Accepted Manuscript

Volume 239, Issue 3, 16 December 2014
ISSN 0377-2217
ELSEVIER
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
Available online at www.sciencedirect.com
ScienceDirect
www.elsevier.com/locate/ejor

A Multi-Item Approach to Repairable Stocking and Expediting in a Fluctuating Demand Environment

Joachim Arts

PII: S0377-2217(16)30423-4
DOI: 10.1016/j.ejor.2016.06.003
Reference: EOR 13755

To appear in: *European Journal of Operational Research*

Received date: 27 May 2015
Revised date: 30 May 2016
Accepted date: 2 June 2016

Please cite this article as: Joachim Arts, A Multi-Item Approach to Repairable Stocking and Expediting in a Fluctuating Demand Environment, *European Journal of Operational Research* (2016), doi: 10.1016/j.ejor.2016.06.003

Highlights

- We present a model for multi-item repairable stocking and expediting
- We accommodate Markov modulated Poisson demand and provide fitting algorithms
- We provide a lower bound via decomposition and column generation
- We show the value of lead time flexibility; cost reductions are around 25%

A Multi-Item Approach to Repairable Stocking and Expediting in a Fluctuating Demand Environment

Joachim Arts[1]

[1]Eindhoven University of Technology, School of Industrial Engineering, j.j.arts@tue.nl

June 11, 2016

Abstract

We consider a single inventory location where multiple types of repairable spare parts are kept for service and maintenance of several different fleets of assets. Demand for each part is a Markov modulated Poisson process (MMPP). Each fleet has a target for the maximum expected number of assets down for lack of a spare part. The inventory manager can meet this target by stocking repairables and by expediting the repair of parts. Expedited repairs have a shorter lead time. There are multiple repair shops (or departments) that handle the repair of parts and the load imposed on repair shops by expedited repairs is constrained. A dual-index policy makes stocking and expediting decisions that depend on demand fluctuations for each spare part type. We formulate the above problem as a non-linear non-convex integer programming problem and provide an algorithm based on column generation to compute feasible near optimal solutions and tight lower bounds. We show how to use the MMPP to model demand fluctuations in maintenance and other settings, including a moment fitting algorithm. We quantify the value of lead time flexibility and show that effective use of this flexibility can yield cost reductions of around 25%.

Keywords: repair, inventory, spare parts, column generation, maintenance, Markov modulated Poisson process

1. Introduction

Service and manufacturing operations rely heavily on the availability of equipment such as aircraft, MRI-scanners, trains, and manufacturing equipment. The owners of such assets need to keep their equipment up and running as efficiently as possible. This is usually done by replacing defective components with ready for use components. The defective component is often expensive. Therefore a defective component is usually repaired and put back on stock. Such components are called *repairables*, and the working method described above is called *repair-by-replacement*. A sufficiently large number of spare repairables are needed to make such a system work, in particular to ensure a sufficiently high level of availability of capital assets. Buying sufficient repairables of all types needed to maintain a fleet of equipment is a major investment decision for firms with capital assets, because repairables are expensive.

The amount of spare repairables to buy is not the only major decision that affects the availability of capital assets. The repair of parts is usually done in house by several repair shops organized according to technical disciplines. For example, large airlines have repair shops for, amongst others, mechanical parts, avionics, and pneumatics. Our research has been instigated by a project we conducted at NedTrain, the maintenance division for rolling stock for the Dutch railways. The model we present in this paper has been used in a case study at NedTrain and was conducted by Van Aspert (2014) (see also Van Aspert (2013)). NedTrain also invests in repairable spare parts and has repair shops for mechanical parts, compressors, pneumatics, low voltage electronics, and high voltage electrical systems amongst others. The repair operations in these shops affect the availability of several fleets of trains. At the operational level, the inventory and repair shop planners coordinate to make sure repair priority is given to parts for which the inventory is most likely to run out in the near future.

The objective of this paper is to present a tractable optimization model that assists decision makers in answering the following questions

1. How many spare parts should we buy of each repairable type?

2. When should we expedite the repair of a given repairable type?

We assume the decision maker has to make these decisions for several fleets of equipment (e.g. local trains and long distance trains), and across parts that use different repair resources (e.g. pneumatics and electronics). The objective of the decision maker is to minimize the costs involved with purchasing or holding repairable spare parts while:

- meeting a service level in the form of a maximum average number of backorders for each fleet, and

- keeping the load imposed on each repair resource due to expedited orders below a set target level.

Note that this stocking problem cannot be resolved for each fleet separately because repairables that belong to different fleets (may) use the same resources for repair. We consider a setting where repair resources are flexible and model this through the possibility to request regular repair or expedited repair when sending a defective part to the repair shop. Expedited repairs have a shorter lead time than regular repairs. Since the flexibility of a repair resource is limited, there is a constraint on the amount of repair work that can be expedited per time unit for each repair resource. We refer to the amount of work that a repair resource handles per time unit as the load. Repairables from different fleets compete for the opportunity to load a repair resource with expedited orders.

Demand for a single type of repairable spare part usually fluctuates over time. These demand fluctuations arise for several reasons such as periodic inspections, usage patterns of equipment over time and the season of year. Slay and Sherbrooke (1988) observe empirically that demand for aircraft parts is non-stationary, and our experience with NedTrain also shows demand for many parts is non-stationary. When the reasons for demand fluctuations are understood, expediting decisions can be made to anticipate these fluctuations and to make effective use of repair resources.

In this article, we provide a mathematical model for the decision problem described above. This model has been conceived with an application at NedTrain in mind. We emphasize however that the applicability of the model and results in this article extend to other companies that maintain their own equipment. We will illustrate the need, as well as the application of the model, using an example that runs throughout this entire article. This example is about a fictitious railway company and is big enough to capture all model facets, yet small enough to inspect results in detail to gain insights and intuition. We finish this introduction by starting this example. The rest of the article is organized as follows. §2 reviews related literature and positions the contribution of this article with respect to existing literature. The mathematical model is provided in §3. The analysis of the model is in §4. Computational results of the model for industrial size instances are provided in §5 and concluding remarks are offered in §6.

Example 1. *The railway company Thomas&Co needs new trains to replace locomotives with pulled carriages. They decide to buy 100 trains from Liam Engineering Inc., and plan to use those for the next 30-40 years on long distance train services. Along with this order of 100 trains, Liam Engineering Inc. offers the possibility to buy (repairable) spare parts at a considerably discounted price. Thomas&Co would like to buy repairable spare parts at this discounted price and is taking this opportunity to decide on the stocking levels of repairables for the new fleet, as well as to reconsider the stocking levels for repairables of other fleets.* ◇

2. Literature review and contribution

Multi-item repairable inventory models are abundant in literature. We refer the reader to the books of Sherbrooke (2004), Muckstadt (2005) and Van Houtum and Kranenburg (2015), and review papers by Guide Jr. and Srivastava (1997), Kennedy et al. (2002) and Basten and Van Houtum (2014) for a broad overview. In this section, we briefly discuss literature with similar modeling assumptions and literature that expounds on or uses similar solution methods as those used in this article. On the modeling side, the main contributions of this article are the fluctuating demand model and the use of a dynamic expediting policy that depends on demand fluctuations. On the analysis side, we decompose the problem per item via a column generation algorithm. Therefore, this section is organized around three main topics: fluctuating demand (§2.1), repair expediting and scheduling policies (§2.2), and decomposition and column generation algorithms (§2.3).

2.1 Fluctuating demand

Demand for repairables that fluctuates over time has been considered before in a series of models developed by the RAND corporation under the name Dyna-METRIC (Hillestad, 1982; Carillo, 1989; Isaacson and Boren, 1993). Initially, these models were based on an extension of Palm's theorem for non-stationary Poisson processes, but these efforts eventually developed into simulation models that do not allow efficient optimization. In the Dyna-METRIC approach, demand is a non-stationary Poisson proces, but the Poisson demand rate is a deterministic function of time. Rather than performing steady-state analysis, the Dyna-METRIC approach is to perform a transient analysis at some particular point in time that is chosen by the modeler. The Dyna-METRIC model does not include the possibility to expedite repair. Demand fluctuations are therefore only buffered by holding inventory.

A similar approach is followed by Lau and Song (2008) with two exceptions: They also model the finite repair capacity using queueing approximations and they evaluate the transient behavior of the system at several points of interest rather than only one. For their extensions to Dyna-METRIC, they take heuristic and approximative approaches.

Our work differs from these contributions because demand fluctuations are modeled by a Markov modulated Poisson process. This resembles practice more closely as the intensity of demand over time behaves as a stochastic process rather than a deterministic function. Additionally, our model deals with these demand fluctuations not only by holding repairable inventory, but also by using the possibility to expedite repair. Our modeling also allows us to evaluate our system exactly and compute tight lower bounds on optimal system performance. The use of the Markov modulated Poisson process to model demand for inventory systems has already been advocated by Song and Zipkin (1993). However, no practical fitting algorithms have been provided for modeling demand. (There are, however, practical fitting algorithms for the MMPP process in the context of communication networks; see e.g. Heffes and Lucantoni (1986); Meier-Hellstern (1987); Yoshihara et al. (2001); Nelson and Gerhardt (2010)). We provide two practical fitting procedures. The first is based on the maintenance and repair setting and uses information from maintenance planning. The second procedure is a moment fitting procedure that fits on demand over the lead-time.

2.2 Expediting and repair scheduling policies

The possibility to either expedite repair or prioritize the scheduling of repairs in the repair shop has been considered many times, mostly under the assumption of fixed given turn-around stock levels (Hausman and Scudder, 1982; Scudder, 1986; Scudder and Chua, 1987; Pyke, 1990; Tiemessen and Van Houtum, 2012; Liang et al., 2013). In these contributions, the repair shop is modeled by a finite server queue. Given a limited capacity, the question becomes: How should limited repair capacity be allocated to repair jobs of various types, i.e., which repair jobs deserve priority?

As observed by Tiemessen and Van Houtum (2012), even for fixed given turn-around stock levels, computing optimal priority rules, or evaluating a given rule, requires computation times that grow exponentially in the number of different repairable types. Also the derivation of structural properties of optimal policies or evaluation of heuristic policies is limited to cases with only two repairable types (e.g Zheng and Zipkin, 1990; Veatch and Wein, 1996; Ha, 1997). Accordingly, most contributions in this area use simulation to study heuristic priority rules. All these authors report that system performance increases substantially by using various priority rules. Hausman and Scudder (1982) and Tiemessen and Van Houtum (2012) both point out that substantial stock reductions should be possible as a result of using an effective priority rule. Under *static* priority rules, the priority of a spare part depends on its type only. Under these relatively simple rules, Sleptchenko et al. (2005) and Adan et al. (2009) have shown numerically that significant reductions in inventory investment are possible compared to simple first come first serve scheduling of repair jobs. More sophisticated priority rules also consider the on-hand inventory and expected future demand in deciding the priority of a part. These *dynamic* priority rules are essentially mechanisms that change the repair lead time of an item based on current on-hand stock and estimated future demand. In this regard, the possibility to schedule repairs can be interpreted as providing lead time flexibility. The expediting policy in our model provides this lead time flexibility, but does not suffer from the tractability issues that dynamic priority queueing models suffer from.

We retain tractability because we assume a rather simple priority rule and refrain from explicitly modeling the queueing behavior that occurs in the repair shop. If the repair shop is external to the company holding inventory, this is a natural modeling choice, but even when the repair shop is internal to the company, this model has merit: In many organizations, the repair shop and inventories are managed separately. Coordination of repair priorities often happens implicitly through lead time agreements between the inventory manager and the repair shop manager. Our model is a first step in explicitly considering the effect of smart priority rules when deciding on turn-around stocks. A simulation study at NedTrain shows that our model is also a good approximation under more sophisticated priority rules (Loeffen, 2012).

The possibility to expedite the repair of a part without considering queueing effects in the repair shop has been considered previously by Verrijdt et al. (1998), but their policy only depends on the on-hand inventory of a part and considers Poisson demand only. Moinzadeh and Schmidt (1991) study the same policy that we use, but in the context of deterministic lead times and Poisson demand. Song and Zipkin (2009) show that the model of Moinzadeh and Schmidt (1991) can be reinterpreted as a special type of queueing network for which a product-form solution exists. This observation allows them to significantly generalize the model of Moinzadeh and Schmidt (1991), but it does not allow expediting policies that somehow depend on demand fluctuations. The expediting policy we propose in this article, does depend on demand fluctuations, and is shown to be optimal under certain conditions described in Arts et al. (2016). (They consider a fully economical model without service levels.) The merit of this rule is that it captures the essential trade-off involved in dynamically scheduling repair of spare parts, while being sufficiently simple to make the problem of deciding inventory levels tractable.

2.3 Decomposition and column generation

Decomposition and column generation is a general technique to deal with optimization problems that have a Lagrangian that can be decomposed. The most straightforward way of dealing with such problems is by manipulating the Lagrange multipliers as suggested by Everett (1963) and later by Fisher (1981). Brooks and Geoffrion (1966) show that one efficient way of finding the best Lagrange multipliers is via setting up a linear program in which each variable corresponds to a solution for each of the parts that compose the Lagrangian. The Lagrange multipliers then correspond to shadow prices (or dual variables) of the linear program. The algorithm we present in this article is based on that idea.

In the context of spare parts inventory optimization, decomposition and column generation has been used as early as in the seminal paper of Sherbrooke (1968) to solve the METRIC model, where the Lagrangian is decomposable per spare part type. Essentially, the technique reduces the original optimization problem that encompasses many types of repairables, to repeatedly solving a single-item inventory problem for each repairable. Usage of this technique for spare part inventory optimization problems has found much recent following, e.g. Kranenburg and Van Houtum (2007, 2008); Alvarez et al. (2013, 2015). In all these papers (including Sherbrooke (1968)), there is one or more service level constraints that need to be achieved by all parts collectively (rather than individually). After moving these service level constraints to the objective by taking the Lagrangian, the best Lagrange multipliers are found via dual variables in a linear programming relaxation of the problem. (See also Dantzig and Wolfe (1960) and Lübbecke and Desrosiers (2005) for a more general and thorough treatment of this technique.)

We use the same technique to find a tight lower bound and a feasible solution for our model. Different from all the papers mentioned in the previous paragraph, different repairable items are not only linked because of a collective service level, but also through the expediting load that they have on one or more repair resources. This is a merit of how our model is set up: Our expediting rule mimics the dynamic priorities given to repairs but allows for tractable analysis through the technique of decomposition and column generation.

2.4 Statement of contributions

Our model captures the following features for the first time in multi-item inventory problems of industrial size:

1. Demand intensity that fluctuates over time as a stochastic process as modeled by the Markov modulated Poisson process

2. The repair of items can be expedited, and the expediting policy depends on what we know about demand fluctuations.

3. Decisions for stocking and expediting repairables affect decision for parts that belong to different fleets of assets because they share the same resource for repair.

With regards to the first item, we also provide two new fitting procedures to model demand with a Markov modulated Poisson process. The first procedure is specific to the repair context and requires information on the maintenance regime of assets. The second is a generic moment fitting procedure that can also be used in different contexts.

With regards to the second and third item: We provide a formulation that yields a tight lower bound on the optimal solution and near optimal feasible solutions. The formulation is tractable because it relies on modeling repair shop flexibility through lead time differentiation. This approach allows us to decompose the problem and use column generation algorithms. The approach has been applied to an industrial case by Van Aspert (2014) and led to a saving potential of 40% compared to the current way of working.

Finally, our numerical work shows that explicitly considering lead time flexibility as a tool to anticipate demand fluctuations can decrease the investment required to meet a certain service level by as much as 25% on average across a large test bed.

3. Model

In this section, we model our problem and illustrate most modeling steps by continuing the example started in the introduction. We start with some notation and preliminaries in §3.1. Then we discuss the control policy we use for each repairable type in §3.2. Fluctuating demand models are discussed in §3.3. We conclude this section by formally stating our optimization problem in §3.4.

3.1 Notation and preliminaries

We consider several fleets of assets for which we keep repairable spare parts on stock. We denote the set of fleets by A and the set of repairable items by I. We refer to each element of I as a stock keeping unit (SKU). The set of SKUs used to maintain fleet $a \in A$ is denoted I_a^A. There is a set of repair resources, C, that are used to repair defective parts. The items that load repair resource $c \in C$ are contained in the set I_c^C. We will assume that I_a^A and I_c^C partition the set of all SKUs, that is $\cup_{c \in C} I_c^C = \cup_{a \in A} I_a^A = I$ and $\cap_{c \in C} I_c^C = \cap_{a \in A} I_a^A = \emptyset$. This assumption is not essential to the analysis, but it considerably simplifies notation and presentation.

Each SKU $i \in I$ faces Markov modulated Poisson demand. This means that demand for SKU i is a Poisson process whose intensity varies with the state of an exogenous Markov process $Y_i(t)$. The Markov process $Y_i(t)$ is irreducible and has a finite state space $\Theta_i = \{1, ..., |\Theta_i|\}$ with generator matrix $\mathbf{Q}_i$ whose elements we denote by $q_i(m,n)$. ($|x|$ denotes the cardinality of x if it is a set, and its absolute value if it is a real number.) For notational convenience, we define $q_i(m) = -q_i(m,m)$ and $q_i^{\max} = \max_{m\in\Theta_i} q_i(m)$. When $Y_i(t) = y$, the intensity of Poisson demand at time t is given by $\lambda_i(y) \geq 0$; $\boldsymbol{\lambda}_i = (\lambda_i(1), ..., \lambda_i(|\Theta|))$, $\lambda_i(y) > 0$ for at least one $y \in \Theta_i$ and $\lambda_i^{\max} = \max_{y\in\Theta_i} \lambda_i(y)$. We denote demand for SKU i in the time interval $(t_1, t_2]$ given $Y_i(t_1) = y$ as $D_i^y(t_1, t_2)$. Note that $Y_i(t_1)$ provides information about the distribution of demand in the interval $(t_1, t_2]$, $t_2 > t_1$. We assume that $Y_i(t)$ can be observed directly for all $i \in I$ and provides a form of aggregated advance demand information. Example 3, shows an example of how demand might fluctuate over time. We address how to model such demand and provide examples in §3.3.

There exists a regular and an expedited repair option for each SKU $i \in I$. The expedited repair lead time for SKU i is deterministic and denoted by ℓ_i. The expedited repair lead time may represent things such as the transport time and the repair time or a lead time agreed upon with an external company that provides emergency repair service. We also refer to using the expedited repair mode as expediting repair. The regular repair lead time of SKU i consists of the emergency repair lead time ℓ_i, and a random component of length L_i. The random variable L_i has an exponential distribution with mean $1/\mu_i$. L_i models such things as the time that a part waits for resources to become available in the repair shop or the lead time difference between regular and emergency repair lead times as contracted with an external repair shop. The assumption that L_i has an exponential distribution, seems rather restrictive, but numerical evidence in Arts et al. (2016) suggests that it is not a very strong assumption at all as the performance of the system seems rather insensitive to the exact distribution of L_i for a fixed mean. The inventory manager knows for each repair order of SKU i when L_i has lapsed, and the remaining lead time of an order is ℓ_i.

Of each SKU i, we already own S_i^{LB} parts. The main decision variables are the total number of parts to own for each SKU. This is denoted by S_i for SKU $i \in I$ and is also referred to as the turn-around stock. For each SKU $i \in I$ there is an acquisition price C_i^a for buying additional spare repairables.

Each repair of an SKU $i \in I_c^C$ part, imposes a 'load' of u_i on repair resource $c \in C$. We use the term 'load' for u_i, but the interpretation of u_i can vary broadly. To illustrate this, consider the following two examples:

- Repair is performed by an external repair shop and the repair lead time may be shortened in exchange for an increased price for the repair. However, there is a maximum target on the amount of money that can be used for requesting expedited lead times from external parties. In this case, the repair resource c might be this annual target for expedited repairs expenses and u_i is the additional cost of an expedited repair over a regular repair.

- Repairs are conducted by a repair shop within the company. This repair shop can expedite the repair of certain parts upon request, as long as the load imposed on the repair shop by expedited repairs is limited. Manpower is the bottleneck in the repair shop. The load imposed on the repair shop u_i could then be man hours required for the repair of a SKU $i \in I_c^C$ part.

For each repair resource $c \in C$, there is maximum $\mathcal{E}_c^{\max}$ on the load this repair resource is allowed to experience due to expedited repair orders.

Table 1 summarizes the notation we have introduced so far as well as notation we will introduce later. Now we return to our example to put all this notation in some perspective.

Example 2. *Thomas&Co already has a fleet of 200 trains that are used for services with many stops. This fleet is called* $\textsc{Village}$*, while the fleet of 100 trains they are about to buy is called* $\textsc{City}$*. Now* $A = \{\textsc{City}, \textsc{Village}\}$*. All mechanical repairs are done in an internal repair shop, while the repair of climate and airconditioning units is outsourced to an external company. Therefore,* $C = \{\textsc{Outsource}, \textsc{Mechanic}\}$*. Manpower is the bottleneck in the internal repair shop so* u_i *is measured in man hours if* $i \in I^C_{\textsc{Mechanic}}$*. If* $i \in I^C_{\textsc{Outsource}}$*, then* u_i *is measured in EUROS. Thomas&Co has gathered all this data as shown in Table 2. Note that from Table 2 we can also read that* $I^C_{\textsc{Mechanic}} = \{2, 3, 5, 6\}$*,* $I^C_{\textsc{Outsource}} = \{1, 4\}$*,* $I^A_{\textsc{Village}} = \{1, 2, 3\}$*, and* $I^A_{\textsc{City}} = \{4, 5, 6\}$*. The data not shown in Table 2 is that* $\ell_i = 2$ *and* $\mathbb{E}[L_i] = 3$ *for all* $i \in I$*. In the next example, we will consider demand data.* ◇

3.2 Control policy

Let $X_i(t)$ be the number of parts of SKU i that have been sent to regular repair and have not yet completed the exponential phase of their repair at time t. The replenishment policy for each SKU is to place a replenishment order whenever a demand occurs. Such a policy is also called a base-stock policy, order-up-to policy, or $(S-1, S)$ policy in literature, e.g. Muckstadt (2005). For the expediting policy, we propose to expedite whenever $X_i(t)$ exceeds some threshold that depends on $Y_i(t)$, i.e. replenishment orders are expedited at time t if $X_i(t) \geq T_i(y)$ when $Y_i(t) = y$. Thus the control policy for any SKU i can be described by the turn-around stock S_i and a vector $\mathbf{T}_i = (T_i(1), T_i(2), \cdots, T_i(|\Theta_i|))$ containing the expediting thresholds for each modulating state. The stochastic process $X_i(t)$ depends on $\mathbf{T}_i$ and so we will write this explicitly: $X_i^{\mathbf{T}_i}(t)$. Figure 1 gives a graphical representation of the control policy for any SKU i. The combined policy is denoted by $(S_i, \mathbf{T}_i)$.

The $(S_i, \mathbf{T}_i)$ policy can be re-interpreted as a state dependent *dual-index policy* as has also been noted in Arts et al. (2016). This interpretation is perhaps the most natural way to think of the system: We place a repair order each time a demand occurs. If the number of parts in inventory (minus backorders) and in the pipeline that will arrive to inventory within the expedited lead time ℓ_i is below $S - T_i(Y_i(t))$, then we expedite the repair of this part to avoid a “likely” backorder. The sum of parts in inventory (minus backorder) and in the pipeline that will arrive within ℓ_i time units is often called the emergency inventory position. $S_i - T_i(Y_i(t))$ can then be interpreted as the state dependent *expedite-up-to-level.* Usually

Table 1: Overview of notation

		Sets
I	:	Set of all SKUs.
A	:	Set of all fleets.
C	:	Set of all types of repair shop resources.
I_a^A	:	Set of SKUs used to maintain fleet $a \in A$.
I_c^C	:	Set of SKUs that load repair resource $c \in C$.
Θ_i	:	Set of modulating states of the Markov modulating chain of demand for SKU $i \in I$
		Input Parameters
$\lambda_i(y)$	:	Demand intensity for SKU $i \in I$ when $Y_i(t) = y \in \Theta_i$
$\boldsymbol{\lambda}_i$	:	The vector $(\lambda_i(1), \lambda_i(2), \cdots, \lambda_i(\lvert\Theta_i\rvert))$
$\lambda_i^{\max}$	:	$\max_{y \in \Theta_i} \lambda_i(y)$ for SKU $i \in I$
$\mathbf{Q}_i$	:	Generator matrix of the modulating process $Y_i(t)$ of SKU $i \in I$
$q_i(m,n)$	:	The element of row m column n of $\mathbf{Q}_i$, $i \in I$
$q_i(m)$	:	$-q_i(m,m)$
$q_i^{\max}$	:	$\max_{m \in \Theta_i} q_i(m)$
ℓ_i	:	The (deterministic) expedited repair lead time of SKU $i \in I$
μ_i^{-1}	:	Mean of the additional regular repair lead time, $\mathbb{E}[L_i]$; (the mean regular repair lead time is $\ell_i + \mu_i^{-1}$)
S_i^{LB}	:	Lower bound on the size of the turn-around-stock for SKU $i \in I$
C_i^a	:	Acquisition costs for SKU $i \in I$
u_i	:	Resource load associated with the repair of SKU $i \in I$
$\mathcal{B}_a^{\max}$	:	The maximally allowed mean number of backorders over all SKUs $i \in I_a^A$ for $a \in A$.
$\mathcal{E}_c^{\max}$	:	The maximally allowed mean resource loading resulting from repair expediting over all items $i \in I_c^C$ for expediting resource $c \in C$.
		Decision variables
S_i	:	Size of the turn-around-stock for SKU $i \in I$
$T_i(y)$	:	Expediting threshold for SKU $i \in I$ when $Y_i(t) = y \in \Theta_i$
$\mathbf{T}_i$	:	The vector $(T_i(1), T_i(2), \cdots, T_i(\lvert\Theta_i\rvert))$
		Output of model
$X_i^{\mathbf{T}_i}(t)$	:	The number of parts of SKU $i \in I$ in regular repair at time t and not arriving to inventory before time $t + \ell_i$ under an expediting policy with thresholds $\mathbf{T}_i$.
$B_i^{(S_i,\mathbf{T}_i)}(t)$	:	Number of backorders of SKU i at time t under policy $(S_i, \mathbf{T}_i)$;
$D_i^y(t_1,t_2)$	:	Demand for SKU $i \in I$ in the interval $(t_1, t_2]$ given $Y_i^{t_1} = y \in \Theta_i$
$Y_i(t)$	:	Modulating chain of the demand process of part $i \in I$ at time t
L_i	:	Additional regular repair lead time; has exponential distribution with mean μ_i^{-1}
$\mathcal{B}_i^{(S_i,\mathbf{T}_i)}$	:	Expected number of backorders of SKU $i \in I$, $\lim_{t\to\infty} \mathbb{E}[B_i^{(S_i,\mathbf{T}_i)}(t)]$
C_P^{LB} (C_P^{UB})	:	Lower (upper) bound for optimal solution to problem (P) in (8)-(13)
C_{BENCH}^{LB}	:	Lower bound for the optimal cost of a benchmark instance
$\mathcal{E}_i^{\mathbf{T}_i}$	:	Expected number of repairs of SKU $i \in I$ that are expedited per unit time $\sum_{y \in \Theta_i} \lambda_i(y) \mathbb{P}(X_i^{\mathbf{T}_i} \geq T_i(y) \mid Y_i = y) \mathbb{P}(Y_i = y)$

Table 2: Input data for Thomas&Co

SKU#	Description	C_i^a (kEURO)	Fleet	Repair Resource	u_i	S_i^{LB}
1	Climate unit	30	Village	Outsource	500	2
2	Electro motor	45	Village	Mechanic	16	1
3	Break set	5	Village	Mechanic	4	5
4	Airconditioning unit	10	City	Outsource	500	0
5	Electro motor	30	City	Mechanic	16	0
6	Break set	2	City	Mechanic	4	0

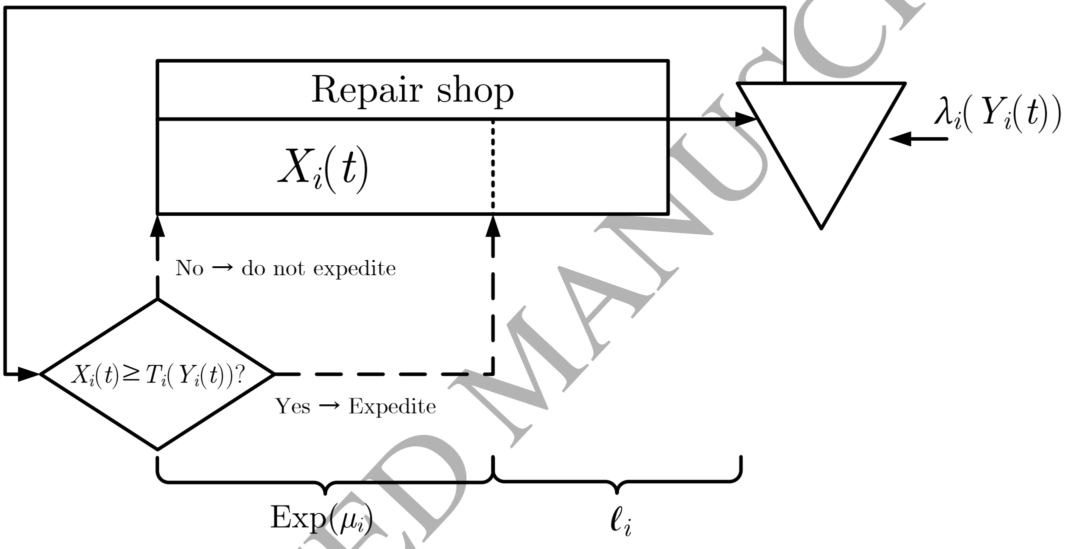

Figure 1: A graphical representation of the model for a single item.

$S_i - T_i(y_1) \geq S_i - T_i(y_2)$ if $\lambda_i(y_1) \geq \lambda_i(y_2)$, i.e., the expedite-up-to level increases with instantaneous demand intensity.

The state dependent dual-index policy we propose is actually optimal under a linear backordering and expediting cost structure as shown in Theorem 2 of Arts et al. (2016). Furthermore, their numerical study shows that the performance of a state dependent dual-index policy is rather insensitive to the assumption that L_i has an exponential distribution, i.e., both the performance evaluation error and optimality gap for similar systems where L_i has a different distribution with the same mean are small (within 2.76% and 0.70% respectively over a large test bed).

Under a dual-index policy with parameters $(S_i, \mathbf{T}_i)$, $(X_i^{\mathbf{T}_i}(t), Y_i(t))$ is a Markov process with state-space

$$\mathcal{S}_i = \left\{ (x,y) \middle| x \in \left\{0, \ldots, \max_{k \in \Theta_i} T_i(k)\right\}, \quad y \in \Theta_i \right\}.$$

The Markov process $(X_i^{\mathbf{T}_i}(t), Y_i(t))$ has three types of transitions:

1. Demand occurs: transitions from (x, y) to $(x+1, y)$ with intensity $\lambda_i(y)$ if $x < T_i(y)$
2. Part in regular repair finishes first part of its lead-time: transitions from (x, y) to $(x-1, y)$ with intensity $x\mu_i$
3. Demand rate changes: transitions from (x, y) to (x, y') with intensity $q_i(y, y')$ if and $y \neq y'$.

The joint steady state distribution of $(X_i^{\mathbf{T}_i}(t), Y_i(t))$ can be determined from these transition intensities by solving the (linear) balance equations of this Markov process. We denote the corresponding steady-state random variables by dropping the time index t. The expected number of backorders and expedited repair per time unit of a given SKU can be determined from the distribution of $(X_i^{\mathbf{T}_i}, Y_i)$ as follows: Let $B_i^{(S_i,\mathbf{T}_i)}(t)$ denote the number of backorders of SKU i at time t under dual-index policy $(S_i, \mathbf{T}_i)$. The dynamics of $B_i^{(S_i,\mathbf{T}_i)}(t)$ are given by

$$B_i^{(S_i,\mathbf{T}_i)}(t+\ell_i) = \left(D_i^{Y_i(t)}(t, t+\ell_i) - \left(S_i - X_i^{\mathbf{T}_i}(t)\right)\right)^+, \tag{1}$$

and so the expected number of backorders of SKU i in steady state, $\mathcal{B}_i^{(S_i,\mathbf{T}_i)}$, satisfies:

$$\mathcal{B}_i^{(S_i,\mathbf{T}_i)} = \lim_{t\to\infty} \mathbb{E}\left[B_i^{(S_i,\mathbf{T}_i)}(t+\ell_i)\right] = \mathbb{E}_{Y_i}\mathbb{E}_{X_i^{\mathbf{T}_i}}\left[\left(D_i^{Y_i}(t, t+\ell_i) - S_i + X_i^{\mathbf{T}_i}\right)^+ \middle| Y_i\right]. \tag{2}$$

Equation (2) can be evaluated after noting that the probability mass function of $D_i^y(t, t+\ell_i)$ can be computed by numerical inversion of a generating function. Details of this are provided in Appendix A.

Next consider the expected number of repairs that are expedited per time unit of SKU i, and denote it by $\mathcal{E}_i^{\mathbf{T}_i}$. (Note that $\mathcal{E}_i^{\mathbf{T}_i}$ depends on the expediting threshold $\mathbf{T}_i$ only, not on the turn-around stock level S_i.) We have:

$$\mathcal{E}_i^{\mathbf{T}_i} = \sum_{y\in\Theta_i} \lambda_i(y)\mathbb{P}(X_i(\mathbf{T}_i) \geq T_i(y)|Y_i = y)\mathbb{P}(Y_i = y). \tag{3}$$

$\mathcal{B}_i^{(S_i,\mathbf{T}_i)}$ and $\mathcal{E}_i^{\mathbf{T}_i}$ can be computed in many ways. In §5, we use value iteration (see e.g. Tijms (2003)) to compute $\mathcal{B}_i^{(S_i,\mathbf{T}_i)}$ and $\mathcal{E}_i^{\mathbf{T}_i}$ to a precision of 10^{-8}.

3.3 Markov Modulated demand models and fitting

Fitting a MMPP demand model to data has not received much attention in the literature. Fitting procedures do exist, but these are geared primarily to applications of queueing models in telecommunication systems (e.g. Heffes and Lucantoni, 1986; Meier-Hellstern, 1987; Yoshihara et al., 2001; Nelson and Gerhardt, 2010). Using Markov modulated demand in the context of inventory problems has been advocated by Song and Zipkin (1993) and Zipkin (2000). However, practical algorithms to fit MMPP demand models to data have not been provided in the literature. In this section, we provide two fitting techniques.

The first fitting procedure in §3.3.1 is specific for the maintenance context in this article. The second fitting procedure in §3.3.2 is a moment fitting procedure, that we believe can also be useful outside of the setting considered in this article.

3.3.1 Fitting based on maintenance strategy and installed base

The fitting procedure we describe is best understood by first considering an example.

Example 3. *For the SKUs in Table 2, Maintenance engineers at Thomas&Co are asked to assess what the demand will behave like over the next 30-40 years. From past experience, they know that break sets need to be replaced on each train approximately every year and so they expect a relatively steady demand of* $200/50 = 4$ *for SKU 3 and* $100/50 = 2$ *parts per week for SKU 6. (We work with a year of 50 weeks.) An airconditioning unit (SKU 4) is estimated to fail due to random causes about once every 5 years. Over the entire fleet, this means that demand due to failure maintenance will be about* $\frac{1}{5}100/50 = 0.4$ *parts per week. Additionally, the maintenance engineers expect that the airconditioning units of the entire* City *fleet will need to be overhauled roughly every 4 years. They warn that this will lead to peaks in demand during overhaul periods. How high this peak will be, depends on the length of the overhaul period. Currently, revision periods are planned to last a year. For SKU 5, the* City *electro motor, random failures occur around once every 10 years so they expect a relatively steady demand of* $\frac{1}{10}100/50 = 0.2$ *per week. Electro motors require overhaul every 7 or so years, so here too, maintenance engineers insist that inventory will be needed to deal with peak demand during overhaul periods. Similar estimates are also available for SKUs 1 and 2: SKU 1 and 2 fail due to random causes once every 4 and 8 years respectively and need to be replaced and overhauled every 4 and 6 years respectively.* ◇

Example 3 illustrates how an understanding of maintenance can improve the understanding of how demand for certain repairables fluctuates. This understanding can then be modeled as the modulating chain for demand. Suppose that demand for repairables behaves as described in Example 3: Demand is relatively steady over some period, until demand peaks because of a revision period in which parts are overhauled preventively. Then a simple MMPP that models demand is the following. Let N_a denote the number of equipment in fleet a and consider an SKU $i \in I_a^A$. Let λ_i^{ran} denote the intensity with which any piece of equipment in the fleet fails randomly (i.e. not due to wear out). Wear out failures do not occur because all repairables in the fleet are overhauled during revision periods. The time between revision periods is a random variable M_i for SKU i. (M_i is not deterministic because the time between revision periods is decided upon based on the condition of the fleet.) Once the revision period starts, it lasts R_i time units and all repairables in the fleet are expected to be replaced and revised during this period. R_i is also a random variable. If we approximate M_i and R_i by exponential random variables a

MMPP demand model is given by:

$$\mathbf{Q}_i = \begin{pmatrix} -\mathbb{E}[M_i]^{-1} & \mathbb{E}[M_i]^{-1} \\ \mathbb{E}[R_i]^{-1} & -\mathbb{E}[R_i]^{-1} \end{pmatrix}, \quad \boldsymbol{\lambda}_i^{\mathrm{T}} = \begin{pmatrix} \lambda_i^{\mathrm{ran}} N_a \\ \lambda_i^{\mathrm{ran}} N_a + N_a/\mathbb{E}[R_i] \end{pmatrix}, \tag{4}$$

where $\boldsymbol{\lambda}_i^T$ is the transpose of $\boldsymbol{\lambda}_i$. Rather than using the exponential distribution for R_i and M_i, it is possible to use any phase type distribution if appropriate. The restriction of modeling R_i and M_i by phase type distributions is rather weak because phase type distributions are dense in the class of all non-negative distributions (Tijms, 2003).

Example 4. *Thomas&Co decide to use* (4) *to model their demand. This yields (time units are weeks):*

$$\mathbf{Q}_1 = \begin{pmatrix} -\frac{1}{200} & \frac{1}{200} \\ \frac{1}{50} & -\frac{1}{50} \end{pmatrix}, \qquad \mathbf{Q}_3 = 0, \qquad \mathbf{Q}_5 = \begin{pmatrix} -\frac{1}{350} & \frac{1}{350} \\ \frac{1}{50} & -\frac{1}{50} \end{pmatrix},$$

$$\mathbf{Q}_2 = \begin{pmatrix} -\frac{1}{400} & \frac{1}{400} \\ \frac{1}{50} & -\frac{1}{50} \end{pmatrix}, \qquad \mathbf{Q}_4 = \mathbf{Q}_1, \qquad \mathbf{Q}_6 = 0,$$

and

$$\boldsymbol{\lambda}_1^T = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \boldsymbol{\lambda}_2^T = \begin{pmatrix} \frac{1}{2} \\ \frac{9}{2} \end{pmatrix}, \boldsymbol{\lambda}_3^T = 4, \boldsymbol{\lambda}_4^T = \begin{pmatrix} \frac{2}{5} \\ \frac{12}{5} \end{pmatrix}, \boldsymbol{\lambda}_5^T = \begin{pmatrix} \frac{1}{5} \\ \frac{11}{5} \end{pmatrix}, \boldsymbol{\lambda}_6^T = 2.$$

where $\boldsymbol{\lambda}_i^T$ *is the transpose of* $\boldsymbol{\lambda}_i$. ◇

3.3.2 Fitting based on moments of demand over expected lead time

One of the drawbacks of the stationary Poisson demand model is that it has only one parameter and so fixing the mean demand per period, also fixes the variance of demand per period. In practice one often wishes to fit a demand model on the mean and variance of demand that were found in past observations. The MMPP demand model can be chosen so as to coincide with the mean and variance of demand over some fixed period, provided that the variation coefficient (variance divided by mean) is greater than 1.

Suppose that we know the mean and variance of demand over some standard period and we wish to model our MMPP demand process such that it has the same mean and variance. Without loss of generality, we scale time such that the length of the standard period is one time unit. The following proposition provides a two-state MMPP with the required mean and variance of demand over a time unit.

Proposition 1. *If the mean and variance of demand over one time unit are given by* μ *and* σ^2 *respectively, then the MMPP specified below matches this mean and variance for any* $\kappa \geq 2$*:*

$$\mathbf{Q} = \begin{pmatrix} -\beta & \beta \\ \alpha\beta & -\alpha\beta \end{pmatrix}, \qquad \boldsymbol{\lambda} = (0, \lambda), \tag{5}$$

with

$$\alpha \geq \kappa \frac{\sigma^2 - \mu}{\mu^2}, \qquad \lambda = (1+\alpha)\mu, \tag{6}$$

and β the unique solution to the attractive fixed point equation

$$\beta = f(\beta), \quad \text{with} \quad f(\beta) = \frac{\mu\sqrt{2\alpha e^{-(\alpha+1)\beta}(\sigma^2-\mu) + 2\alpha(\mu-\sigma^2) + \alpha^2\mu^2} + \alpha\mu^2}{(\alpha+1)(\sigma^2-\mu)}. \tag{7}$$

Since (7) is an attractive fixed point equation, a β that satisfies (7) can be determined to arbitrary precision by setting $\beta_0 = 1$ and iteratively computing $\beta_{i+1} = f(\beta_i)$ until $|\beta_{i+1} - \beta_i| < \varepsilon$ for some desired accuracy ε. Appendix B provides the proof of Proposition 1 as well as several figures of the fitted demand distribution that this procedure provides for different $\kappa \geq 2$.

3.4 Optimization problem

The objective of the manager is to minimize the investment he is about to make in buying repairable spare parts. The constraints are to keep the total expected backorders for each fleet $a \in A$ below $\mathcal{B}_a^{\max}$ and to keep the total expected resource loading due to expedited repair orders below $\mathcal{E}_c^{\max}$ for each repair resource $c \in C$. A backorder for a part renders some equipment down. If an expedited repair mode is available for SKU $i \in I$, it is unacceptable that any particular backorder for SKU $i \in I$ lasts longer than ℓ_i. To ensure this never happens, it suffices to ensure that $T_i(y) \leq S_i$ for each $i \in I$ and $y \in \Theta_i$. Combining all this leads to the following formal statement of our optimization problem which we call P:

$$
\begin{aligned}
(P) \qquad & \min_{\{S_i, \mathbf{T}_i | i \in I\}} && \sum_{i \in I} C_i^a \left(S_i - S_i^{LB}\right) && && (8)\\
& \text{subject to} && \sum_{i \in I_a^A} \mathcal{B}_i^{(S_i, \mathbf{T}_i)} \leq \mathcal{B}_a^{\max} && \forall a \in A && (9)\\
& && \sum_{i \in I_c^C} u_i \mathcal{E}_i^{\mathbf{T}_i} \leq \mathcal{E}_c^{\max} && \forall c \in C && (10)\\
& && S_i^{LB} \leq S_i && \forall i \in I && (11)\\
& && T_i(y) \leq S_i && \forall i \in I, \forall y \in \Theta_i && (12)\\
& && S_i, T_i(y) \in \mathbb{N}_0 && \forall i \in I, \forall y \in \Theta_i. && (13)
\end{aligned}
$$

We denote the optimal costs to problem (P) by C_P. In the next section, we construct a feasible solution with cost C_P^{UB} for problem (P) as well as a lower bound, C_P^{LB}, on the optimal cost of problem (P).

Example 5. *Thomas&Co would like to adhere to the goals of having $\mathcal{B}_{\text{VILLAGE}}^{\max} = 1$ and $\mathcal{B}_{\text{CITY}}^{\max} = 0.5$. For expediting the repair of climate and airconditioning units (*OUTSOURCE *repair resource) there is a weekly budget of 180 EUROS, $\mathcal{E}_{\text{OUTSOURCE}}^{\max} = 180$. (Note that the 'loads' for each SKU $i \in I$ are provided in Table 2 as discussed in Example 2.) For expediting the repair for the internal repair shop that handles mechanical repairs, the agreement with the repair shop manager is to keep requests for expedited repair orders below the nominal load of 20 man hours per week on average, $\mathcal{E}_{\text{MECHANIC}}^{\max} = 20$.* ◇

4. Analysis

The analysis will proceed by giving an algorithm to construct a lower bound for problem (P) in §4.1. In 4.2, we show how to find a good feasible solution for problem (P) based on the lower bound constructed in §4.1.

4.1 Constructing lower bounds with column generation

To obtain a lower bound for problem (P), we first reformulate it to an integer linear program and then relax the integrality constraints. We refer to this problem as the master problem (MP). To this end, we introduce the set K_i of all dual-index policies k for item i that respect constraints (11)-(13) of problem (P). Policy $k \in K_i$ has base-stock level and expediting thresholds $(S_i^k, \mathbf{T}_i^k)$. We also introduce the decision variable $x_i^k \in \{0,1\}$ that indicates whether policy k is chosen for item i. If we relax the integrality constraint on x_i^k, we obtain the master problem:

$$\begin{array}{llll} (MP) & \displaystyle\min_{\{x_i^k | i \in I, k \in K_i\}} & \displaystyle\sum_{i \in I} C_i^a \left(S_i^k - S_i^{LB} \right) x_i^k & \qquad (14) \\ & \text{subject to} & \displaystyle\sum_{i \in I_a^A} \sum_{k \in K_i} \mathcal{B}_i^{(S_i^k, \mathbf{T}_i^k)} x_i^k \leq \mathcal{B}_a^{\max} \quad \forall a \in A & \qquad (15) \\ & & \displaystyle\sum_{i \in I_c^C} \sum_{k \in K_i} u_i \mathcal{E}_i^{\mathbf{T}_i^k} x_i^k \leq \mathcal{E}_c^{\max} \quad \forall c \in C & \qquad (16) \\ & & \displaystyle\sum_{k \in K_i} x_i^k = 1 \quad \forall i \in I & \qquad (17) \\ & & x_i^k \geq 0 \quad \forall i \in I, \forall k \in K_i. & \end{array}$$

Since K_i is an infinite set, (MP) is an infinite dimensional linear program. The way to solve (MP), is to introduce a restricted master problem (RMP) in which we replace K_i with a finite subset K_i^{res} and solve (RMP) to optimality. Then we consider whether we can improve the solution to (RMP) by adding policies $k \in K_i \setminus K_i^{\text{res}}$ to K_i^{res}. To see if such policies exist for SKU i, we need to solve a sub-problem. (This sub-problem is also called the column generation problem or pricing problem.) We let p_a denote the dual variable of (RMP) corresponding with fleet $a \in A$ for constraint (15), ρ_c denote the dual variable of (RMP) corresponding with repair resource $c \in C$ for constraint (16) and v_i denote the dual variable of (RMP) corresponding with SKU i for constraint (17). If $i \in I_a^A \cap I_c^C$, then the sub-problem for SKU

i is given by:

$$
\begin{aligned}
(SUB(i)) \quad & \min_{\{(S_i,\mathbf{T}_i)\}} && C_i^a\left(S_i - S_i^{LB}\right) - p_a\mathcal{B}_i^{(S_i,\mathbf{T}_i)} - \rho_c u_i \mathcal{E}_i^{\mathbf{T}_i} - v_i \\
& \text{subject to} && S_i^{LB} \leq S_i \\
& && T_i(y) \leq S_i \qquad \forall y \in \Theta_i \qquad (18) \\
& && S_i, T_i(y) \in \mathbb{N}_0 \qquad \forall y \in \Theta_i. \qquad (19)
\end{aligned}
$$

If a feasible solution to $(SUB(i))$ exists with a negative objective value, then the objective of (RMP) can be improved by adding this solution to K_i^{res} and solving (RMP) with this larger set K_i^{res}. An optimal solution to (RMP) is also an optimal solution for (MP) if the optimal objective of $(SUB(i))$ is non-negative for each $i \in I$. Since (MP) is a relaxation of (P), we have also found a lower bound for problem (P) that we denote by C_P^{LB}.

Note that all policies that yield a negative objective for $(SUB(i))$ can improve the solution of (RMP), so we do not need to solve $(SUB(i))$ to optimality each time we obtain new dual variables from the restricted master problem. We do need to solve $(SUB(i))$ to optimality to verify that an optimal solution to (RMP) is also optimal for (MP). The next section treats heuristic and exact methods to solve $(SUB(i))$.

Finally, we note that $SUB(i)$ can be interpreted also as an economical problem where p_a denotes the cost of a backorder per time unit and ρ_c is the cost of expediting a part. A very similar problem is studied also in Arts et al. (2016).

4.1.1 Solving the sub-problem

The optimization problem $(SUB(i))$ is almost identical to the single-item problem discussed in Arts et al. (2016). The main differences are that:

- $(SUB(i))$ assumes a state dependent dual-index form for the control policy for each item
- The expediting thresholds in $(SUB(i))$ are restricted to be below S_i rather than any number in $\mathbb{N}_0 \bigcap \{\infty\}$.

However, the methods from Arts et al. (2016) can be applied almost immediately by observing that the form of the policy we assume is actually optimal as shown in Theorem 1 of Arts et al. (2016) and that constraint (18) can be accommodated by setting the constant M in Arts et al. (2016) equal to S_i. The exact and heuristic methods Arts et al. (2016) can easily be adapted to solve $(SUB(i))$ by restricting the search over S_i to be above S_i^{LB}.

4.2 Constructing a good feasible solution

Several methods have been suggested to find a good feasible solution based on a lower bound of the type constructed in the previous section. Kranenburg and Van Houtum (2007) and Kranenburg and Van Houtum (2008) suggest rounding the fractional solution obtained from solving (MP) and then performing a local search to find a good feasible solution. More recently, Alvarez et al. (2013) and Alvarez et al. (2015) suggest solving the final version of (RMP) after all columns have been generated as an integer linear program. Because they found very good results compared to local search algorithms, we also take that approach. To speed up the solution process we use the feasibility pump heuristic (Fischetti et al., 2005) and stop the solution of the integer linear program as soon as a feasible solution with optimality gap[1] of less than 0.5% is found or 1 minute has elapsed (whichever occurs first). This results in a feasible solution to (P) that is also an upper bound. We denote the cost of this solution by C_P^{UB}.

Alvarez et al. (2013) and Alvarez et al. (2015) report that this approach is computationally feasible with a commercial solver such as CPLEX. Our approach works well with the GLPK open source solver, even though the performance of this solver is consistently lagging in benchmarks[2].

Example 6. *For the instance of Thomas&Co, we find a lower bound on the optimal cost of* $C_P^{LB} = 846.39$ *kEURO. (Note that since all prices of parts are integer multiples of 1000 EURO, 846 kEURO is also a lower bound on the optimal costs of acquiring new repairable parts.) We also found a feasible solution with cost* $C_P^{UB} = 883$ *kEURO. This solution is shown in Table 3. The solution in Table 3 is further characterized by* $\sum_{i\in I^A_{\text{VILLAGE}}} \mathcal{B}_i^{(S_i,\mathbf{T}_i)} = 0.9156$, $\sum_{i\in I^A_{\text{CITY}}} \mathcal{B}_i^{(S_i,\mathbf{T}_i)} = 0.4749$, $\sum_{i\in I^C_{\text{OUTSOURCE}}} \mathcal{E}_i^{\mathbf{T}_i} = 176.23$, *and* $\sum_{i\in I^C_{\text{MECHANIC}}} \mathcal{E}_i^{\mathbf{T}_i} = 19.82$. *The optimality gap* $(C_P^{UB} - C_P^{LB})/C_P^{LB} \cdot 100\% = 4.3\%$.

Table 3: Feasible solution for the Thomas&Co instance of problem (P)

SKU#	S_i	$T_i(1)$	$T_i(2)$	$\lambda_i(1)$	$\lambda_i(2)$	$S_i - T_i(1)$	$S_i - T_i(2)$	C_i^a	$\mathcal{B}_i^{(S_i,\mathbf{T}_i)}$	$\mathcal{E}_i^{\mathbf{T}_i}$
1	19	19	11	1	5	0	8	30	0.4379	172.05
2	5	3	0	$\frac{1}{2}$	$\frac{9}{2}$	2	5	45	0.4773	8.95
3	10	10	-	4	-	0	-	5	0.0004	4.83
4	12	12	12	$\frac{2}{5}$	$\frac{12}{5}$	0	0	10	0.1339	4.18
5	2	1	0	$\frac{1}{5}$	$\frac{11}{5}$	1	2	30	0.3381	5.44
6	9	9	-	2	-	0	-	2	0.0028	0.60

The feasible solution that we find has intuitive properties: The expedite-up-to-level $S_i - T_i(y)$ *increases with demand intensity* $\lambda_i(y)$*. Thus, we buffer periods with high demand intensity by expediting repair of*

[1] Observe that this optimality gap is with respect to the integer linear programming formulation with a finite number of columns, *not* with respect to the original optimization problem.

[2] See for example the MIPLIB2010 (Koch et al., 2011) benchmark accessible via the benchmark site of Hans Mittelmann: `http://plato.asu.edu/bench.html`

parts. Furthermore, the expediting loads $\mathcal{E}_i^{\mathbf{T}_i}$ increase with the price C_i^a of a part (within a repair resource group). Expediting and inventory are both mechanisms to cope with demand uncertainty. As inventory buffers become more expensive (C_i^a increases), expediting repair becomes a better alternative to buffer demand uncertainty. Finally we observe that the expected number of backorders of a part increases with its price, except for cheap items with low expected backorders (SKUs 3 and 6). Since Thomas&Co care about the total expected number of backorders, it does not matter what parts generate most backorders. Thus investing in cheap parts to avoid backorders is more cost efficient and the solution reflects this.

5. Computational results

We discuss the questions we would like to answer, and the test bed we use in §5.1. We present and discuss the numerical results in §5.2.

5.1 Objectives and test bed

The objectives of this numerical study are to:

1. Determine whether the algorithm to find a feasible solution to (P) is effective, i.e., determine whether it finds solutions that are close to optimal;

2. Determine whether the algorithm to find a feasible solution to (P) is efficient, i.e., determine whether it finds a feasible solution within reasonable time;

3. Determine how much stock investments can be reduced by differentiating between regular and expedited lead times and how this reduction is achieved.

To answer these questions, we set up a large test bed of instances. The order of magnitude of problem parameters for our test instances are based on observations made at NedTrain and other maintenance companies we have observed over the years. We introduce the notation $U(a,b)$ for a uniform random variable on the interval (a,b). An overview of how instances in the test bed are generated is shown in Table 4. The total number of instances in the test bed is $3^5 2^3 = 1944$. For each combination of parameters 1,2,3,4,5,9, and 10 in Table 4, we generate two instances randomly as follows:

- For each SKU $i \in I$, we generate a Markov modulated Poisson demand process with $\mathbf{Q}$ generated as shown under 7 in Table 4, and $\boldsymbol{\lambda}$ generated by one of the two option shown under 8 in Table 4. (This is why two instances are generated.);

- Each SKU $i \in I$ is assigned uniformly at random to a repair resource set I_c^C for $c = 1, \ldots, |C|$;

- For each SKU $i \in I$, we generate an acquisition price from $U(100, 1000)$;

- For each SKU $i \in I$, set $u_i = 1$;

- The total load faced by a repair resource $c \in C$ is given by $\sum_{i \in I_c^C} \sum_{y \in \Theta_i} u_i \lambda_i(y) \mathbb{P}(Y_i = y)$. We generate $\mathcal{B}_a^{\max}$ such that a fraction $\xi \in \{0.2, 0.1, 0.5\}$ of the total repair load may be expedited; see 10 in Table 4. (Note that $\xi = 0.2$ is equivalent to saying that at most 20% of all repairs can be expedited.) Since the expedited and additional regular repair lead time are identical for all parts within an instance, this implies that the expected lead time over all parts together in a feasible solution is at least $\xi \ell_i + (1 - \xi)(\ell_i + \mathbb{E}[L_i]$.

- The total demand intensity across all parts within a fleet $a \in A$ is given by $\sum_{i \in I_a^A} \sum_{y \in \Theta_i} \lambda_i(y) \mathbb{P}(Y_i = y)$. The mean backorder constraints $\mathcal{B}_a^{\max}$ are set as a fraction $\nu \in \{0.05, 0.02, 0.01\}$ of the total demand intensity of that fleet so that the backorder targets are aligned with the mean total demand; see 9 in Table 4.

Table 4: Parameters for test bed instances

	Parameter	Values
1	Number of fleets $\|A\|$	1,2,4
2	Number of repair resources $\|C\|$	1,2,4
3	Number of SKUs per fleet $\|I_a^A\|$	20,50,100
4	Mean of additional regular repair lead time, $\mathbb{E}[L_i]$	2,4
5	Expedited repair lead time, ℓ_i	1,2
6	Acquistion cost for SKU $i \in I$, C_i^a	$U(100, 1000)$
7	Modulating chain generator for SKU $i \in I$, $\mathbf{Q}_i$	$\begin{pmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{pmatrix}$ with $q_1 = [U(200, 400)]^{-1}$, $q_2 = [U(5, 50)]^{-1}$
8	Demand intensity vector for SKU $i \in I$, $\boldsymbol{\lambda}_i$	$\begin{pmatrix} U(0.01, 0.1) \\ U(0.5, 1.5) \end{pmatrix}$, $\begin{pmatrix} U(0.01, 0.5) \\ U(1, 2) \end{pmatrix}$,
9	Upper bound on backorders for fleet $a \in A$, $\mathcal{B}_a^{\max}$	$\nu \sum_{i \in I_a^A} \sum_{y \in \Theta_i} \lambda_i(y) \mathbb{P}(Y_i = y)$ for $\nu = 0.05, 0.02, 0.01$
10	Upper bound on expediting load for resource $c \in C$, $\mathcal{E}_c^{\max}$	$\xi \sum_{i \in I_c^C} \sum_{y \in \Theta_i} u_i \lambda_i(y) \mathbb{P}(Y_i = y)$ for $\xi = 0.2, 0.1, 0.05$

Note that lead times in the original instance are equal across all SKUs. Of all the repair orders placed, a fraction of no more than ξ only has the expedited lead time ℓ_i while a fraction of at least $(1 - \xi)$ has expected lead time $\ell_i + \mathbb{E}[L_i]$. This means that the mean lead time experienced by a random order is at least $\xi \ell_i + (1 - \xi)(\ell_i + \mathbb{E}[L_i])$. Now for each *original* problem instance, we create a *benchmark* instance that is identical except that there is only one fixed lead time of $\xi \ell_i + (1 - \xi)(\ell_i + \mathbb{E}[L_i])$ (so no possibilities to expedite) [3].

Any cost difference between the optimal costs for these instances must therefore be ascribed to the flexibility of expediting when needed in response to demand fluctuations, but not to any mean lead time difference. This comparison will therefore shed light on objective 3 above of the numerical study.

Now for each generated original instance, we compute a feasible solution with cost C_P^{UB} as described in §4.2 and compare it to the lower bound C_P^{LB} that is obtained via the method described in §4.1:

$$\%GAP = \frac{C_P^{UB} - C_P^{LB}}{C_P^{LB}} \cdot 100\%. \tag{20}$$

This allows us to determine how effective our algorithm is at finding near optimal solutions.

Next we investigate the relative difference with the benchmark instance. We denote the lower bound on the optimal objective of the benchmark instance by C_{BENCH}^{LB}. We compare C_P^{UB} with C_{BENCH}^{LB}:

$$\%VAL = \frac{C_{BENCH}^{LB} - C_P^{UB}}{C_{BENCH}^{LB}} \cdot 100\%. \tag{21}$$

$\%VAL$ will reveal how much stock investments can be reduced by using expediting for smart lead time differentiation.

The algorithms described in §4.1-4.2 were programmed as a single threaded application in C with GLPK as the solver of both linear and integer linear programs. All computations were carried out on a PC running Windows (32 bit) with Intel Core Duo 2.33 GHz CPU and 4 GB of RAM.

5.2 Results

Table 5 shows the results of the computational experiment. For each of the parameters in Table 4 that has several settings, we computed the mean and maximum $\%GAP$ and $\%VAL$ as well as the mean and maximum computation time in seconds for each of the settings. We will now discuss objective 1-3 as stated in the previous subsection.

The average optimality gap of our feasible solution is very small at 0.67% but optimality gaps of up to 6.76% do occur. The optimality gap seems to increase with the number of fleets and repair resources. This is not surprising, because (MP) has $|I|+|A|+|C|$ constraints and the same number of basic variables in an optimal solution. Because of constraint (17), there is a basic variable for each $i \in I$. Therefore, there will be at most $|A|+|C|$ SKUs for which the optimal solution to (MP) is fractional. This explains why the optimality gap increases with both $|A|$ and $|C|$. Somewhat surprisingly, the optimality gap does not seem to decrease significantly with $|I_a^A|$. This is different form other multi-item spare parts problem where the optimality gap typically does decrease with the number of SKUs considered, (e.g Kranenburg

[3] The benchmark instance can be solved with the same algorithm as the original instance by not allowing expediting and making the static lead time equal to the required $\xi\ell_i + (1-\xi)(\ell_i + \mathbb{E}[L_i])$ of the original instance.

Table 5: Summary of computational results

		Optimality gap		Benchmark saving		Computation time (s)	
		(%GAP)		(%VAL)		CPU time (s)	
Parameter	Values	avg	max	avg	max	avg	max
Number of fleets, $\lvert A\rvert$	1	0.39	3.00	25.0	48.1	31	154
	2	0.56	6.76	25.0	49.1	76	313
	4	1.05	5.49	24.8	48.2	152	522
Number of	1	0.40	4.54	25.1	49.1	76	473
repair resources, $\lvert C\rvert$	2	0.55	3.57	25.1	48.2	85	511
	4	1.04	6.76	24.7	48.1	98	522
Number of SKUs	20	0.64	5.49	24.7	49.1	38	157
per fleet, $\lvert I_a^A\rvert$	50	0.68	4.71	25.0	47.9	80	282
	100	0.68	6.76	25.1	47.6	141	522
Fraction of total demand	0.05	0.72	6.76	24.9	48.2	78	462
per time unit that may be	0.02	0.68	5.49	24.8	48.0	88	502
backordered, ν	0.01	0.61	5.26	25.1	49.1	93	522
Fraction of total demand	0.2	0.88	6.76	32.2	49.1	86	452
per time unit that may be	0.1	0.62	4.71	24.3	38.8	88	502
expedited, ξ	0.05	0.50	3.72	18.3	31.4	85	522
Expedited repair	1	0.70	6.76	28.5	49.1	59	305
lead time, ℓ_i	2	0.63	5.26	21.4	42.7	113	522
Random demand	$\begin{pmatrix} U(0.01, 0.5) \\ U(1, 2) \end{pmatrix}$	0.72	5.26	28.2	49.1	111	522
intensity vector	$\begin{pmatrix} U(0.01, 0.1) \\ U(0.5, 1.5) \end{pmatrix}$	0.61	6.76	21.7	41.4	62	281
Additional regular repair	2	0.69	4.74	20.9	39.4	82	502
lead time, $\mathbb{E}[L_i]$	4	0.64	6.76	29.0	49.1	91	522
Total		0.67	6.76	24.9	49.1	86	522

and Van Houtum, 2007, 2008; Alvarez et al., 2013, 2015). This can be explained by the fact that we put a time limit of 1 minute on the integer linear programming solver.

The computation times of finding a feasible solution are 86 seconds on average and at most 522 seconds, which is quite acceptable given the size of the problems. It is also convenient that the computation time seems to scale linearly in the number of SKUs. Over 95% of the computation time for solving (MP) to optimality is spent in solving $(SUB(i))$. This task can also be parallelized on modern multi-core processors so that the computation time can be further reduced by a factor roughly equal to the number of cores on a processor.

The value of using expediting to influence repair lead times of repairables is quite valuable with an average benefit of 24.9% and even benefits of up to 49.1%. As was to be expected, the benefits increase with the fraction of total demand that can be expedited and decrease with the expedited lead time. But even the opportunity to expedite 5% of demand leads to average savings of as much as 18.3% compared to static lead times.

The solutions to the instances generally exhibit the same behaviour seen in Example 6: The expedite-up-to-levels $S_i - T_i(y)$ increase with demand intensity $\lambda_i(y)$ so that periods with high demand are buffered via expediting rather than inventory investments. Further it is generally true that expediting loads increase with the price of a part C_i^a because expediting buffers are cheaper than inventory buffers as C_i^a increases. There are parts that do not satisfy these general trends because of the integrality requirements. (This can be checked by observing that the non-integer lower bound solution does satisfy these trends.)

6. Conclusion

This article presented an efficient and effective algorithm to determine near optimal turn-around stock levels for a large group of repairable items that are used in the maintenance of several fleets of equipment. The use of expediting to influence the repair lead time of repairables was shown to be quite effective in reducing the stock investment needed to meet service levels for several fleets of equipment. This reduction is achieved mainly by expediting expensive parts during high demand fluctuations so that the required service level can be achieved with less investment in an expensive inventory buffer.

Acknowledgements

This research was funded by NedTrain and the Netherlands Organization for Scientific research. NedTrain also provided input for a real life case study conducted by Martijn van Aspert based on this model. We also thank the referees and the editor for constructive comments that improved the readability of the manuscript.

References

J. Abate and W. Whitt. Numerical inversion of probability generating functions. *Operations Research Letters*, 12:245–251, 1992.

I.J.B.F. Adan, A. Sleptchenko, and G.J. Van Houtum. Reducing costs of spare parts supply systems via static priorities. *Asia-Pacific Journal of Operational Research*, 26(4):559–585, 2009.

E.M. Alvarez, M.C. van der Heijden, and W.H.M. Zijm. The selective use of emergency shipments for service-contract differentiation. *International Journal of Production Economics*, 143:518–526, 2013.

E.M. Alvarez, M.C. van der Heijden, and W.H.M. Zijm. Service differentiation in spare parts supply through dedicated stocks. *Annals of Operations Research*, 231(1):283–303, 2015.

J.J. Arts, R.J.I. Basten, and G.J. Van Houtum. Repairable stocking and expediting in a fluctuating demand environment: Optimal policy and heuristics. *Operations Research*, accepted/in press, 2016.

R.J.I. Basten and G.J. Van Houtum. System-oriented inventory models for spare parts. *Surveys in Operations Research and Management Science*, 19(1):34–55, 2014.

R. Brooks and A. Geoffrion. Finding Everett's Lagrange, multipliers by linear programming. *Operations Research*, 14(6):1149–1153, 1966.

M.J. Carillo. Generalizations of Palm's theorem and Dyna-METRIC's demand and pipeline variability. *RAND report*, R-3698-AF, 1989.

G.B. Dantzig and P. Wolfe. Decomposition principle for linear programs. *Operations Research*, 8:101–111, 1960.

H. Everett. Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Operations Research*, 11(3):399–417, 1963.

W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.

M. Fischetti, F. Glover, and A. Lodi. The feasibility pump. *Mathematical Programming*, 10(1):91–104, 2005.

M.L. Fisher. The Lagrangian relaxtion method for solving integer programming problems. *Management Science*, 27(1):1–18, 1981.

V.D.R. Guide Jr. and R. Srivastava. Repairable inventory theory: Models and applications. *European Journal of Operational Research*, 102:1–20, 1997.

A.Y. Ha. Optimal dynamic scheduling policy for a make-to-stock production system. *Operations Research*, 45:42–53, 1997.

W.H. Hausman and G.D. Scudder. Priority scheduling rules for repairable inventory systems. *Management Science*, 28(11):1215–1232, 1982.

H. Heffes and D.M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 4(6):856–868, 1986.

R.J. Hillestad. Dyna-METRIC: Dynamic Multi-Echelon Technique for Recoverable Item Control. *RAND report*, R-2785-AF, 1982.

K.E. Isaacson and P.M. Boren. Dyna-METRIC version 6: An advanced capability assessment model. *RAND report*, R-4214-AF, 1993.

W.J. Kennedy, J.D. Patterson, and L.D. Fredendall. An overview of recent literature on spare parts inventories. *International Journal of Production Economics*, 76:201–215, 2002.

T. Koch, T. Achterberg, E. Andersen, O. Bastert, T. Berthold, R.E. Bixby, E. Danna, G. Gamrath, A.M. Gleixner, S. Heinz, A. Lodi, H. Mittelmann, T. Ralphs, D. Salvagnin, D.E. Steffy, and K. Wolter. MIPLIB 2010. *Mathematical Programming Computation*, 3:103–163, 2011.

A.A. Kranenburg and G.J. Van Houtum. Effect of commonality on spare part provisioning costs for capital goods. *International Journal of Production Economics*, 108:221–227, 2007.

A.A. Kranenburg and G.J. Van Houtum. Service differentiation in spare parts inventory management. *Journal of the Operational Research Society*, 59:946–955, 2008.

H.C. Lau and H. Song. Multi-echelon repairable item inventory system with limited repair capcity under non-stationary demands. *International Journal of Inventory Research*, 1(1):67–92, 2008.

W.K. Liang, B. Balcıoğlu, and R. Svaluto. Scheduling policies for a repair shop problem. *Annals of Operations Research*, 211:273–288, 2013.

N. Loeffen. Repair shop and inventory control for spare parts: min-max versus a lead time interface agreement. Master's thesis, Eindhoven University of Technology, 2012. URL `http://alexandria.tue.nl/extra2/afstversl/tm/Loeffen_2012.pdf`.

M.E. Lübbecke and J. Desrosiers. Selected topics in column generation. *Operations Research*, 53(6): 1007–1023, 2005.

K.S. Meier-Hellstern. A fitting algorithm for Markov-modulated Poisson processes having two arrival rates. *European Journal of Operational Research*, 29:370–377, 1987.

K. Moinzadeh and C.P. Schmidt. An $(S-1, S)$ inventory system with emergency orders. *Operations Research*, 39(3):308–321, 1991.

C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

J.A. Muckstadt. *Analysis and Algorithms for Service Part Supply Chains*. Springer: Berlin, 2005.

B.L. Nelson and I. Gerhardt. On capturing dependence in point processes: Matching moments and other techniques. *Working Paper*, 2010. URL `http://users.iems.northwestern.edu/~nelsonb/Publications/GerhardtNelsonSurvey.pdf`.

D.F. Pyke. Priority repair and dispatch policies for reparable-item logistics systems. *Naval Research Logisitics*, 37(1):1–30, 1990.

G.D. Scudder. Scheduling and labour assignment policies for a dual-constrained repair shop. *Intenational Journal of Production Research*, 24(3):623–634, 1986.

G.D. Scudder and R.C.H. Chua. Determining overtime policies for a repair shop. *Omega*, 15(3):197–206, 1987.

C.C. Sherbrooke. METRIC: A multi-echelon technique for recoverable item control. *Operations Research*, 16(1):122–141, 1968.

C.C. Sherbrooke. *Optimal inventory modeling of systems: Multi-echelon techniques*. Wiley, 2 edition, 2004.

F.M. Slay and C. Sherbrooke. The nature of the aircraft component failure process. Technical Report IR701R1, Logistics Management Institute, Washington D.C., 1988.

A. Sleptchenko, M.C. van der Heijden, and A. van Harten. Using repair priorities to reduce stock investment in spare part networks. *European Journal of Operational Research*, 163:733–750, 2005.

J.S. Song and P. Zipkin. Inventory control in a fluctuating demand environment. *Operations Research*, 41(2):351–370, 1993.

J.S. Song and P. Zipkin. Inventories with multiple supply sources and networks of queues with overflow bypasses. *Management Science*, 55(3):362–372, 2009.

H.G.H. Tiemessen and G.J. Van Houtum. Reducing costs of repairable inventory supply systens via dynamic scheduling. *International Journal of Production Economics*, 143:478–488, 2012.

H.C. Tijms. *A First Fourse in Stochastic Models.* John Wiley & Sons, 2003.

M. Van Aspert. Deciding on turn-around stock level using advanced demand models and expediting repair policies. Master's thesis, Eindhoven University of Technology, 2013. URL `http://alexandria.tue.nl/extra2/afstversl/tm/Aspert_van_2013.pdf`.

M. Van Aspert. Markov, dantzig en wolfe zorgen voor theoretische kostenbesparing van 40%. *StatOR*, 2: 14–16, 2014.

G.J. Van Houtum and B. Kranenburg. *Spare parts inventory control under system availability constraints.* Springer, 2015.

M.H. Veatch and L.M. Wein. Scheduling a make-to-stock queue: index policies and hedging points. *Operations Research*, 44:634–647, 1996.

J. Verrijdt, I. Adan, and T. de Kok. A trade off between emergency repair and inventory investment. *IIE Transactions*, 30:119–132, 1998.

T. Yoshihara, S. Kasahara, and Y. Takahashi. Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process. *Telecommunication Systems*, 17(1-2):185–211, 2001.

Y-S. Zheng and P. Zipkin. A queueing model to analyze the value of centralized inventory information. *Operations Research*, 38:296–307, 1990.

P.H. Zipkin. *Foundations of inventory management.* McGraw-Hill, 2000.

A. Determining $\mathbb{P}\{D_i^y(t,t+\ell_i)=k\}$

This section has been adapted from Arts et al. (2016). In this section, we show how $\mathbb{P}\left\{D_i^y(t,t+\ell_i)=k\right\}$ can be determined numerically. To this end, let $p_{y,y'}(k,\ell_e)=\mathbb{P}\left\{D_i^y(t,t+\ell_i)=k|Y_i(t+\ell_e)=y'\right\}$ be the (y,y')-entry of the matrix $\mathbf{P}(k,\ell_e)$. Then the matrix generating function $\widetilde{\mathbf{P}}(z,\ell_e)=\sum_{k=0}^{\infty}\mathbf{P}(k,\ell_e)z^k$ satisfies (e.g. Fischer and Meier-Hellstern, 1992):

$$\widetilde{\mathbf{P}}(z,\ell_e)=\exp\left([\mathbf{Q}-(1-z)\operatorname{diag}(\boldsymbol{\lambda})]\ell_e\right).$$

A plethora of numerical methods to compute the matrix exponential are discussed in Moler and Van Loan (2003). For the numerical work in this paper, we use the scaling and squaring algorithm with a Padé approximation. The probabilities $\mathbb{P}\left\{D_i^y(t,t+\ell_i)=k|Y_i(t+\ell_e)=y'\right\}$ can be obtained from $\widetilde{\mathbf{P}}(z,\ell_e)$ by numerical inversion using the LATTICE-POISSON algorithm of Abate and Whitt (1992) which uses the approximation

$$\begin{aligned}&\mathbb{P}\left\{D_i^y(t,t+\ell_i)=k|Y_i(t+\ell_e)=y'\right\}\\&\approx\frac{1}{2kr^k}\left\{\widetilde{\mathbf{P}}(r,\ell_e)+(-1)^k\widetilde{\mathbf{P}}(-r,\ell_e)+2\sum_{n=1}^{k-1}(-1)^n\operatorname{Re}(\widetilde{\mathbf{P}}(r\exp(n\pi\mathrm{i}/k),\ell_e))\right\},\end{aligned}$$

where $\mathrm{i}=\sqrt{-1}$, $0<r<1$ and $\operatorname{Re}(x)$ denotes the real part of the complex number x. The absolute error in this approximation is bounded by $\frac{r^{2k}}{1-r^{2k}}$ and so by choosing $r=10^{-\gamma/(2k)}$, we obtain an accuracy of approximately $10^{-\gamma}$. Then the needed probability, $\mathbb{P}\left\{D_i^y(t,t+\ell_i)=k\right\}$, can be found by un-conditioning:

$$\mathbb{P}\left\{D_i^y(t,t+\ell_i)=k\right\}=\sum_{y'\in\Theta_i}\mathbb{P}\left\{D_i^y(t,t+\ell_i)=k|Y_i(t+\ell_e)=y'\right\}\mathbb{P}\{Y_i(t+\ell_e)=y'|Y_i(t)=y\}$$

The probabilities $\mathbb{P}\{Y_i(t+\ell_e)=y'|Y_i(t)=y\}$ are found from the transient analysis of $Y_i(t)$. In particular, if we let $r_{y,y'}=\mathbb{P}\{Y_i(t+\ell_e)=y'|Y_i(t)=y\}$ be the (y,y')-th element of the matrix $\mathbf{R}(\ell_e)$, then $\mathbf{R}(\ell_e)=\exp(\ell_e\mathbf{Q})$.

B. Proof of Proposition 1 and examples

We start with some preliminaries. Consider a two state MMPP with generator $\mathbf{R}$ and intensity vector $\boldsymbol{\nu}$ given by

$$\mathbf{R}=\begin{pmatrix}-r_1 & r_1\\ r_2 & -r_2\end{pmatrix},\qquad \boldsymbol{\nu}=(\nu_1,\nu_2).$$

We let N_t denote the number of arrivals this MMPP generates in an interval of length t when it is in steady state. From Heffes and Lucantoni (1986), we have that

$$\mathbb{E}[N_t]=\frac{\nu_1r_2+\nu_2r_1}{r_1+r_2}\tag{22}$$

and

$$\mathbf{Var}[N_t] = \mathbb{E}[N_t] + 2At - \frac{2A}{r_1 + r_2}\left(1 - e^{-(r_1+r_2)t}\right) \tag{23}$$

with

$$A = \frac{r_1 r_2 (\nu_1 - \nu_2)^2}{(r_1 + r_2)^3}.$$

Now we start the proof of Proposition 1.

Proof. Let N denote the number of arrivals during one time unit in steady state in the MMPP in the proposition. Using (22), we find that

$$\mathbb{E}[N] = \frac{\lambda}{\alpha + 1}, \tag{24}$$

and equating this with μ and solving for λ yields

$$\lambda = (\alpha + 1)\mu. \tag{25}$$

Substituting (25) with $\mathbb{E}[N] = \mu$ into (23) yields

$$\mathbf{Var}[N] = \mu + \frac{2\alpha\mu^2}{(\alpha + 1)\beta} - \frac{2\alpha\mu^2}{(\alpha+1)^2\beta^2}\left(1 - e^{-(\alpha+1)\beta}\right). \tag{26}$$

Equating (26) with σ^2, and rearranging we obtain

$$(\sigma^2 - \mu)(\alpha + 1)\beta^2 - 2\alpha\mu^2(\alpha + 1)\beta + 2\alpha\mu^2 = 2\alpha\mu^2 e^{-(\alpha+1)\beta}. \tag{27}$$

Applying the quadratic root formula to (27) and simplifying, we find that if there is a $\beta > 0$ that satisfies

$$\beta = \frac{\mu\sqrt{2\alpha e^{-(\alpha+1)\beta}(\sigma^2 - \mu) + 2\alpha(\mu - \sigma^2) + \alpha^2\mu^2} + \alpha\mu^2}{(\alpha + 1)(\sigma^2 - \mu)}, \tag{28}$$

we have a fit. Now we show that such a unique $\beta^* > 0$ does exist provided

$$\alpha \geq \kappa\frac{\sigma^2 - \mu}{\mu^2}, \quad \text{and} \quad \frac{\sigma^2}{\mu} > 1, \tag{29}$$

for some $\kappa \geq 2$.

For convenience define $f : \mathbb{R}_+ \to \mathbb{R}$ as

$$f(\beta) = \frac{\mu\sqrt{2\alpha e^{-(\alpha+1)\beta}(\sigma^2 - \mu) + 2\alpha(\mu - \sigma^2) + \alpha^2\mu^2} + \alpha\mu^2}{(\alpha + 1)(\sigma^2 - \mu)} \tag{30}$$

where $\mathbb{R}_+ = [0, \infty)$ and let α, σ^2 and μ satisfy (29). To show that there is a unique $\beta^* > 0$ that solves (28), it suffices to show that $f(0) > 0$ and that $f'(\beta) < 0$ for all $\beta \in \mathbb{R}_+$. That $f(0) > 0$ can be verified directly and for $f'(\beta)$ we have

$$f'(\beta) = -\frac{\alpha\mu e^{-(\alpha+1)\beta}}{\sqrt{\alpha^2\mu^2 - 2\alpha\left(1 - e^{-(\alpha+1)\beta}\right)(\sigma^2 - \mu)}} < 0. \tag{31}$$

The strict inequality holds because (29) holds. Next we observe that $f'(\beta) > -1$ for all $\beta > 0$ and in particular for β^*, because $f''(\beta) > 0$ for all $\beta > 0$:

$$\begin{aligned} f''(\beta) = {} & \frac{\alpha\mu e^{-\frac{3}{2}(\alpha+1)\beta}\left\{2\alpha(\alpha+1)e^{(\alpha+1)\beta}(\mu-\sigma^2)+(\alpha+1)\alpha^2\mu^2e^{(\alpha+1)\beta}\right\}}{2\left\{\alpha^2\mu^2e^{(\alpha+1)\beta}+2\alpha e^{(\alpha+1)\beta}(\mu-\sigma^2)+2\alpha(\sigma^2-\mu)\right\}^{\frac{3}{2}}} \\ & + \frac{(\alpha+1)\alpha\mu e^{(\alpha+1)\beta}}{\sqrt{\alpha^2\mu^2e^{(\alpha+1)\beta}+2\alpha e^{(\alpha+1)\beta}(\mu-\sigma^2)+2\alpha(\sigma^2-\mu)}} > 0. \end{aligned} \tag{32}$$

The strict inequality again holds because (29) holds. Since $f'(0) = -1$, and $f'(\beta) < 0$ and $f''(\beta) > 0$ for all $\beta > 0$, we conclude that $|f'(\beta)| < 1$ for all $\beta > 0$ and in particular for β^*. This implies that β^* is an attractive fixed point of f. □

The fit provided in Proposition 1 is parameterized by $\kappa \geq 2$. To gain some intuition on the fit provided and the role of the parameter κ, we provide some examples of the distribution of N_1 that this fit generates in Figures 2 and 3.

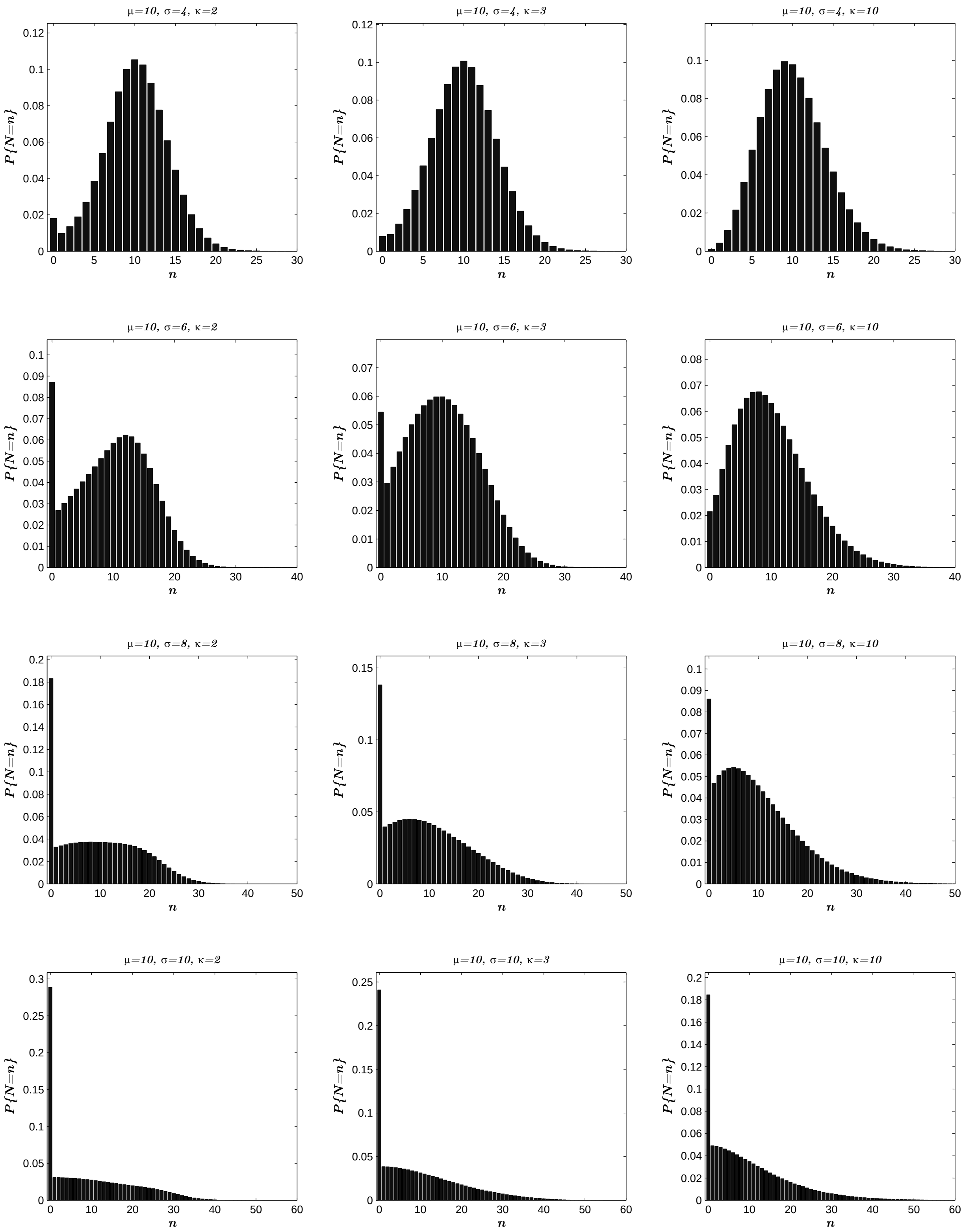

Figure 2: Fitted distributions generated by the procedure in Proposition 1. Standard deviation and the fitting parameter κ are varied as shown in the plots. As κ increases, probability mass shifts away from the origin.

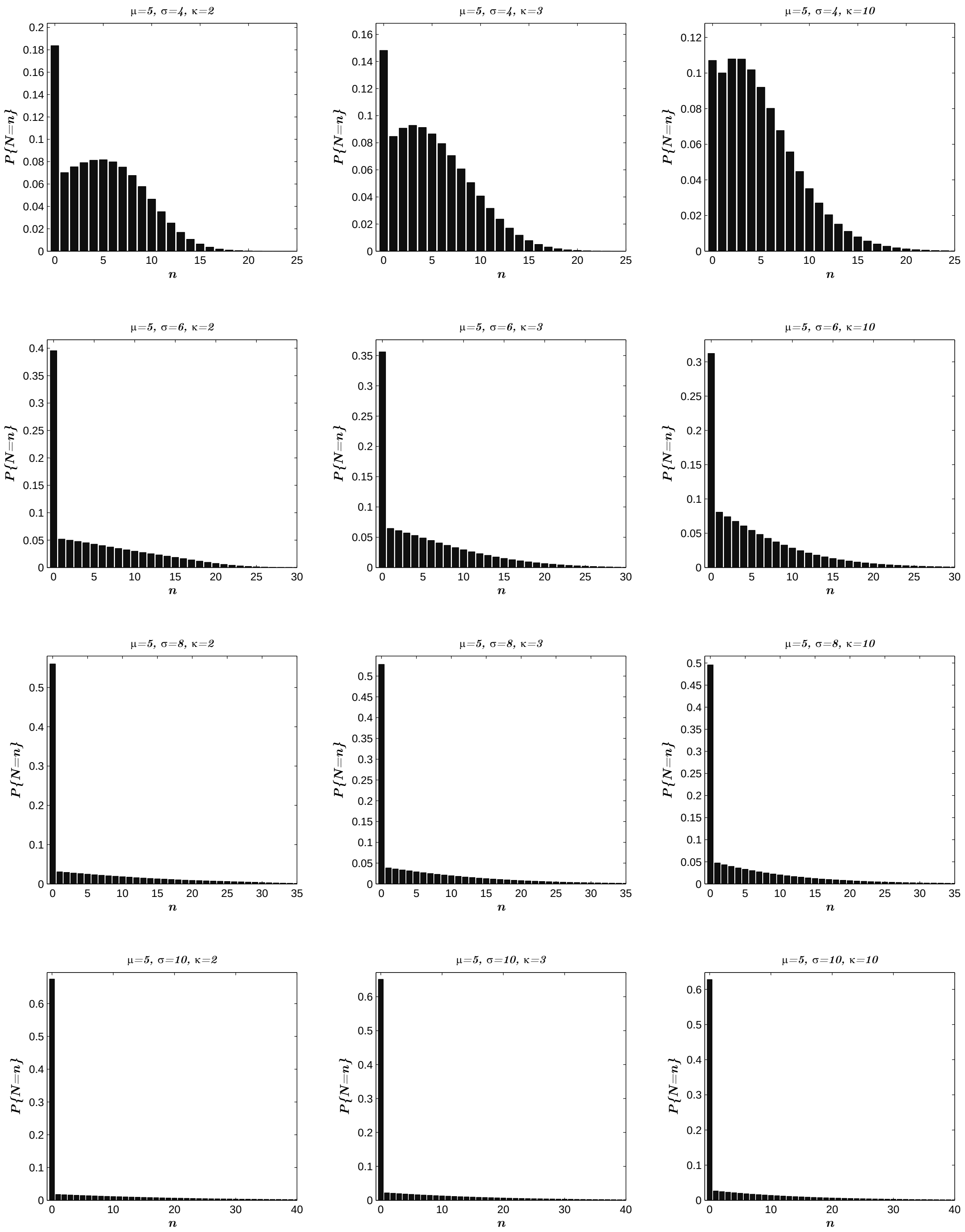

Figure 3: Fitted distributions generated by the procedure in Proposition 1. Standard deviation and the fitting parameter κ are varied as shown in the plots. As κ increases, probability mass shifts away from the origin.