Shipment Consolidation in Discrete Time and Discrete Quantity: Matrix-Analytic Methods

by

Qishu Cai

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Applied Science

in

Management Sciences

Waterloo, Ontario, Canada, 2011

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

Qishu Cai

I understand that my thesis may be made electronically available to the public.

Qishu Cai

ABSTRACT

Shipment consolidation is a logistics strategy whereby many small shipments are combined into a few larger loads. The economies of scale achieved by shipment consolidation help in reducing the transportation costs and improving the utilization of logistics resources.

The fundamental questions about shipment consolidation are i) to how large a size should the consolidated loads be allowed to accumulate? And ii) when is the best time to dispatch such loads? The answers to these questions lie in the set of decision rules known as shipment consolidation policies.

A number of studies have been done in an attempt to find the optimal consolidation policy. However, these studies are restricted to only a few types of consolidation policies and are constrained by the input parameters, mainly the order arrival process and the order weight distribution. Some results on the optimal policy parameters have been obtained, but they are limited to a couple of specific types of policies.

No comprehensive method has yet been developed which allows the evaluation of different types of consolidation policies in general, and permits a comparison of their performance levels. Our goal in this thesis is to develop such a method and use it to evaluate a variety of instances of shipment consolidation problem and policies.

In order to achieve that goal, we will venture to use matrix-analytic methods to model and solve the shipment consolidation problem. The main advantage of applying such methods is that they can help us create a more versatile and accurate model while keeping the difficulties of computational procedures in check.

More specifically, we employ a discrete batch Markovian arrival process (*BMAP*) to model the weight-arrival process, and for some special cases, we use phase-type (*PH*) distributions to represent order weights. Then we model a dispatch policy by a discrete monotonic function, and construct a discrete time Markov chain for the shipment consolidation process.

Borrowing an idea from matrix-analytic methods, we develop an efficient algorithm for computing the steady state distribution of the Markov chain and various performance measures such as i) the mean accumulated weight per load, ii) the average dispatch interval and iii) the average delay per order. Lastly, after specifying the cost structures, we will compute the expected long-run cost per unit time for both the private carriage and common carriage cases.

ACKNOWLEDGMENTS

This thesis would not have been possible without the contributions, assistance, support, and encouragement of many persons. To each of the following, I give my deepest thanks.

- To Professor Jim Bookbinder and Professor Qi-Ming He, my supervisors and mentors, for their encouragement, advice, support, patience, and more;
- To Professor Beth Jewkes and Professor Steve Drekic for their generous comments and suggestions;
- To Amy Xu for her loving support and encouragement.

Most of all, I owe immense thanks to my parents for generously putting up with me for so many years. This thesis is dedicated, with love, to them.

TABLE OF CONTENTS

CHAPTER FIVE

CHAPTER SIX

CHAPTER SEVEN

CHAPTER EIGHT

CHAPTER ONE

INTRODUCTION

1.1 Background to Shipment Consolidation

Shipment consolidation is a logistics strategy used by shippers to gain potential economies of scale in transportation. The basic idea is that instead of shipping individual loads whenever an order arrives, the shipper will hold the outbound shipments for a period of time, and combine several small loads with the same general destination into one larger load, and then dispatch them on the same vehicle. As a result, even though the shipper will incur greater inventory carrying costs, the potential savings in transportation costs due to freight volume discounts or better utilization of vehicles might prevail, and lead to reduced total cost.

A good shipment consolidation strategy must take into consideration the following key factors: the frequency of order arrivals, the weight and size of shipments, the availability and capacity of vehicles, and the service level requirement of customers. Since most of these factors are represented by random variables or stochastic processes, shipment consolidation is generally considered as a stochastic problem. The fundamental question here is: what is the optimal level of accumulation before dispatch in order to achieve the lowest total cost per unit time? Therefore, the shipment consolidation problem can also be classified as an optimization problem, with the objective of minimizing the expected total logistics (transportation plus inventory holding) cost per unit time.

We will now illustrate two simple practical applications of shipment consolidation. Through these examples, we will try to identify some necessary conditions for shipment consolidation to be effective, and demonstrate its benefit. At the same time, we will also make distinctions between the notions of "private carriage" and "common carriage" in shipment consolidation problems.

In our first example, we assume that transportation will be done by the shipper's own fleet; this is commonly referred to as the "private carriage" case. Suppose that a company selling custom-made furniture makes their own deliveries to customers free of charge. They only have one truck available for delivery, but have sufficient warehouse space to hold inventory. Thus, the delivery cost for each trip consists of fuel cost and labor cost, while the inventory holding cost is minimal. It usually takes the factory one day to build an order, but instead of delivering the furniture upon completion, the company chooses to hold the furniture and deliver them on the coming Monday. Therefore, all the orders from the previous week will be delivered on the following Monday.

This is a desirable application of shipment consolidation since the inventory holding cost for the company is almost negligible. It would be costly and inefficient if the company decided to make individual deliveries. By combining the deliveries, economies of scale are achieved, so the total delivery cost will be reduced. From the management perspective, it is also beneficial because, by setting a fixed delivery date, it is easier to schedule the delivery route and personnel. The fixed date can also provide buffer time for the factory if there are multiple orders arriving on the same day. From the customers' perspective, receiving custom-made furniture the next week free of delivery charges is usually acceptable.

In our second example, we consider transportation by an outside trucking company available for hire. This is commonly referred to as the "common carriage" case. This particular example was originally described by Higginson (1993), and states the following. Suppose there are 6000 pounds of roofing material required to be shipped every day. Inventory holding cost for this type of material is estimated by the manufacturer to be \$0.10*/cwt* (hundred pounds). The trucking company hired by the manufacturer to ship this material has offered a freight rate of \$2.95*/cwt.* Therefore, the total transportation and inventory holding cost each day can be calculated as:

Transportation cost: $TC_T = 60\ cwt \times \$2.95 / cwt = \$177 / day$

Inventory holding cost: $TC_H = 60 / 2\ cwt \times \$0.10 / cwt = \$3 / day$

Total cost: $TC = \$180 / day$

According to the previous result, the total weekly transportation and inventory holding cost is \$180 × 5 = \$900 / *week* (assuming five work days per week).

On the other hand, suppose the shipper chooses to use a consolidation strategy, in which they hold five daily shipments and dispatch them together at the end of each week on a 300 *cwt* load. The total weight of this single, larger shipment will enable the shipper to qualify for a volume discount with their carrier. As a result, the freight charge is now determined by the lower "volume rate". Assuming that the applicable volume freight rate is \$2.07 */ cwt*, based on this arrangement, the weekly transportation cost becomes

$$TC_T = 300\ cwt \times \$2.07 / cwt = \$621 / week\ .$$

However, due to the longer holding period, weekly inventory holding cost increases to

$$TC_H = 300 / 2\ cwt \times \$0.10 / cwt \times 5 = \$75 / week\ .$$

Using this shipment consolidation strategy, the shipper will benefit from the freight rate volume discount and reduce the total logistical cost from \$900 / *week* to \$696 / *week*. The shipper will gain a saving of \$204 / *week*, equivalent to almost 24% reduction over that with daily shipments.

1.2 Shipment Consolidation Policies

According to Higginson (1993), shipment consolidation can be carried out by the shipper, the consignee, the carrier, or by a third party such as a freight forwarder. This thesis focuses on shipper-performed consolidation. In this case, shipment consolidation is performed at the shipper's premises. The consolidated load will be transported either by the shipper's own fleet or by a common carrier. The shipper's main concern is to select an appropriate shipment consolidation policy which determines when to terminate the consolidation process and dispatch a load.

In the private carriage case, the shipper usually has its own logistic division and manages a private fleet of trucks. The transportation cost is an internal cost and is subjected to the shipper's logistical capability, such as vehicles, fleet capacity, transportation personnel and facility locations. The shipper will typically incur a fixed charge per shipment that is independent of the weight of the load. The shipper's objective is to consolidate shipments while trying to maintain a relatively low inventory holding cost and satisfactory service level. The primary component of cost saving comes from the reduced total dispatch cost if fewer trips are required to deliver the same amount of product.

In the common carriage case, transportation is by an outside trucking company available for hire. The shipper will be charged by the carrier according to the weight of its shipment. Suppose *c(w)* is the tariff charged by the carrier for a given shipment of weight *w*. The common carrier's tariff function has been defined by Çetinkaya and Bookbinder (2003) as

$$c(w) = \begin{cases} c_N w\ , & w < MWT \\ c_V w\ , & w \geq MWT \end{cases} \quad . \qquad (1.1)$$

Here c_V and c_N are the volume and non-volume freight rates, respectively, and $c_V < c_N$. *MWT* is the minimum weight required to qualify for a volume discount, as specified by the carrier.

In this situation, the shipper will often try to consolidate up to a weight greater than *MWT*. However, in practice, it is not always preferable to do so, because that will prolong the service time even though there simply may not be enough orders to accumulate to that weight level. If the actual consolidated weight, w, is slightly under *MWT*, and $c_N w > c_V MWT$, the shipper will choose to over declare the weight of its load as *MWT*, thus qualifying for the volume discount to lower the total cost. This effect is labeled as the shipment of "phantom freight" (Tyworth, 1987). It is also commonly referred to as the "bumping clause", whereby the actual weight is bumped into a higher weight category in order to receive the volume discount (Çetinkaya and Bookbinder, 2003).

Three types of policies for shipper-performed shipment consolidation have been reported in the logistics literature. They are quantity-based, time-based and time-and-quantity (*TQ*)-based consolidation policies. Newbourne and Barrett (1972) and Pollock (1978) identified them as practical policies initially. Subsequently, these policies have become popular industry practices. Assuming stochastic demand/order arrivals, analytical models have been developed for these policies. A common goal of those models is to examine the expected long-run cost per unit time.

According to Mütlü, Çetinkaya and Bookbinder (2010), “Under a quantity-based policy, customer orders are held/combined until a target load, assuring scale economies, is accumulated”. In a quantity policy model, the main decision variable is the target load Q. We seek the optimal value Q^* of that critical weight (Çetinkaya and Bookbinder, 2003). A consolidation cycle begins immediately after the previous dispatch, and ends upon arrival of the order which causes the cumulative weight to reach Q (or exceed Q for the first time). The cycle length is random, depending on the interarrival times between orders, and the load dispatched will often be greater than Q because of excess.

Similarly, “Under a time-based policy, consolidated shipments are released at periodic intervals; orders that arrive between the release epochs are combined” (Mütlü, Çetinkaya and Bookbinder, 2010). In a time policy model, the shipments are consolidated and dispatched on schedule every T units of time (the unit of time can be hours, days, or weeks, etc. depending on the characteristics of the order frequency). T is the main decision variable and it is also the constant cycle length. Naturally, the load dispatched is random, depending on the weight of the individual orders accumulated.

A TQ-based policy is actually a hybrid between the first two, thus we will refer to it as the “hybrid” policy from now on. It has two parameters: a target load Q and a maximum waiting time T. “Under this policy, a consolidated shipment is released either when the target load is accumulated or when the waiting time of an order exceeds a certain threshold before the target load is consolidated" (Mütlü, Çetinkaya and Bookbinder, 2010). Therefore, in a hybrid policy model, the decision variables are Q and T.

This policy has been regarded as a practical and effective alternative to the previous two classes of policies. "It is aimed at realizing both the scale economies inherent under quantity-based policies and the timely delivery benefits of time-based policies" (Mütlü, Çetinkaya and Bookbinder, 2010). In fact, hybrid policies are widely adopted in real-life for managing day-to-day operations associated with expedited orders.

In addition to the three well-documented consolidation policies, there have been other consolidation policies utilized by the shippers in practice. A majority of these policies take into account both the accumulated weight and the waiting time, but unlike the hybrid policy, the target load may vary over time. For instance, the shipper may aim to consolidate up to 200 *cwt* (hundred pounds) before dispatch within the first two days after the previous dispatch. If that is not achieved, during the next two days, a dispatch will be triggered when the accumulated weight exceeds 100 *cwt* And if that is still not attained, everything will be shipped on the fifth day.

Thus, the target load for these policies is typically a non-increasing step function of the time since last dispatch. Such a policy makes sense intuitively because as time passes and the accumulated weight remains low, it becomes less likely to attain the initial target load, hence the shipper should lower the target. In this thesis, we will refer to such policies as the “general” consolidation policies. It should be noted that the first three classes of policies are essentially special cases of this general policy.

1.3 Benefits and Drawbacks of Shipment Consolidation

The most significant benefit of shipment consolidation is the reduction in transportation cost. These cost savings can be achieved in several ways in different scenarios. The first type of cost saving is the “reduced cost of private carriage due to spreading of fixed transportation charges” (Higginson, 1993). Since private carriers operate their own fleet of vehicles, the total transportation cost for private carrier depends mainly on the distance and time. This cost may include fuel consumption, driver labor hours, vehicle maintenance and depreciation, etc. Thus, for a given distance, most of the transportation cost is fixed whether the vehicle is full or empty. Therefore, a fixed charge per load will be incurred, independent of the weight of the load. By the notion of economy of scale, if the shipper chooses to consolidate the shipments and dispatch larger loads, it will be able to reduce the total as well as the per-unit transportation cost.

The second form of cost saving is the reduced common-carrier freight rates due to freight volume discount. As we mentioned in the previous section, the shipper may choose to consolidate several shipments to increase the total weight of a load beyond the minimum weight required for a discounted freight rate, specified by the common carrier. Even if the shipper cannot attain the minimum volume weight, it may still obtain the volume discount by declaring "phantom freight" under the "bumping clause". The volume rates of common carriers are usually significantly lower than the non-volume rates. Newbourne and Barret (1972) remarked that the average less-than-truckload (*LTL*) freight rate is approximately twice the corresponding truckload rate. Higginson (1993) concluded through statistical sampling of freight rates from the U.S. Rail Uniform Freight Classification and the U.S. Motor Carrier Freight Classification, that the mean carload/truckload rate was approximately 60% of the mean *LTL* rate.

The benefits of shipment consolidation are not limited to financial gains only; consolidation also enhances the utilization of logistical resources and elevates the customer service level. Both private and common carriers handling consolidated shipments benefit from better utilization of vehicles and personnel because of the larger load sizes (Higginson, 1993). This has become an increasingly valuable aspect of shipment consolidation because in the last two decades, government legislators and environmental groups have identified the transportation industry as a main source of green house gas emission and global warming. The industry is under tremendous pressure to reduce its carbon footprint. Shipment consolidation has provided them with a viable option to do so.

From the accounting point of view, especially in the case of private carrier, shipment consolidation would allow the carrier to purchase fewer transportation assets, mainly vehicles. This would improve the liquidity of the company and help the company to stay lean, which is a popular management philosophy.

In terms of improved customer service level, shipment consolidation will result in more direct deliveries on dedicated vehicles, especially in the common carrier case (Higginson, 1993). This is because if a shipper tenders a small load to a common carrier, the carrier will usually consolidate it with other small shipments from other customers in an attempt to make a full truck load. However, these shipments will have different destinations; as a result, the full load will be transported to a local terminal for sorting and reloading on delivery vehicles. This shows that even though the shipper wishes to make frequent small deliveries, probably according to the demand by customers, the result is increased transportation time and less shipper control.

On the contrary, if the shipper chooses to consolidate some shipments into a larger load, the common carrier will be more likely to make a direct delivery right away, upon receiving that load. This will help the shipper to decrease the transportation time, reduce the handling of goods, increase his control over the shipments, and improve his position when negotiating with carriers (Higginson, 1993). The direct shipment of a consolidated load will also enhance service by lowering the chance of damage, loss or pilferage; allow for easier tracing of shipments; and lessen the administrative work relating to claims.

Despite its various advantages, shipment consolidation still has its disadvantages. Firstly, shipment consolidation will increase inventory levels and inventory holding costs. This is because the consolidation process requires the shipper to delay shipments in order to create effective consolidated loads; it also requires the consignee to hold larger safety stocks on their site to compensate for a possibly longer or more uncertain order lead time. Greater inventories also mean additional space is required for storage.

Another problem of shipment consolidation is that orders do not always come frequently or on a regular basis. A consolidation program may require that goods to be held until a minimum weight or volume has been reached, hence there will be additional holding time, depending on the interarrival times between orders. If that pattern is irregular, lengthy and erratic holding times could result, causing prolonged total lead times and order cycle lengths.

To effectively manage the shipment consolidation process, the shipper must keep thus track of customer orders presently waiting for shipment and those expected within a near term in the future. To achieve that will require more complicated administrative work and more frequent communication with customers. The consolidation process itself also requires close coordination between order processing, inventory control and transportation. All these will result in more administrative work and higher costs.

1.4 Literature Review on Shipment Consolidation Problem

In the last three decades, there have been many studies concerning shipment consolidation. Their common goal has been to discover accurate and effective ways to derive the optimal dispatch decisions. Researchers have used different techniques such as simulation, stochastic modeling or empirical analysis to solve this problem.

Some early publications on shipment consolidation focused on simulation-based cost comparison between the different consolidation policies. They include works by Masters (1980), Jackson (1981), Cooper (1984), and Closs and Cook (1987). Based on these preliminary works and through a more extensive simulation study, Higginson and Bookbinder (1994) examined the cost effectiveness of the three commonly used dispatch strategies, namely the time policy, the quantity policy and the time-and-quantity (*TQ*) policy. Their simulation results were based on a large range of the relevant parameters, long-run order arrival rates and maximum holding times. Using these results, they computed the cost per load, cost per hundredweight, and average order delay for each policy; and they made recommendations on how to choose the appropriate policy under different situations.

In his paper about “recurrent” decision approaches to shipment-release timing, Higginson (1995) made a distinction between the “recurrent approaches” and “non-recurrent approaches” to determining the optimal consolidation policy. The word “recurrent” means to re-evaluate the shipment-release question several times within an order accumulation cycle, in order to obtain the current optimal dispatch decision. In contrast, a non-recurrent approach “sets a target time or weight prior to accumulating orders and dispatches when the target is reached” (Higginson, 1995).

Many early analytical models were based on non-recurrent shipment consolidation approaches. These models tended to focus on the quantity policy and have often applied the concept of a deterministic economic shipment quantity (*ESQ*). That was intended to be the target dispatch weight which minimizes the total cost. The *ESQ* was also used to determine the long-run average cost. Examples of such works include Blumenfeld et al. (1985), Burns et al. (1985), Hall (1987), Daganzo (1988), Abdelwahab and Sargious (1990) and Russel and Krajewski (1991). These models could not address the issues of the occasional prolonged consolidation cycle which led to poor customer service. These difficulties are usually the result of variations in the order arrival process and order weight distributions.

Higginson (1995) presented two probabilistic models (for the cases of private carriage and common carriage respectively). By comparing the performance of these models with the non-recurrent dispatching policies, he concluded that the recurrent decision heuristic would outperform the non-recurrent ones when the economic shipment weight is close to vehicle capacity.

Higginson and Bookbinder (1995) proposed a Markovian Decision Process (*MDP*) approach to determine when to release consolidated loads, recurrently. In other words, whenever an order arrives, a choice must be made between dispatching this order plus all previously accumulated orders, or continuing to consolidate until at least the arrival of the next order. They constructed a discrete time Markov chain to represent the *MDP* model, and applied the fixed-weight aggregation technique to define a finite number of states for that Markov chain. Through some small but realistic numerical examples, their work has provided some inspiration on the potential of utilizing Markov chains to help solve shipment consolidation problems.

Other stochastic modeling techniques have been found useful in solving shipment consolidation problems. Gupta and Bagchi (1987) used stochastic clearing system theory to model shipment consolidation. They built a stochastic model for the quantity policy. Bookbinder and Higginson (2002) extended that idea to a hybrid policy and built a probabilistic model.

Çetinkaya and Bookbinder (2003) applied renewal theory to the quantity policy and time policy. They derived analytical expressions and explicit formulas to compute the optimal policy parameters. Even though their work was mainly based on the assumption of a Poisson order arrival process and exponentially distributed order weights, the results can be extended to some other order arrival processes or weight distributions. Recently, Mütlü, Çetinkaya and Bookbinder (2010) have extended the renewal theory model to the hybrid policy, and also extended its scope to cover the integrated inventory/consolidation problem.

Dispatch decisions in shipment consolidation problems depend heavily on the nature of the order arrival process and the weight distribution of orders. The weight of each order and the interarrival times between orders have often been modeled as independent and identically distributed (*i.i.d.*) random variables.

In the previous analytical studies about shipment consolidation, researchers have made different assumptions about the order arrival process. In their simulation model (1994), their *MDP* model (1995), and their probabilistic model (2002), Bookbinder and Higginson assumed that arrivals of orders follow a Poisson process (hence the interarrival times are exponentially distributed). In her two papers about freight consolidation and warehouse strategies, Cooper (1983, 1984) also assumed an exponential distribution for interarrival times between orders. On the other hand, Masters (1980) modeled interarrival times as a uniform distribution, while Ha, Khasnabis and Jackson (1988) used empirical distributions. In general, the majority of past studies have assumed the order arrival process to follow a Poisson process. Consequently, this has been regarded as a reasonable assumption by many researchers, supported by evidence gathered by shippers.

There are greater disparities in the published literature among the assumptions of order weight distribution. Empirical data suggest that the distributions of order weights not only vary among products, but they also vary between the different perspectives of shippers, carriers and purchasers. Therefore, there has not been any attempt to generalize the order weight distribution for shipment consolidation problems. Researchers have made their own assumptions based on the setting of their individual problems. For instance, Masters (1980) modeled order weight as a normal distribution; while Cooper (1984) and Ha, Khasnabis, and Jackson (1988) used truncated normal distributions.

In their series of papers on shipment consolidation, Bookbinder and Higginson (1994, 1995, 2002) have utilized an unshifted gamma distribution to model order weights. Their empirical data came from a medium-size national packaged goods distributor. After comparing the empirical plots with other data sets from the industry, they observed a common skewness towards lower weights in the empirical data. Thus, they used this property to justify their assumption.

Very few attempts have been made to fit theoretical probability distributions to empirical order weights, and those attempts did not provide satisfactory results. Akaah and Jackson (1988) tried to fit empirical data with the normal, uniform, and Poisson distributions. However, only about a quarter of their data sets actually fitted those distributions, and they did not suggest or test any better-fitting distributions.

Areas of concern for conducting analytical studies on shipment consolidation problems have thus been the choice of appropriate distributions for interarrival times and order weights. Previous literature has concentrated on developing and comparing models for only a few types of order arrival processes and order weight distributions. These are highly case sensitive and have limited application values. A more sophisticated model that is applicable to a wider range of order arrival processes and order weight distributions would be a breakthrough in research on shipment consolidation problems. We hope our studies in this thesis will shed some light on this matter.

1.5 Thesis Overview

The main objective of this thesis is to develop a method to evaluate any particular shipment consolidation policy, given an order arrival process and order weight distribution. Our goal is to make this method versatile and accurate so it can be applied to a variety of problem instances. We will also try to make this method computationally efficient and easy to use.

To achieve these goals, we need to find a way to accurately model any combinations of order arrival processes and order weight distributions; we also must be able to incorporate different types of consolidation policies; finally, some efficient algorithms must be developed to speed up the calculation procedures in our model.

The rest of this thesis is arranged as follows. In *Chapter Two*, we will introduce the basic ideas of matrix analytic methods and relate them to our problem. In *Chapter Three*, we will give a formal definition of our model and discuss its performance measures. Then in *Chapter Four*, we will present a Markov chain for our model and demonstrate how to compute its steady state distribution, among other long-run statistics.

Useful model modifications for special cases of interest will be presented in *Chapter Five*. In *Chapter Six*, we will show the formulas for measuring the costs under both private and common carriage. Finally, in *Chapter Seven*, we will present our numerical results, while *Chapter Eight* contains some concluding remarks about our model and future research directions.

CHAPTER TWO

INTRODUCTION TO MATRIX-ANALYTIC METHODS

Since we are going to use matrix analytical methods to model and solve the shipment consolidation problem, we need to provide some background and describe some basic techniques of those methods. Pertaining to this thesis, we will first introduce the concept of a Markovian arrival process (*MAP*) and its extension, the batch Markovian arrival process (*BMAP*). We will also introduce a versatile class of probability distribution called the phase-type (*PH*) distribution. Lastly, we will briefly describe the algorithmic approach for matrix geometric solutions.

2.1 Markovian Arrival Process

The Markovian Arrival Process (*MAP*) is a useful tool for modeling arrival processes. It was first introduced by Neuts (1979) as "an arrival process in which customers arrive at the epochs of transitions of an irreducible *K*-state Markovian process". More specifically, a *MAP* is a counting process that is defined on top of a finite state Markov chain (a.k.a. the underlying Markov chain).

Arrivals are typically associated with the transitions between states in the underlying Markov chain. However, in some cases, arrivals can also occur during the stay in certain states of the underlying Markov chain. For a discrete *MAP*, transitions and arrivals take place in discrete time epochs.

Let us now give a more formal definition of a discrete *MAP*. According to He (2009), if we have only one type of arrival event, we can find a pair of matrices (D_0 , D_1) of order *m*, such that $D_0 = [\ d_{0,ij}\ ]$ and $D_1 = [\ d_{1,ij}\ ]$, where $i, j = 1, 2, \ldots, m$. Each element $d_{0,ij}$ can be interpreted as the transition probability for the underlying Markov chain to go from state *i* to state *j* without any arrival, and $d_{1,ij}$ can be interpreted as the transition probability from state *i* to state *j* with an arrival.

We can also get $D = D_0 + D_1$, which is the transition probability matrix for the underlying Markov chain $\{\ I(t),\ t \geq 0\ \}$, where $I(t)$ is the current state of that underlying chain at time *t*. If we let $N(t)$ be the number of arrivals in the interval $[0, t]$ and $N(0) = 0$, then $\{\ N(t), I(t), t \geq 0\ \}$ is called a *Markovian Arrival Process*. It can be shown that $\{\ (\ N(t), I(t)\ ), t \geq 0\ \}$ is also a Markov chain with transition probability matrix

$$P = \begin{bmatrix} D_0 & D_1 & & \\ & D_0 & D_1 & \\ & & \ddots & \ddots \\ & & & \ddots \end{bmatrix}. \tag{2.1}$$

For each *MAP*, we can define its *steady-state arrival rate* as $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$, where $\boldsymbol{\theta}$ is the steady state distribution vector of *D*, i.e., $\boldsymbol{\theta} D = \boldsymbol{\theta}$ and $\boldsymbol{\theta}\mathbf{e} = 1$, and $\mathbf{e}$ is a column vector of ones (He 2009). At any arbitrary time, if the underlying Markov chain is in state *i*, the arrival rate at the moment is the i^{th} element of $D_1\mathbf{e}$. By conditioning on the state at that epoch, the average arrival rate at the epoch is $\boldsymbol{\theta} D_1 \mathbf{e}$. This explains why λ is called the “steady-state” arrival rate.

Now let us view *MAP* in the context of the shipment consolidation problem. It is often assumed that for a shipper that has a large customer base and frequent orders, the interarrival times of orders are i.i.d.. This is a general assumption about the order arrival process found in much previous research.

However, in some cases, it is necessary to consider the possibility that there exist some correlations between consecutive order arrivals. For instance, in certain industries, purchasing decisions are influenced by factors such as seasonality, actions by competitors, and economic conditions. *MAP* can be used to capture the correlations between consecutive interarrival times.

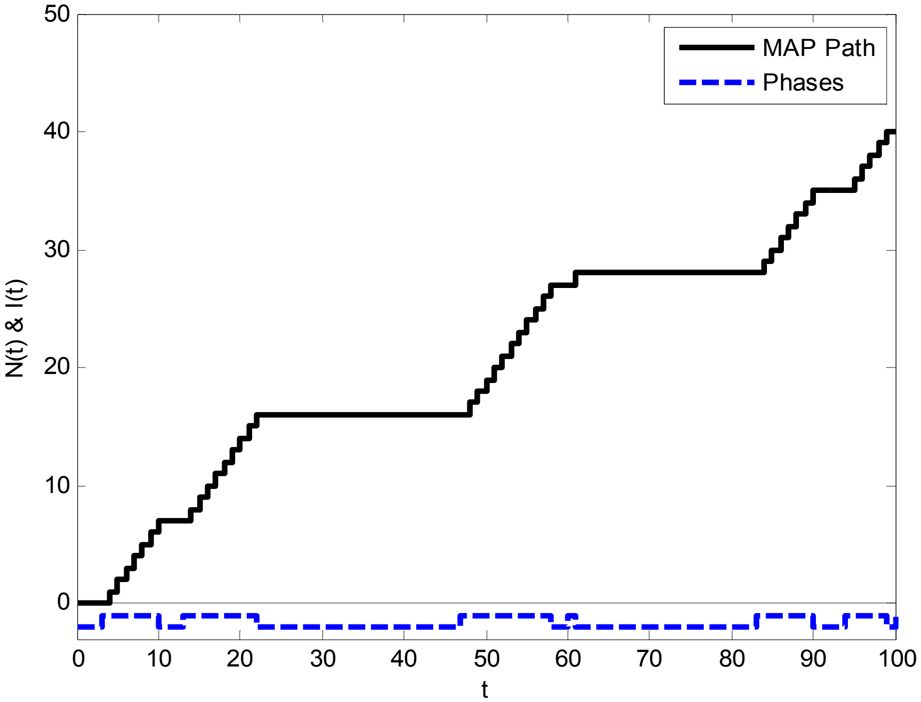
Consider an example of a *MAP* defined as

$$D_0 = \begin{bmatrix} 0 & 0 \\ 0.1 & 0.9 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 0.9 & 0.1 \\ 0 & 0 \end{bmatrix}. \tag{2.2}$$

It can be observed that, if the process is in state 1, there is a high probability (0.9) for an arrival to occur while the process remains in that state, and a low probability (0.1) for the process to transit to state 2 with an arrival. On the other hand, if the process is in state 2, no arrival will occur and there is a high probability (0.9) to remain in state 2.

In this particular example, the expected times between transitions are roughly the same (around 10); these intervals can also be interpreted as the average time spent in each state. Therefore, according to He (2009), this *MAP* has a distinctive ‘bursty’ nature, with alternating ‘busy’ periods (time in state 1 with frequent arrivals) and ‘idle’ periods (time in state 2 with no arrivals). *Figure 2.1* shows the sample paths of $N(t)$ and $I(t)$ in a single realization of this *MAP*.

Figure 2.1: Plot of *MAP* Path and State Changes



Similar to Poisson processes, *MAPs* can be combined or decomposed. To demonstrate the decomposition (otherwise known as "marking") of *MAPs*, let us assume that $\{ (N(t), I(t)), t \geq 0 \}$ is a *MAP* with matrix representation (D_0 , D_1). If for some probabilities $(p , 1 - p)$ we can mark arrivals independently as two types, we can then obtain a marked Markovian arrival process $\{ (N_1(t), N_2(t), I(t)), t \geq 0 \}$ with matrix representation $(D_0, pD_{1,} (1 - p)D_1)$. Breaking it into individual processes, $\{ N_1(t), t \geq 0 \}$ and $\{ N_2(t), t \geq 0 \}$ are two *MAPs* with matrix representations $(D_0 + (1 - p)D_1 , pD_1)$ and $(D_0 + pD_1 , (1-p)D_1)$.

2.2 Batch Markovian Arrival Process

A Batch Markovian arrival process (*BMAP*) is a direct generalization of *MAP*. The idea is based on interpretation of the transition / arrival probabilities. For a *MAP* (D_0 , D_1), elements of D_0 are interpreted as the transition probabilities without an arrival, while elements of D_1 are interpreted as transition probabilities with an arrival. We can define more complicated arrival processes by dividing the elements in D_1, and assigning different meanings to those probabilities.

Recall that for *MAPs,* we only allow at most one order to arrive in each period, but for *BMAPs*, we allow more than one order. When more than one order (but at most N) does arrive in any period, we group them into batches. Thus, to distinguish the arrival of different batch sizes, we break D_1 into N matrices, such that $D_1 = D'_1 + D'_2 + \ldots + D'_N$. Each element in matrix D'_i, for $i = 1, 2, \ldots, N$, represents the transition probabilities of an arrival of batch size i.

Many *BMAPs* are explicitly defined in terms of matrices D_i . Suppose, however, that the batch sizes are independent of the arrival process and there exists a probability p_i for the occurrence of batch size i, such that $\sum_{i=1}^{N} p_i = 1$. We can then simply decompose (D_0 , D_1) into (D_0 , p_1D_1 , p_2D_1 , …, p_ND_1), which becomes a *BMAP*. Note that in this case, N does not necessarily have to be finite; if $N = \infty$, the *BMAP* will have infinitely many batch sizes.

Similar to *MAPs*, the counting process { $N(t)$, $t \geq 0$ } still records the number of arrivals for a *BMAP* within interval [0, t]; { $I(t)$, $t \geq 0$ } still represents the underlying Markov chain. If we express the *BMAP* as (D_0, D_1, D_2, …, D_N), then $D = D_0 + D_1 + D_2 + \ldots + D_N$ is the transition probability matrix of the underlying Markov chain. { ($N(t)$, $I(t)$), $t \geq 0$ } is a Markov chain with transition probability matrix

$$P = \begin{bmatrix} D_0 & D_1 & \cdots & D_N & & \\ & D_0 & D_1 & \cdots & D_N & \\ & & \ddots & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{2.3}$$

The arrival rate of *batches* (not the total number of arrivals) is given by $\hat{\lambda} = \boldsymbol{\theta}\left(\sum_{j=1}^{N} D_j\right)\mathbf{e}$. The arrival rate of a batch of size j is thus $\lambda_j = \boldsymbol{\theta} D_j \mathrm{e}$ (He 2009).

A very important theorem about *BMAPs* states that any stochastic arrival process can be approximated closely by a *BMAP* (Asmussen and Koole, 1993). The latter will therefore be an extremely versatile tool for modeling order arrival processes in shipment consolidation problems.

In conclusion, *MAP* and *BMAP* can both be used to approximate any arrival process to a high precision. They are also useful in terms of modeling special characteristics such as a business cycle, seasonality, and busy/idle periods in order arrival processes. For more details on *MAPs* and *BMAPs*, see, for example Neuts (1979, 1981, 1992), Lucantoni (1991), He and Neuts (1998), Latouche and Ramaswami (1999) and Latouche, Remiche and Taylor (2003).

2.3 Phase-Type Distribution

Phase-type distributions (*PH*-distributions) were first introduced by Neuts (1975) as an extension of the Erlang distribution. Due to the ease of generating explicit matrix geometric solutions (Neuts, 1981), *PH* is a highly versatile class of probability distributions that is widely used in the analysis of queueing models and other complex stochastic models. The class can be divided into two subgroups: continuous and discrete phase-type distributions.

Since we are dealing with discrete time and discrete quantity in this thesis, we will focus on the discrete phase-type distribution. We can think of it as a probability distribution that arises from a system of one or more interrelated geometric distributions occurring in some sequence. A discrete *PH*-distribution is also the distribution of the time until absorption of a Markov chain with finitely many states, where all states are transient except for one absorbing state.

According to Neuts (1981), a discrete *PH*-distribution can be defined by a Markov chain with $m + 1$ states that has the transition probability matrix of the following form

$$P = \begin{bmatrix} S & \mathbf{S}^0 \\ \mathbf{0} & 1 \end{bmatrix}, \tag{2.4}$$

where $\mathbf{S}^0 = \mathbf{e} - S\mathbf{e}$, and $\mathbf{e}$ is a column vector of ones.

Note that S is a substochastic matrix, such that all entries in S are non-negative; $\sum_{j=1}^{m} S_{ij} + \mathbf{S}_i^0 = 1$, for each row i (or this can be expressed as $S\mathbf{e} + \mathbf{S}^0 = \mathbf{e}$); and $I - S$ is nonsingular. The initial probabilities are given by $(\boldsymbol{\beta}, \beta_{m+1})$, with $\boldsymbol{\beta}\mathbf{e} + \beta_{m+1} = 1$.

Based on the definition given by He (2009), a discrete *PH*-distribution can be represented as $Y \sim (\boldsymbol{\beta}, S)$. Its probability mass function is defined as

$$f(t) = \boldsymbol{\beta} S^{t-1} \mathbf{S}^0, \quad t \geq 1 \text{ and } f(0) = \beta_{m+1} ; \tag{2.5}$$

its distribution function is given by

$$F(t) = P\{X \leq t\} = 1 - \boldsymbol{\beta} S^t \mathbf{e} ; \tag{2.6}$$

and its probability generating function is expressed as

$$P(Z) = \beta_{m+1} + z\boldsymbol{\beta}(I - zS)^{-1} \mathbf{S}^0 . \tag{2.7}$$

The k^{th} factorial moments are thus

$$\begin{aligned} P^k(1) &= k!\boldsymbol{\beta} S^{k-1}(I - S)^{-k} \mathbf{e} \\ P^k(1) &= E[X(X-1)\cdots(X-k+1)], \quad k \geq 1 \end{aligned} . \tag{2.8}$$

From the moment functions, we can derive the mean and variance as

$$E[X] = P^{(1)}(1) = \boldsymbol{\beta}(I - S)^{-1} \mathbf{e} ; \tag{2.9}$$

$$\begin{aligned} Var(X) &= P^{(2)}(1) + P^{(1)}(1) - \left(P^{(1)}(1)\right)^2 \\ Var(X) &= 2\boldsymbol{\beta} S(I - S)^{-2} \mathbf{e} + \boldsymbol{\beta} S(I - S)^{-1} \mathbf{e} - \left(\boldsymbol{\beta} S(I - S)^{-1} \mathbf{e}\right)^2 \end{aligned} . \tag{2.10}$$

Neuts (1981) stated that in the analysis of even very simple stochastic models, the increasing complexity of the ensuing conditional probability distributions would constitute a barrier for obtaining any explicit solutions. Similar to the exponential distribution and the Poisson process, the *MAP*, *BMAP* and *PH*-distributions all share the Markov property. That will improve the ease of conditioning, a valuable property for them to be used in stochastic modeling. It will help reduce the difficulty in deriving exact and detailed numerical results about the steady-state properties of a model. Many probability models that have matrix-geometric solutions involve *MAPs*, *BMAPs* or *PH*-distributions in one way or another.

Neuts (1981) also mentioned another advantage for utilizing *MAP*, *BMAP* or *PH*-distributions in stochastic modeling. Due to the growing importance of qualitative modeling, many models endeavor to capture the true nature of stochastic processes, such as fluctuating arrival rates, seasonal patterns of inventories, priority rules, etc, as opposed to imposing restrictive distributional assumptions. General distributions tend to fail or become too complicated and intractable in such models. Fortunately, probability distributions of point processes, which are well represented by *MAP*, *BMAP* and *PH*-distributions, can be used to reflect the qualitative features of such a model, and yet still remain mathematically elementary and computationally tractable.

Perhaps the most notable advantage of employing *PH*-distributions in stochastic modeling is the fact that they are dense in the set of all probability distributions on [0, ∞). This property was first noted by Neuts (1975), and it allows *PH*-distributions to approximate all other positive valued distributions.

Asmussen, Nerman and Olsson (1996) further commented that "due to the denseness, one can view phase-type modeling as a semi-parametric density estimation procedure with a built-in smoothing"; the degree of smoothness was said to be determined by the number of phases *m*. The phases would have no physical interpretation under such applications. However, in some other applications, we may find meaningful probabilistic interpretations for the phases.

The procedure to estimate the parameters of a *PH*-distribution according to some empirical data, or with respect to some other known distribution, is commonly known as "*PH*-fitting". There have been many previous publications concerning *PH*-fitting techniques and their effectiveness (see, for example, Johnson and Taaffe (1990a, 1990b), Asmussen and Nerman (1991), Bobbio and Cumani (1992), Horvath and Telek (2002) and Thummler, Buchholz and Telek (2006)).

The above mentioned advantages of *PH*-distributions indicate that they can be useful in modeling shipment consolidation problems. Since order weights are usually i.i.d. and tend to have arbitrary empirical distributions, *PH*-distributions can be used to approximate them. This allows us to model each problem with a distinctive *PH*-distribution of order weights, as opposed to being restricted by the limited parameter choices of exponential or gamma distributions, which were traditionally used to model order weights.

When the quantity policy is employed in practice, one cannot guarantee that each load dispatched will have the precise weight of the target load. “Excess weight” occurs frequently. Previous studies such as Çetinkaya and Bookbinder (2003) modeled order weights as an exponential distribution. Thus, by applying the memory-less property of exponential distributions, the distribution of excess weight is also shown to be exponential, with the same parameter as the order weight distribution. Under such an assumption, the expected excess weight is thus equal to the expected order weight. This is not true in practice, where the excess weight is usually much smaller than the average order weight.

Our model aims to capture the distribution of excess weight more precisely, and *PH*-distributions will provide an effective solution. If we treat the sequence of order weights as a *PH*-renewal process, according to the memoryless property, the future phases of this process depend only on present phases but not on past phases. Thus, if we can find the distribution of phases at the time when accumulated weight reaches the target load, we can then find an explicit *PH*-distribution for the excess weight.

2.4 Matrix Geometric Solution

In stochastic modeling, we often encounter Markov chains of large dimension, for which we have to solve for their steady state distributions (i.e. obtain the limiting probabilities). Consider a discrete time Markov chain with transition probability matrix P. Even if P is sparse, but has very large dimension, it can be challenging to solve for its steady state distribution $\boldsymbol{\pi}$. However, if P also has some special block structures, we can use an algorithmic approach to compute $\boldsymbol{\pi}$.

For instance, suppose P is defined as

$$P = \begin{bmatrix} A_{0,0} & A_{0,1} & & & \\ A_{1,0} & A_{1,1} & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \end{bmatrix}, \tag{2.11}$$

where $\{ A_0, A_1, A_2, A_{0,0}, A_{0,1}, A_{1,0} \}$ are nonnegative matrices of size m; and we must have $(A_0 + A_1 + A_2)\mathbf{e} = \mathbf{e}$, $A_{1,0}\mathbf{e} + A_{1,1}\mathbf{e} + A_0\mathbf{e} = \mathbf{e}$ and $A_{0,0}\mathbf{e} + A_{0,1}\mathbf{e} = \mathbf{e}$. This type of Markov chain represents what is commonly known as a Quasi Birth-and-Death (*QBD*) process. For more detail on *QBD*, please refer to Neuts (1981) and Latouche and Ramaswami (1999).

When the limiting probabilities $\boldsymbol{\pi}$ exist, we need to solve the linear system $\boldsymbol{\pi}P = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\mathbf{e} = 1$ to find $\boldsymbol{\pi}$. Taking advantage of the diagonal block structure, we expand the first equation as

$$\begin{aligned}
\boldsymbol{\pi}_0 &= \boldsymbol{\pi}_0 A_{0,0} + \boldsymbol{\pi}_1 A_{1,0}; \\
\boldsymbol{\pi}_1 &= \boldsymbol{\pi}_0 A_{0,1} + \boldsymbol{\pi}_1 A_{1,1} + \boldsymbol{\pi}_2 A_2; \\
\boldsymbol{\pi}_n &= \boldsymbol{\pi}_{n-1} A_0 + \boldsymbol{\pi}_n A_1 + \boldsymbol{\pi}_{n+1} A_2, \quad n \geq 2.
\end{aligned} \tag{2.12}$$

According to Neuts (1981), equations (2.12) have a matrix-geometric solution of the form $\boldsymbol{\pi}_n = \boldsymbol{\pi}_1 R^{n-1}$ for $n \geq 1$. Substituting that solution back into the equations, we obtain, for $n \geq 2$, $\boldsymbol{\pi}_1 R^{n-2}(R - A_0 - RA_1 - R^2 A_2) = 0$. A nonnegative matrix R can be found, satisfying equation $R = A_0 + RA_1 + R^2 A_2$; R is usually referred to as the *rate matrix.* Limiting probabilities $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ can be found accordingly by solving the first two equations. A more formal definition of the matrix-geometric solution follows.

Theorem 2.1 The stationary distribution of a *QBD* is given by Neuts (1981) as

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_1 R^{n-1}, \quad n \geq 1 \; ; \tag{2.13}$$

where the *rate matrix R* is the minimal nonnegative solution of the nonlinear equation

$$R = A_0 + RA_1 + R^2 A_2 \ . \tag{2.14}$$

Vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ are the unique positive solutions to

$$\begin{aligned} \boldsymbol{\pi}_0 &= \boldsymbol{\pi}_0 A_{0,0} + \boldsymbol{\pi}_1 A_{1,0}; \\ \boldsymbol{\pi}_1 &= \boldsymbol{\pi}_0 A_{0,1} + \boldsymbol{\pi}_1 (A_{1,1} + RA_2); \\ 1 &= \boldsymbol{\pi}_0 \mathbf{e} + \boldsymbol{\pi}_1 (I - R)^{-1} \mathbf{e}. \end{aligned} \tag{2.15}$$

The matrix-geometric solution is a fundamental result in matrix-analytic methods. The approach of using the rate matrix *R* to compute the steady state distribution has been extended to other Markov chains with special structures such as the *M*/*G*/1 or *GI*/*M*/1 type Markov chains. In this thesis, we will also take advantage of that algorithmic approach. For more about the matrix-geometric solution and matrix-analytic methods in general, see, for example Neuts (1981, 1989a, 1989b), Lucantoni and Ramaswami (1985), Hsu and He (1991), Gail, Hantler and Taylor (1994, 1997) and Latouche and Ramaswami (1993, 1999).

CHAPTER THREE

THE DISCRETE SHIPMENT CONSOLIDATION MODEL

Recall that in Chapter One, we introduced the shipment consolidation problem, and in Chapter Two, we showed some basic techniques of matrix-analytic methods. In this chapter, we will model the discrete version of the shipment consolidation problem using matrix-analytic methods. More specifically, we will discuss how to model the order arrival process and weight distribution using a single *BMAP*, how to define a consolidation policy as a discrete function, and will note which performance measures to record for a shipment consolidation process.

3.1 Model Introduction

The main function of our model is to evaluate a specific shipment consolidation policy by either private or common carriage. Due to the randomness in the order arrival process and order weight, we utilize a Markov chain to mimic the actual shipment consolidation process. To exploit some existing methodologies for discrete time Markov chains, we assume that both the time and the weight of orders are discrete.

At the beginning of each period, orders are received by the shipper who then decides whether or not to dispatch a shipment by the end of that period. This decision is based on the total accumulated weight of all outstanding orders and the total time elapsed since the last dispatch. If the accumulated weight exceeds a threshold or the elapsed time surpasses a certain point, these orders are consolidated into one load and promptly dispatched. After that, a new cycle of accumulation and dispatch commences in the following period with zero initial weight.

To begin constructing our model, we first need to establish the order arrival process and identify the order weight distribution. This can be done through fitting empirical data or approximation by known stochastic processes and probability distributions. Next, we need to define a policy to govern the shipment consolidation process and find a suitable representation for it. Then, we can model the process as a discrete time Markov chain and solve for its steady state distribution. From those results, we can eventually measure the effectiveness of that particular shipment consolidation policy, and compute the cost of having it implemented by either private or common carrier.

3.2 The Weight-Arrival process

Unlike many previous studies which used a Poisson process to model the order arrival process, we choose to use the discrete *BMAP*. This allows us to combine the order weight distribution and the order arrival process through a convolution. (From now on, when we use the term "weight-arrival process," we shall mean this convolution.)

There are several advantages behind our choice of *BMAP*. One worth noting is that *BMAP* lets us model situations in which order weights are correlated with order frequency. Another is that, as mentioned in Section 2.2, the *BMAP* is capable of approximating almost any stochastic arrival process, which we hope will make our model more versatile and accurate.

Without loss of generality, we assume that at most one order can arrive in any period. The weight of orders accumulates according to a *BMAP* representing the weight-arrival process. Recall from Chapter Two, a *BMAP* can be denoted by a matrix-representation (D_0, D_n, $n = 1, 2, \ldots$), where D_0 and D_n are $m \times m$ nonnegative matrices and m is a positive integer. For $n > 0$, the matrices D_n contain the probabilities that an order of weight n will arrive in a period. On the other hand, the matrix D_0 contains the probabilities of no order arriving during that period.

Since the weight-arrival process is defined by a *BMAP*, it has an underlying Markov chain with m states. We denote that process as$\{ I_a(t), t = 0, 1, 2, \ldots \}$, where $I_a(t)$ is the state of this underlying Markov chain at the beginning of period t and $I_a(t) \in \{1, 2, \cdots, m\}$. The transition probability matrix for that chain is given as $D = \sum_{n=0}^{\infty} D_n$. If we assume that the underlying Markov chain is irreducible, then so is *D*. By definition of *BMAP,* the row sums of *D* are all equal to one, so *D* is stochastic.

The intuition behind the underlying Markov chain is that the weight-arrival process alternates between *m* states. For instance, if order-arrival frequencies vary under different business scenarios, then the weight-arrival process may change from one period to the next as the scenario changes. The probability that an order arrives in a given period may vary, depending on the scenario, and so may the order weight distribution. If we denote each entry in matrices D_n, $n = 1, 2, \ldots$ as $[\ D_n\ ]_{ij}$, then each entry represents the probability for an order of weight *n* to arrive. Following this arrival, the scenario changes from *i* to *j*.

D is stochastic and irreducible. By the property of discrete time Markov chains, there exists a steady state distribution for the underlying Markov chain, denoted by the row vector $\boldsymbol{\theta}_{\mathrm{a}} = (\ \theta_{a,1},\ \ \theta_{a,2},\ \ldots,\ \theta_{a,m}\ )$. Consequently, $\boldsymbol{\theta}_{\mathrm{a}}$ is the unique solution to the linear system $\boldsymbol{\theta}_{\mathrm{a}}D = \boldsymbol{\theta}_{\mathrm{a}}$ and $\boldsymbol{\theta}_{\mathrm{a}}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of ones.

To learn more about the weight-arrival process in the long run, let us denote by λ_{wt} the rate at which the weight accumulates. Denote the rate at which orders arrive by λ_{av}. Then we have

$$\begin{aligned} \lambda_{wt} &= \boldsymbol{\theta}_a \left(\sum\nolimits_{n=1}^{\infty} nD_n \right) \mathbf{e} \\ \lambda_{av} &= \boldsymbol{\theta}_a \left(\sum\nolimits_{n=1}^{\infty} D_n \right) \mathbf{e} \end{aligned} \quad . \tag{3.1}$$

The probabilistic interpretation of these two equations can be given as follows. Since $\theta_{a,i}$ = $P\{$ *the BMAP is in state* i *in steady state* $\}$, and ($D_n\mathbf{e}$ $)_i$ = $P\{$ *an order of weight* n *will arrive in the period* $\Big|$ *the BMAP is currently in state* i $\}$, therefore $\boldsymbol{\theta}_{\mathrm{a}}D_n\mathbf{e}$ = $P\{$ *an order of weight* n *arrives in this period* $\}$ in the steady state. Summing up these probabilities will give us $P\{$ *an order of any weight arrives per period in steady state* $\}$, which is equivalent to λ_{av}. Similarly, multiplying $\boldsymbol{\theta}_{\mathrm{a}}D_n\mathbf{e}$ by n and then summing them together will yield the expected weight accumulated in each period in steady state, which equals λ_{wt}.

Now we will present several examples of using *BMAP* to model the weight-arrival process. Some of these cases will lead to interesting model structures and results, so we will keep referring back to them throughout the rest of this thesis.

Example 3.1 Orders arrive according to a discrete *MAP* with a matrix representation (D_0 , D_1). The weight of each order has a general discrete distribution $\{ p_1 , p_2 , \ldots \}$ that is independent of the order arrival process, where p_n, for $n > 0$, is the probability for the order to have weight n. In this case, the weight-arrival process can be modeled as a *BMAP* with matrix representation (D_0 , p_nD_1 , n = 1, 2, …).

Example 3.2 Suppose there is only one state for the underlying Markov chain (m = 1), then D_n , n = 0, 1, 2, … are just the probabilities for an order of weight n to arrive in a period and $\sum_{n=0}^{\infty} D_n = 1$. If we let d_0 = D_0 and $d_1 = \sum_{n=1}^{\infty} D_n$, then the interarrival times between orders are geometrically distributed with parameter d_1, which is the probability that an order of any positive weight to arrive in each period. The distribution of order weights, p_n , can be obtained from a general discrete distribution independent of the arrival process, or computed as p_n = D_n / d_1 if the D_n are all known. Therefore, the weight-arrival process is actually a compound geometric distribution with representation (d_0 , p_nd_1 , n = 1, 2, …).

Example 3.3 Orders arrive according to a discrete *MAP* with a matrix representation (D_0 , D_1). Order weights have an independent discrete *PH*-distribution ($\boldsymbol{\beta}$, S). Thus, the distribution of order weights is given by $p_n = \boldsymbol{\beta} S^{n-1}(I - S)\mathbf{e}, \quad n = 1, 2, \cdots$. Again, the process under consideration can be modeled as a *BMAP* with matrix representation (D_0 , p_nD_1 , n = 1, 2, …).

Example 3.4 Orders arrive according to a *PH*-renewal process whose interarrival times have a common *PH*-distribution ($\boldsymbol{\alpha}$, T). Note that any *PH*-renewal process can also be summarized as the *MAP* representation (D_0 = T, D_1 = $\mathbf{T}^0\boldsymbol{\alpha}$), where

$\mathbf{T}^0 = \mathbf{e} - \boldsymbol{T}\mathbf{e}$. Similar to *Example 3.3*, if order weights are independent discrete *PH*-random variables denoted by ($\boldsymbol{\beta}, S$) then $p_n = \boldsymbol{\beta} S^{n-1}(I - S)\mathbf{e}, \quad n = 1, 2, \cdots$, and the process can be modeled as a *BMAP* with matrix representation (D_0 , $p_n D_1$, n = 1, 2, …).

3.3 Shipment Consolidation Policies

In Chapter One, we described the three commonly used shipment consolidation policies, which are the quantity policy, the time policy and the hybrid policy. We also mentioned that in practice, some shippers use a more general policy which reduces the target load as the waiting time elapses. Since our goal is to create a model capable of evaluating all these policies, we will try to express them in a common form.

Let us represent a shipment consolidation policy by a discrete function $f(j)$, where j is the time elapsed since the last dispatch. At the end of period j, if the accumulated weight reaches or exceeds $f(j)$, all outstanding orders are consolidated and a shipment is dispatched. Therefore, this function relates the two main decision variables of a shipment consolidation policy: the time elapsed since the previous dispatch and the current accumulated weight.

To enable the function $f(\cdot)$ to be a realistic representation of consolidation policies in practice, we make the following assumptions without loss of generality:

- $f(j) = q \geq 0$, for $j \geq j_q$, where j_q is a given positive integer.

- $j_q \geq 2$ and $f(j) > 1$ for $j = 1, 2, \ldots, j_q - 1$.
- $f(j)$ is non-increasing.

These three assumptions are intuitive and typically used. The first translates to the fact that if there has not been a dispatch for the past j_q units of time, the target load will be set to a constant q. This q could be equal to zero, which would then immediately trigger a dispatch. The second assumption, needed due to a technical reason which will be explained in *Theorem 4.1*, states that before the elapsed time reaches j_q, the target load must exceed one. The third assumption suggests that shippers should not raise the target load as the elapsed time increases, since even the previous target has not even been achieved.

The following examples demonstrate how different consolidation policies are represented by the function $f(\cdot)$.

Example 3.5 Quantity policy: Suppose a shipper sends out orders only when the accumulated weight reaches 150 *cwt* (target load Q = 150). This policy can be modeled as

$$f(j) = \begin{cases} Q = 150\,, & j = 1 \\ Q = 150\,, & j \geq j_q = 2 \end{cases}.$$

Example 3.6 Pseudo-time policy: Suppose a shipper dispatches a shipment once every seven days (maximum waiting time T = 7). For modeling reasons (see

Theorem 4.1), we will still set a target load Q as an upper bound for the total accumulated weight. Then if Q is very large, this policy approximates a time policy. If we set Q = 400 *cwt*, which is often the capacity of a truck, it can be expressed as

$$f(j) = \begin{cases} Q = 400\,, & 1 \le j < j_q = T = 7 \\ 0, & j \ge j_q = T = 7 \end{cases} .$$

Example 3.7 Hybrid policy: Suppose a shipper decides that order deliveries will be held back until either T days have passed since the last dispatch, or the total accumulated weight reaches the target load Q. If T is very large, this policy approximates a quantity policy; on the other hand, if Q is very large, similar to *Example 3.6*, this policy approximates a time policy. In a reasonable case where Q = 180 *cwt* and T = 5, the shipment consolidation policy can be expressed as

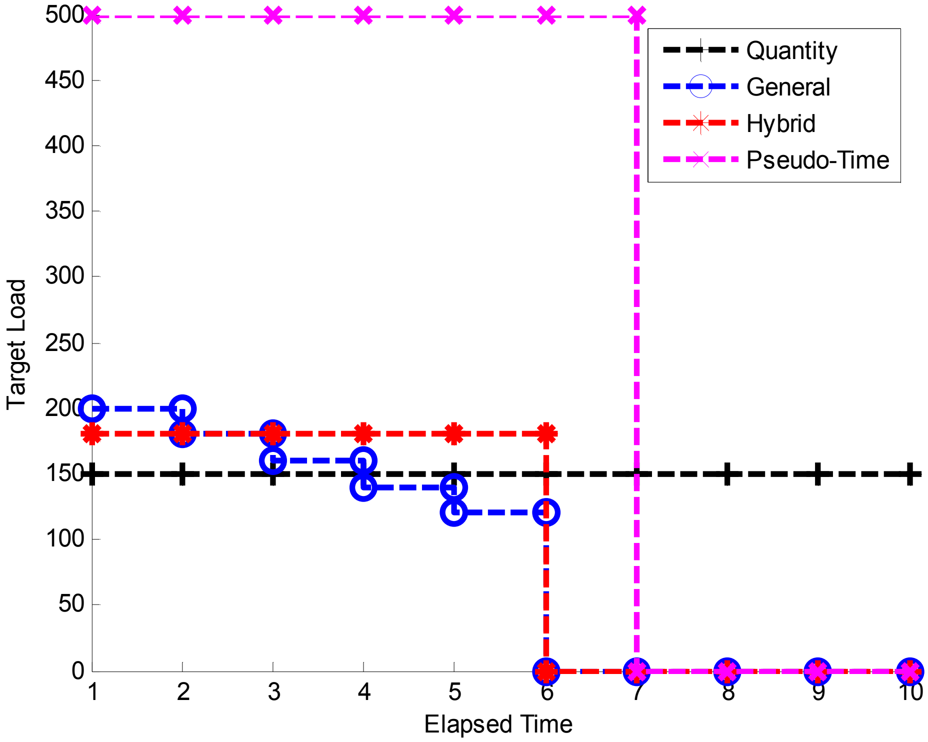
$$f(j) = \begin{cases} Q = 180\,, & 1 \le j < j_q = T = 5 \\ 0, & j \ge j_q = T = 5 \end{cases} .$$

Example 3.8 General policy: Suppose that during the first day after a previous dispatch, the target load is 200 *cwt*, and that target will be reduced by 20 *cwt* for each of the next four days. On the sixth day, if there still has not been a dispatch, all outstanding orders will be consolidated and shipped. This policy can be modeled by the following function

$$f(j)=\begin{cases}200-20(j-1), & 1\le j<j_q=6\\ 0, & j\ge j_q=6\end{cases}.$$

Note that the function for the general policy can be modified to resemble the other three policies and they can all be characterized as "step functions that move downward and to the right". *Figure 3.1* illustrates those functions defined in *Examples 3.5* through *3.8*.

Figure 3.1: Functions Representing Shipment Consolidation Policies



3.4 Performance Measures

The next important step in our model formulation is to identify the performance measures that can be used to evaluate the effectiveness of different consolidation policies. In an analytical model, upon selecting a shipment consolidation policy, we often use criteria that are the expected values of weight per load, excess per load, consolidation cycle length, delay per order, or number of orders per load to measure the performance of that policy.

The expected weight per load can help us determine the vehicle utilization rate for private carrier, and can be used to estimate the transportation cost by common carrier. In the case of a quantity policy, the mean weight of each load is approximately the target load; for a hybrid policy, a load weighing less than the target load can be dispatched if the allowed waiting time for consolidation has run out first. As for the general policy, the weight of each load depends on the elapsed time; but for a time policy, not much can be said about the expected weight per load.

As observed previously, when a dispatch is triggered by a target load, the actual load does not usually weigh exactly that much. In fact the actual load is equal to the effective target load (the value of $f(j)$ at the moment of dispatch) plus some excess. For example, suppose the accumulated weight at the end of period $t - 1$ was 100 *cwt* and in period t, an order of 60 *cwt* had arrived. Given that $f(t)$ = 150, a consolidated load weighing 160 *cwt* would be dispatched at the end of period *t*, and there would be an excess of 10 *cwt*. We will call this excess the “excess over target load”. (In some other shipment consolidation literature, for example, Bookbinder and Higginson (2002) and Çetinkaya and Bookbinder (2003), this excess is referred to as the “overshoot” of a policy.) Ideally, the excess over target load should be close to zero; however, that is often not the case, since the excess is sensitive to the weight-arrival process, especially if the order weight distribution has large variance.

There exists another type of excess for a load, which we will call the “excess over vehicle capacity”. This measures the amount by which a load exceeds the capacity of its transportation medium. It is more crucial for the private carrier since going over the capacity would require the dispatch of another vehicle which will incur a sizeable fixed charge. In the common carriage case, this becomes less important since there usually is no fixed dispatch cost. The two types of excesses we mentioned are recorded over the long run, so again we are interested in finding their expected values.

Let us consider one round of weight accumulation plus dispatch as a *cycle* in the shipment consolidation process. This cycle keeps repeating itself as long as the weight-arrival process and shipment consolidation policy remain the same. Breaking down the shipment consolidation process into many identical consolidation cycles, we would then be able to use renewal-reward theory to analyze it. In order to do so, we need to find the expected cycle length, which is equivalent to the average time between consecutive dispatches.

By its nature, shipment consolidation will cause delays to order deliveries. From the customer service perspective, we want to find out whether this will result in an unacceptable service level. Some consolidation policies like the *time policy* and *hybrid policy* have an inherent advantage in terms of preventing order delays over the others. However, those delays also depend on the weight-arrival process. Therefore, for each consolidation cycle in the long run, we would like to estimate the expected delay per order, which is equal to the average time between an order's arrival and its eventual dispatch.

Another interesting measure is the expected number of orders per load in the long run. This value is particularly useful for the shipper if there is extra handling or processing cost attributed to each order. For instance, in the furniture delivery example in Chapter One, moving furniture into the homes of customers requires time and labor, so the number of orders per load affects the delivery cost for that load. Combined with the weight of the load and the consolidation cycle length, this value can also give us some insights about the weight-arrival process.

So far all we have discussed are examples of non-financial performance measures, and they are independent of the type of carrier. Although each shipment consolidation policy can be employed by either a private or common carrier, the type of carrier has no effect on the weight-arrival process and how the policy is carried out. However, when it comes to financial performance measures, there is a big distinction between private carriage and common carriage.

One of the biggest motivations for shipment consolidation lies in its potential to reduce total logistics cost (transportation plus inventory holding cost); hence we choose cost as our main financial performance measure. Due to the different cost structures of the two types of carrier, we have two different cost functions. Because of the differences in expected cycle length and expected weight per load with respect to different policies, we cannot simply compare their total costs per load. Instead, we will use a standardized measure of expected long-run cost per unit time, denoted as *C(f)$_k$*, for each combination of carrier type *k* and policy *f*. Applying renewal-reward theory by treating each consolidation cycle as a renewal interval, we can obtain *C(f)$_k$* as

$$C(f)_k = \frac{E\left[\text{Transportation Cost per Cycle}\right] + E\left[\text{Inventory Holding Cost per Cycle}\right]}{E\left[\text{Cycle Length}\right]} \quad . \qquad (3.2)$$

Similar cost functions have been used by Çetinkaya and Bookbinder (2003) and Mütlü, Çetinkaya and Bookbinder (2010). We will analyze this cost function in greater detail in Chapter Six. A summary of all the performance measures is shown in *List 3.1*.

List 3.1: Summary of Performance Measures

1. Expected long-run cost per unit time
2. Expected weight per load
3. Excess weight beyond target load
4. Excess weight beyond vehicle capacity
5. Expected consolidation cycle length
6. Expected delay per order
7. Expected number of orders per load

CHAPTER FOUR

A DISCRETE TIME MARKOV CHAIN FOR THE MODEL

For the model defined in Chapter 3, we will now construct a discrete time Markov chain and its transition probability matrix for the shipment consolidation process. Those will allow us to find the steady state distribution of the process and calculate other long-run statistics.

4.1 The Markov Chain of Interest

We begin constructing the Markov chain by defining two system variables:

- Let $W(t)$ be the accumulated weight of all outstanding orders at the beginning of period t (orders arriving during period t are not included).
- Let $J(t)$ be the time elapsed since the last dispatch if that is less than or equal to j_q; otherwise, let $J(t) = j_q + 1$.

Then we can represent the status of the system at the beginning of period t by ($J(t), W(t), I_a(t)$), where $I_a(t)$ is the current state of the underlying Markov chain for the *BMAP* weight-arrival process. At the start of each new consolidation cycle, $J(t) = 1$ and $W(t) = 0$. This implies that there was a shipment dispatched at the end of the previous period. Also note that at the beginning of each period, the total weight includes only those orders accumulated in previous periods.

Recall that if $f(j_q)$ = 0, then all outstanding orders must be shipped when the elapsed time has reached j_q. In this case, the elapsed time cannot increment further but instead must be reset to one, so $J(t)$ can never reach j_q+1. However, for technical reasons (see *Theorem 4.1*), if $f(j_q)$ = 0, we can still have a system state as ($J(t) = j_q+1,\ W(t) = 0,\ I_a(t)$).

The stochastic process { ($J(t),\ W(t),\ I_a(t)$), t = 0, 1, 2, … } has state space:

$$\begin{cases} \{1\}\times\{0\}\times\{1, 2, ..., m\}, & \text{for} \quad J(t)=1; \\ \{j\}\times\{0, 1, 2, ..., \max\{\ 0, f(j-1)-1\}\}\times\{1, 2, ..., m\}, & \text{for} \quad 2\le J(t)=j\le j_q+1. \end{cases}$$

This can be understood by noting that to begin each consolidation cycle, the system state must be initialized as ($J(t)$ = 1, $W(t)$ = 0, $I_a(t)$); while in any other period j of that cycle, the accumulated weight must be non-negative and smaller than the target load of the previous period. Let us call $J(t)$ the “level” variable and ($W(t),\ I_a(t)$) the “phase” variable; we can then refer to the set of phases with $J(t)$ = j as “level j”.

Now we need to prove that { ($J(t)$, $W(t)$, $I_a(t)$), t = 0, 1, 2, … } is indeed a Markov chain. This is readily seen because all system variables satisfy the Markov property, which requires the future states of the system to be independent of the past. In other words, future states depend solely on current states and future events. First we see that { $I_a(t)$, t = 0, 1, 2, … } is a Markov chain, so it automatically satisfies this condition. Then we know that $W(t)$ depends only on the current accumulated weight and future order arrivals, and $J(t)$ depends only on $W(t)$ and $I_a(t)$. Therefore, they all satisfy the Markov property.

4.2 Transition Probability Matrix

Now we can present the transition probability matrix for this discrete time Markov chain.

Theorem 4.1 The process $\{ (J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots \}$ is a Markov chain with transition probability matrix

$$P_{TW} = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q+1 \end{array} \begin{pmatrix} A_{1,1} & A_{1,2} & & & & \\ A_{2,1} & 0 & A_{2,3} & & & \\ \vdots & & \ddots & \ddots & & \\ \vdots & & & \ddots & \ddots & \\ A_{j_q,1} & & & & 0 & A_{j_q,j_q+1} \\ A_{j_q+1,1} & & & & & A_{j_q+1,j_q+1} \end{pmatrix}, \tag{4.1}$$

where $A_{1,1} = \overline{D}_{f(1)}$, $A_{1,2} = \begin{pmatrix} D_0 & D_1 & D_2 & \cdots & \cdots & D_{f(1)-1} \end{pmatrix}$; for $2 \le j \le j_q - 1$,

$$A_{j,1}=\begin{array}{c}0\\1\\\vdots\\f(j)-2\\f(j)-1\\f(j)\\\vdots\\f(j-1)-1\end{array}\begin{pmatrix}\bar{D}_{f(j)}\\\bar{D}_{f(j)-1}\\\vdots\\\bar{D}_2\\\bar{D}_1\\\bar{D}_0\\\vdots\\\bar{D}_0\end{pmatrix},\quad A_{j,j+1}=\begin{array}{c|ccccc}
 & 0 & 1 & \cdots & \cdots\ \cdots & f(j)-1\\
\hline
0 & D_0 & D_1 & \cdots & \cdots & D_{f(j)-1}\\
1 & & D_0 & D_1 & \ddots & D_{f(j)-2}\\
\vdots & & & \ddots & \ddots & \vdots\\
\vdots & & & & \ddots & D_1\\
f(j)-1 & & & & & D_0\\
f(j) & 0 & 0 & \cdots & \cdots & 0\\
\vdots & \vdots & \vdots & \cdots & \cdots & \vdots\\
f(j-1)-1 & 0 & 0 & \cdots & \cdots & 0
\end{array};\ (4.2)$$

for $j = j_q$ and $j = j_q + 1$, if $f(j_q) > 0$, the matrices $A_{j_q,1}$, A_{j_q,j_q+1}, $A_{j_q+1,1}$, and A_{j_q+1,j_q+1} are given by equation (4.2); if $f(j_q) = 0$, then only $A_{j_q,1}$ is given by equation (4.2), while $A_{j_q,j_q+1} = 0$, $A_{j_q+1,1} = \bar{D}_0$, and $A_{j_q+1,j_q+1} = 0$. Note that $\bar{D}_n = \sum_{j=n}^{\infty} D_j$, $n = 0, 1, \cdots$.

Proof The transitions between levels $j = 1, 2, \ldots, j_q$ can be identified based on the following observations:

- The value of $J(t)$ always increases by 1, except for possible transitions to level 1 when a shipment is dispatched.
- The value of $W(t)$ is non-decreasing, except for possible transitions to level 1 when a shipment is dispatched.

- If $J(t) = j$, the initial weight in period t is between 0 and $f(j-1) - 1$, and the ending weight is between 0 and $f(j) - 1$.

The transitions associated with level $j_q + 1$ are based on the fact that for policy $f(j)$, the dispatch quantity is the same for $j \geq jq$. If $f(j_q) = 0$, a shipment must be dispatched once the time elapsed since the last shipment reaches j_q. Thus, there is no transition from level j_q to level $j_q + 1$. The transition probabilities are obtained accordingly. □

There are a few comments we can make about the structure of P_{TW}. First, transitions from level j to level $j + 1$ are governed by the probability matrix $A_{j,j+1}$, while those from level j back to level 1 are governed by $A_{j,1}$. We again divide each level j into sub-levels 0 through $f(j) - 1$, (note that $f(0) = 1$). These sub-levels represent different accumulated weights at the beginning of period j.

Second, the matrix component D_k in $A_{j,j+1}$, where $0 \leq k \leq f(j)$-1, can be interpreted as the probability matrix for the accumulated weight to increase by k units *without* triggering a dispatch in period j. In the block matrix $A_{j,1}$, the component $\overline{D}_n = \sum_{j=n}^{\infty} D_j$, where $n = 0, 1, \ldots, f(j)$, is the probability matrix that the weight of the next order is greater than or equal to n units, and leading to a dispatch.

Third, in Chapter Three, we made assumptions about the shipment consolidation policy function stating that $j_q \geq 2$ and $f(j) > 1$ for $j = 1, 2, \ldots, j_q - 1$. This was necessary because we needed to have at least two levels in the matrix P_{TW} to explicitly represent both level $j = 1$ (whereby the cycle is reset), and level $j = j_q$ (at which the elapsed time reaches its threshold).

Fourth, the formula for P_{TW} given in *Theorem 4.1* does not allow us to model the time policy directly. This is because that policy has an infinitely large target load before the elapsed time threshold is reached, resulting in infinitely many sub-levels for levels 1 through j_q – 1. As a result, we had to use a "pseudo-time" policy to approximate the time policy, where a sufficiently large upper bound Q for the accumulated weight will restrict the size of matrix P_{TW}.

Fifth, that formula for P_{TW} of *Theorem 4.1* is sufficient to model the other classes of policies described in Section 3.3. However, for quantity policies, we can simplify the structures of P_{TW} to improve computational efficiency. We will discuss this in more detail in Chapter Five.

4.3 Steady State Distribution

Denote by row vector $\boldsymbol{\theta}_{TW}$ the steady state distribution of the Markov chain $\{ (J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots \}$. Then according to the properties of discrete time Markov chains, $\boldsymbol{\theta}_{TW}$ is the unique solution of the linear system $\boldsymbol{\theta}_{TW}P_{TW} = \boldsymbol{\theta}_{TW}$ and $\boldsymbol{\theta}_{TW}\mathbf{e} = 1$. For the transition probability matrix P_{TW} defined by *Theorem 4.1*, the following algorithm can efficiently compute its steady state distribution $\boldsymbol{\theta}_{TW}$, if it exists.

Algorithm I

I.1) Compute the matrices

$$\begin{aligned} R_1 &= I; \\ R_j &= R_{j-1} A_{j-1,j}, \quad 2 \le j \le j_q; \\ R_{j_q+1} &= R_{j_q} A_{j_q,j_q+1} (I - A_{j_q+1,j_q+1})^{-1}; \\ P_1 &= \sum_{j=1}^{j_q+1} R_j A_{j,1}. \end{aligned} \tag{4.3}$$

I.2) Solve the linear system $\boldsymbol{\theta}_1 P_1 = \boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_1 \left(\sum_{j=1}^{j_q+1} R_j \right) \mathbf{e} = 1$ for $\boldsymbol{\theta}_1$.

I.3) Calculate $\boldsymbol{\theta}_j = \boldsymbol{\theta}_1 R_j$, for $j = 1, 2, \ldots, j_q + 1$.

Note that, due to the special structure within the matrix A_{j_q+1,j_q+1}, the inversion of $I - A_{j_q+1,j_q+1}$ can be done efficiently even if $m \cdot f(j_q)$ is large. The validity of *Algorithm I* is guaranteed by the finiteness of the number of states and the fact that states in level 1 can be reached from all other states. Utilization of the *R* matrix to obtain the steady state distribution is a classic approach in matrix-analytic methods; it follows the same idea as that for *QBD* in Section 2.4.

Having found a valid algorithm to determine $\boldsymbol{\theta}_{TW}$, we now need to show that $\boldsymbol{\theta}_{TW}$ does exist and is indeed the steady state distribution of the Markov chain.

Theorem 4.2 The steady state distribution of the Markov chain $\{ (J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots \}$ exists and is given by $\boldsymbol{\theta}_{TW}$.

Proof First, it is easy to verify that $\boldsymbol{\theta}_{TW}$ is a solution to the linear system $\boldsymbol{\theta}_{TW}P_{TW} = \boldsymbol{\theta}_{TW}$ and $\boldsymbol{\theta}_{TW}\mathbf{e} = 1$. We check the first equation by substituting the formula from *Algorithm I* into the left-hand side then expand it to derive the right-hand side. We do not need to check the second equation because it is used in the *Algorithm I* to obtain $\boldsymbol{\theta}_{TW}$.

Since the underlying Markov chain $\{ I_a(t), t = 0, 1, 2, \ldots \}$ is irreducible and the process can reach any level j and weight w within the state space, the Markov chain $\{ (J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots \}$ has a single closed set that is recurrent. Consequently, limiting probabilities exist and are unique. This then implies that the solution to the linear system is unique. Thus, $\boldsymbol{\theta}_{TW}$ is the steady state distribution. This completes the proof. □

Let us decompose $\boldsymbol{\theta}_{TW}$ as follows: $\boldsymbol{\theta}_{TW} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{j_q}, \boldsymbol{\theta}_{j_q+1})$, and then $\boldsymbol{\theta}_j = (\boldsymbol{\theta}_{j,0}, \boldsymbol{\theta}_{j,1}, \ldots, \boldsymbol{\theta}_{j,\, max\{0, f(j-1)-2\}}, \boldsymbol{\theta}_{j,\, max\{0, f(j-1)-1\}})$, for $j = 2, 3, \ldots, j_q + 1$. Note that all vectors { $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_{j,w}$, w=0,1,…, $max\{0, f(j-1)-1\}$, for j = 2, 3, …, j_q+1 } are row vectors of size m. As an immediate consequence of *Theorem 4.2*, we see that

$$\boldsymbol{\theta}_a = \sum_{j=1}^{j_q+1} \sum_{w=0}^{\max\{0, f(j-1)-1\}} \boldsymbol{\theta}_{j,w} \ ,$$

which can be used for checking the accuracy in computation.

4.4 Long-Run Accumulated Weight and Excess Weight Distributions

A number of steady state performance measures can be found directly or indirectly from $\boldsymbol{\theta}_{TW}$. First let us denote by O_w the amount of weight over a threshold function $g(j)$ at a dispatch epoch. If we let $g(\cdot) = f(\cdot)$, then O_w represents the long-run excess over target load defined in Section 3.4; if we let $g(j) = Q_o$ for all j, where Q_o is the truck capacity, then O_w represents the long-run excess over vehicle capacity.

Next we denote by W the accumulated weight at the beginning of an arbitrary period when the system is in steady state. Since W is also equivalent to the inventory level for an arbitrary period in the steady state, if we can find the distribution of W, we can then directly calculate the expected inventory holding cost per unit time in steady state.

Similarly, let W_c be the accumulated weight of an arbitrary shipment (i.e., the total weight accumulated during an arbitrary cycle). This can be used to find the mean weight per shipment and the mean weight shipped per unit time. In addition, we can denote by P_S the probability that a shipment takes place in an arbitrary period in the long run. The distributions of O_w, W, W_c and the probability P_S can be obtained from $\boldsymbol{\theta}_{TW}$ in a rather straightforward manner.

Corollary 4.3 For the shipment consolidation model defined in Chapter 3, we have

(i) $P\{\, W = w \,\} = \left(\displaystyle\sum_{j=1:\ w \le f(j-1)-1}^{j_q+1} \boldsymbol{\theta}_{j,w} \right) \mathrm{e}$, $w = 0, 1, 2, \ldots, f(1) - 1$.

(ii) $P_S = \boldsymbol{\theta}_1 \mathbf{e}$.

(iii) $P\{\, W_c = i \,\} = \dfrac{1}{P_S} \displaystyle\sum_{j=1}^{j_q+1} \left(\sum_{w=0:\, i \ge f(j) \text{ and } i \ge w}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} D_{i-w} \mathbf{e} \right)$, $i = 0, 1, 2, \ldots$.

(iv) $P\{\, O_w = i \,\} = \dfrac{1}{P_S} \displaystyle\sum_{j=1}^{j_q+1} \left(\sum_{w=0:\, i+g(j) \ge f(j)}^{\min\{\, i+g(j),\ f(j-1)-1\}} \boldsymbol{\theta}_{j,w} D_{i+g(j)-w} \mathbf{e} \right)$, $i = 0, 1, 2, \ldots$.

(v) $P_o = P\{O_w > 0\} = \dfrac{1}{P_S} \displaystyle\sum_{j=1}^{j_q+1} \left(\sum_{w=0}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} \overline{D}_{\max\{\ 0,\max\{\ 1+g(j),\ f(j)\}-w\}} \mathbf{e} \right)$.

The means $E[W]$, $E[W_c]$, and $E[O_w]$ can be obtained from the distributions accordingly.

Proof Since *W* corresponds to the system state variable *W*(*t*) in the Markov chain, part (i) is obtained by summing the steady state distribution over the two other system state variables *J*(*t*) and $I_a(t)$ for every value of *W*(*t*). For P_S, we have

$$\begin{aligned} P_S &= \sum_{i=0}^{\infty} \left(\sum_{j=1}^{j_q+1} \left(\sum_{w=0:\, i \geq f(j) \text{ and } i \geq w}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} D_{i-w} \mathbf{e} \right) \right) \\ &= \sum_{j=1}^{j_q+1} \left(\sum_{w=0}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} \overline{D}_{\max\{\, f(j)-w, 0\}} \mathbf{e} \right) \qquad (4.4) \\ &= \sum_{j=1}^{j_q+1} \boldsymbol{\theta}_j A_{j,1} \mathbf{e} = \boldsymbol{\theta}_1 P_1 \mathbf{e} = \boldsymbol{\theta}_1 \mathbf{e}. \end{aligned}$$

This equation can be interpreted as follows. Suppose the system is currently in state (*J*(*t*) = *j, W*(*t*) = *w,* $I_a(t)$), which has the steady state probabilities $\boldsymbol{\theta}_{j,w}$. The arrival of an order with weight greater than *max*{ *f*(*j*) – *w,* 0 } will lead to a dispatch in this period, which occurs with probability $\overline{D}_{\max\{\, f(j)-w, 0\}}$. By definition of conditional probability, we can multiply the two sets of probabilities and sum them over all system states to obtain P_S, which incidentally equals the probability of the system returning to level 1.

Parts (iii) through (v) can be found directly from equation (4.4). Conditioning on the event that a shipment takes place, part (iii) is the steady state probability for the shipment to have some specific weight *i*. Parts (iv) and (v) measure the excess weight over the threshold function *g*(*j*). If *g*(*j*) is set to Q_o , the excess is measured against the vehicle capacity, but if *g* (*j*) is equal to the function of the dispatch policy, $f(j)$, excess over target load is measured. Part (iv) is the probability for the shipment to have an excess of weight i; while part (v) is the probability for any size of excess to occur. Note that part (v) is equal to the summation of part (iv) over all values of i. □

Remark 4.1: By definition, immediately after a shipment takes place, the clock for a new consolidation cycle is set to 1 (i.e., $J(t) = 1$). Thus, the probability P_S that a shipment takes place is equal to the probability that the elapsed time since the last shipment is one (i.e., $\boldsymbol{\theta}_1\mathbf{e}$). This gives an intuitive interpretation to part (ii) of *Corollary 4.3*.

4.5 Expected Cycle Length and Order Delays

Next, we construct an *absorbing Markov chain* to investigate the length of a consolidation cycle L_c and the waiting time L_w of an arbitrary order. L_c is of particular interest because it is needed to compute the expected long-run cost per unit time. L_w can be used to determine the customer service level (i.e. average order delay) for each consolidation policy. Define

$$T_c = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q+1 \end{array} \begin{pmatrix} 0 & A_{1,2} & & & & \\ & 0 & A_{2,3} & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 0 & A_{j_q, j_q+1} \\ & & & & & A_{j_q+1, j_q+1} \end{pmatrix}, \mathbf{T}_c^0 = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q+1 \end{array} \begin{pmatrix} A_{1,1} \\ A_{2,1} \\ \vdots \\ \vdots \\ A_{j_q,1} \\ A_{j_q+1,1} \end{pmatrix} \qquad (4.5)$$

and $P_c = (I, 0, \cdots, 0)(I - T_c)^{-1} \mathbf{T}_c^0$.

It is readily seen that $P_{TW} = T_c + (\mathbf{T}_c^0, \mathbf{0}, \ldots, \mathbf{0})$. Let us create an artificial absorbing level 1*'* and treat levels 1 through $j_q + 1$ as transient levels. We can rewrite P_{TW} as

$$P_{TW} = \begin{pmatrix} T_c & \mathbf{T}_c^0 \\ \mathbf{0} & 1 \end{pmatrix}.$$

A consolidation cycle always starts from level 1, and completion of a cycle is equivalent to absorption into level 1*'*. Therefore, the probability for any consolidation cycle to start from level 1 and get absorbed into level 1*'* upon completion is given by $P_c = (I, 0, \cdots, 0)(I - T_c)^{-1} \mathbf{T}_c^0$.

When the system returns to level 1*,* the state of the underlying Markov chain might have changed. The matrix P_c is a stochastic matrix that governs the transitions of that underlying chain $\{ I_a (t), t = 0, 1, 2, \ldots \}$ at the start of any consolidation cycle. In other words, it is the embedded Markov chain for the state of the underlying Markov chain at the beginnings of consolidation cycles. Each transition in P_c coincides with the start of a new consolidation cycle and determines the state of the underlying chain at that moment.

Denote by $\boldsymbol{\eta}_c$ the steady state distribution associated with P_c. Then $\boldsymbol{\eta}_c$ is the unique solution of the linear system $\boldsymbol{\eta}_c \mathrm{P_c} = \boldsymbol{\eta}_c$ and $\boldsymbol{\eta}_c\mathbf{e} = 1$.

Theorem 4.4 The distribution $\boldsymbol{\eta}_c$ is given by $\boldsymbol{\eta}_c = \boldsymbol{\theta}_1 / (\boldsymbol{\theta}_1\mathbf{e})$. In steady state, the distribution of the length of a consolidation cycle L_c has a discrete *PH*-distribution with matrix representation (($\boldsymbol{\eta}_c$, 0, …, 0), T_c). The distribution of $L_w + 1$ also has a discrete *PH*-distribution with matrix representation ($\boldsymbol{\theta}_{TW}$, T_c). In addition, we have

$$\begin{aligned} E[L_c] &= 1/(\boldsymbol{\theta}_1\mathrm{e}); \\ E[L_w] &= \sum_{j=1}^{j_q} j\boldsymbol{\theta}_j\mathrm{e} + j_q\boldsymbol{\theta}_{j_q+1}\mathrm{e} + \boldsymbol{\theta}_{j_q+1}(I - A_{j_q+1,j_q+1})^{-1}\mathrm{e} - 1. \end{aligned} \tag{4.6}$$

Proof Due to the special structure within the matrix T_c, the first row of the inverse of $I - T_c$ can be found explicitly as (R_1, R_2, …, R_{j_q+1}), given by equation (4.3). Immediately, we obtain $P_c = \sum_{j=1}^{j_q+1} R_j A_{j,1}$. The vector $\boldsymbol{\eta}_c$ can be interpreted as the steady state distribution of the underlying Markov chain of the arrival process at dispatch epochs (or at the beginning of a consolidation cycle), which is the unique solution to the linear system $\boldsymbol{\eta}_c \left(\sum_{j=1}^{j_q+1} R_j A_{j,1} \right) = \boldsymbol{\eta}_c$ and $\boldsymbol{\eta}_c \mathbf{e} = 1$.

Existence of the steady state distribution is again guaranteed by the fact that P_c has a single closed set of states. From *Algorithm I*, it is easy to see that $\boldsymbol{\eta}_c = \boldsymbol{\theta}_1 / (\boldsymbol{\theta}_1 \mathbf{e})$, because $P_c = P_1$ and $\boldsymbol{\theta}_1 P_1 = \boldsymbol{\theta}_1$. The initial probability distribution of the Markov chain $\{ (J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots \}$ at the beginning of a consolidation cycle is given by $(\boldsymbol{\eta}_c, 0, \ldots, 0)$, since $J(t) = 1$ at the beginning of any cycle. Thus, the length of a consolidation cycle has a *PH*-distribution.

The mean cycle length is obtained by straightforward simplification of the expression $(\boldsymbol{\eta}_c, 0, \ldots, 0)(I - T_c)^{-1} \mathbf{e}$, which leads to

$$E[L_c] = \boldsymbol{\eta}_c \left(\sum_{j=1}^{j_q+1} R_j \right) \mathbf{e} = (\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 + \ldots + \boldsymbol{\theta}_{j_q+1}) \mathbf{e} / (\boldsymbol{\theta}_1 \mathrm{e}) = 1 / (\boldsymbol{\theta}_1 \mathbf{e}).$$

For L_w, at the beginning of a period in which an order arrives, the system can be in any state according to the steady state distribution $\boldsymbol{\theta}_{TW}$. The sum of the delay of the order plus its period of arrival is equal to the absorption time for the absorbing Markov chain whose initial probabilities are $\boldsymbol{\theta}_{TW}$. Note that the delay is zero if an order arrives and a shipment is dispatched in the same period. By some routine calculations, the mean of the *PH*-distribution is $\boldsymbol{\theta}_{TW} (I - T_c)^{-1} \mathbf{e}$, and the results can be obtained. □

Remark 4.2 i) $E[L_c] = 1/(\boldsymbol{\theta}_1\mathbf{e})$ can be explained intuitively. From *Corollary 4.3*, we know that the probability that the system is at the beginning of a consolidation cycle is equal to P_S = $\boldsymbol{\theta}_1\mathbf{e}$. Thus, if we consider dispatch as an event that may occur in each period with probability P_S, then the time between consecutive dispatches has a geometric distribution and its mean is given by 1 / ($\boldsymbol{\theta}_1\mathbf{e}$).

ii) With the expected weight per cycle $E[W_c]$ (*Corollary 4.3*) and the expected cycle length $E[L_c]$ (*Theorem 4.4*), the average weight shipped per unit time can be obtained as $E[W_c]$ / $E[L_c]$. As indicated in Section 3.2, the weight arrival rate is given by λ_{wt}. Then we must have λ_{wt} = $E[W_c]$ / $E[L_c]$. Such a relationship is useful for checking the accuracy of numerical computations.

4.6 Expected Number of Orders per Load

Finally in this section, we introduce a *terminating Markovian arrival process* (He and Neuts (1998)) to study the number of orders N_c received in a consolidation cycle. First, we decompose matrices T_c and $\mathbf{T}^0{}_c$ defined in equation (4.5) as

$$T_{c,0} = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q+1 \end{array} \left(\begin{array}{cccccc} 0 & A_{1,2,0} & & & & \\ & 0 & A_{2,3,0} & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 0 & A_{j_q,j_q+1,0} \\ & & & & & A_{j_q+1,j_q+1,0} \end{array}\right), \qquad \mathbf{T}_{c,0}^{0} = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q+1 \end{array} \left(\begin{array}{c} A_{1,1,0} \\ A_{2,1,0} \\ \vdots \\ \vdots \\ A_{j_q,1,0} \\ A_{j_q+1,1,0} \end{array}\right); \tag{4.7}$$

$$T_{c,1} = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q+1 \end{array} \left(\begin{array}{cccccc} 0 & A_{1,2,1} & & & & \\ & 0 & A_{2,3,1} & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 0 & A_{j_q,j_q+1,1} \\ & & & & & A_{j_q+1,j_q+1,1} \end{array}\right), \qquad \mathbf{T}_{c,1}^{0} = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q+1 \end{array} \left(\begin{array}{c} A_{1,1,1} \\ A_{2,1,1} \\ \vdots \\ \vdots \\ A_{j_q,1,1} \\ A_{j_q+1,1,1} \end{array}\right),$$

where $A_{j,j+1,0}$ and $A_{j,1,0}$ are obtained respectively from $A_{j,j+1}$ and $A_{j,1}$ by keeping only block D_0, while $A_{j,j+1,1}$ and $A_{j,1,1}$ are obtained respectively from $A_{j,j+1}$ and $A_{j,1}$ by removing the block D_0.

By definition, we must have $A_{j,j+1} = A_{j,j+1,0} + A_{j,j+1,1}$, $A_{j,1} = A_{j,1,0} + A_{j,1,1}$, $T_c = T_{c,0} + T_{c,1}$, and $\mathbf{T}_c^0 = \mathbf{T}_{c,0}^0 + \mathbf{T}_{c,1}^0$. We consider a terminating Markovian arrival process defined by ($T_{c,0}$, $T_{c,1}$, $\mathbf{T}^0{}_{c,0}$, $\mathbf{T}^0{}_{c,1}$), where $T_{c,1}$ and $\mathbf{T}^0{}_{c,1}$ correspond to transitions with order arrivals, and $T_{c,0}$ and $\mathbf{T}^0{}_{c,0}$ correspond to transitions without order arrivals. This Markovian arrival process is called a *terminating process* since we count only the number of order arrivals before or at the time the process enters level one.

Theorem 4.5 Given an initial probability distribution ($\boldsymbol{\theta}_1$ / ($\boldsymbol{\theta}_1\mathbf{e}$), 0, …, 0), the number of orders that occur in a consolidation cycle N_w equals the total number of arrivals in the terminating Markovian arrival process ($T_{c,0}$, $T_{c,1}$, $\mathbf{T}^0{}_{c,0}$, $\mathbf{T}^0{}_{c,1}$). Consequently, in steady state, we have

$$P\{N_c = n\} = \begin{cases} \left(\dfrac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1 \mathrm{e}}, 0, \cdots, 0\right)\left(I - T_{c,0}\right)^{-1}\mathbf{T}_{c,0}^0 \mathrm{e}, & n = 0; \\ \left(\dfrac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1 \mathrm{e}}, 0, \cdots, 0\right)\left(\left(I - T_{c,0}\right)^{-1} T_{c,1}\right)^n \left(I - T_{c,0}\right)^{-1}\mathbf{T}_{c,0}^0 \mathrm{e} & \\ + \left(\dfrac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1 \mathrm{e}}, 0, \cdots, 0\right)\left(\left(I - T_{c,0}\right)^{-1} T_{c,1}\right)^{n-1} \left(I - T_{c,0}\right)^{-1}\mathbf{T}_{c,1}^0 \mathrm{e}, & n \geq 1. \end{cases} \qquad (4.8)$$

The mean number of orders per cycle is given by

$$E[N_c] = \frac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1\mathbf{e}}\left(\sum\nolimits_{j=1}^{j_q} R_j A_{j,j+1,1} + R_{j_q+1} A_{j_q+1,j_q+1,1}\right)\mathbf{e} + \frac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1\mathbf{e}}\left(\sum\nolimits_{j=1}^{j_q+1} R_j A_{j,1,1}\right)\mathbf{e}. \qquad (4.9)$$

Proof The terminating process ($T_{c,0}$, $T_{c,1}$, $\mathbf{T}^0{}_{c,0}$, $\mathbf{T}^0{}_{c,1}$) is a special discrete version of the terminating Markovian arrival process defined in He and Neuts (1998) (See also Latouche et al. (2003)). The distribution of N_c is obtained routinely. First, by using a renewal argument, the moment generating function of N_c can be expressed as

$$E[z^{N_c}] = \left(\frac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1\mathrm{e}}, 0, ..., 0\right)\left(I - T_{c,0} - zT_{c,1}\right)^{-1}(\mathbf{T}^0_{c,0} + z\mathbf{T}^0_{c,1})\mathrm{e}, \quad 0 < z < 1. \qquad (4.10)$$

It is straightforward to obtain equation (4.8) from equation (4.10). The latter equation leads to equation (4.9) by noticing (again) that the first row of the matrix $\left(I - T_{c,0} - T_{c,1}\right)^{-1} = (I - T_c)^{-1}$ is (R_1, R_2, …, R_{j_q+1}). □

Remark 4.3 i) Equation (4.9) can be interpreted intuitively. The sum of the numerators in that equation is the probability that an order arrives in an arbitrary time period, which is also the expected number of order arrivals in that period. Multiplying that sum by the mean cycle length yields the total number of order arrivals in an arbitrary cycle.

ii) By *Theorems 4.4* and *4.5*, the number of order arrivals per unit time is given by $E[N_c]$ / $E[L_c]$. Then we must have $\lambda_{av} = E[N_c] / E[L_c]$, again useful for checking computational accuracy.

CHAPTER FIVE

MODEL MODIFICATIONS FOR SPECIAL CASES OF INTEREST

The Markov chain, algorithm and theorems in Chapter Four are designed for general problem instances. They are applicable in most cases but may be inefficient for some. In this chapter, we will identify particular special cases and demonstrate how to customize the model to boost its efficiency and accuracy.

5.1 A Simplified Algorithm for Quantity Policy Models

For a model with a quantity policy, $f(j) = Q$ for all j. Let us ignore the assumption we made earlier that $j_q \geq 2$; instead, we set $j_q = 1$. Then the Markov chain $\{ (J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots \}$ can be reduced to having two levels: $J(t) = 1$ and 2, and P_{TW} is reduced to

$$P_{TW} = \begin{array}{c} (1,0) \\ (2,0) \\ (2,1) \\ \vdots \\ (2, Q-2) \\ (2, Q-1) \end{array} \begin{pmatrix} \overline{D}_Q & D_0 & D_1 & \cdots & \cdots & D_{Q-1} \\ \overline{D}_Q & D_0 & D_1 & \cdots & \cdots & D_{Q-1} \\ \overline{D}_{Q-1} & & D_0 & D_1 & \ddots & D_{Q-2} \\ \vdots & & & \ddots & \ddots & \vdots \\ \overline{D}_2 & & & & \ddots & D_1 \\ \overline{D}_1 & & & & & D_0 \end{pmatrix} \tag{5.1}$$

Based on this simplified transition probability matrix P_{TW}, a more efficient algorithm can be used to determine its steady state distribution $\boldsymbol{\theta}_{TW} = (\boldsymbol{\theta}_{1,0}, \boldsymbol{\theta}_{2,0}, \boldsymbol{\theta}_{2,1}, \ldots, \boldsymbol{\theta}_{2,Q-1})$.

Algorithm I(Q)

I(Q).1) Compute the matrices

$$\begin{aligned} &H_0 = D_0(I - D_0)^{-1}; \\ &H_w = \left(D_w + \sum_{k=0}^{w-1} H_k D_{w-k} \right)(I - D_0)^{-1}, \quad 1 \le w \le Q-1; \\ &P_1 = \overline{D}_Q + \sum_{w=0}^{Q-1} H_w \overline{D}_{Q-w}. \end{aligned} \tag{5.2}$$

I(Q).2) Solve linear system $\boldsymbol{\theta}_{1,0}P_1 = \boldsymbol{\theta}_{1,0}$ and $\boldsymbol{\theta}_{1,0}\left(I + \sum_{n=0}^{Q-1} H_w\right)\mathbf{e} = 1$ for $\boldsymbol{\theta}_{1,0}$.

I(Q).3) Obtain $\boldsymbol{\theta}_{2,w} = \boldsymbol{\theta}_{1,0}H_w$, for $w = 0, 2, \ldots, Q$-1.

When we use *Algorithm I* for a quantity policy, the calculation of matrices of size $Q{\cdot}m$ is required. In comparison, *Algorithm I(Q)* deals only with matrices of size m, which is independent of the consolidation threshold Q. This implies that, if the algorithms are used to search for the optimal consolidation weight Q^* (which minimizes the expected long-run cost per unit time $C(f)_k$ for carrier type k), *Algorithm I(Q)* has computational advantages.

Corollary 5.1 For a quantity policy system, we have

(i) $P\{\ W = 0\ \} = \boldsymbol{\theta}_{1,0}\,(\ I{+}H_0\ )\ \mathbf{e} = (\ \boldsymbol{\theta}_{1,0}{+}\boldsymbol{\theta}_{2,0}\ )\ \mathbf{e}$;

$P\{\ W = w\ \} = \boldsymbol{\theta}_{1,0}\ H_w\ \mathbf{e} = \boldsymbol{\theta}_{2,\mathrm{w}}\ \mathbf{e}$, $w = 1, 2, \ldots, Q - 1$; and $E[W] = \boldsymbol{\theta}_{1,0}\sum_{w=1}^{Q-1} wH_w\mathbf{e}$.

(ii) $P_s = \boldsymbol{\theta}_{1,0}\left(\overline{D_Q} + \sum_{w=0}^{Q-1} H_w\overline{D_{Q-w}}\right)\mathbf{e} = \boldsymbol{\theta}_{1,0}\mathbf{e}$

(iii) $P\{\ W_c = i\ \} = \boldsymbol{\theta}_{1,0}\left(D_i + \sum_{w=0}^{Q-1} H_w D_{i-w}\right)\mathbf{e}\ /\ P_s$, $i = Q, Q + 1, Q + 2, \ldots$;

and $E[W_c] = \sum_{i=Q}^{\infty} i\ \boldsymbol{\theta}_{1,0}\left(D_i + \sum_{w=0}^{Q-1} H_w D_{i-w}\right)\mathbf{e}\ /\ P_s$.

(iv) $P\{\ O_w = i\ \} = \boldsymbol{\theta}_{1,0}\left(D_{i+Q} + \sum_{w=0}^{Q-1} H_w D_{i+Q_o-w}\right)\mathbf{e}\ /\ P_s$, $i = 0, 1, 2, \ldots$;

and $E[O_w] = \sum_{i=1}^{\infty} i\ \boldsymbol{\theta}_{1,0}\left(D_{i+Q_o} + \sum_{w=0}^{Q-1} H_w D_{i+Q_o-w}\right)\mathbf{e}\ /\ P_s$.

(v) $P\{O_w > 0\} = \boldsymbol{\theta}_{1,0}\left(\overline{D_{1+Q_o}} + \sum_{w=0}^{Q-1} H_w\overline{D_{1+Q_o-w}}\right)\mathbf{e}\ /\ P_s$.

(vi) $E[L_c] = 1\,/\,\boldsymbol{\theta}_{1,0}\mathbf{e}$

Remark 5.1 The proof of *Corollary 5.1* is similar to that of *Corollary 4.3*, hence details are omitted here. It is easy to see that the excess over vehicle capacity is equal to $O_w = \max\{\ 0, W_c - Q_o\}$ for the quantity policy. If we replace Q_o with Q in (vi), then we get the excess over target load. We then have $E[W_c] = Q + E[O_w]$ if $Q_o = Q$, a relationship that can be used for an accuracy check in the computation.

5.2 A Revised Model for the Quantity Policy and PH-Weight Case

In this case, let us consider a quantity policy model $f(j) = Q$ for all $j \geq 1$. The weight-arrival process is given by (D_0, $\boldsymbol{\beta} S^{n-1}(I–S)\mathbf{e}D_1$, n = 1, 2, …), like that defined in *Examples 3.3* and *3.4*. We will use an alternative approach to analyze this case. By taking advantage of the partial memoryless property of the *PH*-distributions, a new discrete time Markov chain can be introduced for the weight process.

The idea is: After every order arrival, stop the clock of the order arrival process and start a fictitious clock for the underlying Markov chain of the *PH*-weight distribution. The fictitious clock is stopped, and the clock of the order arrival process resumes, when the underlying Markov chain of the *PH*-distribution reaches its absorbing state.

More specifically, let { $I_w(t)$, t = 0, 1, 2, … } be the phase of the underlying Markov chain for the *PH*-distribution ($\boldsymbol{\beta}$, S) before absorption. Then a new Markov chain can be constructed as follows.

- If an order arrives, the underlying Markov chain { $I_w(t)$, $t = 0, 1, 2, \ldots$ } is turned on immediately, initialized by **β**, and the underlying Markov chain { $I_a(t)$, $t = 0, 1, 2, \ldots$ } is frozen.
- If the underlying Markov chain { $I_w(t)$, $t = 0, 1, 2, \ldots$ } enters its absorption state, it is terminated, and the underlying Markov chain { $I_a(t)$, $t = 0, 1, 2, \ldots$ } is unfrozen.

Define

$\hat{I}_a(t)$: $\hat{I}_a(t) = I_a(t)$, if the clock of the order arrival process is on; otherwise, $\hat{I}_a(t)$ is the last value of $I_a(t)$ before $I_a(t)$ is frozen.

$\hat{I}_w(t)$: $\hat{I}_w(t) = I_w(t)$, if the clock of the *PH*-weight distribution is on; but $\hat{I}_w(t) = 0$, if the clock of the order arrival process is on.

$\hat{W}(t)$: $\hat{W}(t) = W(t)$, if the clock of the order arrival process is on; otherwise, if the clock of the *PH*-weight distribution is on, $\hat{W}(t)$ increases by one per unit time if $\hat{W}(t-1) < Q - 1$, and becomes 0 if $\hat{W}(t-1) = Q - 1$.

Note that $W(t)$ takes values { 0, 1, 2, …, Q – 1 }. Then the process $\{(\hat{W}(t), \hat{I}_a(t), \hat{I}_w(t)),\ t = 0, 1, 2, \ldots\}$ is a Markov chain with transition probability matrix $P_{TW} = D$ for $Q = 1$, and, for $Q \geq 2$,

$$P_{TW}=\begin{array}{r|cccccc|}
(0,i_a) & D_0 & \begin{pmatrix}0, & D_1\otimes\boldsymbol{\beta}\end{pmatrix} & & & & \\
\begin{matrix}(1,i_a)\\(1,i_a,i_w)\end{matrix} & \begin{pmatrix}0\\0\end{pmatrix} & \begin{pmatrix}D_0 & 0\\ I\otimes\mathbf{S}^0 & 0\end{pmatrix} & \begin{pmatrix}0 & D_1\otimes\boldsymbol{\beta}\\ 0 & I\otimes S\end{pmatrix} & & & \\
\begin{matrix}(2,i_a)\\(2,i_a,i_w)\end{matrix} & \vdots & & \begin{pmatrix}D_0 & 0\\ I\otimes\mathbf{S}^0 & 0\end{pmatrix} & \begin{pmatrix}0 & D_1\otimes\boldsymbol{\beta}\\ 0 & I\otimes S\end{pmatrix} & & \\
\vdots & \vdots & & & \ddots & \ddots & \\
\begin{matrix}(Q-2,i_a)\\(Q-2,i_a,i_w)\end{matrix} & \begin{pmatrix}0\\0\end{pmatrix} & & & & \begin{pmatrix}D_0 & 0\\ I\otimes\mathbf{S}^0 & 0\end{pmatrix} & \begin{pmatrix}0 & D_1\otimes\boldsymbol{\beta}\\ 0 & I\otimes S\end{pmatrix} \\
\begin{matrix}(Q-1,i_a)\\(Q-1,i_a,i_w)\end{matrix} & \begin{pmatrix}D_1\\ I\otimes(S\mathbf{e})\end{pmatrix} & & & & & \begin{pmatrix}D_0 & 0\\ I\otimes\mathbf{S}^0 & 0\end{pmatrix}
\end{array}. \qquad (5.3)$$

At the beginning of each consolidation cycle, the system will be initialized in state (0, i_a, 0). This means the accumulated weight is zero, the clock of the arrival process is turned on, and the clock of the weight distribution is off.

In each of the subsequent periods, with probabilities D_0, no order will arrive and the clock of the arrival process continues while the clock of the weight distribution remains off. On the other hand, with probabilities $D_1 \otimes \boldsymbol{\beta}$, an order will arrive. The accumulated weight then increases by one, the clock of the arrival process stops, and the clock of the *PH*-weight distribution will be turned on.

Once the clock of the weight distribution is on, the accumulated weight will continue to grow with probabilities $I \otimes S$ until it reaches Q – 1, whereby it will stop with probabilities $I \otimes \mathbf{S}^0$, which means the weight of an order has been fully generated and the order arrival process resumes.

When the accumulated weight reaches Q – 1, if the clock of the arrival process is on, the weight will remain unchanged with transition probabilities D_0 (no order arrives). Alternatively, an order will arrive with matrix probabilities D_1, which will then trigger a dispatch and initialize a new consolidation cycle. If the clock of the weight distribution is on, either the accumulated weight stops growing with probabilities $I \otimes \mathrm{S}^0$, or it grows by one more unit and completes the consolidation cycle with probabilities $I \otimes (S\mathbf{e})$.

Let $\boldsymbol{\pi}_{TW}$ be the steady state distribution of P_{TW}, i.e., $\boldsymbol{\pi}_{TW}P_{TW} = \boldsymbol{\pi}_{TW}$ and $\boldsymbol{\pi}_{TW}\mathbf{e} = 1$. We decompose $\boldsymbol{\pi}_{TW}$ as follows: $\boldsymbol{\pi}_{TW} = (\boldsymbol{\pi}_0, (\boldsymbol{\pi}_{1,a}, \boldsymbol{\pi}_{1,w}), (\boldsymbol{\pi}_{2,\mathrm{a}}, \boldsymbol{\pi}_{2,w}), \ldots, (\boldsymbol{\pi}_{Q-2,a}, \boldsymbol{\pi}_{Q-2,w}), (\boldsymbol{\pi}_{Q-1,a}, \boldsymbol{\pi}_{Q-1,w}))$. The steady state distribution $\boldsymbol{\pi}_{TW}$ can be found by using the following algorithm.

Algorithm I(PH)

I(PH).1) Compute the matrices

$$
\begin{aligned}
&X_0 = I; \\
&X_1 = X_0 \begin{pmatrix} 0 & D_1 \otimes \boldsymbol{\beta} \end{pmatrix} \left(I - \begin{pmatrix} D_0 & 0 \\ I \otimes \mathbf{S}^0 & 0 \end{pmatrix} \right)^{-1}, \\
&X_w = X_{w-1} \begin{pmatrix} 0 & D_1 \otimes \boldsymbol{\beta} \\ 0 & I \otimes S \end{pmatrix} \left(I - \begin{pmatrix} D_0 & 0 \\ I \otimes \mathbf{S}^0 & 0 \end{pmatrix} \right)^{-1}, \quad 2 \le w \le Q-1; \qquad (5.4) \\
&P_1 = D_0 + X_{Q-1} \begin{pmatrix} D_1 \\ I \otimes S\mathbf{e} \end{pmatrix}.
\end{aligned}
$$

I(PH).2) Solve the linear system $\boldsymbol{\pi}_0 P_1 = \boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_0 \left(\sum_{j=0}^{Q-1} X_j \right) \mathbf{e} = 1$ for $\boldsymbol{\pi}_0$.

I(PH).3) Determine ($\boldsymbol{\pi}_{i,a}$, $\boldsymbol{\pi}_{i,w}$) = $\boldsymbol{\pi}_0 X_i$, for $i = 1, 2, \ldots, Q$–1.

Compared to *Algorithm I*, *Algorithm I*(*PH*) computes matrices only of the size m_a or $m_a m_b$, where m_a is the number of phases in the underlying Markov chain of the order arrival process, and m_b is the number of phases in the underlying Markov chain of the *PH*-weight distribution. Performance measures can be obtained accordingly. First, similar to *Theorem 4.4*, the expected cycle length between shipments can be found as follows.

Theorem 5.2 The average time between two consecutive entrances to level zero from level Q–1 of the Markov chain $\{(\hat{W}(t), \hat{I}_a(t), \hat{I}_w(t)),\ t = 0, 1, 2, \ldots\}$ is given by $1/(\boldsymbol{\pi}_0 D_1 \mathbf{e})$. Consequently, the expected cycle length of consolidated shipments is $E[L_c] = 1/(\boldsymbol{\pi}_0 D_1 \mathbf{e}) + 1 - Q$.

Proof Following the same approach used in the proof of *Theorem 4.4*, the transition probability matrix of the embedded Markov chain at the end of each consolidation cycle is given by $P_2 = (I - D_0)^{-1} X_{Q-1} \begin{pmatrix} D_1 \\ I \otimes S\mathbf{e} \end{pmatrix}$. Let $\boldsymbol{\eta}_0$ be the invariant vector of P_2, i.e., $\boldsymbol{\eta}_0 P_2 = \boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_0 \mathbf{e} = 1$. Then $\boldsymbol{\eta}_0$ can be interpreted as the probability distribution of the state of the order arrival process at the beginning of a consolidation cycle.

By equation (5.4), it can be shown that $\boldsymbol{\eta}_0(I - D_0)^{-1} = \delta\boldsymbol{\pi}_0$, which leads to $\boldsymbol{\eta}_0 = \boldsymbol{\pi}_0(I - D_0) / (\boldsymbol{\pi}_0(I - D_0)\mathbf{e})$ and $\delta = 1 / (\boldsymbol{\pi}_0(I - D_0)\mathbf{e})$. Similar to *Theorem 4.4*, the average time between two consecutive entrances from level Q–1 to level zero can be obtained as $1/(\boldsymbol{\pi}_0 (I - D_0)\mathbf{e})$. Intuitively, the probability that the Markov chain just entered level zero is $\boldsymbol{\pi}_0(I - D_0)\mathbf{e}$, and thus the result follows.

Since the fictitious time between two consecutive visits to level zero from level $Q - 1$ is exactly $Q - 1$ (i.e., the accumulated weight increases to Q), the expected cycle length is obtained as $E[L_c] = 1/\big(\boldsymbol{\pi}_0(I - D_0)\mathbf{e}\big) - (Q-1)$, which, together with $(D_0 + D_1)\mathbf{e} = \mathbf{e}$, leads to the theorem's result. □

Next, we show that, for this special case, explicit results can be obtained for the excess O_w.

Theorem 5.3 Assume that the excess threshold function is $g(j) = Q_o \geq Q$ for all j. The excess at a dispatch epoch has a phase-type distribution with matrix representation ($\boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}$, S). In addition, we have $E[O_w] = \boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}(I-S)^{-1}\mathbf{e}$, and $P\{O_w > 0\}$ = $\boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}\mathbf{e}$.

Proof Consider a Markovian arrival process with matrix presentation (S, $\mathbf{S}^0\boldsymbol{\beta}$) (i.e., a *PH*-renewal process). If we treat Q as a time, then the excess at Q is the time until the next arrival of that Markovian arrival process. The distribution of phases at Q is $\boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}$. Then the excess at Q has a discrete *PH*-distribution with matrix representation ($\boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}$, S). The expected excess and the probability that an excess occurs are obtained accordingly. This completes the proof. □

By *Theorem 5.3*, the consolidated weight per cycle is equal to Q plus the excess O_w. That excess has a *PH*-distribution ($\boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S$, S). The expected weight per cycle is then given by

$$E[W_c] = Q + \boldsymbol{\beta}(S + \mathrm{S}^0\boldsymbol{\beta})^{Q-1}S(I - S)^{-1}\mathbf{e} . \tag{5.5}$$

To find the accumulated weight at an arbitrary time, we consider the steady state distribution for the process $\{(W(t), I_a(t)), t = 0,1,2,\ldots\}$ (i.e., censoring out all the phases associated with the underlying Markov chain $\{I_w(t), t = 0,1,2,\ldots\}$). Define

$$\begin{aligned}\boldsymbol{\pi}_{TW,a} &= (\,\boldsymbol{\phi}_0 , \boldsymbol{\phi}_1 , \boldsymbol{\phi}_2 , \ldots, \boldsymbol{\phi}_{Q-2} , \boldsymbol{\phi}_{Q-1}\,) \qquad (5.6)\\ &= (\,\boldsymbol{\pi}_0 , \boldsymbol{\pi}_{1,a} , \boldsymbol{\pi}_{2,a} , \ldots, \boldsymbol{\pi}_{Q\text{-}2,a} , \boldsymbol{\pi}_{Q\text{-}1,a}\,) / (\,\boldsymbol{\pi}_0 + \boldsymbol{\pi}_{1,a} + \boldsymbol{\pi}_{2,a} + \ldots + \boldsymbol{\pi}_{Q-2,a} + \boldsymbol{\pi}_{Q-1,a}\,)\mathbf{e}.\end{aligned}$$

By definition, $\boldsymbol{\pi}_{TW,a}$ is the steady state distribution of the accumulated weight $W(t)$ at an arbitrary time. Then we obtain $E[W] = \sum_{i=0}^{Q-1} i\boldsymbol{\phi}_i\mathbf{e}$. In addition, we must have $\boldsymbol{\theta}_a = \boldsymbol{\phi}_0 + \boldsymbol{\phi}_1 + \boldsymbol{\phi}_2 + \ldots + \boldsymbol{\phi}_{Q-2} + \boldsymbol{\phi}_{Q-1}$, which is useful for checking computational accuracy.

5.3 Independent-Weights Model

Consider the weight-arrival process defined in *Example 3.1*. The order weights are independent of the order arrival process. Then we have (note $p_0 = 0$)

$$\overline{D}_0 = D_0 + D_1 \text{ and } \overline{D}_n = (1 - p_1 - p_2 - \ldots - p_{n-1})D_1 \text{, } n = 1, 2, \ldots. \tag{5.7}$$

Denote by w_o the weight of an arbitrary order. First we see that equation (3.1) can be modified as

$$\begin{aligned} \lambda_{wt} &= \lambda_{av} E\left[w_o\right] \\ \lambda_{av} &= \boldsymbol{\theta}_a \left(\sum\nolimits_{n=1}^{\infty} D_n\right)\mathbf{e} \quad . \end{aligned} \tag{5.8}$$

Consequently, construction of the transition probability matrix and all expressions given in *Corollary 4.3* can be reduced to finite summations. The expressions for the average weight per cycle and average overshoot per occurrence can be simplified as shown in equations (5.9) and (5.10).

Although it seems complicated, equation (5.9) can be derived from *Corollary 4.3* and explained intuitively. Conditioned on the fact that a dispatch will occur by the end of the period and the system is currently in level j with accumulated weight w, we first calculate the expected weight at the end of the period if an order has arrived. We then remove the cases in which the weight of the arrived order is not sufficient to trigger a dispatch. Lastly, we add the cases in which no order has arrived but a dispatch still occurred because of the lowered target load.

$$E[W_c] = \frac{1}{P_S} \sum_{j=1}^{j_q+1} \left(\sum_{w=0}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} \big(E[w_o] + w \big) D_1 + \sum_{w=0:\, w \ge f(j)}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} (w) D_0 - \sum_{w=0:\, w < f(j)}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} \left(\sum_{i=1}^{f(j)-w-1} (i+w) p_i \right) D_1 \right) \mathbf{e} . \tag{5.9}$$

$$\begin{aligned} E[O_w] = \frac{1}{P_S} \sum_{j=1}^{j_q+1} \left(\sum_{w=0}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} \big(E[w_o] + w - g(j) \big) D_1 + \sum_{w=0:\, w \ge g(j)}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} (w - g(j))\, D_0 \right) \mathbf{e} \\ - \frac{1}{P_S} \sum_{j=1}^{j_q+1} \left(\sum_{w=0:\, w < g(j)}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} \left(\sum_{i=1}^{g(j)-w-1} (i + w - g(j))\, p_i \right) D_1 \right) \mathbf{e} \end{aligned} \quad . \tag{5.10}$$

By adjusting the expected weight at the end of the period this way, we can obtain the expected weight of a dispatched shipment. Similarly, equation (5.10) can be derived from *Corollary 4.3* and explained intuitively as well.

As demonstrated by equations (4.1), (4.2) and (5.8) through (5.10), and by *Corollary 4.3*, the calculation for various performance measures can be simplified by using the probabilities $\{ p_1 , p_2 , \dots \}$ and $E[w_o]$. The advantage is that all summations are finite (except the calculation of $E[w_o]$). These simplifications are applicable to all formulas in models with independent weights. They are especially useful for checking the accuracy and efficiency of numerical computation.

CHAPTER SIX

EVALUATING A SHIPMENT CONSOLIDATION POLICY FOR PRIVATE OR COMMON CARRIAGE

Although we have established formulas to calculate several performance measures in Chapter Four, we have not yet discussed how to compute the cost to carry out a shipment consolidation policy in the long run. In this chapter, we will describe how to measure such cost by either the private or common carriage. In the end, to assist potential users of our model, we will summarize it as a set of procedures for evaluating a particular shipment consolidation policy.

6.1 The Cost of Shipment Consolidation by Private Carriage

As discussed in Section 3.4, cost is an important performance measure for any shipment consolidation policy. When that policy is employed by a private carrier, its cost structure is different from that of a common carrier. We wish to measure the expected long-run cost per unit time. Let us define this cost in its basic form for the private carriage case as $C(f)_p$. From equation (3.2), we know that

$$C(f)_p = \frac{E[\text{ Transportation Cost per Cycle }] + E[\text{ Inventory Holding Cost per Cycle }]}{E[\text{ Cycle Length }]} .$$

First let us take a look at the transportation cost per cycle. We can break it down into three categories: weight-based costs, order-based costs and dispatch cost. Some typical examples that are weight-based include the costs of packing, staging, loading and unloading products. Together, they are commonly referred to as "shipment-handling cost". These are usually variable costs proportional to the weight of the shipment, so we can assign to them an overall rate of K_W per unit weight. Since every order received will eventually be handled, the choice of shipment consolidation policy and the expected cycle length will be unaffected by this cost. In the long run, the average weight-based costs per unit time are equal to $K_W\lambda_{wt}$, where λ_{wt} is the long-run weight arrival rate defined in Section 3.2.

Order-based costs account for the costs of receiving, processing and managing orders. They usually have standard rates associated with each order; together they can be denoted as K_O. Similar to the weight-based costs, they do not affect the choice of consolidation policy, nor the expected cycle length because every order received will require processing. The average order-related cost per unit time can be obtained as $K_O\lambda_{av}$, where λ_{av} is the long-run order arrival rate of the *MAP* for the weight-arrival process. (recall that λ_{av} was defined in Section 3.2.)

From our introduction of private carriage in Section 1.2, we know that there is a fixed charge for dispatching a shipment, which can be denoted as K_D. This cost will only be charged when a consolidated load is shipped. Since the expected cycle length, denoted by $E[L_c]$, varies depending on the consolidation policy, our model will calculate the long-run average dispatch cost per unit time as $K_D / E[L_c]$.

There is no need to actually compute the expected inventory holding cost per cycle. Instead, as mentioned in Section 4.4, suppose we can find $E[W]$. That is, the expected inventory level for any arbitrary period when the system is in steady state. Its product with the holding cost rate h will thus yield the expected inventory holding cost per unit time as $hE[W]$.

Therefore, we can rewrite equation (3.2) for the private carriage case as

$$C(f)_p = hE\left[W\right] + \frac{K_D}{E[L_c]} + K_W \lambda_{wt} + K_O \lambda_{av}. \tag{6.1}$$

Note that K_D, K_W, K_O and h are all given as cost parameters of the private carriage shipment consolidation problem. Once we have found a *BMAP* representation of the weight-arrival process, we can immediately obtain λ_{av} and λ_{wt} by equation (3.1). For general model instances, we can use *Corollary 4.3* to obtain $E[W]$ and follow *Theorem 4.4* to get $E[L_c]$; for those special cases mentioned in Chapter Five, we can calculate $E[W]$ and $E[L_c]$ by using their modified formulas presented in that chapter.

6.2 The Cost of Shipment Consolidation by Common Carriage

In Section 1.2, we briefly introduced the cost structure of a common carrier. Recall that a common carrier charges transportation cost according to the weight of each shipment. The standard (non-volume) freight rate c_N is usually employed to calculate the transportation cost, but when the weight of a shipment exceeds *MWT*, the minimum weight qualifying for a volume discount, the common carrier will offer that discount by lowering its freight rate to c_V. The transportation cost function $c(w)$ for common carrier, where w is the weight of the shipment, is given by equation (1.1) as

$$c(w) = \begin{cases} c_N w, & w < MWT \\ c_V w, & w \geq MWT \end{cases} .$$

We also introduced the notion of the "bumping clause", which refers to the action taken by the shipper to over-declare the weight of their shipment as MWT, so that firm can qualify for the volume discount. This is only worthwhile when $c_N w > c_V MWT$. Therefore, if the bumping clause is allowed, the transportation cost function $c(\cdot)$ should be updated by an *effective* transportation cost function $\tilde{c}(w)$, defined by Çetinkaya and Bookbinder (2003) as

$$\tilde{c}(w) = \begin{cases} c_N w, & w \leq WBT \ , \\ c_V MWT \ , & WBT \ < w \leq MWT \ , \\ c_V w, & w > MWT \ , \end{cases} \tag{6.2}$$

where $WBT = c_V MWT / c_N$ is the weight at which the bumping clause becomes effective.

It is worth pointing out that unlike the private carriage case, when we calculate transportation cost for common carrier, we do not consider the fixed costs associated with each order or shipment. However, those still often exist in practice, but they are usually incorporated into the freight charges by the common carrier.

Recall that in Section 4.4 we defined W_c as the accumulated weight of an arbitrary shipment in steady state, and in *Corollary 4.3* we provided a formula for the probability density function of W_c. If we substitute the random variable W_c for w in equation (6.2), we can calculate the expected common carriage transportation cost per shipment using equation (6.3).

Note that for the independent-weight cases that have an explicit formula for $E[w_o]$, we can use the idea from equation (5.9) to eliminate the infinite summation in equation (6.3) and rewrite it as equation (6.4).

$$\begin{aligned} &E\left[\tilde{c}\left(W_c\right)\right] \\ &= \sum_{i=0}^{WBT} c_N \cdot i \cdot P\left\{W_c = i\right\} + \sum_{i=WBT+1}^{MWT} c_V \cdot MWT \cdot P\left\{W_c = i\right\} + \sum_{i=MWT+1}^{\infty} c_V \cdot i \cdot P\left\{W_c = i\right\} \\ &= \sum_{i=0}^{WBT} c_N \cdot i \cdot P\left\{W_c = i\right\} + \sum_{i=WBT+1}^{MWT} c_V \cdot MWT \cdot P\left\{W_c = i\right\} + c_V\left[EW_c - \sum_{i=0}^{MWT} i \cdot P\left\{W_c = i\right\}\right] \end{aligned} \quad . \tag{6.3}$$

$$\begin{aligned} &E\left[\tilde{c}\left(W_c\right)\right] \\ &= \frac{1}{P_S} \sum_{j=1}^{j_q+1} \left[\sum_{w=0}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} D_1 \left(\sum_{i=1}^{WBT-w} c_N p_i (w+i) + \sum_{i=WBT-w+1}^{MWT-w} c_V p_i MWT + c_V \left(E\left[w_o\right] + w - \sum_{i=1}^{MWT-w} p_i \cdot (w+i) \right) \right) \right. \\ &\quad + \sum_{w=f(j)}^{WBT} \boldsymbol{\theta}_{j,w} D_0 \left(c_N w\right) + \sum_{w=WBT+1}^{MWT} \boldsymbol{\theta}_{j,w} D_0 \left(c_V MWT\right) + \sum_{w=MWT+1}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} D_0 \left(c_v w\right) \\ &\quad \left. - \sum_{w=0}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} D_1 \left(\sum_{i=1}^{\min(WBT,f(j))-w} c_N p_i (w+i) + \sum_{i=WBT-w+1}^{\min(MWT,f(j))-w} c_V p_i MWT + \sum_{i=MWT-w+1}^{f(j)-w} c_V p_i (w+i) \right) \right] \end{aligned} \quad . \tag{6.4}$$

Even if there is no explicit formula for $E[w_o]$ or the order weights are correlated with the arrival process, in most of these cases, W_c will be finite because individual order weights are finite. This is particularly true in practice because by its nature, shipment consolidation is more effective when the weights of individual orders are relatively small. Otherwise, it makes more sense to ship large orders individually. Therefore, computation of equation (6.3) is feasible in most situations.

For common carriage, inventory holding cost can be calculated in the same way as in the private carriage case. Together with the transportation cost, we can write the common carriage cost function as

$$C(f)_c = hE\left[W\right] + \frac{E\left[\tilde{c}\left(W_c\right)\right]}{E[L_c]}. \tag{6.5}$$

To compute equation (6.5), freight rates c_N and c_V, holding cost rate h and *MWT* are given as parameters, which allow us to obtain *WBT*. We then use equations (6.3) or (6.4) to evaluate $E\left[\tilde{c}\left(W_c\right)\right]$ and *Corollary 4.3* to calculate $E[W]$ and $P\{\ W_c = i\ \}$, for $i = 1, 2, \ldots$.

6.3 Procedures to Evaluate a Shipment Consolidation Policy

Now we are ready to summarize our model as a set of procedures to evaluate the performance of a shipment consolidation policy by either private carriage or common carriage.

Procedure SCP (Shipment Consolidation Policy)

1) Define ($D_0, D_n, n = 1,2,\ldots$); compute $\boldsymbol{\theta}_\mathrm{a}$, λ_{wt} and λ_{av}. [*Section 3.2*]
2) Define a consolidation policy function $f(\cdot)$. [*Section 3.3*]
3) Model the shipment consolidation process by P_{TW}. [*Sections 4.1 - 4.2*]
4) Compute the steady state distribution $\boldsymbol{\theta}_{TW}$. [*Section 4.3*]
5) Obtain the non-financial performance measures: $E[W]$, $E[W_c]$, $E[O_w]$, $E[L_c]$, $E[L_w]$ and $E[N_c]$. [*Sections 4.4 - 4.6*]
6) For private carrier, specify cost parameters K_D, K_W, K_O and h, then calculate $C(f)_p$; for common carrier, specify c_N, c_V, MWT and h, then compute $C(f)_c$. [*Sections 6.1 - 6.2*]

The square bracket at the end of each step indicates the earlier sections in which more details about the step can be found. A detailed summary of the notation that appears in *Procedure SCP* can be found in *Table 6.1*.

Table 6.1: Summary of Notation

Notations:	**Interpretations:**	**Formulas:**
$(D_0, D_n, n=1,2,\ldots)$	*BMAP* weight-arrival process	N/A
$\boldsymbol{\theta}_{\mathrm{a}}$	Steady state distribution for the *BMAP*	N/A
λ_{wt}	Long-run weight arrival rate	Equation (3.1)
λ_{av}	Long-run order arrival rate	Equation (3.1)
P_{TW}	Transition probability matrix for the shipment consolidation process Markov chain	Theorem 4.1
$\boldsymbol{\theta}_{TW}$	Stationary distribution for P_{TW}	Algorithm I
$E[W]$	Expected inventory level per period	Corollary 4.3
$E[W_c]$	Expected weight per shipment	Corollary 4.3
$E[O_w]$	Expected excess per shipment	Corollary 4.3
$E[L_c]$	Expected cycle length	Theorem 4.4
$E[L_w]$	Expected order delay	Theorem 4.4
$E[N_c]$	Expected number of orders per shipment	Theorem 4.5
K_D	Private carrier dispatch cost	N/A
K_W	Private carriage weight related transportation rate	N/A
K_O	Private carriage order related transportation rate	N/A
h	Holding cost rate	N/A
c_N	Common carriage non-volume freight rate	N/A
c_V	Common carriage volume freight rate	N/A
MWT	Minimum weight qualifying for volume discount	N/A
$C(f)_p$	Private carriage long-run average cost per unit time	Equation (6.1)
$C(f)_c$	Common carriage long-run average cost per unit time	Equation (6.5)

Combined with results of the previous chapters, *Procedure SCP* will provide detailed instructions on how to model the shipment consolidation process under a particular policy, the way to obtain the steady state statistics, and how to compute the various performance measures. Be aware that this set of procedures is designed for *general* model instances. For the special cases shown in Chapter Five, substitute the modified formulas, theorems and algorithms accordingly.

CHAPTER SEVEN

NUMERICAL ANALYSIS

After illustrating the theoretical components of our model in the last few chapters, we are now ready to put it through numerical tests. We will run it against a variety of test cases involving different weight-arrival processes and consolidation policies. We will also attempt to search for the optimal consolidation policy parameters for each weight-arrival process. Through these analyses, we hope to find out more about which class of policies is more suitable for a particular weight-arrival process.

7.1 Examples of Weight-Arrival processes and Cost Parameters

We begin our numerical analysis by specifying five different examples of weight-arrival processes. We will use them to evaluate each type of consolidation policy. Every example represents a typical convolution between the order arrival process and the order weight distribution, and it has certain distinctive features. Together, they can cover a fairly broad range of weight-arrival processes.

Example 7.1 Our first example is a typical *BMAP* weight-arrival process. Its matrix representation is

$$D_0 = \begin{pmatrix} 0 & 0.3 & 0 & 0 & 0 \\ 0 & 0.1 & 0.6 & 0 & 0 \\ 0 & 0 & 0.1 & 0.6 & 0 \\ 0 & 0 & 0 & 0.1 & 0.7 \\ 0.8 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad D_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$D_4 = \begin{pmatrix} 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad D_5 = \begin{pmatrix} 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{7.1}$$

In this process, the arrival times and the order weights are correlated; large orders (weight of 4 or 5) occur only if the phase of the underlying Markov chain is either 1 or 2.

Example 7.2 This example is an extension of *Example 3.1*. Orders arrive according to a *MAP* with a matrix representation

$$D_0 = \begin{pmatrix} 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0.1 & 0.2 & 0 & 0 \\ 0 & 0 & 0.1 & 0.4 & 0 \\ 0 & 0 & 0 & 0.1 & 0.5 \\ 0.6 & 0 & 0 & 0 & 0.1 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.4 \\ 0.3 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (7.2)$$

The order weights have a normalized Riemann Zeta distribution (Alexander, Baclawski and Rota, 1993) with probability density function { $p_n = 1/(\zeta_{2.5}n^{2.5})$, n=1, 2, … }, where $\zeta_{2.5}$ = 1.3415 is the normalization factor. The weight distribution is independent of the order arrival process. In this case, the weight-arrival process can also be modeled as a *BMAP* with matrix representation (D_0, p_nD_1, n = 1, 2, …). Note that the standard deviation of the order weights is significantly larger than other cases.

Example 7.3 This example extends *Example 3.2*, in which the underlying Markov chain had only one phase. Here the arrival process is modeled as $D_0 = 0.5$ and $D_1 = 0.5$, i.e. in each period an order can arrive with probability 0.5. The distribution of order weights now follows an empirical distribution given by { p_1=0.45, p_2=0.3, p_3=0.1, p_4=0.1, p_5=0.05 and 0 otherwise }, that is independent of the order arrival process. Therefore, the weight-arrival process now has a *compound geometric* distribution and can be represented by a *BMAP* as ($D_0 = 0.5$, $p_1D_1 = 0.225$, $p_2D_1 = 0.15$, $p_3D_1 = 0.05$, $p_4D_1 = 0.05$, and $p_5D_1 = 0.025$).

Example 7.4 This example corresponds to *Example 3.3* and the order arrival process is the same *MAP* given by equation (7.2). The order weights have a *PH*-distribution with matrix representation

$$\boldsymbol{\beta} = \begin{pmatrix} 0.1 & 0.9 \end{pmatrix}, \quad S = \begin{pmatrix} 0.15 & 0.3 \\ 0.2 & 0.3 \end{pmatrix}. \tag{7.3}$$

It is well-known that *PH*-distributions are light tailed. The distribution of order weights can be expressed as { $p_n = \boldsymbol{\beta} S^{n-1}(I - S)\mathbf{e}, \quad n = 1, 2, \cdots$ }, which can then be used to form a *BMAP* (D_0 , p_nD_1 , $n = 1, 2, \ldots$).

Example 7.5 This example corresponds to *Example 3.4* in which the orders arrive according to a *PH*-renewal process. The interarrival times have a common *PH*-distribution

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}, \quad T = \begin{pmatrix} 0 & 0.5 \\ 0.35 & 0.1 \end{pmatrix}. \tag{7.4}$$

If we transform it into a *MAP* representation, we would have

$$D_0 = T = \begin{pmatrix} 0 & 0.5 \\ 0.35 & 0.1 \end{pmatrix}, \quad D_1 = \mathbf{T}^0 \boldsymbol{\alpha} = \begin{pmatrix} 0.25 & 0.25 \\ 0.275 & 0.275 \end{pmatrix}. \qquad (7.5)$$

Order weights have the same discrete *PH*-distribution defined in *Example 7.4*, which is denoted by { $p_n = \boldsymbol{\beta} S^{n-1}(I - S)\mathbf{e}, \quad n = 1, 2, \cdots$ }. A *BMAP* can again be formed as (D_0 , $p_n D_1$, $n = 1, 2, \ldots$).

In *Table 7.1*, we provide a summary of the five examples we gave above. The table shows their respective long-run order arrival rate λ_{av}, long-run weight arrival rate λ_{wt}, mean order weight $E[w_o]$ and the standard deviation of order weights $std(w_o)$. According to equation (5.8), for the independent weight cases such as *Examples 7.2*, *7.4* and *7.5*, the long-run weight arrival rate can be computed as $\lambda_{wt} = \lambda_{av} E[w_o]$.

Note that we have purposely manipulated the parameters so that their long-run weight arrival rates λ_{wt} are all close to one. Therefore, although these weight-arrival processes may differ in order arrival frequencies and sizes, their overall rates in steady state are roughly the same. This allows us to establish some consistencies in comparing the five processes.

Table 7.1: Summary of Weight-Arrival-Process Examples

	Example 7.1	Example 7.2	Example 7.3	Example 7.4	Example 7.5
Arrival Process	*BMAP*	*MAP*	Compound Geometric	*MAP*	*PH*
Weight Distribution		Riemann Zeta		*PH*	*PH*
λ_{av}	0.3354	0.5347	0.5	0.5347	0.5273
λ_{wt}	1.0354	1.0333	1	1.0444	1.0299
$E[w_o]$	N/A	1.9325	2	1.9533	1.9533
std(w_o)	N/A	12.0100	1.1832	1.3458	1.3458

Also note that for *Examples 7.2* to *7.5*, because the order weight distribution is independent of the order arrival process, we can use the model modifications described in Section 5.3. For *Examples 7.4* and *7.5*, since the order weights follow a *PH*-distribution, we can take advantage of the revised model from Section 5.2 if a quantity policy is applied in these cases.

Now we still need to define a set of private carriage cost parameters to be used for all test cases. Suppose K_D = \$10, h = \$0.1, K_W = \$0 and K_O = \$0. We set the last two costs to be zero because they have no effect on the consolidation policies.

We can define another set of cost parameters for common carriage: Let c_N = \$5, c_V = \$4 (equivalent to 20% quantity discount), *MWT* = 15 *cwt*, and hence *WBT* = 12 *cwt*. In addition, we will set Q_o = 20 as the vehicle capacity constraint. In our subsequent numerical analysis, we shall interpret O_w as the excess over vehicle capacity.

Keep in mind that these weight-arrival processes and the cost parameters are conjured merely for the purpose of numerical analysis, so some of them may not be the most realistic representations of practical cases.

7.2 Evaluating Individual Policies

Now we will choose a particular instance for each consolidation policy and evaluate them individually for all five weight-arrival processes.

Quantity Policy: Suppose we have a particular quantity policy defined as

$$f(j) = \begin{cases} Q = 13, & j = 1 \\ Q = 13, & j \geq j_q = 2 \end{cases} .$$

We can use the simplified algorithm for quantity policy models (Section 5.1) for *Examples 7.1* to *7.3*, and we can use the modified model (Section 5.2) for *Examples 7.4* and *7.5*. The results for performance measure are given in *Table 7.2*.

Table 7.2: Performance Measure Results for Quantity Policy

	Example 7.1	Example 7.2	Example 7.3	Example 7.4	Example 7.5
C(f)$_p$	1.2415	1.1789	1.2836	1.3086	1.2982
C(f)$_c$	5.7071	5.3554	5.5616	5.6383	5.5678
***E*[*W*]**	5.2999	5.5479	5.6157	5.5944	5.5944
***E*[*W$_c$*]**	14.5517	16.5574	13.8500	13.9403	13.9403
***P*{*O$_w$>0*}**	0	0.0849	0	0.0029	0.0029
***E*[*O$_w$*]**	0	2.1784	0	0.0056	0.0056
***E*[*L$_c$*]**	14.0540	16.0239	13.8500	13.3476	13.5355
***E*[*L$_w$*]**	7.5912	8.6534	7.2504	6.9404	7.1012
***E*[*N$_c$*]**	4.7139	8.5679	6.9250	7.1369	7.1369

<u>Pseudo-Time Policy:</u> Suppose a particular pseudo-time policy is defined as

$$f(j) = \begin{cases} Q = 100\,, & 1 \le j < j_q = T = 11 \\ 0, & j \ge j_q = T = 11 \end{cases} .$$

The performance measure results are given in *Table 7.3*.

Table 7.3: Performance Measure Results for Pseudo-Time Policy

	Example 7.1	Example 7.2	Example 7.3	Example 7.4	Example 7.5
$C(f)_p$	1.4268	1.3896	1.4091	1.4313	1.4240
$C(f)_c$	5.5381	5.1484	5.4021	5.2894	5.2126
$E[W]$	5.1771	4.7914	5	5.2220	5.1495
$E[W_c]$	11.3896	11.3492	11	11.4884	11.3290
$P\{O_w>0\}$	0.0390	0.0708	0.0223	0.0354	0.0358
$E[O_w]$	0.0809	5.1293	0.0446	0.2263	0.2323
$E[L_c]$	11	10.9835	11	11	11
$E[L_w]$	5	4.9969	5	5	5
$E[N_c]$	3.6896	5.8728	5.5	5.8816	5.8000

Hybrid Policy: Suppose a particular hybrid policy is defined as

$$f(j)=\begin{cases} Q=30, & 1\le j<j_q=T=14 \\ 0, & j\ge j_q=T=14 \end{cases}.$$

The performance measure results are given in *Table 7.4*.

Table 7.4: Performance Measure Results for Hybrid Policy

	Example 7.1	Example 7.2	Example 7.3	Example 7.4	Example 7.5
$C(f)_p$	1.3867	1.2880	1.3641	1.3922	1.3828
$C(f)_c$	5.4749	5.0353	5.3777	4.9628	4.9062
$E[W]$	6.7221	5.6125	6.4971	6.7760	6.6813
$E[W_c]$	14.4913	14.2180	13.9983	14.6143	14.4112
$P\{O_w>0\}$	0.1481	0.1198	0.0983	0.1249	0.1226
$E[O_w]$	0.2988	3.3169	0.1999	1.0996	1.0986
$E[L_c]$	13.9957	13.7598	13.9983	13.9930	13.9927
$E[L_w]$	6.4981	6.4465	6.4993	6.4972	6.4971
$E[N_c]$	4.6944	7.3573	6.9992	7.4820	7.3780

<u>General Policy:</u> Suppose a particular general policy is defined as

$$f(j) = \begin{cases} 20, & 1 \le j \le 5 \\ 15, & 6 \le j \le 10 \\ 10, & 11 \le j \le 15 \\ 0, & j \ge j_q = 16 \end{cases} .$$

The performance measure results are given in *Table 7.5*.

Table 7.5: Performance Measure Results for General Policy

	Example 7.1	Example 7.2	Example 7.3	Example 7.4	Example 7.5
$\boldsymbol{C(f)_p}$	1.3029	1.2249	1.3282	1.3519	1.3424
$\boldsymbol{C(f)_c}$	5.6077	5.2673	5.4838	5.5444	5.4713
$\boldsymbol{E[W]}$	4.6227	4.3697	4.9062	4.9813	4.9426
$\boldsymbol{E[W_c]}$	12.3173	13.1146	11.9389	12.2324	12.1438
$\boldsymbol{P\{O_w>0\}}$	0.00005	0.0591	0.00009	0.0035	0.0035
$\boldsymbol{E[O_w]}$	0.00007	1.6427	0.00013	0.0068	0.0068
$\boldsymbol{E[L_c]}$	11.8960	12.6920	11.9389	11.7123	11.7911
$\boldsymbol{E[L_w]}$	5.7034	6.1970	5.6814	5.5712	5.6223
$\boldsymbol{E[N_c]}$	3.9901	6.7863	5.9695	6.2625	6.2171

The correctness of the algorithm can be checked by using the following relationships: $\lambda_{wt} = E[W_c] / E[L_c]$ (*Remark 4.2*) and $\lambda_{av} = E[N_c] / E[L_c]$ (*Remark 4.3*). When we applied the pseudo-time policy to *Example 7.2*, the expected cycle length $E[L_c]$ is close but not equal to 11. This is mainly because the order weight distribution is heavy tailed, so the chances of an extremely large order to arrive and force the accumulated weight to exceed its limit are significant. Thus, sometimes dispatch may occur before period 11.

In *Example 7.2*, the probability of a shipment exceeding the vehicle capacity and the expected weight of such excess are significantly higher than that in the other examples. This is most likely due to the fact that this *MAP* has a higher standard deviation of order weights. On the other hand, *Example 7.3*, which has the smallest standard deviation of order weights, has the lowest "excess measures" for all policies. Therefore, our numerical results are consistent with the intuition that a more volatile weight-arrival process is more likely to cause vehicle capacity to be exceeded.

7.3 Optimal Policy Parameters

Our previous results are insufficient to find the optimal policy for each weight-arrival process. To obtain that, we will first try to find the optimal policy parameters for each class of policies, and then compare their respective costs.

Since our current model is only capable of evaluating a single policy instance, more extensive research is required to reveal the optimality properties of this problem. For now, we will simply examine a certain range of policies and find the local optima. We will do so by enumerating these policies and use our model to compute their costs.

We tested all quantity policies for $Q \in [2, 50]$ against the five weight-arrival processes; their results are plotted in *Figure 7.1*. We also tested pseudo-time policies for $T \in [2, 30]$; their results are presented in *Figure 7.2*. We used mesh plots (*Figure 7.3*) to show the results for the range of hybrid policies where $Q \in [2, 30]$ and $T \in [2, 20]$.

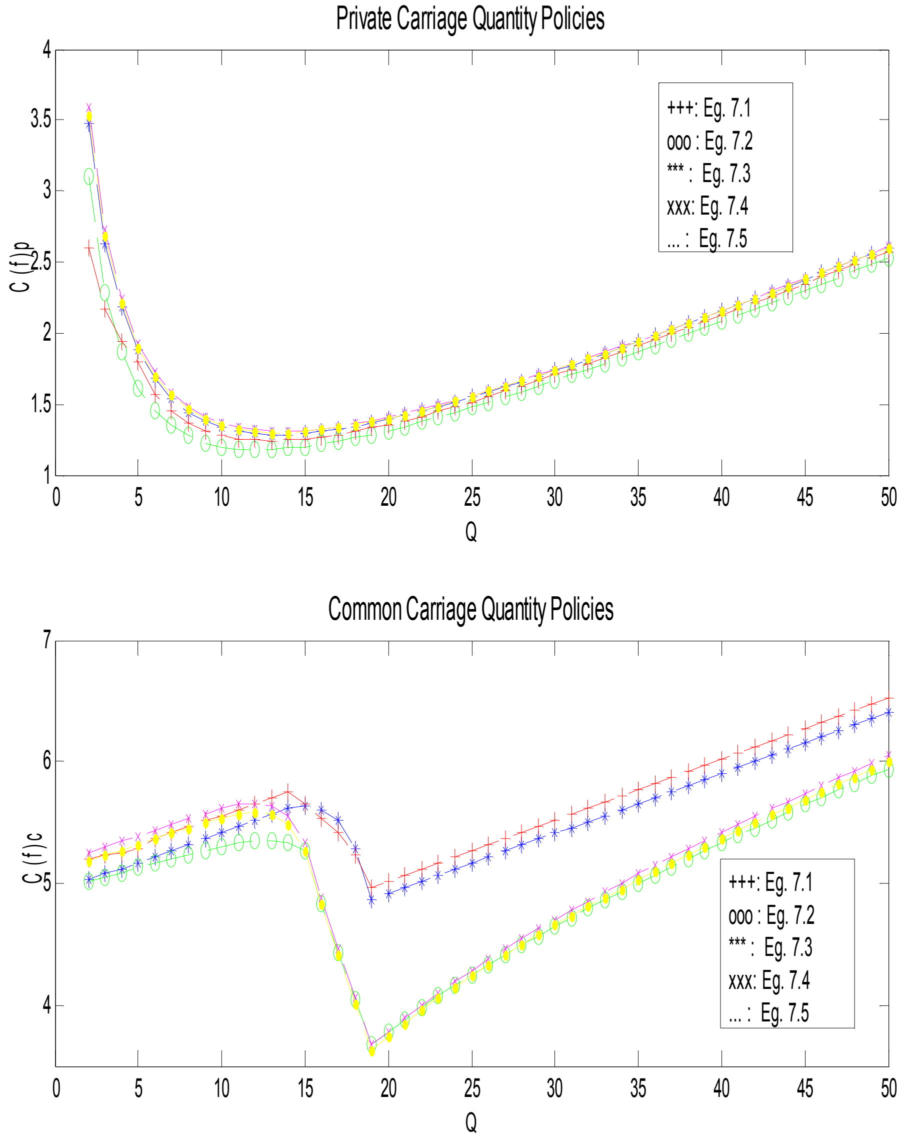
Unfortunately, there are too many feasible general policies even for a very small range, thus the method of enumeration becomes very inefficient. We need to develop a more sophisticated search algorithm to find the optimum. As a result, we did not perform such a search for the general policy.

The policies with the lowest costs are recorded in *Table 7.6*. The ones that gave the lowest cost overall are highlighted in grey. From *Table 7.6*, we see that under both private carriage and common carriage, the quantity policy outperformed the other two classes of policies for our five weight-arrival processes.

Table 7.6: Summary of Lowest-Cost Policies

			Example 7.1	Example 7.2	Example 7.3	Example 7.4	Example 7.5
Quantity Policies	***Private Carriage***	Q^*	13	12	13	14	13
		$C^*(f)_p$	1.2415	1.1773	1.2836	1.3081	1.2982
	Common Carriage	Q^*	19	19	19	19	19
		$C^*(f)_c$	4.9689	3.6700	4.8603	3.6776	3.6286
Pseudo-Time Policies	***Private Carriage***	T^*	14	15	14	14	14
		$C^*(f)_p$	1.3873	1.3111	1.3643	1.3931	1.3837
	Common Carriage	T^*	2	27	2	22	22
		$C^*(f)_c$	5.2289	4.5493	5.0500	4.4422	4.4082
Hybrid Policies	***Private Carriage***	(Q^*, T^*)	(13, 20)	(12, 20)	(13, 20)	(14, 20)	(14, 20)
		$C^*(f)_p$	1.2594	1.1863	1.2894	1.3126	1.3031
	Common Carriage	(Q^*, T^*)	(2, 2)	(19, 20)	(2, 2)	(19, 20)	(19, 20)
		$C^*(f)_c$	5.6910	4.5730	5.0130	4.1532	4.1331

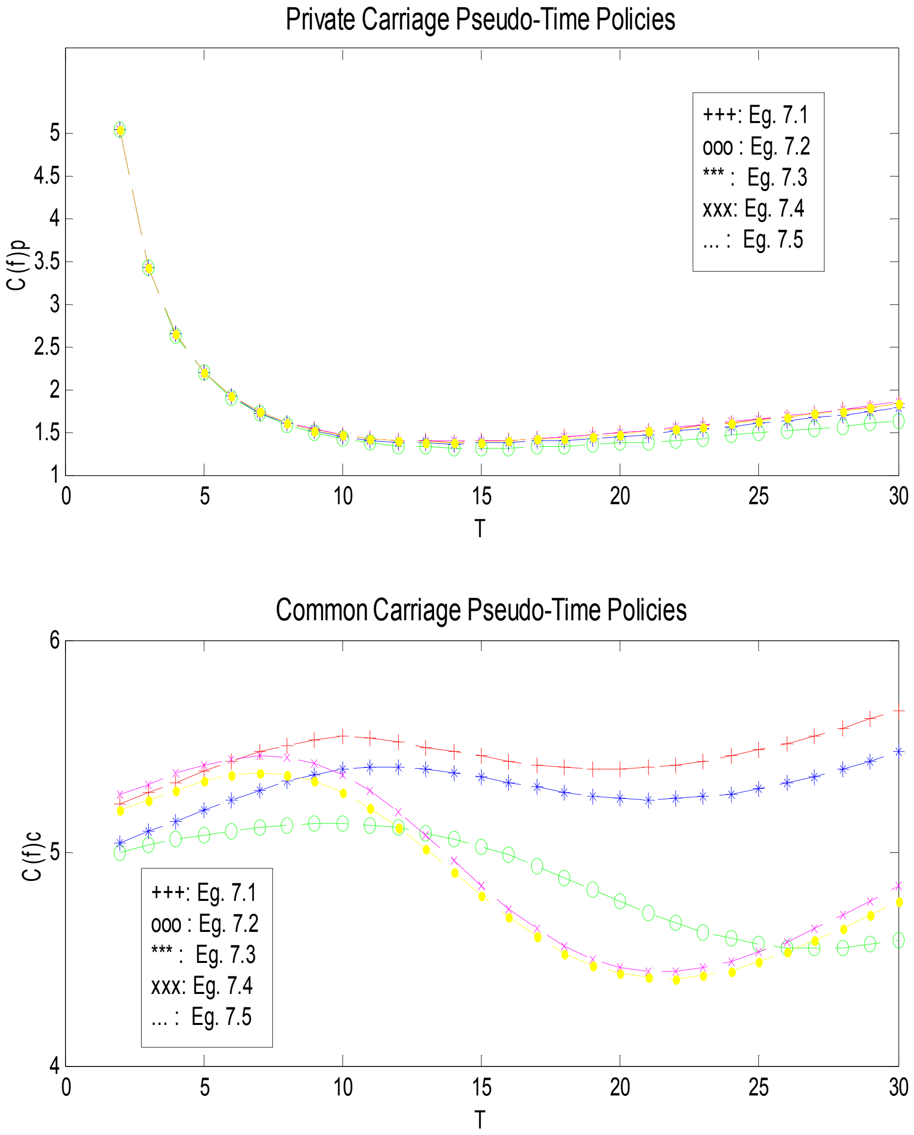
Figure 7.1: Quantity Policy Cost Functions



Private Carriage Quantity Policies

Common Carriage Quantity Policies

Figure 7.2: Pseudo-Time Policy Cost Functions



Private Carriage Pseudo-Time Policies

Common Carriage Pseudo-Time Policies

Figure 7.3-1: Hybrid Policy Cost Functions for Example 7.1

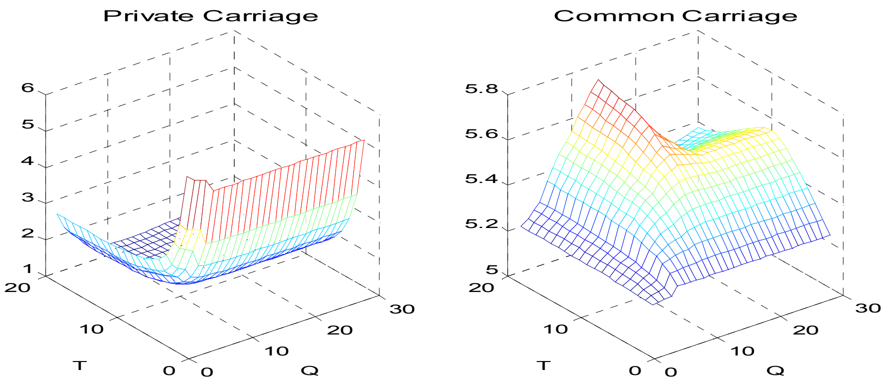


Figure 7.3-2: Hybrid Policy Cost Functions for Example 7.2

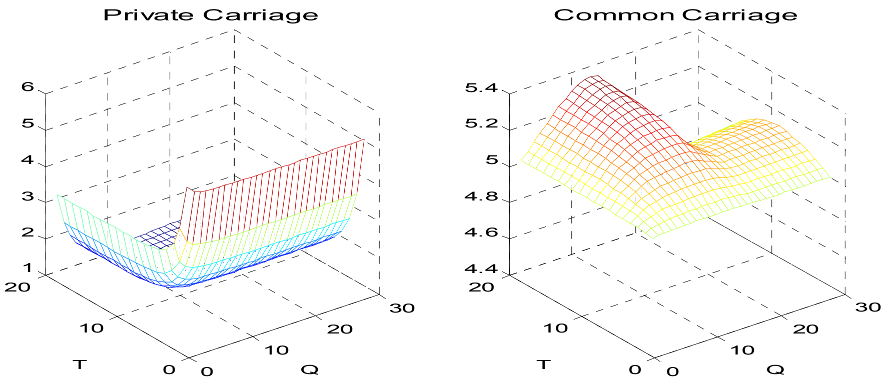


Figure 7.3-3: Hybrid Policy Cost Functions for Example 7.3

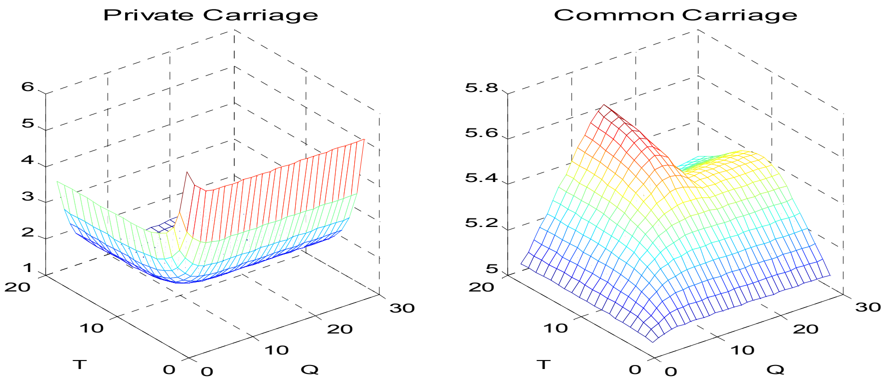


Figure 7.3-4: Hybrid Policy Cost Functions for Example 7.4

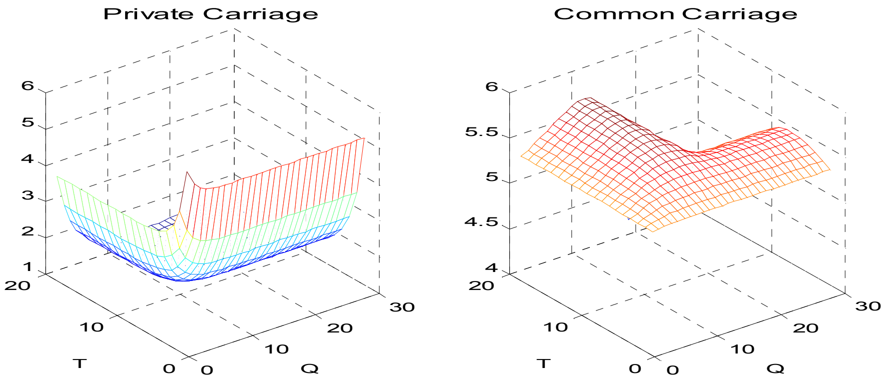
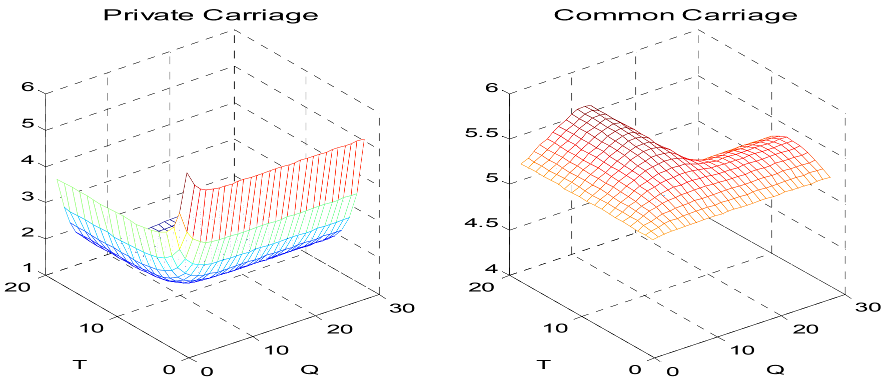


Figure 7.3-5: Hybrid Policy Cost Functions for Example 7.5



From our numerical results, it can be observed that the private carriage cost function $C(f)_p$ is relatively smooth for the quantity, pseudo-time and hybrid policies. The graphs also appear to support the hypothesis that $C(f)_p$ is unimodular for all three classes of policies. The best policies we obtained for private carriage are fairly close to one another. Therefore, to decide which policy is more preferable, we need to take other performance measures into consideration.

In terms of common carriage, there are more noticeable differences in the results. For the quantity policy, there is a steep drop in cost once the target load is more than *MWT*. For a pseudo-time policy, a milder decrease in cost occurs when the maximum waiting time is long enough to allow sufficient accumulation of order weights to obtain the quantity discount. The shapes of the mesh plots for the hybrid policy indicate that lower cost is achieved when the target load is above *MWT* and the maximum waiting time is long enough for orders to accumulate until that level so that the quantity discount can be rewarded.

After evaluating our test cases, we shifted our focus to analyze the compound geometric weight-arrival process. We generated many more examples of this process and found the costs for the optimal quantity, pseudo-time and hybrid policies. We also calculated the costs for a variety of manually selected general policies in each case. All of our test results indicate that quantity policy dominated over the other classes of policies. Our numerical experiments led us to believe that for all cases with the compound geometric weight-arrival process, the optimal policy is a quantity policy. More theoretical research is required to prove this hypothesis.

Although the quantity policy appears to be the least expensive one in all of our previous test cases, we have not yet compared it with the optimal general policy. In fact, it is not uncommon to discover that a certain general policy can be better that the optimal quantity policy. To show that a general policy can indeed out-perform the other policies, consider the following order arrival process

$$D_0 = \begin{pmatrix} 0.0307 & 0 \\ 0.1 & 0.8 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.8693 & 0.1 \\ 0 & 0.1 \end{pmatrix}, \tag{7.6}$$

Consider the private carriage case with cost parameters K_D = \$2 and h = \$0.1. Suppose the order weights are independent of the order arrivals and they follow the same Riemann Zeta distribution as that in *Example 7.2*. The optimal quantity, pseudo-time and hybrid policies and their corresponding costs are shown in the table below. By a manual search, a general policy is found to be better than the respective optima for the other three classes of policies.

	Policy	$C^*(f)_p$
General policy	$f(j) = 6$, for $j = 1, 2, \ldots, 5$ $f(j) = 4$, for $j \geq 6$.	0.3970
Optimal quantity policy	$Q^* = 10$	0.4139
Optimal pseudo-time policy	$T^* = (9, 300)$	0.4361
Optimal hybrid policy	$(Q^*, T^*) = (12, 20)$	0.4095

We found a few other examples in which a certain general policy outperforms the optima for the other three classes. Suppose the order arrival process is defined by the following *MAP*

$$D_0 = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.2 & 0.3 \\ 0.2 & 0 & 0 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.8 \end{pmatrix}. \tag{7.7}$$

With the same order weight distribution and cost parameters, we found that a general policy again has a smaller expected long-run cost per unit time than the best policies from the other three classes.

	Policy	$C^*(f)_p$
General policy	$f(j) = 5$, for $j = 1, 2, \ldots, 4$ $f(j) = 4$, for $j \geq 5$.	0.3265
Optimal quantity policy	$Q^* = 4$	0.3357
Optimal pseudo-time policy	$T^* = 10$	0.3671
Optimal hybrid policy	$(Q^*, T^*) = (4, 25)$	0.3346

Both examples demonstrated that for certain weight-arrival processes, it is possible to have a general policy as the optimal policy. Intuitively, this makes sense because if the accumulation process becomes idle prior to reaching the target load, it would be better to ship out early instead of racking up the inventory cost. Based on these results, we intend to determine under what circumstances (in terms of the weight-arrival process) would the general policy be preferable.

CHAPTER EIGHT

CONCLUDING REMARKS

8.1 Thesis Summary

Let us recap what we have been able to achieve through our studies on the discrete shipment consolidation problem. First and foremost, by utilizing matrix-analytic methods, more specifically through the use of *BMAPs*, *MAPs* and *PH*-distributions, we made it possible to model a greater variety of problem instances in terms of the weight-arrival process.

Modeling the weight-arrival process by a *BMAP* enabled us to take into consideration the potential correlations between the order arrival process and the order weight distribution, as well as any correlations between consecutive arrivals. The *BMAP* has also been proven to be a powerful modeling tool because of its ability to approximate any arbitrary arrival process from empirical data. In addition, the convolution between a *MAP* and an independent weight distribution can be easily expressed as a *BMAP*. Therefore, when the order arrival process and the order weight distribution are uncorrelated, they can be modeled by a *MAP* and a discrete positive distribution, respectively.

Our general model can thus accommodate any discrete weight distribution. When that distribution cannot be expressed in a closed form, one can employ the more versatile *PH*-distribution. We showed that the *PH*-distribution can approximate almost any distribution of non-negative weights. This improves upon the previous research, in furnishing a more general way to model the vast range of order weight distributions. Together with the *MAP* and *BMAP*, the *PH*-distribution gives us greater flexibility when we need to model an arbitrary weight-arrival process.

Another feature worth mentioning is that our modeling approach provides a set of procedures to evaluate a variety of performance measures. Beside the two widely studied performance measures, i.e. average cost per unit time and mean cycle length (see, for example, Higginson, (1993), Çetinkaya and Bookbinder (2003) and Mütlü, Çetinkaya, and Bookbinder (2010)), we showed how to obtain other measures such as expected excess over target load / vehicle capacity, expected order delay and expected number of orders per shipment. Our model provides a wider range of evaluation tools for a shipment consolidation policy.

Note that some previous research on shipment consolidation used renewal theory to construct the common carriage cost function (see, for example Çetinkaya and Bookbinder (2003)). This approach is often hindered by integrals that are difficult to evaluate. As a result, the solution needs be obtained through approximation. Fortunately, matrix-analytic methods enable us to obtain an explicit and relatively easy-to-compute formula for the cost function of common carriage (see equation (6.4)). A precise solution can thus be calculated from this formula.

In general, our model can determine the performance measures and costs of any consolidation policy to a high precision, while still remaining efficient. In addition, some useful relationships between results have been established so we can check them effectively (see remarks 4.2, 4.3 and 5.1). The set of procedures presented in Chapter Six is a guideline on how to apply our model in practice.

Our third achievement is that our model is capable of evaluating almost any consolidation policy that can be found in practice. Beside the quantity, time and hybrid policies which were studied by others before, we found a way to evaluate the more-general policies sometimes used by practitioners. This is perhaps the most important and novel feature of our research.

8.2 Further Comments

Although we have made some progress and contributions in solving the shipment consolidation problem in this thesis, there are still some difficulties that remain to be addressed. First, we have not yet determined which class of consolidation policies is more likely to contain the optimal policy for different weight-arrival processes. Even if we could study the class of all general policies, we lack an efficient search algorithm to locate the optimal one. This is a crucial problem because our ultimate goal is to be able to find that optimal policy efficiently and accurately.

The preceding limitations are even more striking when it is realized that we restricted our model to discrete time and discrete quantity. The discrete nature of those parameters limits the precision of our model solutions. If we try to improve this precision by switching to smaller units, we will create challenges in computation by enlarging the state space of the Markov chain. Consider the size of the target load, first in cwt and then in lbs. That change increases the number of states by a factor of 100. The computation required will thus increase significantly.

However, the following should be kept in mind. In practice, the target weight is limited by the vehicle capacity. The state space of our Markov chain is thus usually a reasonable size, although the preceding comparison between cwt and lbs is still relevant. But eventually, we hope to extend the matrix-analytic methodology to the case of continuous quantity.

Although the *BMAP* and *PH*-distribution can be used to approximate arbitrary weight-arrival processes and distributions, we did not present algorithms to do that. Given an empirical order weight distribution, one must fit it by a *PH*-distribution before using our model. Therefore, methods are required to fit *BMAPs* or *PH*-distributions. This fitting process itself can be complicated and time consuming (e.g. Asmussen and Nerman, 1991; Horvath and Telek, 2002). The quality of the fit will of course directly impact the accuracy of solutions to the model.

8.3 Future Research

We have considered the target load to be a function only of elapsed time since the last dispatch. What if that target load were also determined by the phase of the underlying Markov chain, i.e. the dispatch decision were contingent on the business scenario or seasonality? In that case, we would need to revamp our model and potentially take a new approach to the problem.

Our current model is a useful tool for more systematic studies of the shipment consolidation problem. The next important questions are to determine the optimality properties of the cost functions. More specifically, does there truly exist an optimal policy in terms of cost? Does the class of optimal policies depend on the weight-arrival process?

We have observed certain numerical trends and properties that shed some light on the issue of optimality. For example, of all the compound geometric weight-arrival processes we tested, the optimal policy always turned out to be a quantity policy. In those cases, there seems to be a direct link between the form of the optimal policy and the steady state distribution of the process. The next step in our research is to explore this relationship and give a formal proof.

We tested a variety of weight-arrival processes that are more general than the compound geometric. The identity of the optimal policy became more ambiguous. The quantity policy was still best for Examples 7.1 – 7.5 (Table 7.6). However, in

Section 7.3, other more-general policies were found to outperform the best quantity policy for different instances of weight-arrival processes.

Ideally, we would like to derive the form of the optimal shipment consolidation policy. We would then need a search strategy to find its optimal parameters in a particular case. More extensive research and a formal proof are warranted if we wish to generalize the optimality properties to an arbitrary weight-arrival process.

REFERENCES

Abdelwahab, W.M. and M. Sargious (1990), Freight rate structure and optimal shipment size in freight transportation, *The Logistics and Transportation Review*, vol. 6, no. 3, pp. 271-292.

Akaah, I.P. and G. Jackson (1988), Frequency distributions of customer orders in physical distribution systems, *Journal of Business Logistics*, vol. 9, no. 2, pp. 155-164.

Alexander, K.S., K. Baclawski and G.C. Rota (1993), A stochastic interpretation of the Riemann zeta function, *Proceedings of the National Academy of Sciences of the USA*, vol. 90, pp. 697-699

Asmussen, S. and G. Koole (1993), Marked point processes as limits of Markovian arrival streams, *Journal of Applied Probability*, vol. 30, pp. 365-372.

Asmussen, S. and O. Nerman (1991), Fitting phase-type distributions via the EM algorithm, *Symposium I Anvendt Statistik*, January 21-23, 1991 (ed. K. Vest Nielsen), 335-346. UNI-C, Copenhagen.

Asmussen, S., O. Nerman and M. Olsson (1996), Fitting phase-type distributions via the EM algorithm, *Scandinavian Journal of Statistics*, vol. 23, pp. 419-441.

Blumenfeld, D.E., L.D. Burns, J.D. Diltz and C.F. Daganzo (1985), Analyzing tradeoffs between transportation, inventory, and production costs on freight networks, *Transportation Research B*, vol. 19, no. 5, pp. 361-380.

Bobbio, A. and A. Cumani (1992), ML estimation of the parameters of a PH distribution in triangular canonical form. *Computer Performance Evaluation*, pp. 33-46, G. Balbo and G. Serazzi, eds., Elsevier Science Publishers.

Bookbinder, J.H. and J.K. Higginson (2002), Probabilistic modeling of freight consolidation by private carriage, *Transportation Research E*, vol. 38, pp. 305-318.

Burns, L.D., R.W. Hall, D.E. Blumenfeld and C.F. Daganzo (1985), Distribution strategies that minimize the transportation and inventory costs, *Operations Research*, vol. 33, no. 3, pp. 469-490.

Çetinkaya, S. and J.H. Bookbinder (2003), Stochastic models for the dispatch of consolidated shipments, *Transportation Research B*, vol. 37, pp. 747-768.

Çetinkaya, S. (2004), Coordination of inventory and shipment consolidation decisions: A review of premises, models, and justification, Chapter 1 in *Applications of Supply Chain Management and E-Commerce Research in Industry*, J. Geunes et al., eds., New York: Springer.

Closs, D.J. and R.L. Cook (1987), Multi-stage transportation consolidation analysis using dynamic simulation, *International Journal of Physical Distribution and Materials Management*, vol. 17, no. 3, pp. 28-45.

Cooper, M.C. (1983), Freight consolidation and warehouse location strategies in physical distribution systems, *Journal of Business Logistics*, vol. 4, no. 2, pp. 53-74.

Cooper, M.C. (1984), Cost and delivery time implications of freight consolidation and warehouse stretegies, *International Journal of Physical Distribution and Materials Management*, vol. 14, no. 6, pp. 47-67.

Daganzo, C.F. (1988), Shipment consolidation enhancement at a consolidation center, *Transportation Research B*, vol. 22, no. 2, pp. 103-124.

Gail, H.R., S.L. Hantler and B.A. Taylor (1994), Solutions of the basic matrix equation for *M/G/*1 and *G/M/*1 type Markov chains, *Stochastic Models*, vol. 10, pp. 1-43.

Gail, H.R., S.L. Hantler and B.A. Taylor (1997), Non-skip-free *M/G/*1 and *G/M/*1 type Markov chains, *Adv. Appl. Probab.*, vol. 29, pp. 733-758.

Gupta, Y.P. and P.K. Bagchi (1987), Inbound freight consolidation under Just-In-Time procurement: application of clearing models, *Journal of Business Logistics*, vol. 8, no. 2, pp. 74-94.

Ha, K.H., S. Khasnabis and G. Jackson (1988), Impact of freight consolidation on logistics system performance, *Journal of Transportation Engineering*, vol. 114, no. 2, pp. 173-193.

Hall, R.W. (1987), Consolidation strategy: inventory, vehicles and terminals. *Journal of Business Logistics*, vol. 8, no. 2, pp. 57-73.

He, Qi-Ming (2009), *Lecture Notes on Matrix-Analytic Methods*, University of Waterloo.

He, Qi-Ming and M.F. Neuts (1998), Markov chains with marked transitions, *Stochastic Processes and their Applications*, vol. 74, no. 1, pp. 37-52.

Higginson, J.K. (1993), Modeling shipper costs in physical distribution analysis, *Transportation Research A*, vol. 27, no. 2, pp. 113-124.

Higginson, J.K. (1995), Recurrent decision approaches to shipment-release timing in freight consolidation, *International Journal of Physical Distribution and Logistics Management*, vol. 25(5), pp. 19-32.

Higginson, J.K. and J.H. Bookbinder (1994), Policy recommendations for a shipment consolidation program, *Journal of Business Logistics*, vol. 15, no. 1, pp. 87-112.

Higginson, J.K. and J.H. Bookbinder (1995), Markovian Decision Processes in shipment consolidation, *Transportation Science*, vol. 29, pp. 242-255.

Horvath, A. and M. Telek (2002), PhFit: A general purpose phase type fitting tool. *Tools 2002*, pp. 82-91, London, England, Springer, LNCS 2324.

Hsu, G-H and Qi-Ming He (1991), The distribution of the first passage time for the Markov processes of *GI/M*/1 type, *Stochastic Models*, vol. 7, No. 3, pp. 397-417.

Jackson, G.C. (1981), Evaluating order consolidation strategies using simulation, *Journal of Business Logistics*, vol. 2, no. 2, pp. 110-138.

Johnson, M.A. and Taaffe, M.R. (1990a), Matching moments to phase distributions: nonlinear programming approaches, *Stochastic Models*, vol. 6, pp. 259-281.

Johnson, M.A. and Taaffe, M.R. (1990b), Matching moments to phase distributions: density function shapes, *Stochastic Models*, vol. 6, pp. 283-306.

Latouche, G. (1987), A note on two matrices occurring in the solution of quasi-birth-and-death processes, *Stochastic Models*, vol. 3/2, pp. 251-257.

Latouche, G. and V. Ramaswami (1993), A logarithmic reduction algorithm for quasi-birth-and-death process, *Journal of Applied Probability*, vol. 30, pp. 650-674.

Latouche, G. and V. Ramaswami (1999), *Introduction to Matrix Analytic Methods in Stochastic Modelling*, ASA & SIAM, Philadelphia, USA.

Latouche, G., Marie-Ange Remiche and P. Taylor (2003), Transit Markov arrival processes, *The Annals of Applied Probability*, vol. 13, no.2, pp. 628-640.

Lucantoni, D.M. (1991), New results on the single server queue with a batch Markovian arrival process, *Stochastic Models*, vol. 7, pp. 1-46.

Lucantoni, D.M. and V. Ramaswami (1985), Efficient algorithms for solving the non-linear matrix equations arising in phase type queues, *Stochastic Models*, vol. 1, pp. 29-51.

Maters, J.M. (1980), The effects of freight consolidation on customer service, *Journal of Business Logistics*, vol. 2, no. 1, pp. 55-74.

Mütlü, F., S. Çetinkaya and J.H. Bookbinder (2010), An Analytical Model for Computing the Optimal Time-and-Quantity-Based Policy for Consolidated Shipments, *IIE Transactions*, vol. 42, no. 5, pp. 367-377.

Neuts, M.F. (1975), Probability distributions of phase type, *In Liber Amicorum Prof. Emeritus H. Florin*, pp. 173-206, University of Louvain.

Neuts, M.F. (1979), A versatile Markovian point process, *Journal of Applied Probability*, vol. 16, pp. 764-779.

Neuts, M.F. (1981), *Matrix-Geometric Solutions in Stochastic Models – An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore.

Neuts, M.F. (1989a), *Structured Stochastic Matrices of M/G/*1 *type and their Applications*, Marcel Dekker, New York.

Neuts, M.F. (1989b), The joint distribution of arrivals and departures in quasi-birth-and-death processes, *Stochastic Models*.

Neuts, M.F. (1992), Models based on the Markovian arrival process, *IEICE Trans. Commun. E75-B,* 1255-1265.

Newbourne, M.J. and C. Barrett (1972), Freight consolidation and the shipper, *Transportation and Distribution Management*, vol. 12, no. 2 to 6.

Pollock, T. (1978), A management guide to LTL consolidation, *Traffic World*, April 3, pp. 29-35.

Russell, R.M. and L. Krajewski (1991), Optimal purchase and transportation cost lot sizing for a single item, *Decision Science*, vol. 22, pp. 940-952.

Thummler A., P.Buchholz and M. Telek (2006), A novel approach for phase-type fitting with the EM algorithm, *IEEE Transactions on Dependable and Secure Computing*, vol. 3, No.3, pp. 245-258.

Tyworth, J.E., J.L. Cavinato and C.J. Langley, Jr. (1987), *Traffic Management: Planning, Operations, and Control*, Addison-Wesley Publishing Company, Reading, Massachusetts.

Ülkü, M.A., and J.H. Bookbinder (2006), Policy analysis in shipment consolidation, in *Proceedings of the 26th Turkish National OR/IE Conference*, Kocaeli, Turkey, July, 2006, pp. 9-12.