + \lambda I) p = m_k(p) + \frac{\lambda}{2} p^T p \tag{A.63}$$

Since $\bar{m}(p) \geq \bar{m}(p^*)$, we have

$$m_k(p) \geq m_k(p^*) + \frac{\lambda}{2}((p^*)^T p^* - p^T p) \tag{A.64}$$

Because $\lambda(\Delta_k - \|p^*\|) = 0$ and therefore $\lambda(\Delta^2 - (p^*)^T p^*) = 0$, we have

$$m_k(p) \geq m_k(p^*) + \frac{\lambda}{2}(\Delta_k^2 - p^T p) \tag{A.65}$$

Hence, from $\lambda \geq 0$, we have $m_k(p) \geq m_k(p^*)$ for all p with $\|p\| \leq \Delta_k$. Therefore p^* is a global minimizer of problem A.59. For the converse, we assume that p^* is a global solution of problem A.59 and show that there is a $\lambda \geq 0$ that satisfies conditions A.60. In the case $\|p^*\| < \Delta_k$, p^* is an unconstrained minimizer of m_k and so

$$\begin{aligned} &\nabla m_k(p^*) = B_k p^* + \nabla f_k = 0 \\ &\nabla^2 m_k(p^*) = B_k \text{ positive semi definite} \end{aligned} \tag{A.66}$$

And therefore properties A.60 hold for $\lambda = 0$. Assume for the remainder of the proof that $\|p^*\| = \Delta_k$. Then the second condition in conditions A.60 is immediately satisfied and p^* also solves the constrained problem

$$\begin{aligned} &{}_k(p) \\ \min \quad &s.t. : \quad m \\ &\|p\| = \Delta_k \end{aligned} \tag{A.67}$$

By applying optimality conditions for constrained optimization to this problem, we find that there is a λ such that the Lagrangian function defined by

$$L(p, \lambda) = m_k(p) + \frac{\lambda}{2}(p^T p - \Delta^2) \tag{A.68}$$

has a stationary point at p^*. By setting $\nabla_p L(p^*, \lambda) = 0$, we obtain

$$B_k p^* + \nabla f_k + \lambda p^* = 0 \Rightarrow (B_k + \lambda I)p^* = -\nabla f_k \tag{A.69}$$

so that conditions A.60 holds. Since $m_k(p) \geq m_k(p^*)$ for any p with $p^T p = (p^*)^T p^* = \Delta_k^2$, we have for such vectors p that

$$m_k(p) \geq m_k(p^*) + \frac{\lambda}{2}((p^*)^T p^* - p^T p) \tag{A.70}$$

By substituting the expression for ∇f_k from equation A.69 into the last expression, we obtain

$$\frac{1}{2}(p - p^*)^T (B_k + \lambda I)(p - p^*) \geq 0 \tag{A.71}$$

The third condition in conditions A.60 follows. It remains to show that $\lambda \geq 0$. Because the first and the third condition are satisfied by p^*, we have from lemma 20 that p^* minimizes $\bar{m}_k$, therefore expression A.64 holds. Suppose that there are only negative values of λ that satisfy the first and the second properties in conditions A.60. Then from expression A.64 that $m_k(p) \geq m_k(p^*)$ whenever $\|p\| \geq \|p^*\| = \Delta$. Since we already know that p^* minimizes m_k for $\|p\| \leq \Delta$, it follows that $B_k p = -\nabla f_k$ and B_k is positive semi definite. Therefore the first and third conditions are satisfied by $\lambda = 0$, which contradicts our assumption that only negative values of λ can satisfy the conditions. □

The key features of this theorem are as follows. The second condition in A.60 is a complementarity condition that states that at least one of the non negative quantities λ and $(\Delta_k - \|p^*\|)$ must be zero. Hence, when the solution lies strictly inside the trust region we must have $\lambda = 0$ and $B_k p^* = -\nabla f_k$ with B_k positive semi definite. In the other cases we have $\|p^*\| = \Delta$ and λ is allowed to take a positive value. Note that

$$\lambda p^* = -B_k p^* = -\nabla m_k(p^*)$$

Strategies for solving subproblem A.54 are based on finding approximate solutions to it, which achieve at least as much reduction in m_k as the reduction achieved by the so-called Cauchy point. This point is simply the minimizer of m_k along the steepest descent direction $-\nabla f_k$, subject to the trust region bound. Descriptions of such approximate techniques may be found in [68].

Although in principle we seek the optimal solution of the subproblem A.54, it is enough for purposes of global convergence to find an approximate solution p_k that lies within the trust region and gives a sufficient reduction in the model. The sufficient reduction may be quantified in terms of the Cauchy point, which we denote by p_k^C and define in terms of the following simple procedure.

Cauchy point calculation

Find the vector p_k^S that solves a linear version of problem A.54

$$p_k^S = _{p \in I\!R^n} (f_k + \nabla f_k^T p) \text{ s.t.: } \|p\| \leq \Delta_k$$

Calculate the scalar $\tau_k > 0$ that minimizes $m_k(\tau p_k^S)$ subject to satisfying the trust region bound that is

$$\tau_k = _{\tau > 0} (m_k(\tau p_k^S)) \text{ s.t.: } \|\tau p_k^S\| \leq \Delta_k$$

set $p_k^C = \tau_k p_k^S$

Figure A.4: Pseudocode of the algorithm for the cauchy point approximation

It is easy to write down a closed form definition of the Cauchy point. For a start, the solution for p_k^S is simply

$$p_k^S = -\frac{\Delta_k}{\|\nabla f_k\|} \nabla f_k$$

To obtain τ_k explicitly, we consider the cases of $\nabla f_k^T B_k \nabla f_k \leq 0$ and $\nabla f_k^T B_k \nabla f_k > 0$ separately. For the former case, the function $m_k(\tau p_k^S)$ decreases monotonically with τ whenever $\nabla f_k \neq 0$ therefore τ_k is simply the largest value that satisfies the trust region bound, namely $\tau_k = 1$. For the other case $m_k(\tau p_k^S)$ is a convex quadratic in τ, so τ_k is either the unconstrained minimizer of this quadratic $\frac{\|\nabla f_k\|^3}{\Delta_k \nabla f_k^T B_k \nabla_k}$ or the boundary value 1. The Cauchy step p_k^C is inexpensive to calculate, mo matrix factorizations are required, and is of crucial importance in deciding if an approximate solution of the trust region sub problem is acceptable. Specifically, a trust region method will be globally convergent if its steps p_k give a reduction in the model m_k that is at least some fixed positive multiple of the decrease attained by the Cauchy step. Formal proofs may be found in [68]. Since the Cauchy point p_k^C provides sufficient reduction in the model function m_k to yield global convergence and since the cost of calculating it is so small, we can look for a better approximate solution of problem A.54. Details may be found in [68].

Now let us describe a technique that uses the characterization A.58 of the subproblem solution, applying Newton's method to find the value of λ which matches the given trust region radius Δ_k in problem A.54. The approximate methods cited above make no serious attempt to find the exact solution of the subproblem A.54. They do, however, make some use of the information in the model Hessian B_k and nice global convergence properties. When the problem is relatively small (n is not too large) it may be worthwhile to exploit the model more fully by looking for a closer approximation to the solution of the subproblem. In this section, we describe an approach for finding a good approximation at the cost of a few factorizations of the matrix B_k. Essentially, the algorithm tries to identify the value of λ for which problem A.58 is satisfied by the solution of problem A.54. The characterization of theorem 19 suggests an algorithm for finding the solution p. Either $\lambda = 0$ satisfies the first and the third expression with $\|p\| \leq \Delta_k$ or else we define

$$p(\lambda) = -(B_k + \lambda I)^{-1} \nabla f_k$$

for λ sufficiently large that $(B + \lambda I)$ is positive definite and seek a value $\lambda > 0$ such that

$\|p(\lambda)\| = \Delta_k$. This problem is one dimensional root finding problem in the variable λ.

An affine scaling trust region approach to bound constrained nonlinear systems

In this section we present a generalization of the trust region idea for unconstrained systems of nonlinear equations to the bound constrained problem A.49 and describe a method which enforces the bounds generating strictly feasible approximations to the solution. Given $x_k \in int(\Omega)$ and a search direction p_k, we look along p_k for the next approximation x_{k+1} within Ω. Let $b(p_k)$ be the stepsize along p_k to the boundary, that is

$$b(p_k) = \begin{cases} \infty & if \Omega = I\!R^n \\ \\ min_i \Lambda_i(p_i) & if \Omega \subset I\!R^n \end{cases} \tag{A.72}$$

where, for each $i = 1, 2, \ldots, n$, $\Lambda_i(p_i)$ is given by

$$\Lambda_i(p_k) = \begin{cases} max\{\frac{l_i - (x_k)_i}{(p_k)_i}, \frac{u_i - (x_k)_i}{(p_k)_i}\} & if (p_k)_i \neq 0 \\ \\ \infty & if (p_k)_i = 0 \end{cases} \tag{A.73}$$

It is clear that if $b(p_k) > 1$, then $x_k + p_k$ is within Ω; otherwise a step back along p_k will be necessary to stay within Ω. Let $\theta \in (0, 1)$ be a fixed constant, $\zeta(p_k)$ be given

by

$$\zeta(p_k) = \begin{cases} 1 & if b(p_k) > 1 \\ \\ max\{\theta, 1 - \|p_k\|\}b(p_k) & otherwise \end{cases} \tag{A.74}$$

and $\alpha(p_k) = \zeta(p_k)p_k$. Then to ensure that the new iterate is strictly feasible with respect to the box constraints, we set $x_{k+1} = x_k + \alpha(p_k)$. Now consider the problem of choosing the search direction p_k. In the context of unconstrained nonlinear systems, if x_k is a very good approximation of a solution, Newton method can be applied and p_k is set equal to the solution p_k^N of the Newton equation. However to define a robust iterative process, Newton method can be incorporated into a globally convergent trust region scheme. In the classical trust region approach, a region around the current iterate x_k is defined. Within such a region, the following quadratic model

$$m_k(p) = \frac{1}{2}\|F_k^{'}p + F_k\| = \frac{1}{2}\|F_k\|^2 + F_k^T F_k^{'}p + \frac{1}{2}p^T F_k^{'} F_k^{'}p$$

is trusted to an adequate representation of the merit function

$$f(x) = \frac{1}{2}\|F(x)\|^2$$

Then the search direction p_k is the vector solution of the subproblem

$$\min_p\{m_k(p) : \|p\| \leq \Delta_k\} \tag{A.75}$$

for some given trust region radius $\Delta_k > 0$. Since the Newton step p_n^N is the global minimum of $m_k(p)$, it is the solution of the above trust region subproblem if $\|p_k^N\| \leq \Delta_k$. When the nonlinear system is constrained, we must take into account that the requirement of strict feasibility can lead to reductions on the chosen step p_k. In particular, if the step direction points to a nearby constraint, an excessively small fraction of p_k should be taken to stay within Ω and this may preclude the convergence of the sequence x_k to a solution of A.49. To prevent this occurrence, the affine scaling mapping has been proposed by Bellavia et al. [10]. Consider the gradient $F^{'}(x)F(x)$ of the merit function f and let $v(x)$ be the vector function with components $v_i(x)$ with $i = 1, \ldots, n$ given by

$$\begin{aligned} v_i(x) &= x_i - u_i \quad && if (F^{'T}(x)F(x))_i < 0, u_i < \infty \\ v_i(x) &= x_i - l_i \quad && if (F^{'T}(x)F(x))_i \geq 0, l_i > -\infty \\ v_i(x) &= -1 \quad && if (F^{'T}(x)F(x))_i < 0, u_i = \infty \\ v_i(x) &= 1 \quad && if (F^{'T}(x)F(x))_i \geq 0, l_i = -\infty \end{aligned} \tag{A.76}$$

Further let $D(x)$ be the diagonal scaling matrix such that

$$D(x) = diag(|v_1(x)|^{(} - \frac{1}{2}), |v_2(x)|^{(} - \frac{1}{2}), \ldots, |v_n(x)|^{(} - \frac{1}{2})) \tag{A.77}$$

Then, let $D_k = D(x_k)$ and consider the elliptical trust region defined by

$$\|D_k p\| \leq \Delta_k$$

Namely, instead of the ball trust region above sub problem, consider the following elliptical trust region sub problem

$$\min_p \{m_k(p) : \|D_k p\| \leq \Delta_k\} \tag{A.78}$$

For this sub problem the Cauchy point is the point that minimizes m_k along the scaled steepest descent direction $d_k = -D_k^{-2} F_k^{'T} F_k$ subject to satisfying the trust region bound, that is

$$p_k^C = \tau_k d_k = -\tau_k D_k^{-2} F_k^{'T} F_k$$

where

$$\tau_{k\,\tau>0} \{m_k(\tau d_k) : \|\tau D_k d_k\| \leq \Delta_k\}$$

The relevance of the used scaling matrix depends on the fact that the scaled steepest descent direction d_k is angled away from the approaching bound. Consequently the bounds will not prevent a relatively large stepsize along d_k from being taken. Thus the bounds have been handled implicitly by the diagonal matrix D_k. For global convergence purpose, it is enough to find vector p_k such that $\alpha(p_k)$ gives a sufficient reduction in the quadratic model m_k. This sufficient reduction can be quantified in terms of the Cauchy point p_k^C. Well known convergence results show that the trial step $\alpha(p_k)$ is required to give a reduction in the model m_k that is at least some fixed multiple of the decrease attained by the Cauchy step at each iteration. Then, taking into account the given constraints, we test if the following condition

$$\rho_k^C(p_k) = \frac{m_k(0) - m_k((p_k))}{m_k(0) - m_k(\alpha(p_k^C))} \geq \beta_1$$

is verified for some fixed constant $\beta_1 \in (0, 1]$. The last condition does not necessarily guarantee a good agreement between the model function m_k and the objective function f. Thus, we require that p_k satisfies the following standard condition

$$\rho_k^f(p_k) = \frac{f(x_k) - f(x_k + \alpha(p_k))}{m_k(0) - m_k(\alpha(p_k))} \geq \beta_2$$

where $\beta_2 \in (0, 1)$.

Scaled trust region Newton method

Let $x_0 \in int(\Omega)$, $\Delta_0 > 0$ be given. Choose $\theta \in (0,1)$, $\beta_1 \in (0,1]$, β_2 and β_3 such that $0 < \beta_2 < \beta_3 < 1$, δ_1 and δ_2 such that $0 < \delta_1 < 1 < \delta_2$
$k = 0$
repeat
  Compute the matrix D_k;
  repeat
    Find $p_k =_{\|D_k p\| \leq \Delta_k} m_k(p)$
    Compute the Cauchy point p_k^C
    Compute $\alpha(p_k)$ and $\alpha(p_k^C)$
    if $(\rho_k^C < \beta_1)$
    then set $p_k = p_k^C$
    Set $\Delta_k^* = \Delta_k$
    Set $\Delta_k = \delta_1 \Delta_k$
  until $(\rho_k^f(p_k) \geq \beta_2)$
  Set $x_{k+1} = x_k + \alpha(p_k)$, $\Delta_k = \Delta_k^*$
  if $(\rho_k^f(p_k) \geq \beta_3)$
  then $\Delta_{k+1} = \delta_2 \Delta_k$
  else $\Delta_{k+1} = \Delta_k$
  $k = k + 1$
until $x_{k+1} = x_k$

Figure A.5: Pseudocode of the algorithm for the trust region approach

Bibliography

[1] Airports Council International (ACI). Worldwide and Regional Forecasts, Airport Traffic 2005-2020, (2005).

[2] Alfredsson, P. and J. Verrijdt. Modeling emergency supply flexibility in a two echelon inventory system. *Management Science* 45, (1999) 1416–1431.

[3] Archibald, T.W., Sassen S.A. E., Thomas L.C. An optimal policy for a two depot inventory problem with stock transfer. *Management Science* 43, (1997) 173183

[4] Axsater,S.. Modelling emergency lateral transshipments in inventory systems. *Management Science* 6, (1990) 1329-1338.

[5] Axsater, S. Continuous review policies for multi-level inventory systems with stochastic demand. *S.C. Graves, A.H.G. Rinnooy Kan, and P.H. Zipkin, editors, Handbook in OR and MS* 4, (1993) 175-197

[6] Axsater,S.. Scaling down multi-echelon inventory problems. *Int. J. Production Economics* 71, (2001) 255–261.

[7] Axsater, S. Evaluation of unidirectional lateral transshipments and substitutions in inventory system. *European Journal of Operational Research* 149, (2003) 438-447.

[8] Axsater,S.. Inventory control. *Springer Verlang* (2006).

[9] Balsamo,S.. Product form queueing networks. *Lecture notes in computer science, Berlin, Performance evaluation white book, Vol. LNCS, Ed Springer-Verlag* 1796, (2000) 377–401.

[10] Bellavia,S.,M. Macconi, B. Morini. An affine scaling trust region approach to bound constrained nonlinear systems. *Applied Numerical Mathematics* 44, (2003) 257-280.

[11] Bellavia,S.,M. Macconi, B. Morini. STRSCNE: A scaled Trust-Region Solver for Constrained Nonlinear Equations. *Computational Optimization and Applications* 28 Issue 1, (2004) 31–50

[12] Bertsekas, D.P.. Nonlinear programming. *Athena Scientific* second edition (2003).

[13] Cachon, G. Competitive supply chain inventory management. *Tayur et al.* (1999)

[14] Cesaro, A., Pacciarelli, D.. Evaluation of peaked lateral transshipment in inventory system subject to a service constraint. *MSOM 2007 conference* (June 18–19,2007).

CesaroPacciarelliPOMS Cesaro, A., D. Pacciarelli. Optimal stock allocation in single echelon inventory systems subject to a service constraint. Orlando May 2009

[15] Cesaro, A., Pacciarelli, D.. Performance assessment for single echelon airport spare part management. *Technical Report University Roma Tre* (2009).

[16] Chao, X., M. Miyazawa, M. Pinedo. Wueueing networks. *Wiley* (1999)

[17] Chevalier, P., N. Tabordon. Overflow analysis and cross-trained servers. *Int. J. Production Economics* 85, (2003) 47–70.

[18] Chevalier, P., J. F. Macq, J. C. Van Den Schrieck. Modeling the Variability in Supply Chains with the Peakedness. *MSOM 2007 conference* (June 18–19,2007).

[19] Chiou, C.C. A simulation study on the effectiveness of transshipment rules and inventory control policies. *37-th International Conference Computers Industrial Engineering* (2007)

[20] Chiou, C.C.. Transshipment problems in supply chain systems: review and extensions. *Supply Chains, Theory and Application, Vedran Kordic* (2008).

[21] Clark, A.J., H.E. Scarf. Approximate solutions to a simple multi-echelon inventory problem. *K. J. Arrow, S. Karlin, and H. Scarf, editors, Studies in Applied Probability and Management Science, Stanford University Press* (1962).

[22] Cohen, M.A., P.R. Kleindorfer, H.L. Lee. Optimal stocking policies for low usage items in multi-echelon inventory systems. *Naval Research Logistics Quarterly* 33, (1986) 17-38.

[23] Cohen, M.A., P.R. Kleindorfer, H.L. Lee, Multi-item service constrained (s, S) policies for spare parts logistics systems. *Naval Research Logistics* 39, (1992) 561-577.

[24] Dada, M. Inventory systems for spare parts. *Ph.D. Dissertation, Sloan School of Management, MIT* (1984)

[25] Dada, M.. A two echelon inventory system with priority shipments. *Management Science* 38, (1992) 1140-1153.

[26] Deloitte. The service revolution in global manufacturing industries, *Deloitte Research* (2006)

[27] Li, D., X.L. Sun. Towards strong duality in integer programming. *J. Glob. Optim* 35, (2006) 255–282.

[28] Li, D., J. Wang, X.L. Sun. Computing exact solution to nonlinear integer programming: Convergent Lagrangian and objective level cut method. *J. Glob. Optim* 39, (2007) 127–154.

[29] Li, D., X.L. Sun. Nonlinear integer programming. *Springer's INTERNATIONAL SERIES* (2006)

[30] ENAV website. http://www.enav.it (2009)

[31] Everett, H. Generalized Lagrange multiplier method for solving problems of optimum allocation of resources, *Operations Research* 11, (1963) 399-417.

[32] Evers, P. T. Hidden benefits of emergency transshipments. *Journal of Business Logistics* 18, (1997) 55-77

[33] Evers, P. T. Filling customer orders from multiple locations: A comparison of pooling methods. *Journal of Business Logistics* 20, (1999) 121-140

[34] Federgruen, A., P. Zipkin. Approximations of dynamic, multilocation production and inventory problems. *Management Science* 30 (1984), 69-84.

[35] Fisher, M.L. The Lagrangian relaxation method for solving integer programming problems. *Management Science* 27, (1981).

[36] Fisher, M.L. An applications oriented guide to Lagrangian relaxation. *Interfaces* 15 (1985), 10-21

[37] Geoffrion, A.M. Lagrangian relaxation and its uses in integer programming. *Mathematical programming study* 2, (1974) 82-114.

[38] Grahovac, J., A. Chakravarty. Sharing and Lateral Transshipment of Inventory in a Supply Chain with Expensive Low-Demand Items. *Management Science* 47, (2001) 579–594

[39] Graves, S. C. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science* 31, (1985) 1247-1256

[40] Gross, D., H. D. Kahn, and J.D. Marsh. Queuing models for spares provisioning. *Naval Research Logistics Quarterly* 24 (1977)521-536.

[41] HeidelBerger, P., P. Shahabuddin, V. Nicola. Bounded relative error in estimating transient measures of highly dependable non-Markovian systems. *ACM Transactions on Modeling and Computer Simulation* 4, (1994) 137-164.

[42] Held, M.H., P. Wolfe, H.D. Crowder. Validation of subgradient optimization. *Mathematical Programming* 6, (1974) 62-88.

[43] Huiskonen, J.. Maintenance spare parts logistics: special characteristics and strategic choices. *Int. J. Production Economics* 71, (2001) 125–133.

[44] Iversen, V.B.. Overflow approximations for non random input. *ITU-T E.524, Telecommunication Standardization Sector of ITU, still valid* (1999).

[45] Iversen, V.B.. Teletraffic Engineering Handbook. *ITU-D SG 2/16 and ITC 2002/09/06* (2002).

[46] Kellerer, H., U. Pferschy, D. Pisinger, Knapsack Problems. *Springer, Berlin* (2004)

[47] Kelly,F.. Reversibility and Stochastic networks. *Wiley, Chirchester* (1994).

[48] Kennedy, W.J., J. Wayne Patterson, and L. D. Fredendall. An overview of recent literature on spare parts inventories. *Int. J. Production Economics* 76, (2002) 201–215.

[49] Kranenburg, A.A., Van Houtum, G.J.. A multi-item spare parts inventory model with customer differentiation. *Working Paper WP 110, Research School Beta, Technische Universiteit Eindhoven* (2004).

[50] Kranenburg, A.A. Spare Parts Inventory Control under System Availability Constraints *Technische Universiteit Eindhoven* (2006)

[51] Kranenburg, A.A., G.J.J.A.N. van Houtum. Effect of commonality on spare parts provisioning costs for capital goods. *International Journal of Production Economics*, 108, (2007) 221-227.

[52] Kranenburg, A.A., G.J.J.A.N. van Houtum Service differentiation in spare parts inventory management. *Journal of the Operational Research Society*, 59, (2008) 946-955.

[53] Kranenburg, A.A., G.J.J.A.N. van Houtum. A new partial pooling structure for spare parts networks. *European Journal of Operational Research*, 199, (2009) 908-921.

[54] Kukreja, A., C.P. Schmidt, D.M. Miller. Stocking decisions for low-usage items in a multilocation inventory system *Management Science* 47, (2001) 1371–1383.

[55] Kukreja, A., C.P. Schmidt. 2005. A model for lumpy demand parts in a multi-location inventory system with transshipments. *Computers and Operations Research* 32, (2005) 2059–2075.

[56] Kukzura, A., D. Bajaj. A method of moments for the Analysis of a switched communication networks performance. *IEEE Transaction on communications* 25 (1977) 185-193.

[57] Kutanoglu, E.. Insights into inventory sharing in service parts logistics. *Computers and Industrial Engineering* 54, (2008) 341–358.

[58] Land, A. H. , A. G. Doig. An Automatic Method for Solving Discrete Programming Problems *Econometrica* 28 (1960), 497-520.

[59] Lee, H.L. A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Science* 33, (1987) 1302-1316

[60] Lee, Y.H., J. W. Jung, Y.S. Jeon. An effective lateral transshipment policy to improve service level in the supply chain. *International Journal of Production Economics* 106, (2007) 115-126.

[61] Matsumoto, J., Y. Watanabe. Individual traffic characteristics of queueing systems with multiple poisson and overflow inputs. *IEEE Transactions on communications* 33 (1985) 1–9.

[62] Muckstadt, J.A. Are multi-echelon inventory methods worth implementing in systems with low demand rate items? *Management Science* 26, (1980) 483-494

[63] Muckstadt, J.A. Analysis and algorithms for service parts supply chain. *Springer series in operations research and financial engineering* (2005)

[64] Muharremoglu, A. , J.N. Tsitsiklis. Echelon base stock policies in uncapacitated serial inventory systems. *http://web.mit.edu/jnt/www/publ.html* (2001)

[65] Nahmias, S. Managing reparable item inventory systems: A review. *Schwarz editor, Multi-Level Production/Inventory Control Systems: Theory and Practice* 16, (1981) 253-278

[66] Needham, P.M. , P.T. Evers. The influence of individual cost factors on the use of emergency transshipments. *Transportation Research E* 34, (1998) 149-160.

[67] Nelson, R.. The mathematics of product form queueing networks. *ACM Computing Surveys* 25, (1993) 339–369

[68] Nocedal, J., S.J. Wright. Numerical optimization. *Springer Series in Operations Research and Financial Engineering* (2000)

[69] Ozdemir, D., Yucesan, E., Herer, Y.T. Multi-Location Transshipment Problem with Capacitated Production and Lost Sales *Proceedings of the Winter Simulation Conference, 2006.* 3-6 Dec. 2006 1470-1476.

[70] Reijnen, I.C., T. Tan, and G.J. Van Houtum. Inventory planning for spare parts networks with delivery time constraints. *working paper, Research School Beta, Eindhoven University of Technology.* Available at http://beta.ieis.tue.nl/

[71] Robinson, L.W. Optimal and approximate policies in multiperiod, multilocation inventory models with transshipments. *Operations research* 38, (1990) 278-295

[72] Sherbrooke, C.C. METRIC: A multi-echelon technique for recoverable item control. *Operations Research* 16, (1968) 122-141.

[73] Sherbrooke, C.C.. Multiechelon inventory systems with lateral supply. *Naval Research Logistics* 39, (1992) 29-40

[74] Sherbrooke, C.C.. Optimal inventory modeling of systems: multi-echelon techniques. *Wiley* (2004).

[75] Slay, F.M. VARI-METRIC: An approach to modeling multi-echelon resupply when the demand process is Poisson with a gamma prior.*Report AF301-3, Logistics Management Institute*, (1984)

[76] Silver, E.A. Inventory allocation among an assembly and its repairable subassemblies. *Naval Research Logistics Quarterly* 19, (1972) 261-280.

[77] Silver, E.A., Peterson, R. Decision systems for inventory management and production. *Wiley*, (1998)

[78] Sleptchenko, A. Integral Inventory Control in Spare Parts Networks with Capacity Restrictions. *PhD dissertation, Universiteit Twente, Enschede, the Netherlands* (2002)

[79] Tagaras, G. Effects of pooling on the optimization and service levels of two-location inventory systems. *IIE Transactions* 21, (1989) 250258.

[80] Tagaras, G., M.A. Cohen. Pooling in two-location inventory systems with nonnegligible replenishment lead times. *Management Science* 38 (1992) 10671083

[81] Tagaras, G. Pooling in multi-location periodic inventory distribution systems. *Omega* 27, (1999)

[82] Taragas, G., D. Vlachos. Effectiveness of stock transshipment under various demand distributions and nonnegligible transshipment times. *Production And operations Management* 11, (2002) 183–198.

[83] Tayur, S. Magazine M., Ganeshan R. Quantitatives models for supply chain management *International Series in Operations Research and Management Science, Kluwer Academic Publishers* 17 (1998).

[84] Tijms, H.C.. A First Course in Stochastic Models. *Wiley* (2003).

[85] Thonemann, U.W., A.O. Brown, W. H. Hausman. Easy quantification of improved spare parts inventory policies. *Management Science* 48, (2002) 1213-1225.

[86] Van Utterbeeck F., H. Wong, D. Van Oudheusden, D. Cattrysse. The effects of resupply flexibility on the design of service parts supply systems. *Transportation Research Part E* 45, (2009) 72-85.

[87] Van Houtum, G.J., W.H.M. Zijm. On the relation between cost and service models for general inventory systems. *Statistica Neerlandica* 54, (2000) 127-147

[88] Virtamo, J. Queuing theory (course). http://www.netlab.tkk.fi/opetus/s383143/kalvot/english.shtml

[89] Vliegen, I.M.H., G.J.J.A.N. Van Houtum. Approximate evaluation of order fill rates for an inventory system of service tools. *International Journal of Production Economics* 118, (2009) 339-351.

[90] Wang, J., R. Srinivasan. Unreliable production-inventory system with hyper-exponential renewal demand processes. *Applied Mathematics and Computation* 166, (2005) 475–488.

[91] Wong, H.,D. Catrysse,D. Van Houdeusden, Inventory pooling of repairable spare parts with non-zero lateral transshipment time and delayed lateral transshipments. *European Journal of Operational Research* 165 (2005) 207–218.

[92] Wong, H.,D. Catrysse,D. Van Houdeusden. Stocking decisions for repairable spare parts pooling in a multi-hub system. *Int. J. Production Economics* 93 (2005) 309–317

[93] Wong, H.,G.J. Van Houtum, D. Catrysse,D. Van Houdeusden. Simple, efficient heuristics for multi-item multi-location spare parts systems with lateral transshipments and waiting time constraints. *Journal of the Operational Research Society* 56 (2005), 1419-1430.

[94] Wong,H., G.J. Van Houtum, D. Cattrysse, D. Van Oudheusden. Multi-item spare part systems with lateral transshipments and waiting time constraints. *European Journal of Operational Research* 171, (2006) 1071-1093

[95] Wong,H., D. Van Oudheusden, D. Cattrysse. Two echelon multi-item spare parts systems with emergency supply flexibility and waiting time constraints. *IEE Transactions* 39, (2007) 1045-1057.

[96] Wong, H., D. Van Oudheusden, D. Cattrysse. Cost allocation in spare parts inventory pooling *Transportation Research Part E: Logistics and Transportation Review* 43, (2007) 370-386.

[97] Yanagi, S. , M. Sasaki. An approximation method for the problem of a repairable item inventory system with lateral supply. *IMA Journal of Mathematics Applied in Business and Industry* 3, (1992) 305-314.