NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

SHIPMENT CONSOLIDATION IN BUSINESS LOGISTICS MANAGEMENT

by

James K. Higginson

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Management Sciences

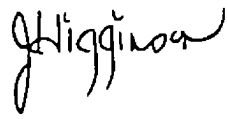
Waterloo, Ontario, Canada, 1992

ISBN 0-315-75758-2

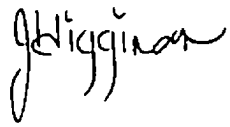
Canada

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.



I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.



The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

SHIPMENT CONSOLIDATION IN BUSINESS LOGISTICS MANAGEMENT

ABSTRACT

Shipment consolidation refers to the active intervention by management to combine many small shipments into fewer, larger loads. These larger loads benefit from reduced per–unit transportation cost, faster and more direct transportation, and improved employee and equipment utilization.

Shipment consolidation has received serious academic attention only in the last ten years. Published research generally has examined the effect on distribution costs and delivery time performance of varying consolidation parameters. Determination of practical decision rules for a consolidation program is lacking.

We seek to provide a better understanding of the interaction of the basic components of a consolidation program. We first give an overview of shipment consolidation, including types, benefits, disadvantages, and prerequisites of this strategy. The related literature also is critiqued.

Our research focuses on the question, "When should customer orders be dispatched as a consolidated load?" There are three commonly–used shipment–release policies: a) time policy, where each order is dispatched at a pre–determined date; b) quantity policy, where all orders are held until a minimum consolidated weight is reached; and c) time–and–quantity policy, where all orders are held until a predetermined time, unless a minimum weight or volume is accumulated first. We examine the implications of these policies on mean per–unit cost and mean order delay of a consolidation program via computer simulation.

After selecting a shipment–release policy, values for policy parameters must be determined. We discuss two types of analytical methods for determining when consolidated loads should be released. **Non-sequential approaches** treat the shipment–release question as a "one–time" decision. The following non–sequential approaches are examined: simple rules–of–thumb and heuristics, the economic shipment quantity concept, stochastic analysis of order characteristics, bulk–service queueing models, and stochastic clearing systems.

Sequential approaches make the shipment–release decision whenever an order arrives. Sequential approaches discussed in the thesis include marginal analysis and Markov decision processes.

Further research topics relating to shipment consolidation are identified. Appendices include a glossary of physical distribution terms and a discussion on modeling shipper distribution costs.

ACKNOWLEDGEMENTS

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 Dr. J.H. Bookbinder, my supervisor and mentor, for his encouragement, advice, support, patience, and more;
- To Dr. H. Armitage, Dr. C. Daganzo, Dr. Y. Gerchak, and Dr. M.J. Magazine, for their comments, criticisms, and suggestions as members of my examining committee;
- To the Social Sciences and Humanities Research Council, the Ontario Graduate Scholarship Program, the Canadian Transportation Research Forum, the Faculty of Engineering, and the Department of Management Sciences, for their generous financial support;
- Regarding Appendix C: to Mr. G.F. MacKay for his comments about transportation and inventory–holding costs, to Dr. F.A. Haight for his continued interest and encouragement, and to two anonymous referees for their suggestions;
- To the prototype railways and the model railroad industry for providing an interesting and enjoyable diversion which is close enough to my academic studies that information exchange occurs, yet far enough away that a relaxing escape results; thanks are especially due to Mr. W.M. Skelton and Mr. W. Young of Requiring Some Assembly Inc. in Waterloo.

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

LIST OF TABLES

LIST OF FIGURES

Chapter 1
SHIPMENT CONSOLIDATION: PROBLEM STATEMENT

1.1 Introduction

This chapter introduces our research in shipment consolidation. Section 1.2 presents an illustration of the cost savings available from this strategy. We emphasize that this example views shipment consolidation in a very simplistic manner; other consolidation techniques are possible and are discussed in Section 1.3.

This is followed by a statement of our research problem and an outline of the scope of our work. A final section defines the physical distribution system configurations that are relevant to this thesis.

1.2 An Illustrative Example

Suppose that, at the end of each day, 6000 pounds of glass fibre roofing material is shipped from Boston to Ann Arbor. Assuming a constant production rate, an inventory–holding cost of ten cents per hundred pounds ("hundredweight" or "cwt.") per day, and a transportation charge of $2.95 per hundredweight, the total daily transportation and inventory–holding cost for this shipment is:

transportation cost TC_T:	6000 lbs. x 2.95/cwt.	= $ 177
inventory–holding cost TC_R:	6000/2 lbs. x 0.10/cwt.	= 3
		$ 180

The total weekly transportation and inventory–holding cost is $180 per day x 5 days = $900 per week.

Assume that the shipper decides to combine five daily shipments into one 30,000–pound load dispatched at the end of each week. The weekly inventory–holding cost now increases to TC_R = 30,000/2 lbs. x 0.10 per cwt. x 5 days = $75 per week.

However, by making only one shipment per week, the total weight of this load is sufficiently large to be classified as a "carload shipment"[1] (if transported by rail) or a "truckload shipment" (if transported by motor carriage). As a result, freight charges are now calculated under lower "volume rates". Based on a truckload shipment, the weekly transportation cost in this example becomes TC_T = 30,000 lbs. x $2.07/cwt. = $621, where $2.07 is the applicable volume freight rate per hundred pounds.

The total weekly transportation and inventory–holding cost is now $696, a saving of $204, or almost 23%, over that with daily shipments.

1.3 Discussion

The above example illustrates the basic idea of shipment consolidation. Active planning by management to aggregate several small shipments into one large load can significantly reduce distribution costs. As discussed in the next chapter, consolidation also can result in faster delivery, reduced damage and loss, and improved utilization of equipment, thus providing a valuable competitive tool.

In our example, we assumed that transportation was done by a "common", or for–hire, carrier. Under common carriage, a shipment that weighs less than a specified minimum weight will be transported under non–volume freight rates (such as less–than–carload (LCL) or less–than–truckload (LTL) rates). By combining several

[1] Logistical terms used in this thesis are defined in Appendix A.

LCL or LTL shipments, a shipper often can increase the weight of this consolidated load to a point above this minimum weight, thus qualifying for the lower volume (carload or truckload) rates. [This thesis will refer to LCL and LTL shipments (rates) as "non–volume shipments (rates)", carload and truckload shipments (rates) as "volume shipments (rates)", and the minimum weight at which volume rates apply as "minimum volume weight". The latter avoids confusion with several other "minimum weights" that exist in physical distribution. Note that terms such as "volume shipments (rates)" refer to weight, not physical size or physical volume.]

Alternately, the shipper might use his own vehicles. Because the cost of operating a private vehicle largely depends on distance rather than on load size, the dispatch of one vehicle each week instead of one each day could reduce transportation cost by as much as 80%.

Shipment consolidation is not limited to one product or to one destination. For example, if several small orders, none qualifying for the lower volume freight rate, were destined for different customers in the same geographical region, the shipper could make one consolidated shipment under volume rates to a central facility. There, the load would be disaggregated for local delivery to individual customers. Distribution system configurations relevant to our research are discussed later in this chapter.

Moreover, suppose that the roofing material manufacturer also sends small loads to a customer in Buffalo. By combining these shipments with those to Ann Arbor, the resulting consolidated weight may be sufficiently large to qualify as a volume shipment, even though a portion of the load will be removed in Buffalo before the vehicle reaches its final destination.

Clearly, shipment consolidation can take many forms and can encompass many techniques. This chapter discusses the scope and goals of our research, and defines the forms of shipment consolidation that are relevant to this work.

1.4 Problem Statement and Research Scope

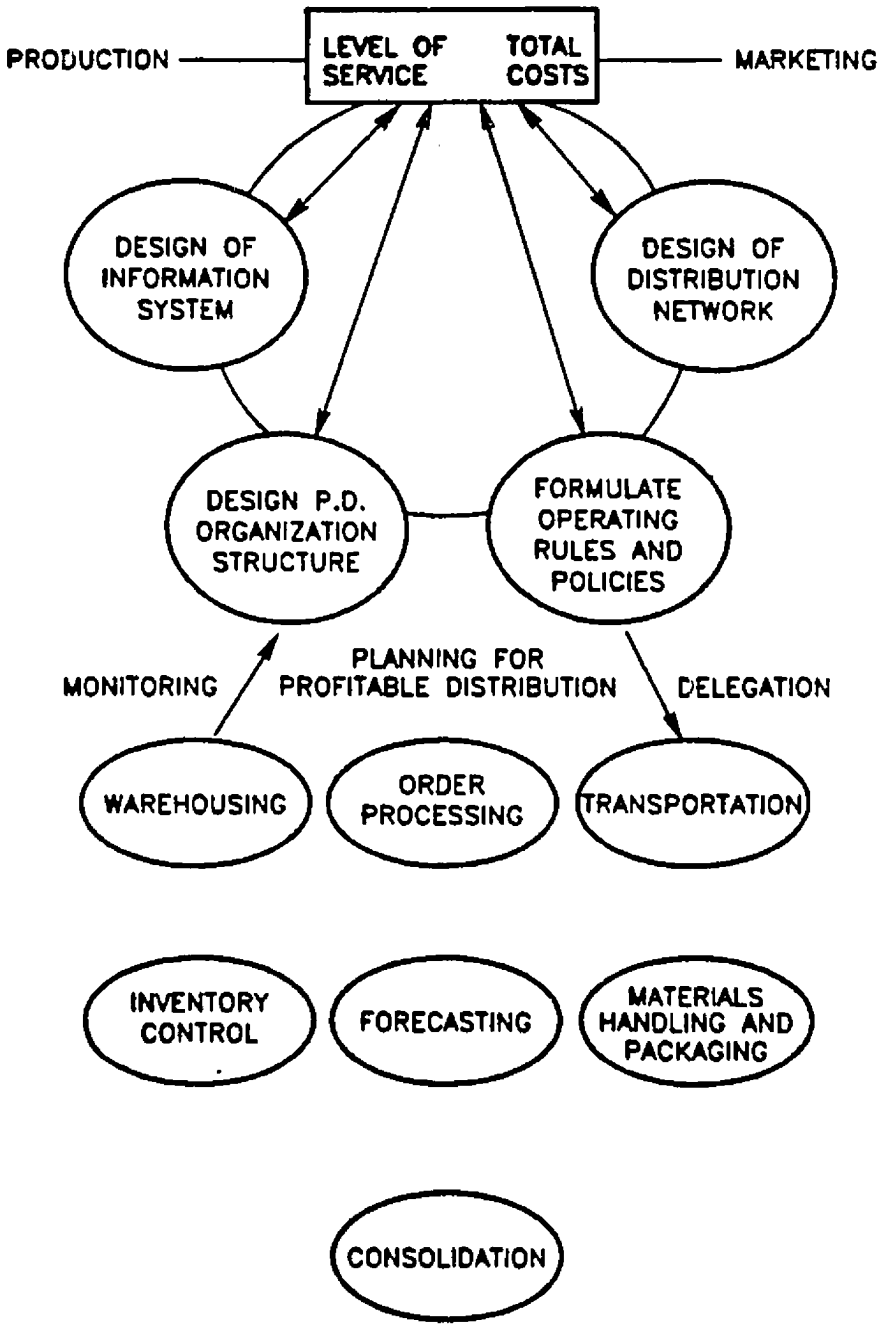
Logistics has been defined as "all organizational activities related to the movement and storage of items from the time of raw material acquisition to the point of final consumption" (Ballou [1992]). One function of a logistics system is the distribution of out–going items. Outbound shipment consolidation is the focus of our research.

The general goal of a physical distribution system is the attainment of acceptable levels of customer service and total cost. It was seen in our introductory example that shipment consolidation can make an important contribution toward the achievement of this goal.

Often, however, the impact of shipment consolidation on other areas of the distribution system is not fully recognized. Because shipment consolidation will influence and be influenced by most components of the distribution system, it must be regarded as a separate function contributing equally to the objectives of the entire system. Figure 1–1 presents our view of the position of shipment consolidation in the physical distribution system.

Only in the last ten years has shipment consolidation received serious academic attention. It is our feeling that much of this lack of research is due to two

Figure 1–1
Position of Shipment Consolidation In The Logistics System
(adapted from Firth et al. [1988])



reasons:

- viewing consolidation too narrowly; as a result, the problem is seen as too small or too simple; and/or
- considering consolidation to be "just" a part of another functional area, such as warehousing or vehicle routing; as a result, the problem is ignored, considered of lesser importance, or treated only in passing when dealing with another topic.

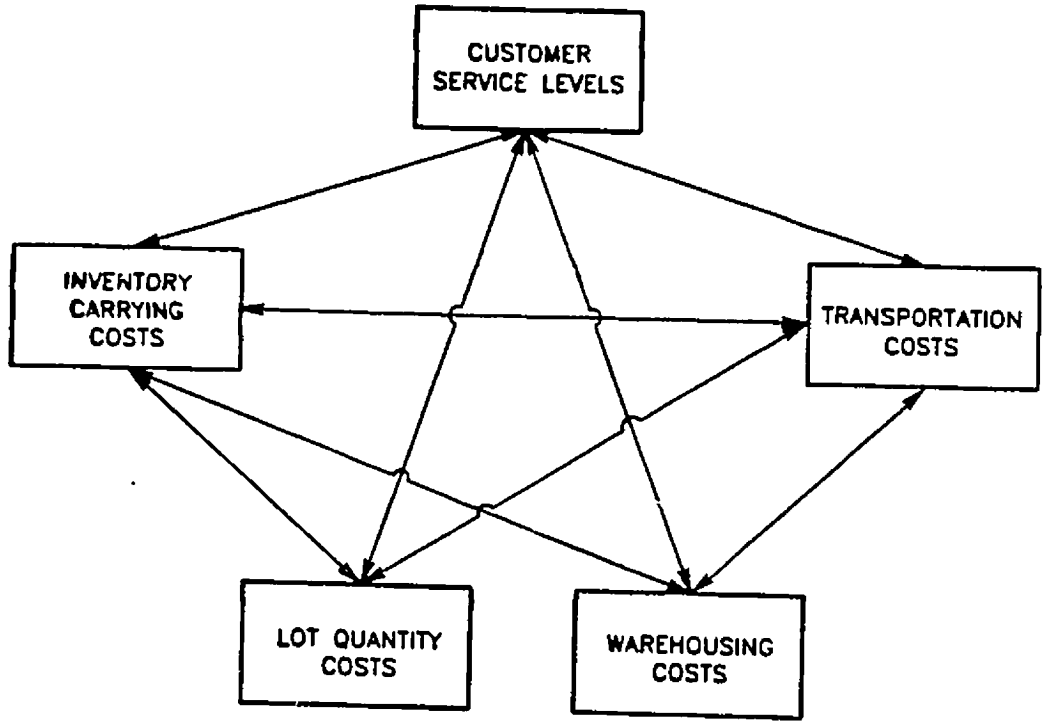
A proper definition of the role of shipment consolidation in the distribution system is needed. One goal of the present research is to examine the interaction of shipment consolidation with other components of the physical distribution system. Our research begins by recognizing that the six major costs of a logistics system are (Lambert [1976], Lalonde and Lambert [1977], Stock and Lambert [1987]):

a) transportation costs;
b) inventory–holding costs;
c) facility (ie., warehousing and terminal) costs;
d) lot–quantity costs;
e) order processing and information costs; and
f) cost of lost sales.

These costs are illustrated in Figure 1–2.

Lot–quantity costs are defined as costs that will change as a result of a change in the logistics system, such as production costs related to set–up, inspection, and scheduling, and costs of purchasing in various quantities (Lalonde and Lambert [1977]). These have been examined in the production management literature. Both lot–quantity costs and order processing and information costs (discussed by Lambert, Bennion, and Taylor [1987]) are relevant to consolidation when either the shipper or consignee can influence their occurrence by combining shipments. An example of this occurs when a consignee consolidates in–coming orders from a single origin to obtain

Figure 1–2
Costs In The Logistics System
(adapted from Lalonde and Lambert [1977])



quantity discounts and reduce the frequency of ordering. Our research, however, assumes that all customer orders are independent, and the shipper has no control over their size or timing. In the problem examined in this thesis, consolidation will not yield any incremental benefits or costs to the shipper from purchase discounts or less frequent set–ups, inspections, purchase discounts, or ordering. Thus, lot–quantity costs and order processing and information costs will not be included in our analysis.

The cost of lost sales is difficult to calculate because it should include both the contribution to profit forgone due to a stockout and the present value of future contributions lost from that customer. Our research treats customer service as an unquantified objective rather than as a cost, and considers it through such measures as mean time that orders are delayed.

The other three logistical costs imply that shipment consolidation can be viewed as being of three primary thrusts: transportation–based, inventory–based, and facility–based. Each of these classifications will have a direct impact on system objectives, consolidation procedures, and distribution costs. We define, for example, a transportation–based consolidation strategy as one that seeks to improve the efficiency, effectiveness, and opportunities for shipment consolidation by altering transportation–related variables, listed in Table 1–1, of the distribution system. Such a strategy attempts to recognize the inherent relationships and tradeoffs between the transportation and consolidation functions. Similar definitions for the other two consolidation classifications can be derived, and some issues faced in one of these classifications also will be relevant when considering one or both of the others.

Table 1–1
Logistical Variables Relevant to Shipment Consolidation

Transportation variables:

- frequency of shipment
- weight of shipment
- volume of shipment
- vehicle–dispatch and shipment–release rules
- mode of shipment
- classification of freight
- destination and distance of shipment
- vehicle route
- in–transit arrangements
- delivery and pick–up services
- delivery time
- delivery time variance

Inventory variables:

- inventory–holding cost
- purpose of inventory
- replenishment quantity
- replenishment frequency
- grouping of items for joint replenishment
- inventory management parameters

Facility/Warehouse/Terminal variables:

- number of facilities
- location of facilities
- function of facilities
- loading, unloading, and handling time
- holding or storage time

Note: not all variables will be applicable in a given shipment consolidation situation

Our research focuses on transportation–related variables. Our goals are:

- to examine transportation–related logistical variables that influence, and are influenced by, shipment consolidation;
- to study the impact of shipment–release policies on the effectiveness of a shipment consolidation program;
- to propose quantitative decision aids for determining optimal or near–optimal shipment–release parameters; and
- to suggest areas relating to shipment consolidation that warrant further study and research.

This thesis approaches these objectives as follows. The remainder of this chapter defines the physical distribution system configurations and types of shipment consolidation that are relevant to our research.

Chapter 2 provides a general background to shipment consolidation. A brief history of the strategy is given, and the various types of consolidation are discussed. The potential benefits and problems with shipment consolidation are outlined.

Chapter 3 reviews the related literature. We comment there that much of the literature is descriptive. Thus, although factors important to the success of a shipment consolidation program have been identified and studied, insight into the calculation of threshold values of many of these factors is lacking. Our research seeks to answer this shortcoming.

The remainder of the thesis focuses on shipment–release timing; that is, when should orders held as a consolidated load be dispatched? Chapter 4 discusses three common shipment–release policies, and compares their cost and customer service performances through a simulation model.

After a shipment–release policy has been selected, values for policy parameters must be determined. Chapter 5 discuss deterministic non–sequential approaches to this problem, while Chapter 6 proposes stochastic non–sequential methods. Chapters 7 and 8 examine sequential approaches. Lastly, topics relating to shipment consolidation that warrant further research are identified in Chapter 9.

Appendix A defines logistics terms used in this thesis, and our mathematical notation is listed in Appendix B. Appendix C provides an in–depth discussion on modeling shipper distribution costs.

Shipment consolidation can take many different forms, and can be carried out by the shipper, the consignee, the carrier, or by a third party such as freight forwarder or shippers' association. The remainder of this chapter defines the physical distribution system configurations that will be considered in our research.

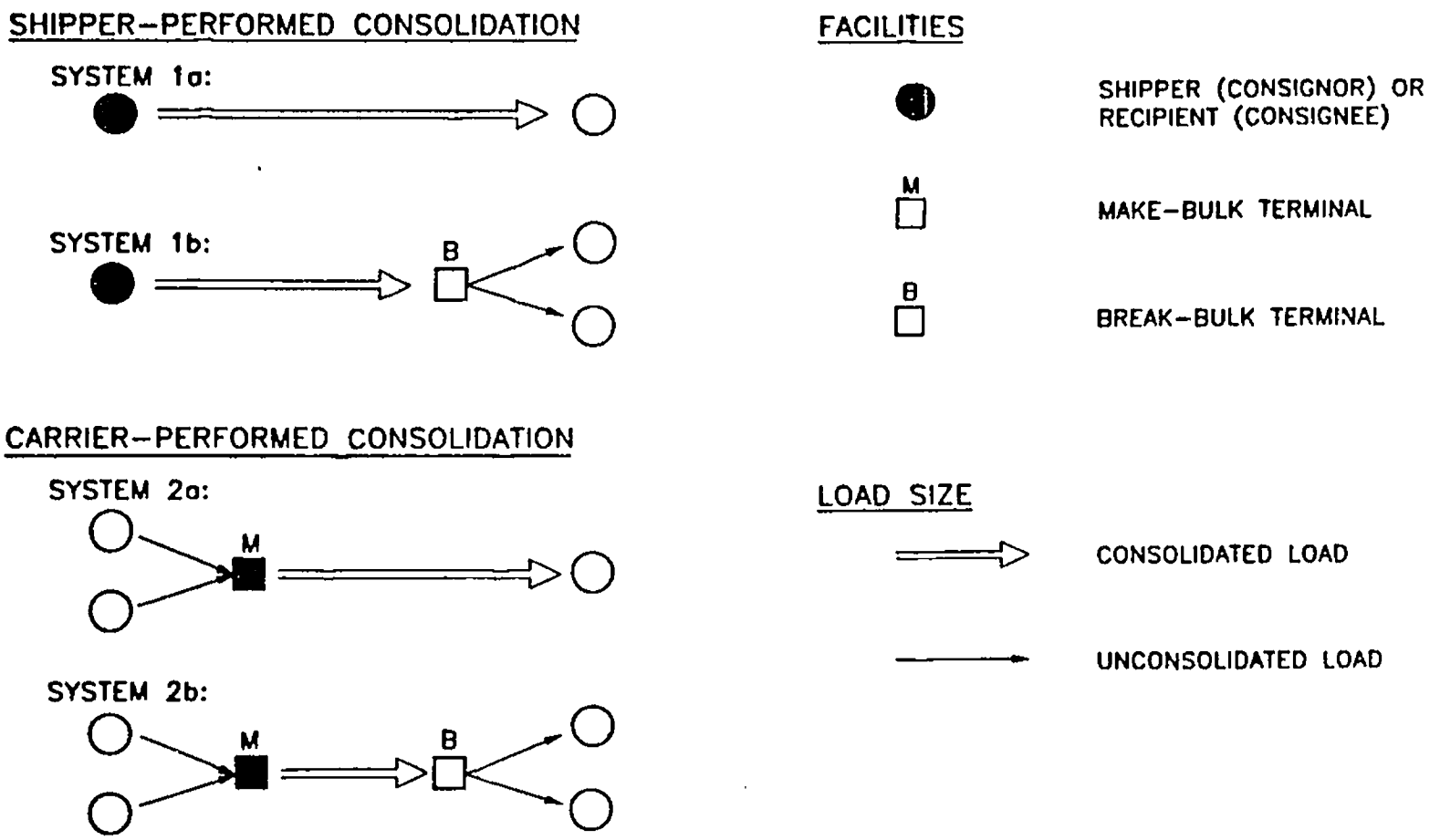
1.5 Physical Distribution System Configurations

Figure 1–3 illustrates four basic physical distribution system configurations. Other system designs are discussed by, for example, Sheahan [1982] and Tyworth, Cavinato, and Langley [1987].

Our four distribution system configurations can be classified by the party performing the shipment consolidation function and the form of carriage employed. Consequently, we can define six shipment consolidation strategies, summarized and discussed below:

- Systems 1a/P and 1b/P: shipper–performed consolidation using private carriage;
- Systems 1a/C and 1b/C: shipper–performed consolidation using common carriage; and

Figure 1-3
Physical Distribution System Configurations



- Systems 2a and 2b: carrier–performed consolidation using common carriage.

		party performing carriage	
		shipper	carrier
party performing shipment consolidation	shipper	Systems 1a/P and 1b/P	Systems 1a/C and 1b/C
	carrier	not available	Systems 2a and 2b

Shipper-performed consolidation: Under these systems, shipment consolidation is performed at the shipper's premises. The consolidated load moves either directly to the consignee (Systems 1a/P and 1a/C) or to a break–bulk terminal located near the consignee(s) for sorting and local delivery (System 1b/P and 1b/C). The choice of System 1a over System 1b will largely depend on the volume of shipments destined to each consignee.

Carrier-performed consolidation: Here, the carrier picks up individual shipments, then consolidates these into larger loads at a local make–bulk facility. Depending on customer shipment volume, this consolidated load is transported by the carrier either directly to the consignee (System 2a) or to a break–bulk terminal to be reloaded into smaller vehicles for local delivery (System 2b).

This thesis examines shipper–performed consolidation. Although many of the models we develop can be applied without major modifications to carrier–performed consolidation, specific problem characteristics can be very different. For example, a common carrier will have a wider variety of shipments, received from a greater number

of origins and going to a larger set of destinations. As well, the common carrier typically will have better access to break–bulk and make–bulk consolidation terminals. Lastly, the cost structures will be different; for example, inventory–holding cost would be irrelevant to a common carrier. These factors result in a much broader problem than studied in this thesis, thus we do not examine carrier–performed consolidation.

Our research considers only the outbound linehaul stage, illustrated in Figure 1–3 by double–line arrows. Local pickup and delivery stages (single–line arrows in Figure 1–3) are not considered, and extension of methods discussed in this thesis to the coordination of consolidation policies for the pickup, linehaul, and delivery stages is proposed as later research.

As well, our analytical work in later chapters is limited to transportation and inventory–holding costs. Incorporating costs related to shipment–handling and other distribution functions would not be difficult, but would require a more stringent definition of the distribution system and generally would not change our methods or conclusions. Modeling of physical distribution costs is discussed in Appendix C.

Chapter 2
BACKGROUND TO SHIPMENT CONSOLIDATION

2.1 Introduction

This chapter provides a background to shipment consolidation. We first give a brief history of the strategy, then outline the various types of consolidation. Lastly, we discuss its potential benefits and problems.

2.2 Historical Background to Shipment Consolidation

Shippers and carriers have long recognized the potential benefits of shipment consolidation. As early as 1920, concern over the high costs and inefficiencies of handling less–than–carload and less–than–truckload freight led to the introduction of innovative carrier services, such as containerization and rail piggyback, priced lower to encourage consolidation by shippers. However, due to the depressed economic climate, a lack of commitment by railroads, and rate regulation that reduced the ability of carriers to offer rate discounts, most of these services did not survive past the mid–1930's.

Even before the First World War, shipment consolidation was practised through rail "in–transit arrangements". These services allowed a shipper to stop a partially loaded boxcar at a point between origin and destination to finish loading the car. Similarly, outbound consolidation could be accomplished by sending a full car toward a final destination, but stopping it enroute to allow consignees at intermediate points to remove portions of the load. Under both strategies, the entire shipment travelled under one volume freight rate (plus a fixed charge per stop) rather than at the sum of

two non–volume rates. Although less important because of deregulation of the transportation industry, common carrier in–transit arrangements are still available today, having been expanded to encompass most products carried by rail or motor freight.

During the 1950's and 1960's, greater population, competition, and demand and supply led to an increased interest in controlling distribution costs. In 1966, Sutton identified shipment consolidation as a valuable method for reducing transportation costs, and encouraged the implementation of ongoing inbound and outbound consolidation programs. Nevertheless, Newbourne and Barrett noted in 1972 that many major shippers did not consolidate outbound shipments simply because all terms of sale were *F.O.B. origin*.

Since the 1950's, there has been a trend in industry toward smaller order quantities, best seen in the Just–in–Time concept of the 1980's. Unfortunately, costs of transportation, whether by common or private carrier, encourage large volume shipments and, in the extreme, result in the embargo of very small shipments. This has contributed greatly to the so–called "small order problem": moving small orders through the distribution system is both costly and inefficient (Lambert, Bennion, and Taylor [1987]).

Greater awareness of logistics costs, especially those related to the small shipments required by a JIT system, has led to the realization that shipment consolidation can be a valuable strategy for controlling distribution costs while maintaining good customer service. In 1983, for example, the Ford Motor Company inaugurated a "pool–car" consolidation system by which Ford's suppliers ship

components to seven central locations for consolidation and transportation by rail to Ford plants. Ford reported savings of $4 million per year from reduced transportation time, in–transit inventory, inventory carrying costs, and receiving–area congestion. A similar system was implemented by General Motors for consolidating Just–In–Time shipments from suppliers in the Detroit area to their plant in Flint, Michigan. Referring to this program, Ansari and Heckel [1987] commented that "the density of suppliers in the Detroit area apparently minimizes the identification and coordination problems associated with implementation of the consolidation concept". In 1984, the National Starch and Chemical Corporation designated three employees as "consolidation coordinators" to identify opportunities for shipment consolidation. Savings were estimated to be in excess of $1 million per year.

A further impact on shipment consolidation resulted from the deregulation of the Canadian and American transportation industries in the 1980's. Deregulation has given shippers the ability to negotiate freight rates with common carriers rather than having to accept fixed tariffs. As a result, some authors (for example, Ackerman [1990]) have claimed that shipment consolidation is irrelevant because cost savings from consolidation can be obtained through negotiation between shipper and carrier. This statement is inaccurate for several reasons. Besides ignoring private carriage, it fails to recognize that the larger loads resulting from consolidation will reduce common carrier costs per ton–mile, thus giving the shipper additional negotiating leverage and greater savings to be gained from negotiation. Moulton [1990] has noted that "consolidation programs, at least in central Canada, are still a sound and viable

marketing strategy despite the unregulated environment"[2]. Thus, transportation deregulation has not made consolidation irrelevant, but has changed the focus of consolidation from a "stand–alone" program for reducing transportation costs to an important component of an integrated distribution strategy for achieving cost and customer service goals.

Other reasons for the growing importance of shipment consolidation are listed in Table 2–1.

2.3 Types of Shipment Consolidation

A general classification of the types of shipment consolidation is presented in Figure 2–1. Boundaries between categories are not exact, and this diagram does not represent the only possible schema.

Bowersox, Closs, and Helferich [1986] define **facility consolidation** as the rationalization of warehouses, terminals, and supply points to improve inventory availability or reduce carrying costs with little or no change in aggregate stock levels. Facility consolidation influences and is influenced by factors (such as location and number of facilities) relevant to shipment consolidation, but it does not imply or require the use of shipment consolidation.

We define **shipment consolidation** as the active intervention by management to combine several small shipments into fewer, larger loads to achieve operating efficiencies, reduce logistical costs, and improve customer service. Cooper [1983] and

---

[2] Transportation deregulation has had a greater impact on common carrier rates in the United States than in Canada. Deregulation occurred in the U.S. earlier than in Canada; as well, rate regulation always has been less in Canada than in the U.S.

Table 2–1
Reasons for Increased Interest in Shipment Consolidation by Shippers

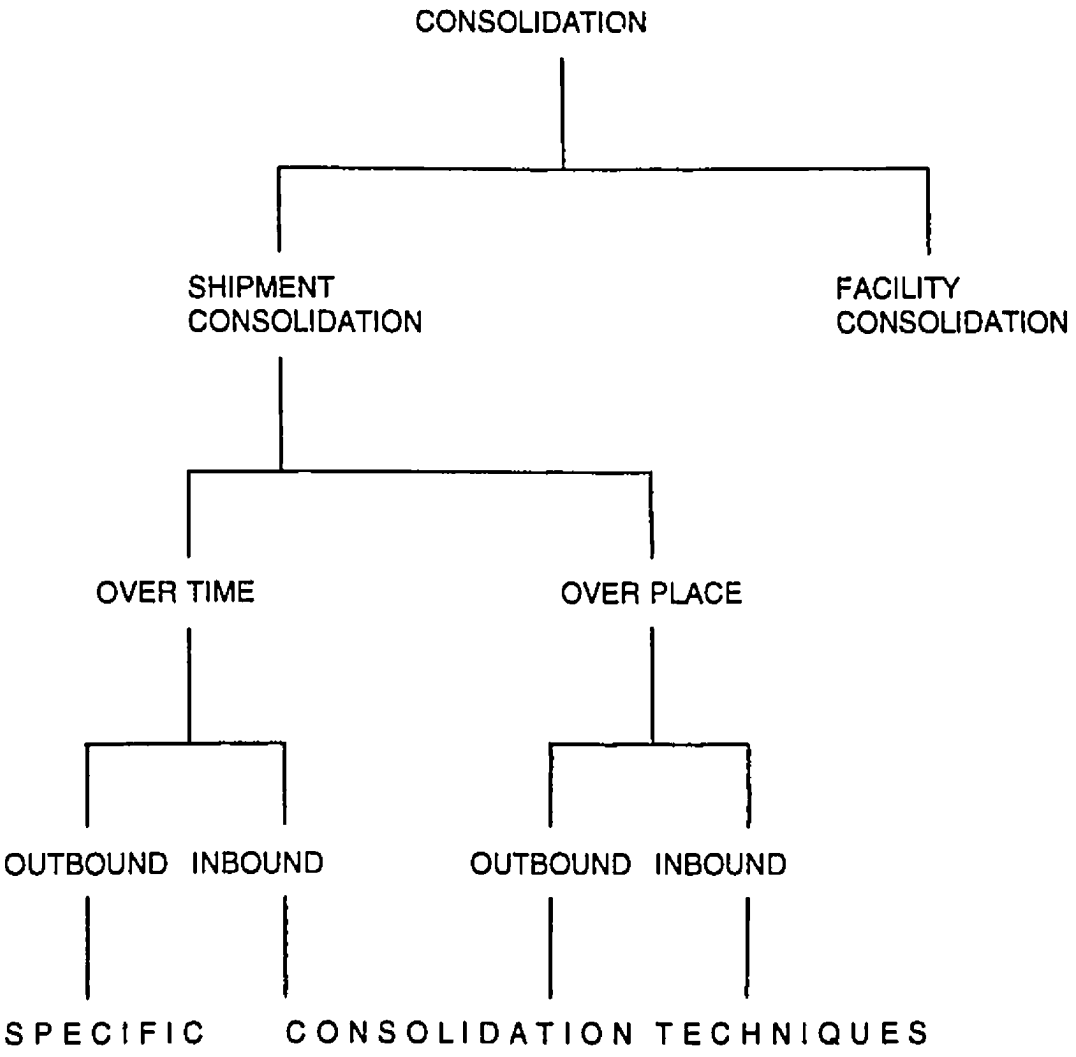
Transportation reasons:

- growing spread between volume and non–volume rates, encouraging large quantity shipments
- dramatic escalation in both volume (ie., large quantity) and non–volume (ie., small quantity) freight rates, forcing closer monitoring and control of distribution costs
- greater carrier ability, due to deregulation, to offer innovative transportation and consolidation services
- greater management awareness of the costs of transportation and the availability of carrier services

Business reasons:

- desire by shippers and manufacturers to serve markets that cannot be reached economically by any other means
- competitive necessity

Figure 2–1
Types of Shipment Consolidation



Min and Cooper [1990] refer to shipment consolidation as "freight consolidation".

Many authors have noted that shipments can be consolidated over time or over place. **Consolidation over time** refers to the accumulation of items produced, ordered, or used at different times, at a common location, for eventual shipment to the purchaser as a single load. Brennan [1981] calls this strategy "temporal consolidation", while Hall [1987] titles it "inventory consolidation".

Consolidation over place refers to the aggregation of items in a common location for immediate or near–immediate shipment as a consolidated load. Hall [1987] divides consolidation over place into "vehicle consolidation" (use of a single vehicle for pick–up and/or delivery of items at various points along its route) and "terminal consolidation" (collection of items of different sizes, orders, or origins at a single location for sorting, combining, and loading onto vehicles for delivery to different locations).

Shipment consolidation can be viewed as outbound or inbound. The major difference is the degree of control that the consignee exercises over the consolidation of inbound shipments. Control over demand for items in a shipment is not usually present in outbound consolidation, and motivates the consignee to alter the size or timing of purchases to coordinate shipments. However, if consolidation is performed at the discretion of a consignor or carrier, outbound consolidation exists but inbound consolidation does not, even for goods transported under *FOB origin* terms. Because the consignee can exercise control in inbound consolidation but typically not in outbound consolidation, we disagree with Schuldenfrei and Shapiro's [1980] comment that "the inbound problem is really the outbound problem in reverse".

Each form of shipment consolidation involves specific approaches. Temporal consolidation techniques aimed at increasing the probability that vehicles are loaded to capacity include holding shipments until a minimum weight is obtained and limiting delivery to specific markets to certain days of the week. Pooled distribution, which uses a third party to consolidate shipments from and to many different organizations, is an example of consolidation over place. Weart [1984] suggests seven opportunities for shipment consolidation, while Tyworth, Cavinato, and Langley [1987] give ten. Sheffi [1986] lists six consolidation techniques, however Min and Cooper [1990] feel that these cannot be considered true classifications because some may occur simultaneously.

Other types of shipment consolidation have been proposed. Brennan [1981] considers **spatial consolidation** to be concerned with determining which facilities or routes are to be aggregated for purposes of performing shipment consolidation. This appears to be an extension of the Bowersox et al. [1986] definition of facility consolidation to include consolidation over place. As well, Brennan defines **product consolidation** to deal with deciding which items destined for a specific customer should go through a terminal and which should be transported directly.

Newbourne and Barrett [1972] divide outbound shipment consolidation into **pooling** and **consolidation proper**. Pooling refers to the aggregation into larger loads of shipments so small that it would not be economically feasible to transport, directly, any of them individually. Consolidation, on–the–other–hand, deals with shipments that, if not consolidated with others, could still be shipped directly but would not qualify for volume rates. Thus, the objective of consolidation proper is to transform

smaller shipments into those large enough to qualify for carload or truckload rates, whereas pooling increases the size of combined loads to achieve more efficient distribution, however under non–volume rates.

Newbourne and Barret [1972] also distinguish between **dock consolidation** and **order consolidation**. Dock consolidation refers to consolidation done at point of shipment. It is a reactive strategy: all shipments in the shipping area or loading docks are held for, at most, a few hours while personnel decide which shipments can be consolidated at that time. Order consolidation is proactive. Here, information regarding orders ready for shipment and forecasts of those expected in the future are used to develop specific consolidation lists. Although some orders may be held a few days to await consolidation, only those to be consolidated are delayed.

Lastly, it is important to recognize that shipment consolidation can be carried out by the consignor or consignee, by the carrier, or by a third party such as freight forwarder or shippers' association. Benefits, costs, and other variables will be viewed, analyzed, or modeled differently under each approach. Indeed, we feel that the greatest source of confusion to persons new to the shipment consolidation literature stems from the absence of a clear indication of the party performing the consolidation.

2.4 Benefits and Disadvantages of Shipment Consolidation

Benefits of Shipment Consolidation

Respondents to a survey conducted by Jackson [1985] felt that reduced transportation cost was the major reason for consolidating shipments. This and other advantages resulting from shipment consolidation are summarized in Table 2–2 and

Table 2–2
Potential Benefits and Disadvantages of Shipment Consolidation

Potential benefits:

- reduced per–unit transportation costs resulting from:
  –spreading of per–shipment fixed transportation costs over larger load sizes
  –decreasing per–unit freight rates as load weight increases
- improved customer service due to:
  –reduced handling of goods, resulting in less chance of damage, loss, or pilferage
  –decreased transportation time because of more direct delivery on dedicated vehicles
- improved utilization of transportation equipment and employees
- increased availability of goods due to reduced transportation charges

Potential disadvantages:

- longer, possibly more erratic, order cycle time due to delaying shipments for consolidation
- increased inventory levels and costs:
  –purchaser must hold additional safety stocks to guard against a possibly longer order cycle
  –shipper must hold items until an efficient consolidated load is available or until customer service considerations force shipment release
- potentially higher freight charges and administrative costs through improper management
- distribution system changes, such as new delivery routes and/or consolidation facilities

discussed below.

Reduced costs of private carriage due to spreading of fixed transportation charges: Appendix C, which discusses the modeling of distribution costs, notes that total transportation cost for private carriage depends mainly on distance or time. Thus, for a given distance, most of the transportation cost is fixed whether the vehicle is full or empty. Shipping fewer but larger loads through consolidation results in reduced total and per–unit fixed transportation costs.

Reduced costs due to decreasing common carrier freight rates as load weight increases: As noted earlier, consolidating several non–volume shipments may increase the weight of a load to a point above the common carrier minimum volume weight, thus entitling the shipper to lower volume rates. Newbourne and Barrett [1972] remark that the average less–than–truckload freight rate is approximately twice the corresponding truckload rate. We applied statistical analysis to a combined random sample of 164 freight rates from the U.S. Rail Uniform Freight Classification and the U.S. Motor Carrier Freight Classification, and found that the mean carload/truckload rate was approximately 60% of the mean LCL/LTL rate.

If the minimum volume weight is not attained, a non–volume load still may be shipped under volume rates by applying the "bumping clause". This option allows the shipper to declare heavier weights than actually exist to push the weight of the load into a higher weight bracket, thus qualifying for a lower freight rate. In effect, the shipper is paying to ship non–existent items in return for a lower total freight charge, which is why this practice is sometimes referred to as "phantom freight" or an "over–

declared shipment". The impact of the bumping clause has been ignored by most analytical research in physical distribution; Russell and Krajewski [1992] is one exception. We will apply this concept throughout the thesis.

Lastly, shippers who can generate frequent volume shipments can usually negotiate for discount rates below the non–volume rates. Conversely, loads below certain weights often cannot be transported economically even at non–volume rates, and must move at higher rates, usually on a per–shipment basis rather than on a weight basis.

Improved equipment utilization by carriers: Both private and common carriers handling consolidated shipments benefit from improved utilization of transportation equipment and employees because of the larger load sizes. As well, they face less variability in costs than if handling only non–volume shipments (Pollock [1978]).

More direct delivery on dedicated vehicles: Under both private and common carriage, shipments consolidated by the shipper can go directly from their origin to their destination. If, however, a shipper tenders a small load to a common carrier, the carrier will consolidate it with other small shipments from other consignors. This load will not be delivered directly to the consignee, but will be transported to a local terminal for sorting and reloading on delivery vehicles. The result is increased transportation time and less shipper control. Conversely, by consolidating shipments into larger loads, a shipper can: i) reduce transportation time; ii) eliminate some pickup, delivery, and terminal costs; iii) increase his control over the transporting of his shipments; and iv) improve his position when negotiating with carriers.

Reduced handling of goods: A consolidated load will experience less handling. This results in less chance of damage, loss, or pilferage, easier tracing of shipments, and reduced administrative work relating to claims.

Public benefit from increased availability of goods: Lastly, the U.S. Interstate Commerce Commission (I.C.C.) noted that the public benefits from shipment consolidation because "many articles, which formerly could not move because the higher rates resulted in sale prices that the public would not pay" now can be transported economically as a consolidated load.

Disadvantages of Shipment Consolidation

Disadvantages of shipment consolidation are listed in Table 2–2. The most commonly mentioned problems are:

- longer and/or more erratic order cycle length;
- increased inventory levels and costs; and
- increased administrative work and costs.

Longer and/or more erratic order cycle length: Shipment consolidation generally requires that some items be held until a maximum age or minimum weight or volume has been reached. This additional holding time may exceed the reduction in transportation time from shipping a consolidated load. The possibly longer and more uncertain order cycle is seen by practitioners as the major disadvantage of consolidation (Jackson [1985]). Newbourne and Barrett [1972], however, claim that the holding delay partially will be offset by savings in transportation time, resulting in a net reduction in total order cycle length.

Increased inventory levels and costs: Shipment consolidation increases inventory levels and holding costs. The consignor must delay shipments to create effective consolidated loads, while the consignee must hold larger safety stocks to compensate for a possibly longer or more uncertain order cycle. Moreover, consolidation presupposes that there is space available for holding those orders being accumulated.

Increased administrative work and costs: Effective generation of consolidation combinations requires knowledge of customer orders presently waiting for shipment and those expected within the next few days. This increases administrative time, work, and complexity, and requires close coordination between order processing, inventory–control, and transportation.

Newbourne and Barrett noted that a change in billing method may be required with common carrier because *freight-collect* terms cannot be used if a consolidated load is destined for two or more consignees. In this case, the shipper must prepay the freight charges and bill the consignees for their portions. It also may be difficult to obtain the necessary cooperation between departments involved in the consolidation program (Weart [1984]).

Potentially higher shipment-handling costs: If consolidation involves large numbers of small quantity orders, shipment–handling costs may increase due to frequent pickup, delivery, and dock and terminal handling. Newbourne and Barrett [1972] contend that this additional handling will be reflected in increased worker productivity, rather than in increased cost, if a dock or terminal was not previously being operated at peak efficiency.

Potentially higher freight charges through improper management: Under certain circumstances, shipment consolidation may result in higher per unit freight charges than would have occurred without consolidation. For example, if a consolidated load consists of many different types of items, the shipper may choose to use "freight, all kinds" (FAK) rates. Because FAK rates are, in theory, the long–run average rate applicable if large quantities of low–rate freight and high–rate freight were combined, FAK rates penalize items that would move under low freight rates if shipped alone. In this case, it may be cheaper to consolidate low–rate items and high–rate items separately. Also, if consolidation causes considerable delay in the loading or unloading of common carrier vehicles, demurrage or detention charges may accrue.

Other considerations: Lastly, shipment consolidation may force changes to the distribution system, such as the realignment of vehicle routes and the establishment of new facilities for the collection and handling of shipments.

The benefits and problems of shipment consolidation have been recognized by many authors, and tradeoffs between these advantages and disadvantages have been investigated in the literature. The next chapter reviews this literature.

Chapter 3
SURVEY OF SHIPMENT CONSOLIDATION LITERATURE

3.1 Introduction

Only since the early 1980's has shipment consolidation received much academic attention. Ballou [1976] suggested that development of mathematical solution methods for problems in shipment consolidation had been hampered by a lack of understanding of regulated transportation. Cooper [1983] also noted that research in consolidation across time and customers was lacking.

Hall [1987] has categorized shipment consolidation literature as four types:

- articles, usually in the trade press, that describe the advantages of consolidation;
- general discussions in textbooks of consolidation and the resulting reductions in transportation cost;
- case studies outlining how a specific organization carries out the consolidation function; and
- descriptions of computer simulation models designed to study the relationship between various components of a hypothetical consolidation program.

It is our experience that most academic literature on shipment consolidation falls into one of two areas:

- analytical discussions of the impact of shipment consolidation on costs (typically transportation and inventory–holding) and cost tradeoffs; or
- simulation studies investigating the relationships between variables, usually customer service and cost, of a consolidation strategy.

This chapter presents a review of the shipment consolidation literature.

3.2 Survey Papers

A detailed survey of research in shipment consolidation prior to 1980 was prepared by Jackson [1980]. Unfortunately, this work is not readily available.

Min and Cooper [1990] reviewed 26 analytical studies appearing in the academic press after 1980. They classify each work using Brennan's consolidation taxonomy ("temporal", "product", or "spatial") and identify the type of analysis (e.g., stochastic, deterministic, optimization, simulation) used. Strengths and weaknesses of each paper are noted, and general suggestions for further research are made.

Min and Cooper's article is a good starting point to academic research on shipment consolidation, and most of the papers they reference are readily accessible. Unfortunately, their survey appears more concerned with classifying, rather than reviewing, the literature. Although they present a methodological classification of conceptual and mathematical approaches, the actual reviews are very brief, a situation made worse by the misclassification of some papers. Moreover, ignoring articles in the trade press is an unfortunate omission that limits the value of their work as an general introduction to the subject.

3.3 Analytical Studies

Studies focusing on potential cost savings from shipment consolidation are not uncommon. Distribution textbooks often include illustrations of the effect of consolidation on transportation costs; see, for example, Ballou [1992] and Tyworth, Cavinato, and Langley [1987].

To compare the effect on costs of shipment consolidation and direct less–than–truckload shipping, Shuster [1979] developed a model of three shippers and one motor carrier using actual shipper costs and carrier rates. Consolidation was shown to reduce both shipper and carrier costs, and to increase the productivity of pickup, delivery, and shipment–handling. As well, carrier gross margins improved by consolidating less–than–truckload shipments that otherwise would have been transported individually at a loss due to higher pick–up, delivery, and handling costs.

Burns, Hall, Blumenfeld, and Daganzo [1985] compared direct shipping from one supplier to many destinations with "peddling", a form of vehicle consolidation where a vehicle delivers to multiple destinations during each delivery run. They concluded that whereas the optimal shipment quantity for direct shipping may or may not equal vehicle capacity, total peddling cost will be minimized when vehicles are dispatched full. As a result, the cost advantage of vehicle consolidation over direct shipping increases with vehicle capacity unless the distance is such that the optimal direct shipment size equals or exceeds capacity. Schwarz [1989] also compared the effectiveness of direct shipping and consolidation strategies.

Blumenfeld, Burns, Diltz, and Daganzo [1985] analyzed the tradeoffs between transportation, inventory, and production set–up costs when shipping direct or via a consolidation terminal. They concluded that when transportation and production are not coordinated and inbound and outbound consolidation are independent, each link of the distribution network (origin to terminal, and terminal to destination) can be treated independently with respect to optimal shipment size and minimum cost. As a result, the least cost strategy can be found by minimizing cost on each link separately.

Hall [1987] noted that both the number of stops per vehicle route and the frequency of delivery can increase the number of items available for consolidation. By modifying the economic order quantity model to include local delivery distance and the cost of making a stop, he concluded that the optimal number of stops and the time between vehicle dispatches depends on the ratio of inventory–holding cost to stop cost. However, delivery vehicles always should be loaded to capacity to maximize the potential number of stops per load. Hall also discusses routing and terminal strategies used in outbound shipment consolidation systems.

Daganzo [1987, 1988a] analyzed vehicle consolidation where collections were made from one or more origins before delivery to multiple destinations. He developed an analytical expression for the minimum cost of such a strategy given the inventory–holding cost, vehicle operating cost, and approximate average distance between each origin–destination pair. The number of break–bulk terminals and the number of vehicle stops were most influenced by the number of origins and destinations and a calculated value reflecting the relative cost "goodness" of a vehicle consolidation tour. The value of this latter parameter is based on per–item inventory and transportation costs, assuming a full vehicle, and the additional distance that must be traveled because more than one stop is required. Daganzo [1990] considers the synchronization at a transportation terminal of scheduled inbound and outbound movements, while Daganzo [1991] examines shipment consolidation under various distribution system designs.

Inbound shipment consolidation was studied by Russell and Krajewski [1991]. Their model includes common carrier volume discounts and supplier quantity

discounts, and evaluates the tradeoffs via a mixed–integer linear program. Quantity discounts and transportation rate weight–breaks are shown to affect optimal inbound consolidation policies. For example, the lack of quantity discounts leads to less consolidation and more products ordered singly. When order weights are small, quantity discounts dominate transportation discounts because minimum shipment weights are difficult to attain; if product weights are high, transportation discounts encourage larger order quantities covering more items.

3.4 Simulation Studies

Masters [1980] studied the effects of "order structure variables" (annual throughput volume and mean order weight) and "consolidation decision variables" (system objective, number of consolidation points, and maximum holding time) on transportation cost, mean delivery time, and delivery time variance. He concluded that consolidation reduced transportation costs and increased mean delivery time, but did not necessarily increase delivery time variance. He also noted that order structure variables influenced consolidation system performance more than did consolidation decision variables, although maximum holding time was of some importance.

Jackson [1981] examined the effect on average order cost, mean order cycle time, and variance of order cycle time from varying the number of consolidation points, maximum holding time, and shipment–release strategy. He found that the most important variable, in terms of cost and mean and variance of order cycle time, was volume of orders. Low–volume systems tended to experience costs similar to those of direct less–than–truckload shipping because orders could not be held sufficiently

long to qualify for volume discounts. Low–volume systems also suffered greater cycle time and variance due to the delay for possible consolidation. Thus, the number of consolidation points, maximum holding time, and shipment release strategy should be set to ensure a sufficient volume of orders for consolidation.

To compare the impact on distribution costs and transportation time from direct shipping versus warehousing, Cooper [1983, 1984] simulated four distribution system configurations: a) direct shipment from plant to customers; b) direct shipment from warehouses to customers; c) consolidated shipment from plant to consolidation points (break–bulk terminals); and d) consolidated shipment from warehouses to customers. As with Masters [1980] and Jackson [1981], Cooper found that consolidation reduced total cost and increased mean order cycle time over that of direct shipments, but did not necessarily result in greater order cycle time variance. Annual system volume, mean order size, and shipping cost per hundredweight had the greatest impact on total cost and on location of warehouses and consolidation points. In turn, location had the greatest effect on delivery time.

Buffa [1986b] recognized that inbound shipment consolidation must include consideration of inventory–grouping methods. He compared the total cost of receiving shipments individually to that of consolidating inbound orders with common review times. His model included consolidation costs from shipment–handling and storage, and adjustments to item prices to reflect ordering in groups. He concluded that although consolidation costs and some inventory costs increased, inbound consolidation of shipments from vendors located in the same region can be cost–effective through use of an appropriate order–grouping technique. Order–grouping

techniques in inbound consolidation are examined by Buffa [1986a, 1987], Buffa and Munn [1989], and Russell and Krajewski [1991]. Inbound consolidation also is discussed by Myers, Fanelli, and Boger [1987].

Buffa [1987] investigated the transportation time and cost factors that contribute most significantly toward the success of an inbound consolidation strategy. He found that four factors (freight rate per mile, minimum volume weight, difference between volume and non–volume rates, and penalty associated with delays in average transportation time) played particularly important roles in the cost effectiveness of an inbound consolidation strategy. He also noted that cost savings from inbound consolidation are positively affected by shipping distance, as well as by demand, weight, and value of the items shipped.

Buffa [1988] compared inbound consolidation with an "independent inventory strategy", which ignores transportation factors when setting order quantities, and an "inventory–transport strategy", under which order points and quantities depend on both transportation and inventory factors. Using order characteristics based on empirical data and an order–grouping technique suggested by Chakravarty [1981], Buffa found that an inbound consolidation resulted in higher order costs, lower but more variable transportation costs, and lower holding costs.

Closs and Cook [1987] commented that most consolidation research assumed a single consolidation point and ignored flat–rate freight charges resulting when a load is less than a specified weight. They simulated a three–echelon distribution system with minimum transportation charges and shipment dispatch based on time and accumulated quantity. Data from an electronics manufacturer was used to test three

scenarios: a) the company's current consolidation strategy of daily shipments, with loads of less than 50 pounds sent by courier; b) an increase in the number of consolidation centres within the three–echelon system; and c) the use of outbound consolidation points between the factory and the present consolidation centres (thus increasing the number of echelons to four). Their model included transportation, order–processing, and shipment–handling costs, but not inventory–holding costs.

Closs and Cook's model is important because it recognizes interactions between echelons of the distribution system. They noted that small volumes in the collection and delivery stages frequently may result in flat–rate transportation charges that overwhelm savings from consolidated linehaul. Thus, system throughput volume may restrict the number of consolidation levels that is cost–justifiable. Because minimum freight charges were found to have a major impact, Closs and Cook suggest reevaluating consolidation policies whenever there are changes in the freight rate differential between shipment sizes.

Bagchi [1988] examined the effectiveness of inbound shipment consolidation through regional make–bulk centres in a Just–in–Time environment. Purchase cost and delivery time were compared under normal procurement procedures and under a Just–In–Time (JIT) system, where JIT distribution was simulated by reducing the weight of each shipment and increasing the frequency of consolidation. Delivery time was found to decrease when many vendors in an area were served by a consolidation terminal. Bagchi concluded that shipment consolidation was an effective strategy in a Just–In–Time environment if there was adequate order volume, vendor concentration, or delivery frequency.

Ha, Khasnabis, and Jackson [1988] analyzed the effect of special delivery requirements (i.e., priority shipping) and number of consolidation points. They found that special delivery requirements, which forced some shipments to be made before the stated holding time had elapsed, resulted in greater transportation and inventory–holding cost and mean delivery time, but reduced delivery time variance. The number of consolidation points influenced both mean delivery time and delivery time variance, but it affected neither transportation cost nor total cost at a statistically significant level.

3.5 Other Shipment Consolidation Literature

Jackson [1985] reviewed the benefits from and reasons for shipment consolidation via a survey of 52 firms in various industries. His most important findings are related to shipment–dispatch rules; these are discussed in Chapter 4. As well, he noted that customer service demands and insufficient throughput volume were the most commonly–cited reasons for not consolidating shipments.

Probably the most comprehensive discussion of the operational aspects of shipment consolidation is a series of five articles by Newbourne and Barrett [1972] and a later book by Newbourne [1976]. Both are invaluable to any study of how shipment consolidation is carried out in practice.

Pollock [1978] and Sheahan [1982] focus on general cost and customer service aspects of the strategy. Shelley [1982] proposes a method, based on percentage of total load and estimated number of delivery stops, for costing the individual shipments that make up a consolidated load. Weart [1984] provides brief reviews of three commercially–available software packages to help users in developing consolidated

shipment lists. Bookbinder and Barkhouse [1992] discuss a logistics information system for coordinating inbound and outbound consolidated shipments.

We add that some authors have studied shipment consolidation indirectly. Several papers by Powell (for example, Powell [1985]; Powell and Humblet [1986]) apply bulk–queue analysis to passenger vehicle holding strategies (where vehicles are held until the load is sufficiently large) and cancellation strategies (where a scheduled vehicle is cancelled because the load is too small). Some classical stochastic models bear resemblance to the dispatch process in shipment consolidation; examples of these models include the machine replacement problem and the level–crossing problem. Lastly, research on vehicle routing implicitly assumes shipment consolidation; explicit recognition and analysis of this is a topic that warrants further research.

3.6 Shipment Consolidation Literature: In Retrospect

Research on shipment consolidation in the academic literature has yielded several results:

- The practice, benefits, problems, and complexities of shipment consolidation now are receiving overdue recognition.
- Variables important to the consolidation process have been identified; see Table 1–1. As well, factors important to the success of a shipment consolidation program have been identified; see Table 3–1.
- Given these process variables and success factors, we can conclude that three conditions must exist before a successful shipment consolidation program can be implemented; see Table 3–2.

Table 3–1
Factors Important to the Success of
A Shipment Consolidation Program

Factors of major importance:

- distribution system throughput volume (Masters [1980]; Jackson [1981]; Cooper [1983, 1984]; Jackson [1985]; Closs and Cook [1987])
- customer order frequency and order weight (Newbourne and Barrett [1972]; Masters [1980]; Cooper [1983, 1984])
- geographical location of order destination (Newbourne and Barrett [1972])
- management policies regarding timing of shipment–dispatch, order–release, and priority shipping (Newbourne and Barrett [1972]; Masters [1980]; Jackson [1981]; Ha, Khasnabis, and Jackson [1988])
- management support (Newbourne and Barret [1972], Ackerman [1990])

Factors of lesser importance:

- number of consolidation points and rules used to assign customers to these points (Jackson [1981]; Ha, Khasnabis, and Jackson [1988])
- product class and freight rate (when shipping by common carrier) (Cooper [1983, 1984])
- mode of transportation (Newbourne and Barrett [1972])

Table 3-2
Necessary Conditions for the Implementation of
A Shipment Consolidation Program
(partially adapted from Ackerman [1990])

- Order patterns and volume must be sufficient to allow aggregation of customer orders to common destinations within a reasonable length of time.
- Management must be willing to accept less than ideal customer service levels to allow shipments to common destinations to be consolidated.
- Transportation savings from the use of consolidation must be sufficient to justify its use.

There are some shortcomings of the shipment consolidation literature:

- Much of this literature is descriptive. Many articles give only a general overview or are almost entirely devoted to illustrating the potential cost savings resulting from shipping consolidated loads. Examples: Pollock [1978]; Schuldenfrei and Shapiro [1980]; Sheahan [1982]; Weart [1984]; Ballou [1987].

- Much work has not recognized that consolidation can be performed by a number of different parties, thus assumptions relating to this aspect usually are not clearly stated.

- Several studies have examined the relationships between policy variables of an existing consolidation program (Masters [1980]; Jackson [1981]; Cooper [1983; 1984]; Ha, Khasnabis, and Jackson [1988]). However, attempts to determine optimal or near–optimal parameters for either an existing or proposed consolidation strategy have been limited.

  Similarly, factors important to the success of a consolidation program have been identified, but insight into the calculation of threshold values of these factors is lacking.

- Quantitative analysis aimed at determining optimal or near–optimal consolidation parameters has been hampered by some limiting assumptions; for example, ignoring the impact of fixed transportation costs or carrier discounts, or basing analysis on "orders" or "items" rather than on weight or volume. Examples: Brennan [1981]; Blumenfeld, Burns, Diltz, and Daganzo [1985]; Burns, Hall, Blumenfeld, and Daganzo [1985].

The present research addresses these oversights.

Chapter 4
SHIPMENT-RELEASE POLICIES IN SHIPMENT CONSOLIDATION

4.1 Introduction

Management decisions in a shipment consolidation program can be generalized as follows:

- *What* will be consolidated? Which customer orders will be consolidated and which will be shipped individually?
- *When* will customer orders be released? What event(s) will trigger the dispatch of a consolidated vehicle load?
- *Where* will the consolidation be done? Should the consolidation be done at the factory, on a vehicle, at a warehouse or terminal, etc.?
- *Who* will perform the consolidation? Should the consolidation be done by the manufacturer, shipper, customer, carrier, third party, etc.?
- *How* will the consolidation be done? Which specific consolidation techniques will be used?

This thesis deals with the "When" question, often referred to as **shipment-release timing**. Shipment–release timing seeks to determine how long customer orders should be held and/or what quantity should be accumulated before a consolidated load is released. Figure 4–1 presents a diagram of this problem in terms of inputs, decision factors, and outputs, for a newly–arrived customer order.

The impact of shipment–release timing on the size and frequency of consolidated loads, order–cycle time, and distribution cost has been recognized by many authors. Table 3–1 notes that shipment–release timing is considered a major management decision in a consolidation program. Table 4–1 summarizes research on shipment–release timing.

Figure 4–1
The "When" Decision in Shipment Consolidation

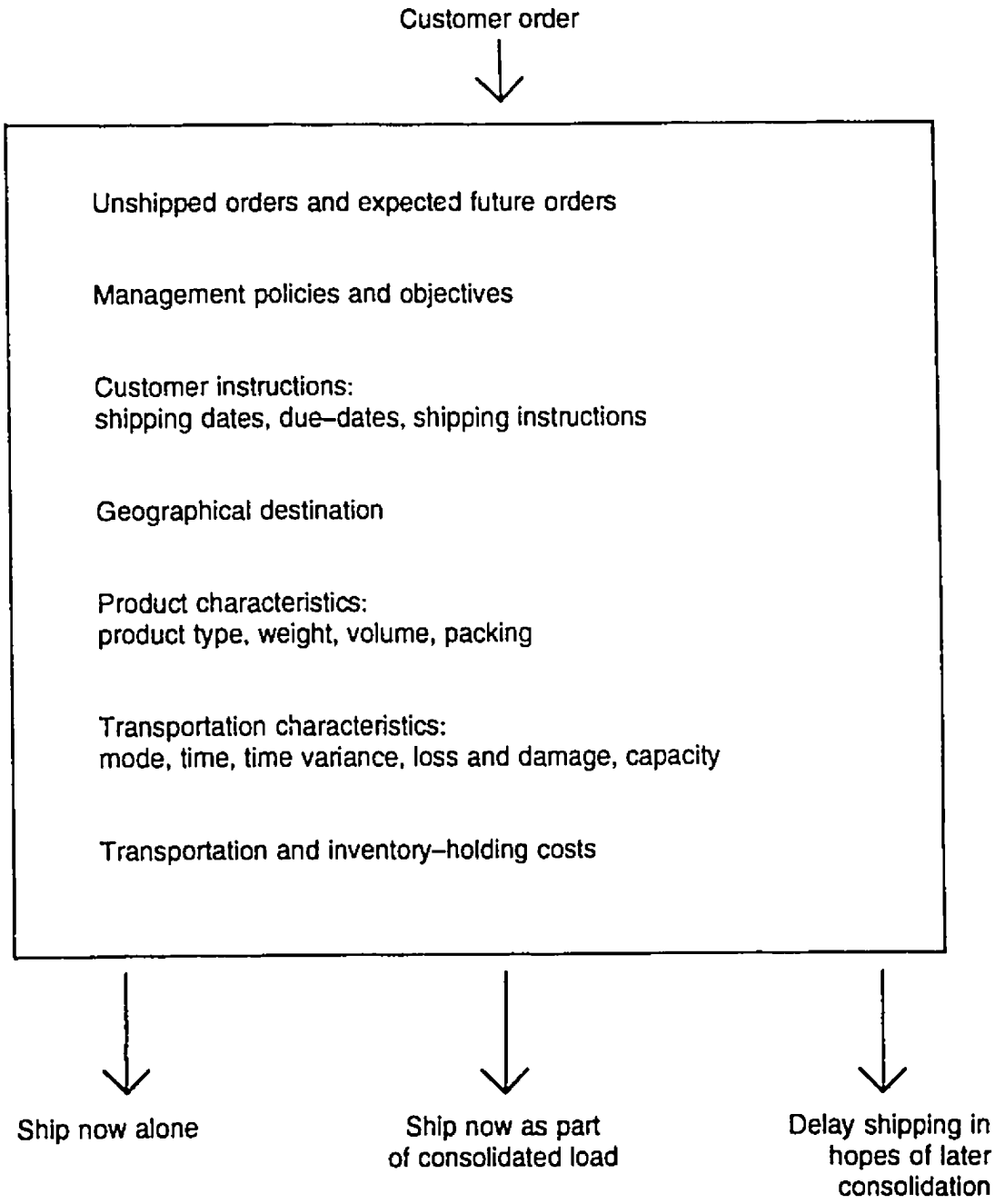


Table 4–1
Literature Summary: Shipment–Release Timing

Newbourne and Barrett [1972]:

method: descriptive

conclusion: optimal holding time is approximately three days, and orders that are not consolidated within that time probably will remain unconsolidated even if held longer

comments: conclusion is supported only by the comment "for reasons involving statistics and common business practices"

Masters [1980]:

method: simulated a shipment consolidation system using maximum order holding time of 1, 4, and 7 days

conclusion: longer order holding time increases mean delivery time and reduces transportation costs

comments: good examination of effect of several consolidation variables; does not suggest optimal policy levels

Jackson [1981]:

method: simulated a shipment consolidation system using various order release rules, including maximum holding time of 1, 4, and 7 days

conclusion: longer order holding times increase the probability that an order will be consolidated and decreases transportation cost, thus orders should be held for maximum time possible

comments: main purpose of study is to examine effect of varying predefined consolidation variables; suggestion of optimal policies is limited to those situations investigated

Cooper [1983, 1984]:

method: simulated a shipment consolidation system using maximum order holding time of 1 and 4 days

conclusion: longer order holding time increases mean delivery time and reduces transportation costs

comments: purpose of study was not to investigate order holding time rules; conclusions are situationally–bound

Table 4–1 (continued)

Jackson [1985]:

method: survey of consolidation practices of 52 firms

conclusion: sixty per cent of respondents reported that their maximum order holding time was 3 days

comments: provides insight into shipment consolidation practice; sample size is limited and decision rules are not proposed

Ansari and Heckel [1987]:

method: discussion of heuristic used by Hewlett–Packard for determining the frequency of consolidated Just–in–Time deliveries

conclusion: frequency of deliveries must consider interaction between freight costs and inventory costs

comments: logic is sound, but mathematics is flawed and results are not optimal

Brennan [1981], Burns et al. [1985], Blumenfeld et al. [1985]:

method: quantitative analysis to determine optimum number of orders to ship

conclusion: optimal number of orders to be shipped can be calculated using economic order quantity analysis to trade off transportation and inventory–holding costs to determine minimum cost strategy

comments: excellent theoretical analysis; practical application is limited by assumptions to general strategic planning

Determining when shipments should be dispatched can be viewed as a two–stage process. First, management must select a **shipment-release policy**; that is, a general approach for guiding the decision. Second, the chosen shipment–release policy must be operationalized. The result of that stage is a set of **shipment-release parameters**.

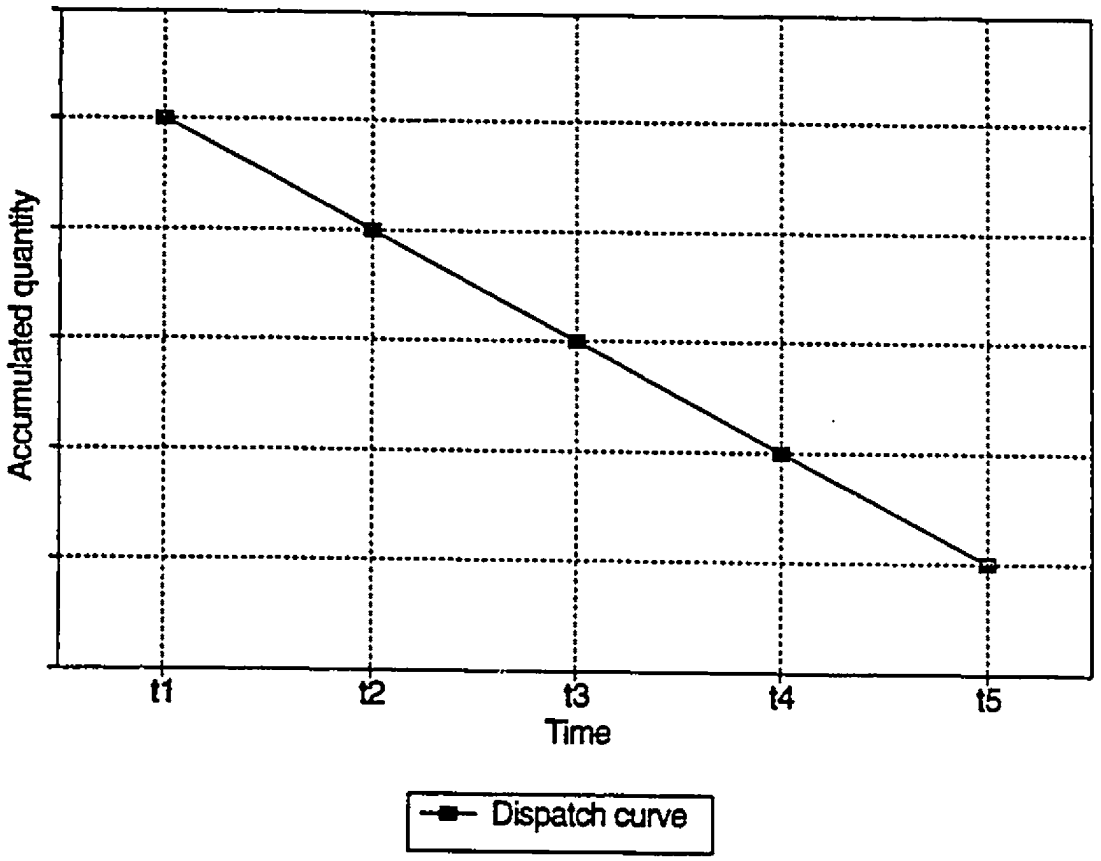
The decision when to dispatch a consolidated load may be based on a large variety of factors. This thesis examines a special class of shipment–release policies, those based only on elapsed time and accumulated quantity. Chapters 5 through 8 then propose methods for determining shipment–release parameters for these policies.

4.2 Time– and Quantity–Based Shipment–Release Policies

In their study of train dispatching, Beckmann, McGuire, and Winsten [1956] discuss a general release policy based on time and number of railcars waiting (ie., accumulated quantity). Their "dispatching policy" yields the decreasing curve reproduced in Figure 4–2. When the accumulated quantity (as a function of time) reaches this curve, a train is dispatched. Clearly, this curve need not be a straight line.

Three special cases of this dispatching policy are commonly used in deciding when to release a consolidated shipment. A **time policy** dispatches each order at a pre–determined shipping date, whether or not it is consolidated. This approach sometimes is referred to as "scheduled shipping". Under a **quantity policy**, all orders for a particular destination are held and shipped when a minimum consolidated weight is reached. Lastly, a **time-and-quantity policy** holds all orders for a particular

Figure 4–2
General Time– And Quantity–Based Release Policy
Suggested by Beckmann, McGuire, and Winsten [1956]



destination until the earliest of: i) a pre–determined shipping date, or ii) the accumulation of a minimum weight or volume.If the latter occurs first, the orders are dispatched before the pre–specified release date.

The focus of a time policy often is customer service, with the shipping date set to meet customer requirements. Conversely, the target consolidated weight of a quantity policy usually is determined by considerations of cost. Intuitively, a time–and–quantity policy might be expected to exhibit the best characteristics of both a time policy and a quantity policy. This is not necessarily true, as will be seen later.

A time policy, as well as the time component of a time–and–quantity policy, can be approached in two ways (Jackson [1985]). With a **maximum holding time** approach (or **oldest order** approach), the shipping date is determined by the order that has been waiting longest. Thus, the time component dictates that a consolidated load be dispatched when the oldest order reaches a certain age. The next accumulation cycle begins with the arrival of the first order after this dispatch.

The time component of a **scheduled shipping and volume** approach may or may not consider the age of the oldest order. For example, the accumulation cycle may begin immediately after the dispatch of a consolidated load or when the first order of a new cycle arrives. If the quantity component is ignored, beginning a new cycle immediately upon load dispatch is similar to releasing shipments at set intervals, with the possibility of cancelling a load if the accumulated weight is too small to be economical. Powell [1985] and Powell and Humblet [1986] studied this case through bulk–queuing theory for passenger vehicle dispatching.

A survey by Jackson [1985] reviewed the use of these three policies by practitioners. A time policy was found to be the most frequently applied, being used by 36% of respondents. However, the differences in percentage–usage between the three polices was not large.

Other authors have made brief statements regarding the impact of shipment–release policy when investigating the effect of varying consolidation parameters. Typically, however, these studies have focused on factors other than shipment–release timing, and results are interdependent. For example, although Cooper [1983, 1984] included both one– and four–day holding times in her simulation, her goal was the comparison of direct shipping versus use of warehouses, not the testing of shipment–release policies. Jackson [1981] compared a time policy with a time–and–quantity policy. Unfortunately, he also varied the number of consolidation points and the system volume, thus his results regarding shipment–release timing are intertwined with those for other parameters. Other differences between his and our study are discussed later.

In this chapter, we present a simulation model designed expressly to examine the effect on mean per–unit cost and mean order delay of the three shipment–release policies. The results of this simulation also will be used in Chapter 7 to test a sequential decision model for timing the dispatch of consolidated loads.

4.3 Simulation Comparison of Shipment–Release Policies

Our simulation was a simple discrete–event model. Customer orders, each weighing a random amount and arriving at a random time, were generated and

accumulated until a target weight or time was reached. All waiting orders then were dispatched, statistics updated, and the accumulation cycle re–started. Because the system emptied with each dispatch, output collection could begin immediately without determination of a steady state. The next sections discuss some modeling considerations.

Simulation Parameters

Our simulation assumed common carrier transportation. Section 2.4 discussed the concept of phantom freight: the ability to declare heavier weights than actually exist, to push the weight of a load into a heavier weight bracket and qualify for a lower freight rate. The minimum weight (which we will call "WBT") at which this strategy is cost–effective equals the minimum volume weight ("MWT") times the ratio of the volume freight rate f_V and the non–volume freight rate f_N, $f_V \leq f_N$; that is, WBT = MWT (f_V/f_N), WBT≤MWT. At this weight, WBT f_N = MWT f_V. This concept is discussed in greater detail in Section 5.3.

Vehicle capacity restrictions were ignored, consistent with the popular assumption that additional common carrier vehicles always are available. The minimum volume weight (MWT) was set at MWT=20000 pounds, with volume rate f_V and non–volume rate f_N of f_V=$2.25 per cwt. and f_N=$3.00 per cwt. This gives a WBT weight of 15000 pounds.

Both a quantity policy and a time–and–quantity policy require specification of a target accumulated weight. We arbitrarily selected target weights of 12500 lbs.,

15000 lbs., 17500 lbs., 20000 lbs., and 22500 lbs., then used the economic shipment weight (ESW) formula to calculate the corresponding order arrival rate.

A mathematical expression for the economic shipment weight (ESW) is derived and discussed in Section 5.3. This concept, similar to the economic order quantity, seeks the load size that minimizes the per–order sum of transportation and inventory–holding costs:

$$ESW = \frac{2 \hat{a} F_L E[W]}{r_w}$$

where â is the order arrival rate, F_L is the sum of all fixed costs associated with a vehicle load, E[W] is the expected weight per customer order, and r_w is the inventory–holding cost per unit weight per time period.

Inventory–holding cost was set at r_w=\$0.25 per cwt. per day, and a fixed cost per load of F_L=\$30 per dispatch was levied. These cost parameters were constant throughout the simulation. Setting the above ESW expression equal to the selected target weights determined the scenarios we investigated:

target weight	comment	order arrival rate
12500 lbs.	less than WBT	3.26 orders per day
15000 lbs.	equal to WBT	4.69 orders per day
17500 lbs.	between WBT and MWT	6.38 orders per day
20000 lbs.	equal to MWT	8.33 orders per day
22500 lbs.	greater than MWT	10.55 orders per day

For the time policy and the time–and–quantity policy, an oldest order approach was used to set the maximum holding time: consolidated loads were dispatched when the order waiting the longest had reached a specified maximum delay. These times were arbitrarily selected after considering the expected time to accumulate the target

weight given the arrival rate. We tested maximum holding times of 0.75 days, 1.0 days, 1.5 days, and 2.0 days.

Modeling Customer Order Arrivals and Weights

Our simulation assumed that order arrivals follow a Poisson process with an arrival rate of â=3 orders per day. Masters [1980] modeled interarrival times between orders as a uniform distribution, while Cooper [1984] used an exponential distribution, and Jackson [1981] and Ha, Khasnabis, and Jackson [1988] used empirical distributions.

Very little guidance exists in the academic literature as to appropriate frequency distributions of customer–order weights. Such a distribution will vary between and among products, shippers, carriers, and purchasers.

Masters [1980] modeled customer order weight as a normal distribution with coefficient of variation (CV) of 1/2. Cooper [1983, 1984] and Ha, Khasnabis, and Jackson [1988] used truncated normal distributions with CV=1. Jackson [1981] and Closs and Cook [1987] employed empirical data.

Fitting of theoretical probability distributions to empirical order weights was attempted by Akaah and Jackson [1988]. Unfortunately, their analysis was limited to the normal, uniform, and Poisson distributions. Although only ten of their forty sets of weights fit one of these distributions, no better–fitting distributions were suggested or tested.

Jackson's [1981] simulation used empirical order weights from a medium–size national package goods distributor. His paper includes a frequency histogram showing

the percentage of orders for ten weight groups; this graph is reproduced as Figure 4–3. Unfortunately, it is not sufficiently exact to be of detailed use. For example, his percentages add to about 105%. Moreover, based on Figure 4–3, we estimated a mean weight of between 1500 and 1900 pounds, depending on our reading of weights from the plot. This differs from Jackson's stated mean of 1300 pounds.

Of many possible theoretical distributions, we chose an unshifted gamma distribution (denoted as Ga(α,β)) for modeling order weight. Our reasoning, which included visual comparison with Figure 4–3, is summarized in Table 4–2. To avoid computational difficulties, we considered only integer values of the shape parameter α. From inspection of Figure 4–3, we set this value at α=2. Larger values of α result in a flatter distribution, while α=1 yields an exponential distribution, which lacks the positive skewness of Jackson's plot. Such skewness appears intuitively satisfying. Basic properties of the gamma distribution are given in Table 4–3.

For simplicity, we selected a mean order weight of 2000 pounds; this is not too different from Jackson's mean of 1300 pounds. The expected value of the gamma distribution is $E[W]=\alpha\beta$, where α and β are shape and scale parameters respectively. Thus, a mean of 2000 pounds and α=2 yields β=1000 and standard deviation of $\sigma=\beta\sqrt{\alpha}$ =1414 pounds. The mode is $\beta(\alpha-1)$=1000 pounds, which is slightly higher than Jackson's mode in Figure 4–3.

Some simpler alternatives to the gamma distribution might be considered. We noted in Section 2.4 that some orders may be too small for cost–efficient consolidation. For example, it may be preferable to ship these orders individually by courier, parcel post, or other means. Then, it may be justified to delete the extreme left weight

Figure 4–3
Empirical Customer Order Weight Distribution
Used by Jackson [1981]

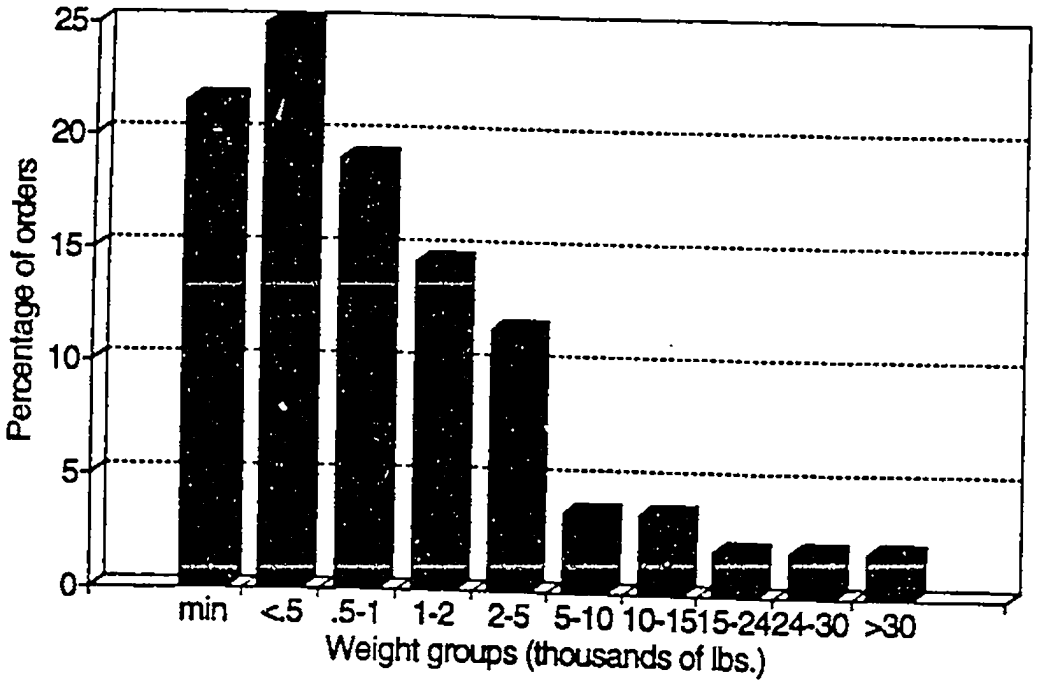


Table 4–2
Possible Theoretical Probability Distributions
For Modeling Customer Order Weight

Gamma distribution

Advantages:

- very flexible in terms of shape and moments
- allows a quasi–Normal distribution without truncation at zero and with positive skewness; such skewness appears intuitively satisfying and is supported by Jackson [1981]
- related to several other distributions, including the exponential, n–Erlang, chi–square, and Weibull distributions

Disadvantages:

- may be analytically difficult to work with; approximations to the distribution may be required

Normal distribution

Advantages:

- justifiable due to Central Limit Theorem
- very flexible in choice of mean and standard deviation

Disadvantages:

- must be truncated to avoid values less than or equal to zero
- can be analytically difficult to work with

Lognormal distribution

Advantages:

- often used to represent "quantities that are the products of a large number of other quantities" (Law and Kelton [1991])
- density takes on shapes similar to gamma distribution

Disadvantages:

- can be analytically difficult to work with; no closed form for the cumulative distribution function

Poisson distribution

Advantages:

- easy to work with, and past research is well documented

Disadvantages:

- discrete distribution: order weight probably should be modeled using a continuous distribution
- standard deviation felt to be too small to be realistic in modeling of order weight

Table 4–2 (continued)

Geometric distribution

Advantages:

- easy to work with

Disadvantages:

- interpretation of distribution is not intuitively satisfying to the modeling of order weights
- discrete distribution
- standard deviation felt to be too large, frequently resulting in huge extreme values requiring truncation

Exponential distribution

Advantages:

- easy to work with
- may be useful if weights are edited (discussed in text)

Disadvantages:

- mode equal to lower bound not felt to be realistic in modeling of order weight
- standard deviation felt to be too large

Table 4–3
Basic Properties of the Gamma Distribution

Probability density:

$$Ga(x|\alpha,\beta,s) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)}(x-s)^{\alpha-1}\, e^{-(x-s)/\beta} \qquad x \geq s$$

$$= \frac{1}{\beta\,\Gamma(\alpha)}[(x-s)/\beta]^{\alpha-1}\, e^{-(x-s)/\beta} \qquad x \geq s$$

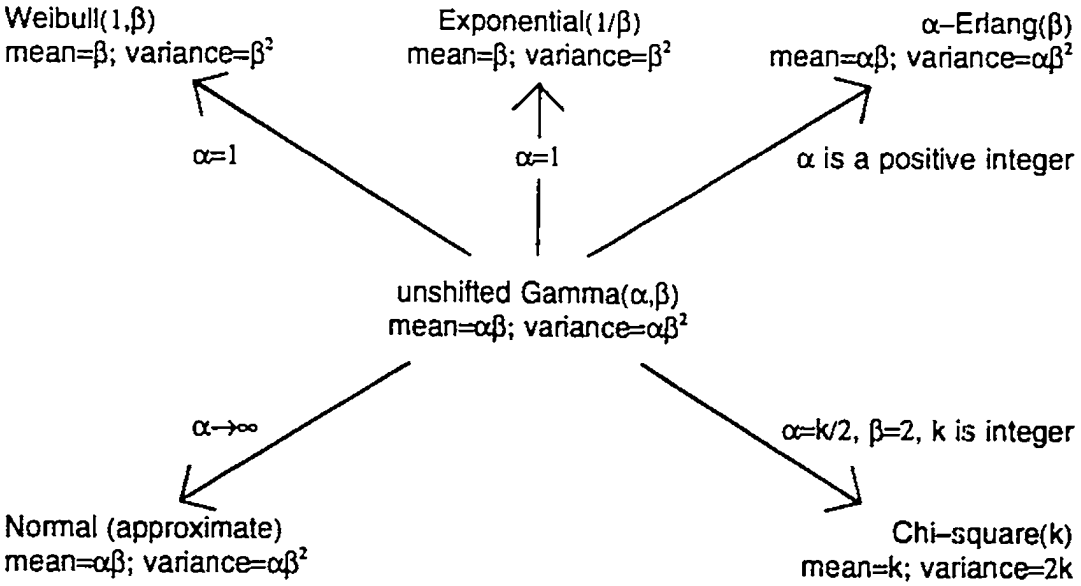
Cumulative distribution:

$$Pr\{X \leq x|\alpha,\beta,s\} = \frac{1}{\beta\,\Gamma(\alpha)}\int_0^{x-s}[(t-s)/\beta]^{\alpha-1}\, e^{-(t-s)/\beta}\, dt \qquad x \geq s;\ t \geq s$$

$$= \frac{1}{\Gamma(\alpha)}\int_0^{(x-s)/\beta}(t-s)^{\alpha-1}\, e^{-(t-s)}\, dt \qquad x \geq s;\ t \geq s$$

where $\alpha>0$ is a shape parameter, $\beta>0$ is a scale parameter, and s is a location (shift) parameter. The mean is $\alpha\beta+s$, the variance is $\alpha\beta^2$, and the mode is $\beta(\alpha-1)+s$ if $\alpha \geq 1$, 0 otherwise.

Important relationships:



category of Figure 4–3. Doing so would yield an exponential–like order weight distribution.

The Weibull distribution takes on shapes similar to the gamma distribution. However, evaluation of the gamma function for non–integer arguments would be required to yield a skewed distribution similar to that of Figure 4–3. Moreover, the Weibull distribution often lacks the long tail of the gamma distribution. This deficiency may not be appropriate for modeling order weights in some situations.

In summary, we chose the gamma distribution to model order weight largely because it allows a large variety of distribution shapes. As well, its close relationship to other probability distributions (see Table 4–3) can simplify some calculations.

4.4 Simulation Results

Output data was generated using the method of batch means (see Law and Kelton [1991]). Batches of customer order were generated prior to the simulation, then independently processed according to a time policy, a quantity policy, and a time–and–quantity policy. Undispatched orders remaining at the end of the simulation were deleted so that only complete consolidation cycles were considered.

Significance testing of the differences between policies was done using a paired–t confidence interval. Most differences were found to be statistically significant at the 90% confidence level. The main exception was the comparison of a quantity policy and a time–and–quantity policy. For arrival rates of â≥8.33, differences were not significant for per–unit cost with maximum holding time of 1.5 days, and for both per–unit cost and mean delay with holding time of 2.0 days. Under these conditions, the

time restriction of a time–and–quantity policy is more active than the quantity component in determining the timing of load dispatches. As well, some differences were not significant when the plot of results for one policy crossed that of another (for example, per–unit cost of a time policy versus time–and–quantity policy with arrival rate â=8.33 and holding time of 1.5 days).

Figures 4–4 to 4–7 illustrate results relating to per–cwt. cost, while figures 4–8 to 4–11 give results relating to mean order delay. Because a quantity policy is not affected by maximum holding time, per–cwt. cost and mean order delay are identical on these graphs for this strategy. Figures 4–12 through 4–17 restate our simulation results in terms of order arrival rate, rather than holding time as used in Figures 4–4 through 4–11. Figures 4–4 through 4–17 are placed at the end of this chapter to avoid disrupting the text.

Figures 4–4 to 4–7 do not always show a quantity policy as the lowest per–unit cost strategy. This is because freight rate weight breaks (discounts for larger load weights) were not considered when selecting the target weights of this policy. Although the economic shipment weight formula used to derive the target weights yields the minimum cost weight, when weight breaks are considered, a lower cost weight may exist. For example, with freight rate weight breaks, the deterministic per–unit cost with arrival rate â=6.38 would be about $2.58 per cwt., as compared to the mean cost of $2.74 actually reported in Figures 4–4 to 4–7.

Figures 4–4 through 4–11 illustrate the important interaction between order arrival rate and maximum holding time. Small order arrival rates coupled with small holding times may yield different preferred policies than will large order arrival rates

and large holding times. Moreover, no one policy yields the lowest per–cwt. cost for all order arrival rates. Thus, there is no one best order–release policy, although a time–and–weight policy consistently yielded the smallest mean delay per order. The choice and performance of a shipment–release strategy is heavily dependent on the order arrival rate and the length of time customers are willing to wait for shipment. This is especially important when the order arrival rate is small.

The following paragraphs discuss these results with regard to specific policies. Following this, we make more detailed comments as to the preferred policy under various combinations of order arrival rate and maximum holding time.

Quantity Policy

The cost–performance of a quantity policy, relative to one that considers holding time, depends on the selected order holding time. A quantity policy will be more expensive than a time policy if the latter has a holding time long enough to accumulate loads sufficiently large to benefit from volume transportation discounts. In Figure 4–4, the holding time of 0.75 days is so short that time–based strategies cannot do this, and are outperformed cost–wise by a quantity policy. As the maximum holding time increases, a time policy produces cost results comparable to a quantity policy because the holding time now is sufficiently long that a time policy can accumulate load sizes comparable to that of a quantity policy. A quantity policy also had the smallest coefficient of variation of per–load cost in our simulation.

Time Policy

Our simulation shows the danger of a pure time policy. A short holding time coupled with a small order arrival rate will result in frequent small loads. Large holding times will produce excessive load sizes, with benefits from transportation savings overwhelmed by inventory costs. In both cases, the per–unit cost will be larger than that of other policies. Moreover, because the maximum holding time has a direct effect on mean order delay, a poor choice can yield excessive delays, as seen in Figures 4–10 and 4–11 for large arrival rates with holding times of 1.5 and 2.0 days.

A time policy consistently yielded the largest coefficient of variation of per–load cost, sometimes twice as much as that of a time–and–quantity policy, and as large as five times that of a quantity policy. As well, a time policy always produced loads at least as large as (and frequently larger than) those of a time–and–quantity policy. The difference in load size between the two policies was not large (though statistically significant) for small arrival rates and short holding times, but was dramatic for large arrival rates and long holding times, where a time policy produced loads as much as 44% bigger. The difference in load sizes depends on both order arrival rate and maximum holding time, however the latter has a much greater impact than the former.

With small arrival rates and small holding times, the larger loads of a time policy resulted in lower per–unit cost than did a time–and–quantity policy. However, with large arrival rates and large holding times, the cost performance of a time policy suffered because transportation savings from volume loads were overwhelmed by inventory–holding costs.

Time-and-Quantity Policy

It might be expected that a time–and–quantity policy would exhibit the best features of both a quantity policy and a time policy. Our results show that this may or may not be true, depending on the basis of comparison. When the holding time is small, the target weight is not operative because there is insufficient time to reach that weight. A time–and–quantity policy then performs much like a time policy, resulting in higher per–unit cost than does a quantity policy. When the maximum holding time is large, the target weight is reached before the time limit is reached, and the strategy acts like a quantity policy. This is seen by comparing per–unit cost for a quantity policy and a time–and–quantity policy for arrival rates of â≥8.333 and holding times of 1.5 and 2.0 days (Figures 4–6 and 4–7 respectively): the differences between the two policies for these cases are not statistically significant.

Intuitively, a time–and–quantity policy should never be more expensive than a time policy, nor should it be cheaper than a quantity policy. Our results agree with this thinking on a cost *per-load* basis; this is not true on a *per-cwt.* basis. For example, for small order arrival rates, a time policy yields higher per–load cost than does a time–and–quantity policy. However, the load produced by a time–and–quantity policy is never larger than that of a time policy because the former policy is constrained by a target weight; a time policy will allow shipments greater than the target weight. Thus, with small arrival rates and short holding times, shipments under a time–and–quantity policy often will not be sufficiently large to qualify for volume freight rates, while the larger loads accumulated by a time policy may. As a result, on a cost per–cwt. basis, a time–and–quantity policy may be more expensive than a time policy.

In terms of mean order delay, however, a time–and–quantity policy outperforms the other two strategies, chiefly for the same reasons that it performs poorly with regard to per–unit cost. When the order arrival rate is small, the time portion of the strategy is active, and loads are dispatched without waiting for the target quantity to be attained. When the holding time is large, the target quantity is attained first, thus avoiding the excessive waiting times of a time policy.

4.5 Comparison of Our Simulation Results With Those in the Literature

The impact on cost and delivery time from changes in maximum holding time were noted by Masters [1980] for a time policy and by Cooper [1983, 1984] for a time–and–weight policy. Both concluded that increasing the holding time reduces transportation costs and increases mean delivery time, although the increase in the latter is a fraction of the increase in holding time. These conclusions agree with ours.

Jackson [1981] examined the impact on shipment cost and order cycle time from changes in the maximum holding time and system volume. He also compared the cost and time performance of the scheduled shipping approach (time policy) and the scheduled–shipping–and–volume approach (time–and–quantity policy). His simulation model differs from ours in several respects. As noted previously, Jackson varied such factors as the number of consolidation points and the system volume. Direct shipping without consolidation also was tested. Transportation time, based on empirical results in Piercy [1977], was included in his calculation of order delay. Like Masters [1981] and Cooper [1983, 1984], however, inventory–holding costs were not considered.

Jackson found that low volume systems suffer from higher transportation costs and longer, more variable order delays. This agrees with our results when the order arrival rate is small. As well, we agree with his conclusion that longer holding times reduce transportation cost, but add that if the dispatch of a shipment is determined solely by target weight or quantity, longer holding times have no effect on transportation cost.

Jackson noted that the combination of long holding times and high system volumes (high order arrival rates) results in lower costs. This is true if inventory–holding costs are ignored, as Jackson did. If they are considered, long holding times may cause transportation savings to be overwhelmed by inventory–holding costs, as seen in Figure 4–16 for a time policy.

We also add that, for any strategy that includes a target consolidated weight, if either the arrival rate or the holding time is sufficiently large that most shipments move under the lowest volume rate, further increases in either parameter will not change per–cwt. cost appreciably. This can be seen by comparing per–cwt. cost for holding times of 1.5 days (Figure 4–6) and 2 days (Figure 4–7) with order arrival rates of 8.333 and 10.546.

Jackson also noted that a time strategy was cheaper and slower than a time–and–weight strategy for high volume systems, but comparable for low volumes. Our results agree. However, for small arrival rates, the difference in mean order delay between the two policies becomes considerably larger as the maximum holding time increases.

Powell [1985] compared vehicle dispatch strategies through analysis of batch arrivals/bulk service queues. He included a service measure calculated as:

$$Wq_{95} = Wq + 1.645\ (var(Wq))^{1/2}$$

where Wq_{95} is the 95th percentile of waiting time, Wq is the mean waiting time, and the constant 1.645 is derived by assuming that the distribution of waiting times is approximately normal. His assumptions differ from ours (for example, order arrivals occurred in batches, and scheduled vehicles could be cancelled if the load was too small), so direct comparison of results is not viable. However, he found that shorter maximum holding times were more expensive than longer ones, and that a weight strategy dominated a time strategy according to his Wq_{95} service measure. This latter conclusion agrees with our results only for long holding times, or for short holding times with very large arrival rates.

In general, the results of our simulation support other research in illustrating that the selection of a shipment–release policy can be very complex. The next section states some specific comments regarding the choice of such a strategy, while Section 4.7 makes some general conclusions.

4.6 Which Policy Is Best?

Given certain combinations of order arrival rate, minimum–cost quantity, and maximum holding time, which policy is preferred? Figures 4–12 through 4–17 restate our simulation results in terms of order arrival rate, while Table 4–4 summarizes Figures 4–4 through 4–17 in terms of dominated policies. As this table shows, for

Table 4–4
Summary of Simulation Results: Recommended and Dominated Shipment–Release Policies
(quantities in parentheses are values of E[%], discussed in text)

holding time	order arrival rate (orders per day) â=3.26	â=6.38	â=10.55
0.75 days	No clear choice (0.245)	No clear choice (0.479)	No clear choice (0.703)
1.0 days	No clear choice (0.326)	No clear choice (0.638)	Quantity policy dominates time policy (0.938)
1.5 days	Time policy dominates quantity policy (0.489)	Quantity policy dominates time policy (0.957)	Time policy is dominated by both other policies (1.407)
2.0 days	Time policy dominates quantity policy if target weight equals minimum volume weight (0.652)	Quantity policy dominates time policy if target weight equals minimum volume weight (1.276)	Time policy is dominated by both other policies (1.876)

Definitions:

For a particular combination of arrival rate â and holding time, policy P_1 dominates policy P_2 when both the cost and delay performances of policy P_1 are at least as good as those for policy P_2, and one of these performances is better for P_1 than for P_2.

"No clear choice" means that the choice of shipment–release policy under this combination of arrival rate and holding time will depend on management objectives of cost and customer service.

many combinations of arrival rate and holding time, the best policy will depend on management's objectives with regard to cost and customer service.

More exact decision rules suggesting the best policy would be helpful. Developing such rules is difficult: parameters may take a wide range of (not necessarily optimal) values. For example, management may prefer a smaller–than–optimal target weight, trading higher per–unit cost for improved customer service.

We tested the following quantitative decision approach. First, given the cost parameters in the simulation, we determined the minimum–cost consolidated weight for each arrival rate. We next calculated the percentage of this minimum–cost weight that would be expected to be accumulated with each holding time; that is:

E[percent of minimum–cost weight accumulated in holding time]

= T_{MAX} â E[W] / minimum–cost weight

where â is the order arrival rate, T_{MAX} is the maximum holding time, and E[W] is the expected weight of an order. We will denote the value of this measure as E[%].

Values of E[%] for the twelve arrival–rate/holding–time combinations simulated are given in parentheses in Table 4–4. Comparing these values to the summary of results given in that table yields the following observations:

if 0≤E[%]≤0.703:

> there is no clear choice as to preferred policy because no policy dominates any other policy with regard to both per–unit cost and mean order delay; thus, preferred policy will depend on management objectives regarding cost and customer service

if 0.938≤E[%]≤1.276:

> time policy is not recommended because it is dominated by both a quantity policy and a time–and–quantity policy; differences between a quantity and a time–and–quantity are sufficiently large that the

preferred policy will depend on management objectives regarding cost and customer service

if E[%]≥1.407:

either quantity or time–and–quantity policy is recommended: performance is similar or statistically the same for both

Obviously, clearer definition of boundaries is required. We also stress that because E[%] was calculated using the minimum–cost shipment weight, these conclusions are only valid if the selected target weight equals this minimum–cost weight.

The following example illustrates use of the E[%] decision approach.

Example: Suppose that the minimum–cost shipment weight is found to be 20000 pounds (this calculation is discussed in detail in Section 5.3). Management estimates that the maximum waiting time acceptable to its customers is T_{MAX}=3 days, and that the arrival rate and mean order weight are â=4 orders per day and E[W]=3000 pounds respectively.

The resulting value of E[%] is 1.8. From the above observations, we conclude that either a quantity policy or a time–and–quantity policy should be adopted (a quantity policy may be preferred simply because it is easier to use). Indeed, if maximum holding time was reduced to T_{MAX}=2.345 days, E[%] would equal 1.407, and, according to the above observations, the choice of shipment–release policy would not change.

If the volume of orders (as reflected in â) decreased so that the E[%] value fell below 0.703, management then should reconsider its choice of shipment–release policy in light of its goals regarding the trade–off between cost and customer service. ■

4.7 Conclusions

Two questions can be asked of our simulation. First, how would the results differ if we had simulated private carrier instead of common carrier? With private carrier, cost savings from consolidation occur from spreading of fixed costs, mainly those related to transportation, over larger loads. However, as the load size grows, the marginal change in per–unit transportation cost decreases. Thus, policies that hold orders for long periods so as to accumulate the common carrier's minimum volume may not be as effective under private carrier. Of course, results will depend on cost parameters and the minimum–cost shipment quantity.

Second, we noted that carrier volume discounts were not considered when setting the target weights of a quantity and a time–and–quantity policy. How would our results be affected if these discounts were included when determining target quantities? Analysis in Section 5.3 shows that the target weight should equal the minimum volume weight (20,000 pounds in our simulation) for all arrival rates less than 8.33 (recall that the target weight was 20,000 pounds for $\hat{a}$=8.33). Thus, results for $\hat{a}$=8.33 and $\hat{a}$=10.55 would not change. For all arrival rates, a quantity policy would result in the lowest per–cwt. cost, as seen in all figures for $\hat{a} \geq 8.33$. For a quantity policy, mean order delay would increase for all $\hat{a} < 8.33$ because the target weight would be larger. The amount of change in delay performance would be greatest for small arrival rates, but would not change the relative rankings of the three policies: a quantity policy yields the largest mean order delay even with the smaller target quantities simulated. The rankings might change for arrival rate $\hat{a}$=6.38 because a quantity policy produces a mean delay that is more comparable to the other policies.

From our simulation results, we can make the following conclusions regarding the relative cost and delay performance of the three shipment–release policies.

Quantity policy

Cost: A quantity policy will yield the lowest per–unit distribution cost, assuming that the lowest–cost quantity is selected. If the lowest per–unit cost quantity is not selected, the performance of this policy relative to that of the time policy will depend on the holding time and the order arrival rate. For example, a weight strategy will be bettered by a time strategy if the latter has a holding time long enough to accumulate large loads.

Delay: A quantity policy may or may not outperform a time policy, depending on the holding time selected, but will never perform better than a time–and–quantity policy.

Comments: Of the three shipment–release policies, a quantity policy is probably the easiest to use. A quantity policy is a discrete–time policy; the "state of the system" is checked only when a new order arrives. The time and time–and–quantity policies are analogous to a continuous–time review system.

Time policy

Cost: A time policy can be very dangerous cost–wise. A short holding time and small order arrival rate will result in frequent small loads and increased cost. Large holding times will produce excessive loads, with benefits from transportation savings overwhelmed by inventory costs. As well, this policy yields the largest variation in per–order cost.

Delay: A time policy may or may not outperform a quantity policy, depending on the holding time selected, but will never perform better than a time–and–quantity policy.

Time-and-quantity policy

Cost: On a per–unit basis, a time–and–quantity policy will never be cheaper than a quantity policy, and may be more expensive than a time policy because a time–and–quantity policy produces load sizes no larger, and frequently smaller, than those of a time–and weight policy.

Delay: A time-and–quantity policy will perform as well or better than both a time policy and a quantity policy for the same reasons that it performs poorly with regard to cost.

From these conclusions, we can make the following statements regarding the choice of a shipment–release policy:

- the selection of a shipment–release policy can be very complex, with the performance of a release policy highly dependent on the relative values of order arrival rate and maximum waiting time;
- knowledge of the level of service required by customers is crucial in selecting a shipment–release policy;
- customer service and order arrival rate must be examined simultaneously, because the value of one can eliminate or reduce significantly the importance of the other in the decision.

After a shipment–release policy has been selected, values for policy parameters must be determined. Approaches to this are discussed in the next four chapters.

Figure 4–4
Comparison of Shipment–Release Policies: Mean Cost per Cwt.
Holding Time = 0.75 Days

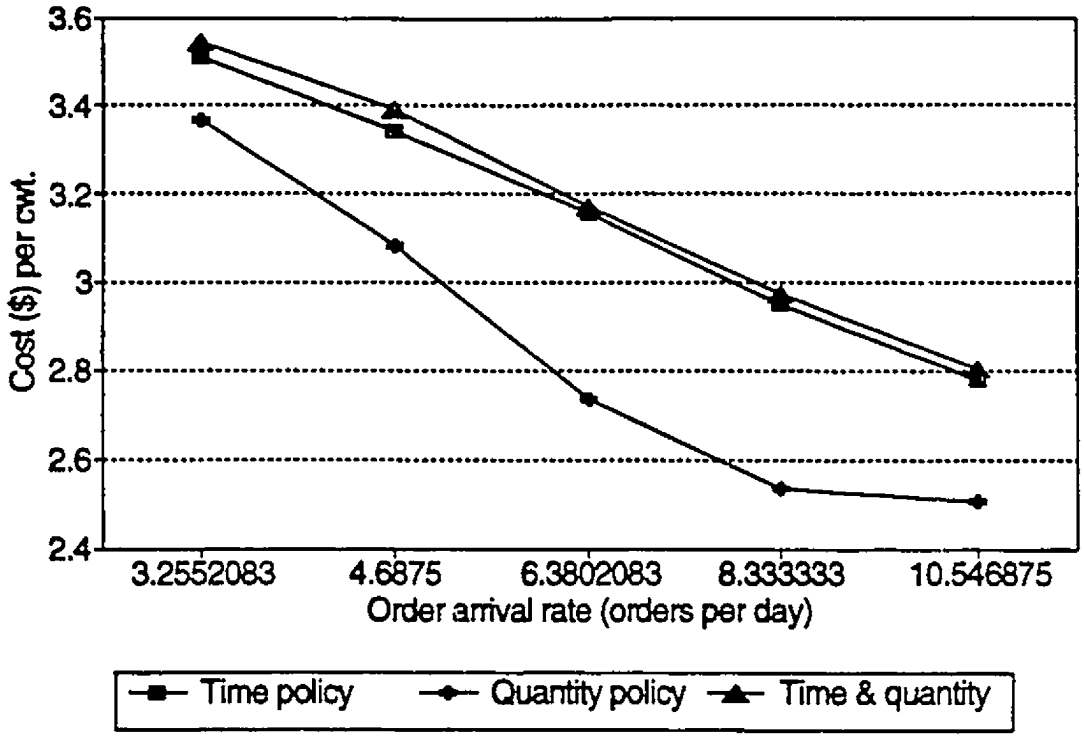


Figure 4–5
Comparison of Shipment–Release Policies: Mean Cost per Cwt.
Holding Time = 1.0 Days

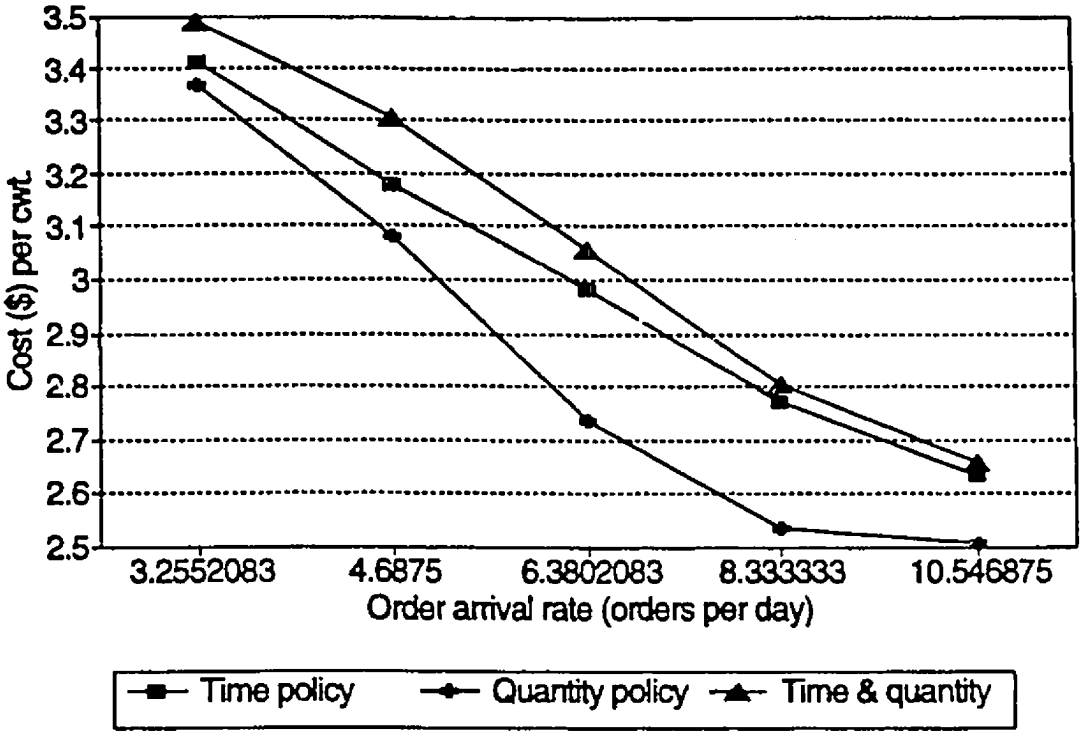


Figure 4–6
Comparison of Shipment–Release Policies: Mean Cost per Cwt.
Holding Time = 1.5 Days

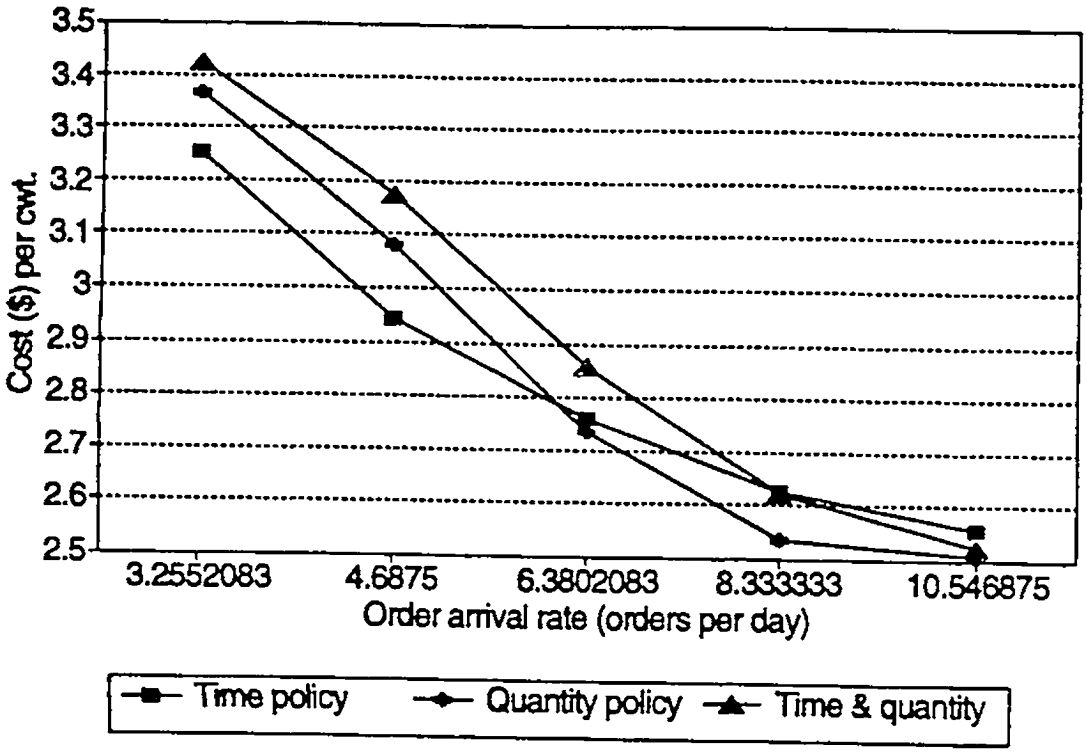


Figure 4–7
Comparison of Shipment–Release Policies: Mean Cost per Cwt.
Holding Time = 2.0 Days

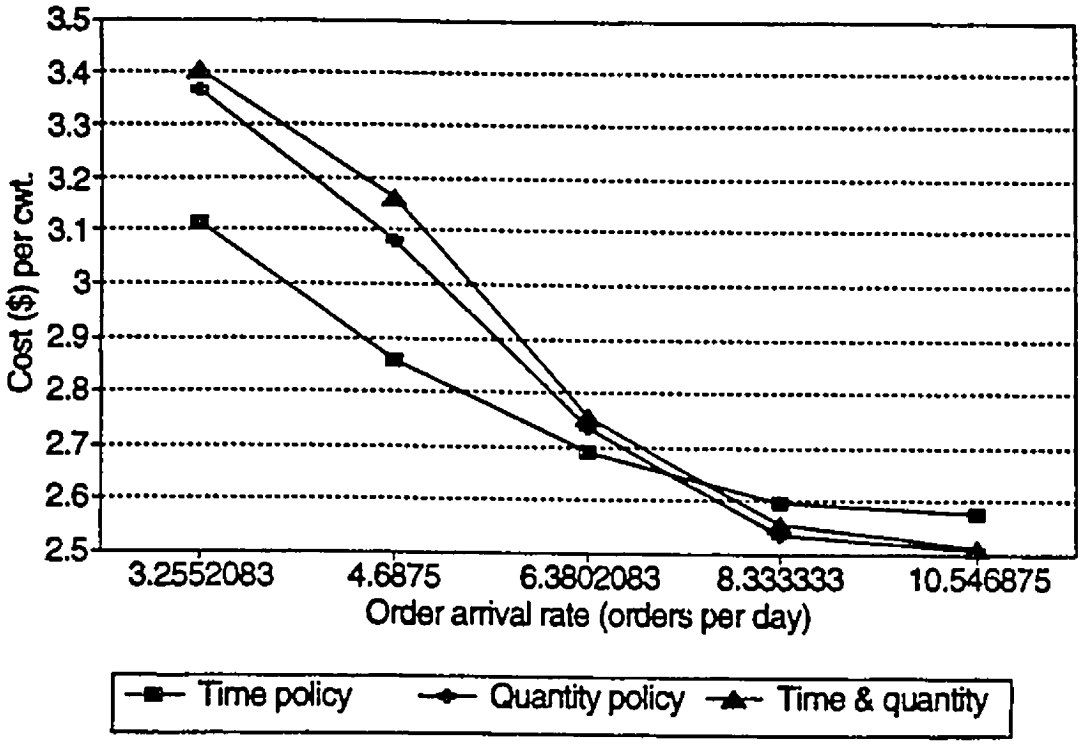


Figure 4–8
Comparison of Shipment–Release Policies: Mean Order Delay
Holding Time = 0.75 Days

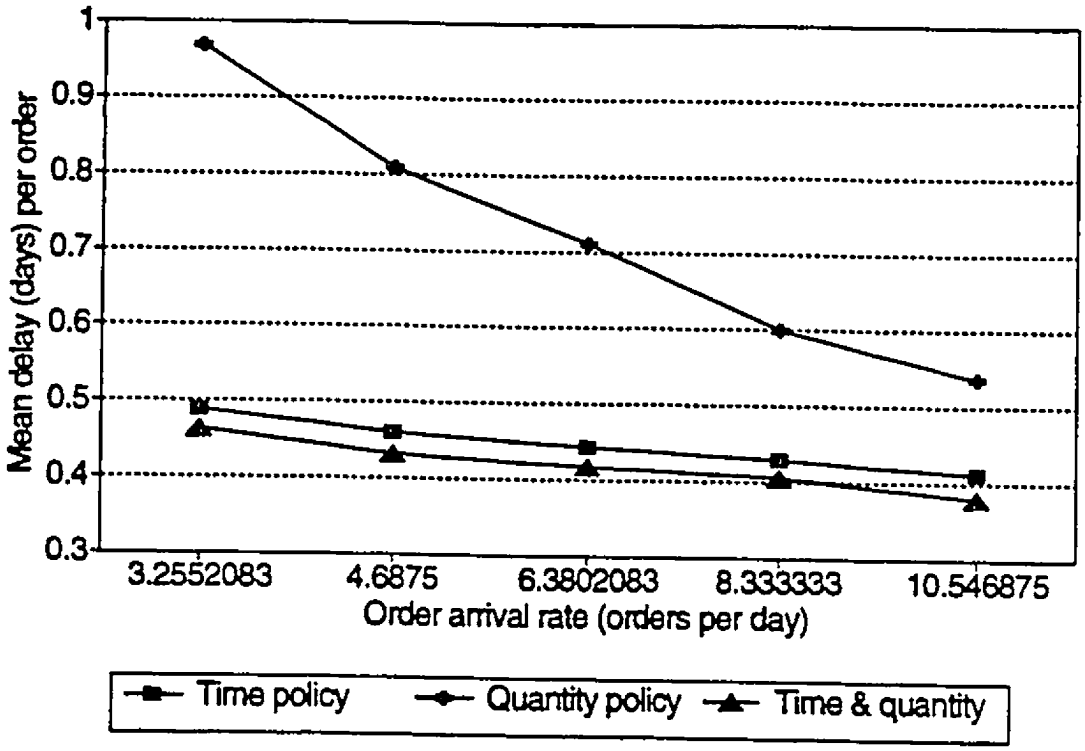


Figure 4–9
Comparison of Shipment–Release Policies: Mean Order Delay
Holding Time = 1.0 Days

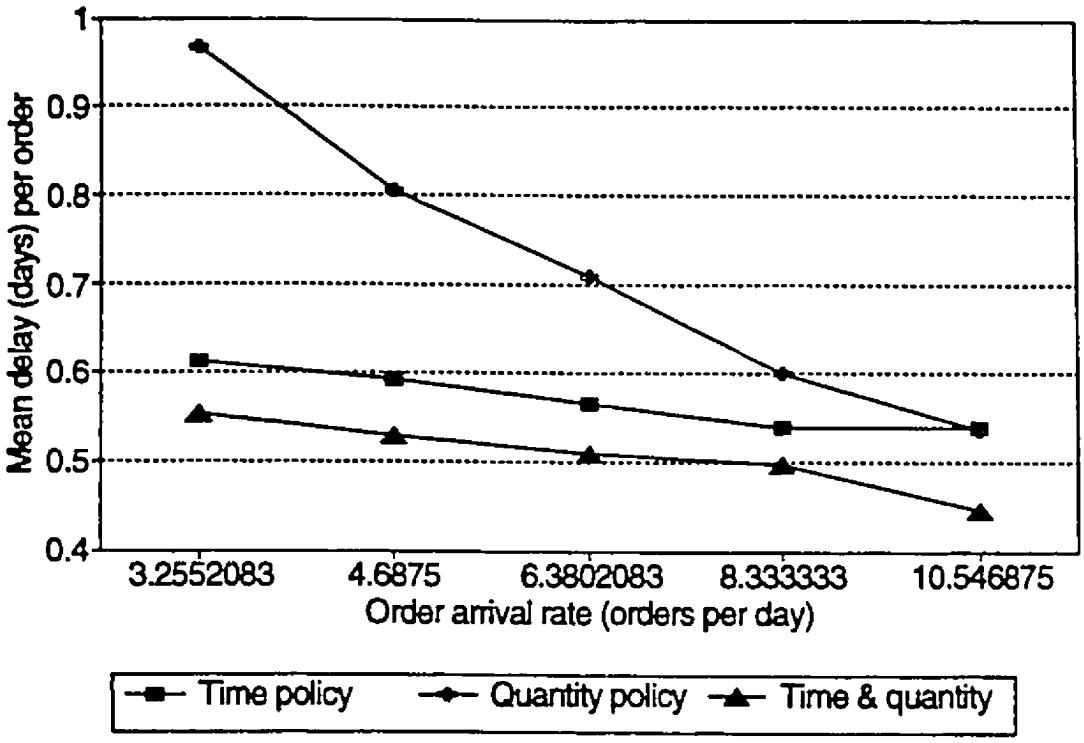


Figure 4–10
Comparison of Shipment–Release Policies: Mean Order Delay
Holding Time = 1.5 Days

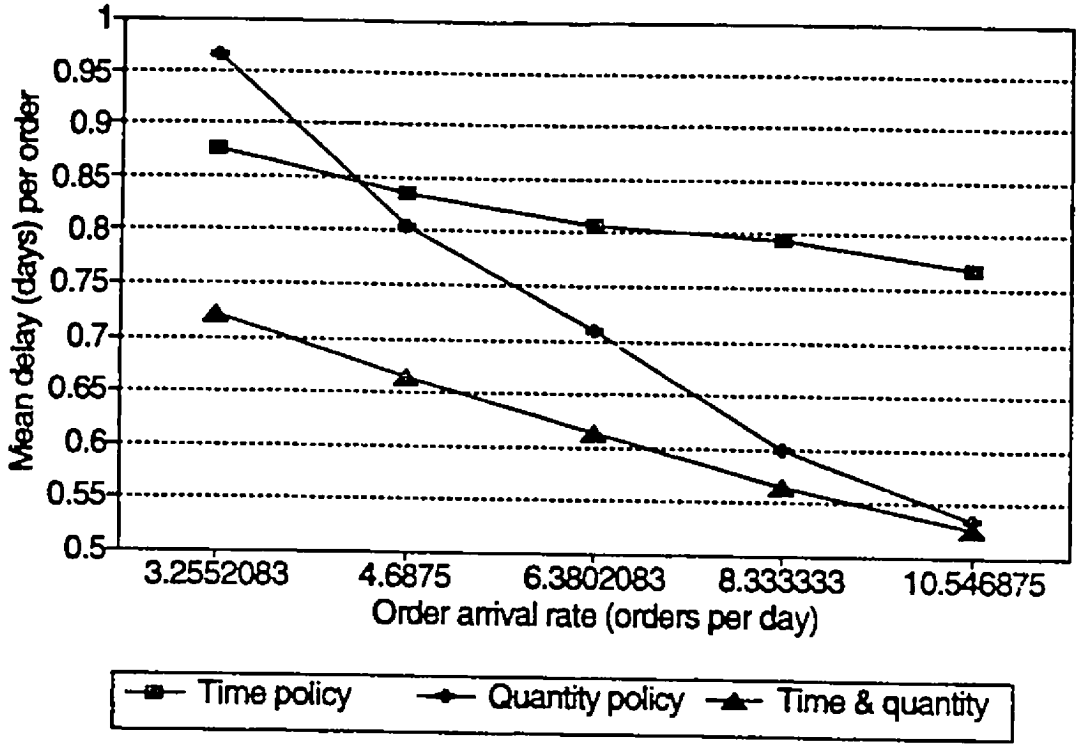


Figure 4–11
Comparison of Shipment–Release Policies: Mean Order Delay
Holding Time = 2.0 Days

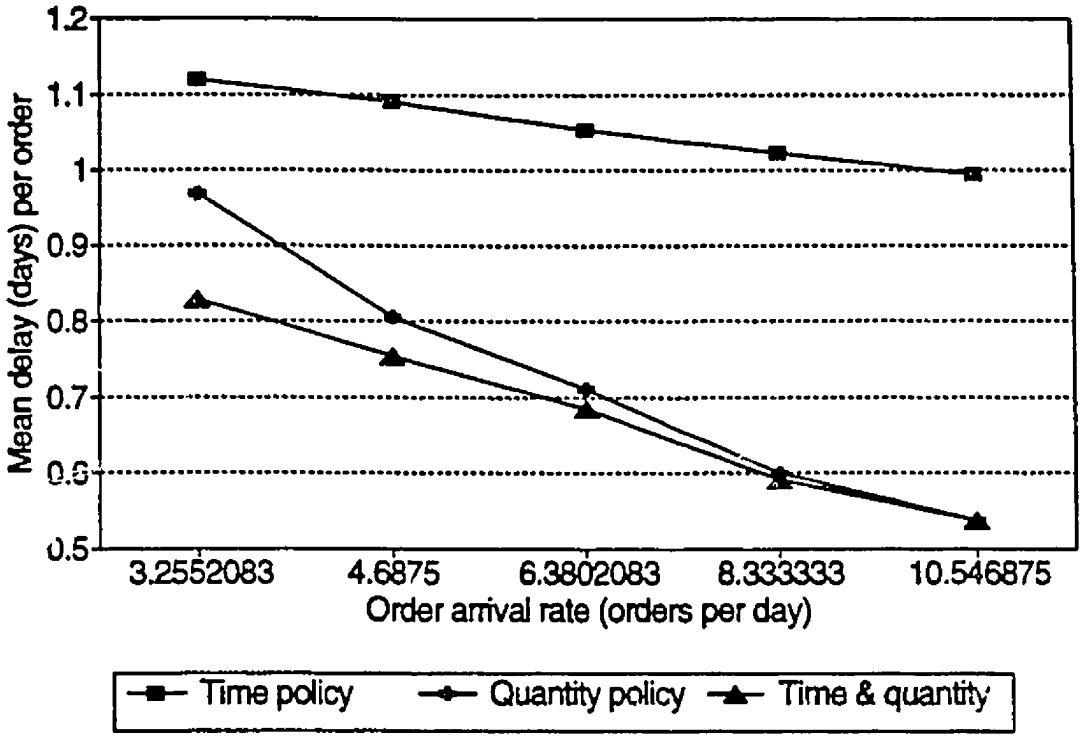


Figure 4–12
Comparison of Shipment–Release Policies: Mean Cost per Cwt.
Arrival Rate = 3.25 Orders Per Day

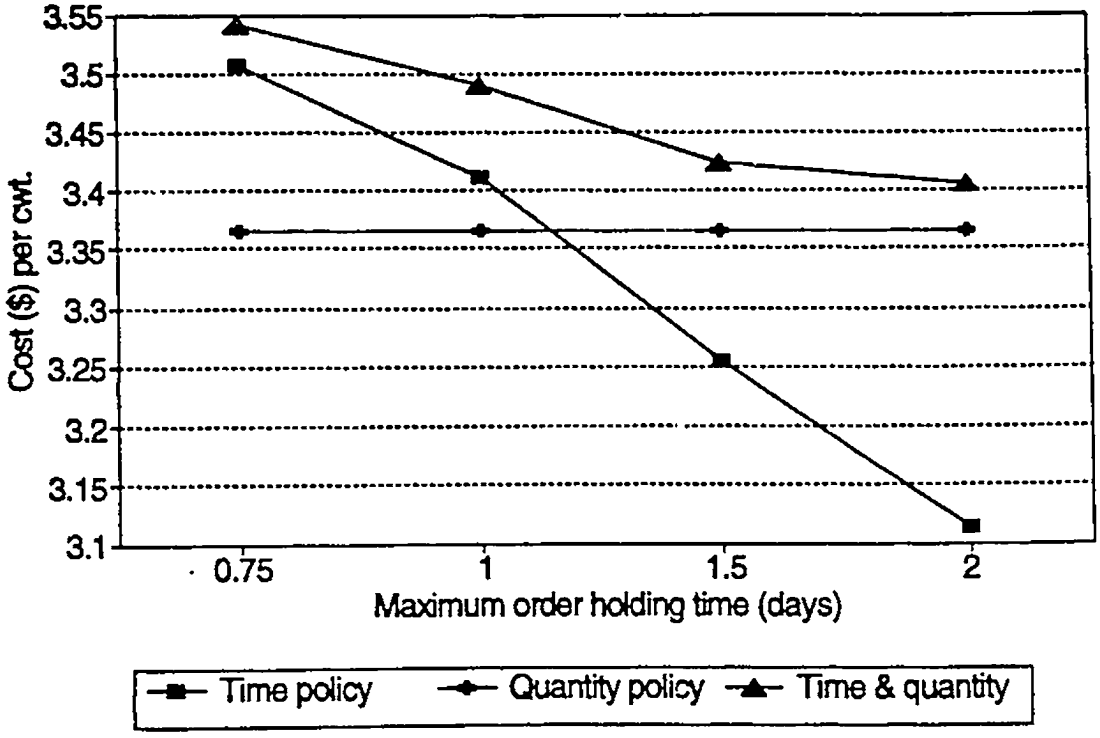


Figure 4–13
Comparison of Shipment–Release Policies: Mean Order Delay
Arrival Rate = 3.25 Orders Per Day

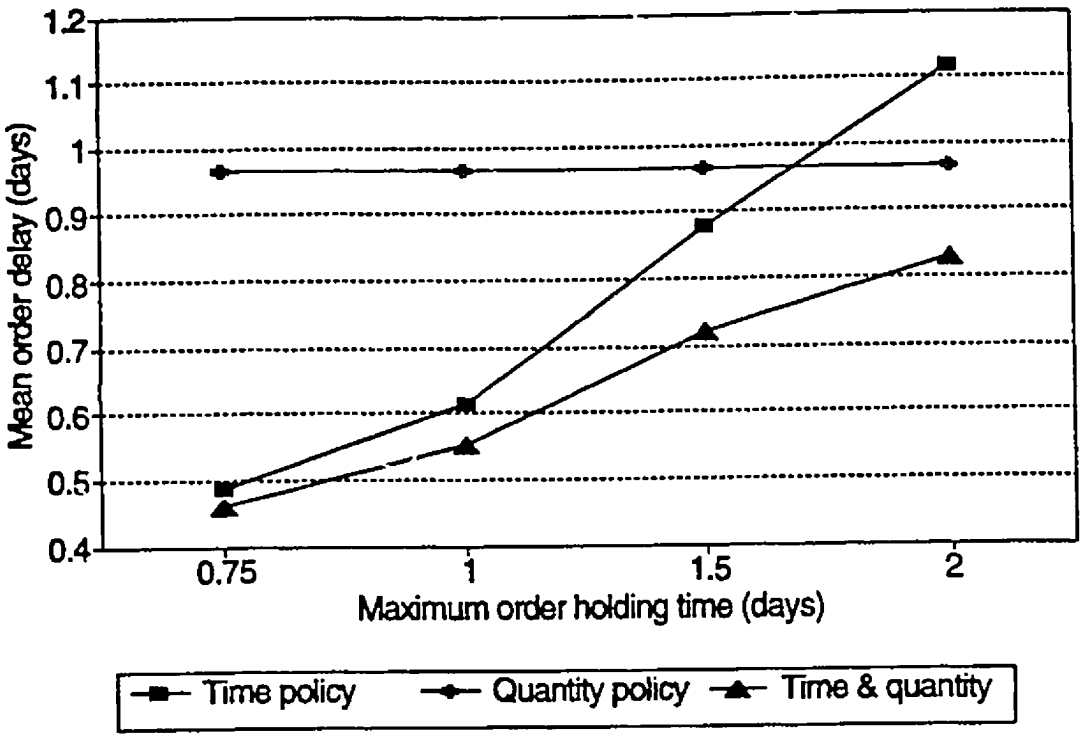


Figure 4–14
Comparison of Shipment–Release Policies: Mean Cost per Cwt.
Arrival Rate = 6.38 Orders Per Day

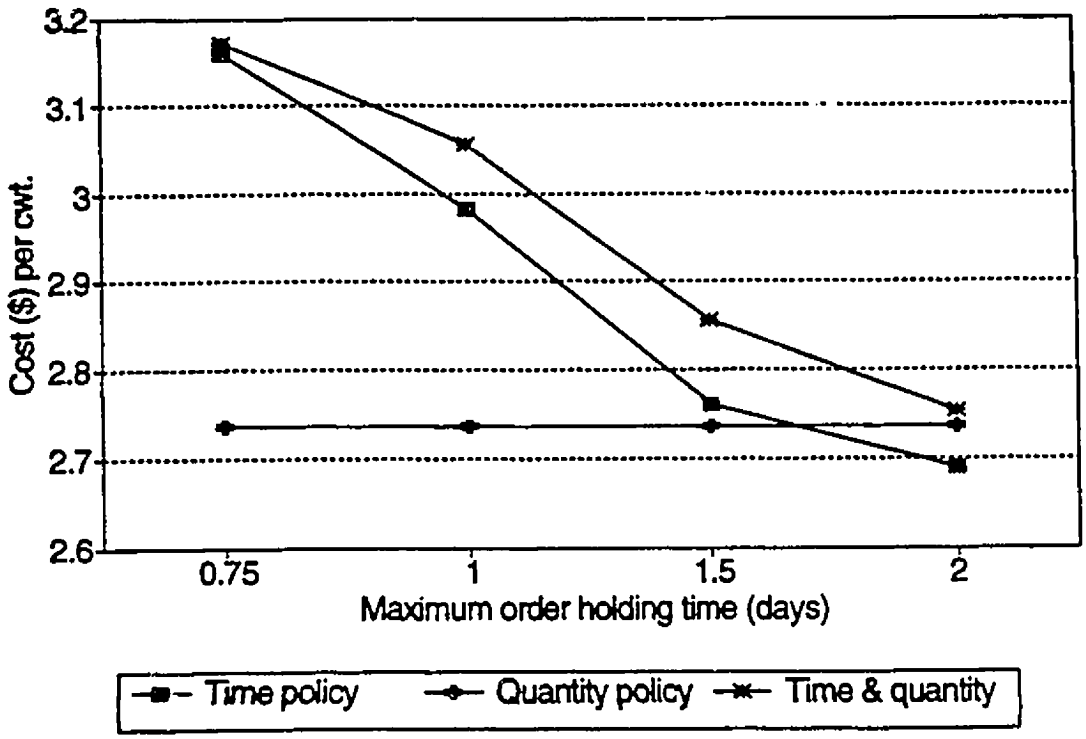


Figure 4–15
Comparison of Shipment–Release Policies: Mean Order Delay
Arrival Rate = 6.38 Orders Per Day

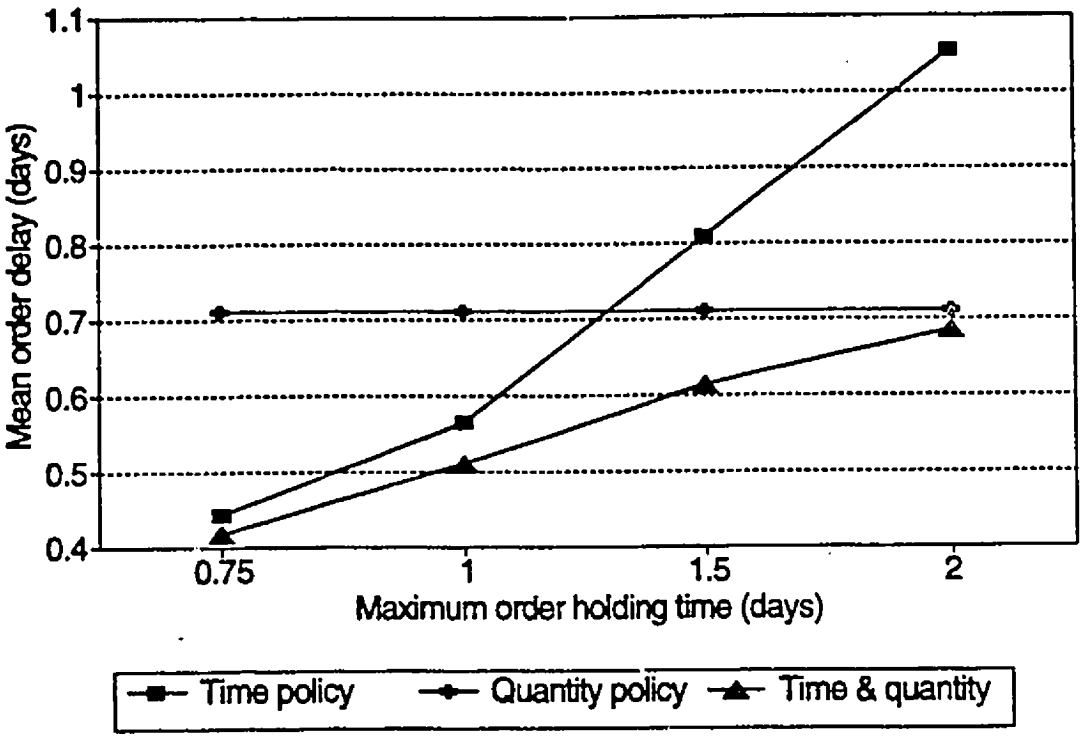


Figure 4–16
Comparison of Shipment–Release Policies: Mean Cost per Cwt.
Arrival Rate = 10.55 Orders Per Day

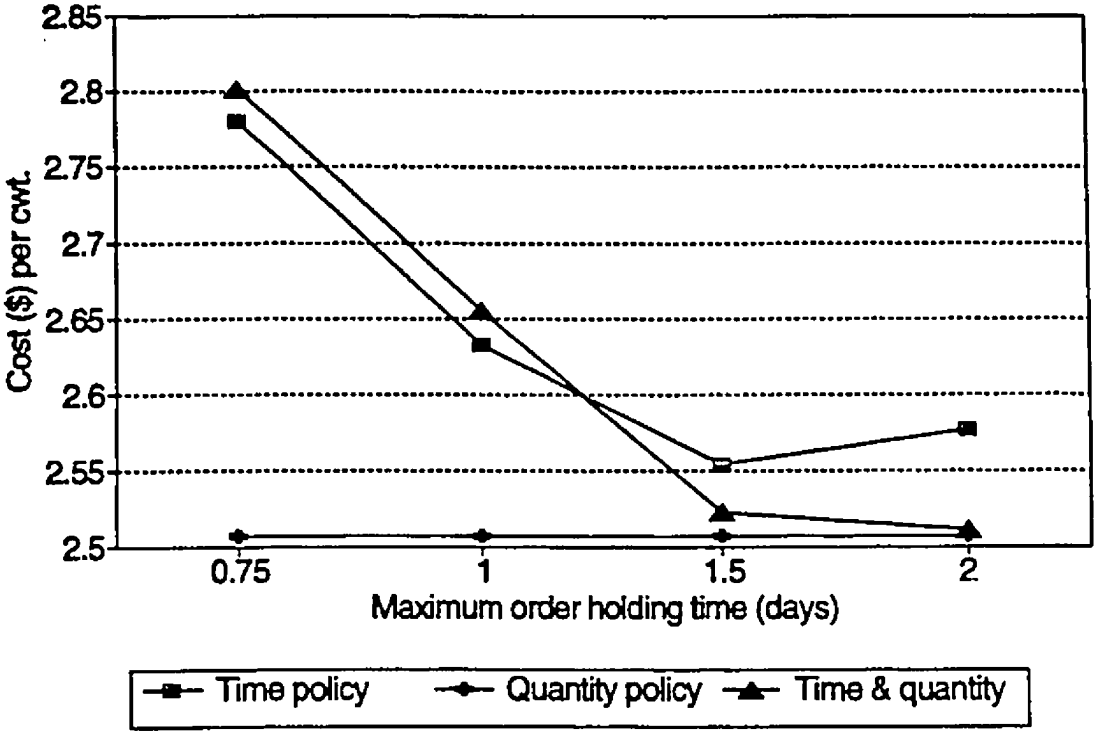
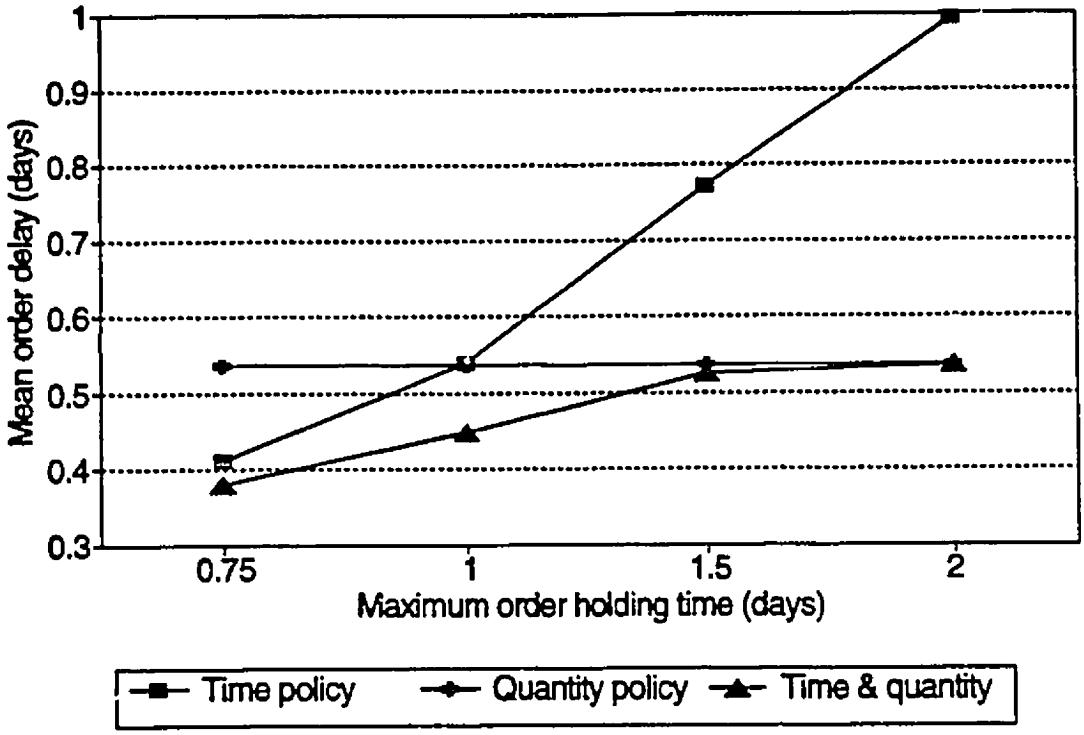


Figure 4–17
Comparison of Shipment–Release Policies: Mean Order Delay
Arrival Rate = 10.55 Orders Per Day



Chapter 5
NON-SEQUENTIAL APPROACHES FOR SETTING SHIPMENT-RELEASE PARAMETERS:DETERMINISTIC MODELS

5.1 Approaches To Setting Shipment–Release Parameters

The previous chapter examined the implications of three time– and/or quantity–based shipment–release policies on the mean cost and mean delay performance of a shipment consolidation program. We now seek to operationalize these policies by determining values for policy parameters.

We feel that there are two general types of analytical methods for determining how long customer orders should be held pending consolidation. **Non-sequential approaches** treat the shipment–release question as a "one–time" decision. This decision, which may be made once per cycle, once per week, once per year, or over any other time–frame, results in a set of targets or limits, such as maximum consolidated weight or maximum holding time.

Sequential approaches make the shipment–release decision as each order arrives. Use of a sequential approach seeks to answer the question, "do we dispatch a consolidated load now or do we wait for the next order to arrive?". Thus, shipments can be dispatched only at the time a customer order arrives. Other types of sequential policies exist that allow shipment–release at times other than the arrival of an order, however we do not study these models in this thesis. Note also that our use of the term "sequential" refers only to the series of "ship now or continue to consolidate?" decisions. This definition of a sequential decision process differs from that found in stochastic analysis.

Under a quantity–based shipment–release policy, both non–sequential and sequential methods function as discrete–time review systems. Non–sequential approaches are akin to a perpetual review or continuous–time system if a time policy or time–and–quantity policy is used. Sequential methods based only on time do not exist: use of a maximum holding time limit implies the existence of guidelines set by a non–sequential approach. Thus, combination non–sequential/sequential methods are possible. Indeed, to eliminate the possibility of lengthy order delays resulting from such a discrete–time review, use of a sequential approach within the continuous–time guidelines of a non–sequential approach may be preferred.

In general, a sequential method would be preferred when there is high uncertainty; for example, if order arrivals or order weights are extremely random or variable (ie., frequent bunching of arrivals followed by longer times without order arrivals), if the range of order sizes is large, or if the items to be consolidated are of a perishable nature. A non–sequential approach would be more efficient with perfect knowledge or when maximum holding time constraints are very important.

This thesis discusses the following non–sequential approaches for setting shipment–release parameters:

Chapter 5: deterministic models

- rules–of–thumb and simple heuristic methods
- Economic Shipment Quantity (ESQ) concept

Chapter 6: stochastic models

- probabilistic analysis of expected performance
- single–server bulk–service queue analysis
- stochastic clearing system analysis

Chapters 7 and 8 examine sequential methods for determining how long orders should be held for consolidation.

5.2 Rules–of–Thumb and Simple Heuristic Methods

Rules–of–thumb and other heuristic methods give satisfactory, but not necessarily optimal, solutions or courses of action. A good heuristic should be efficient, rapid, robust, easy to apply, easy to understand, and extendable to various settings.

Newbourne and Barrett [1972] state that the optimal holding time is approximately three days, and orders not consolidated within that time probably will remain unconsolidated if held longer. They add that "specific factors may shift this figure upward or downward somewhat, but usually only within a fairly narrow range". Their conclusion is supported only by the comment, "for reasons involving statistics and common business practices", however 60% of respondents to a survey by Jackson [1985] reported that their maximum holding time also is three days.

A major problem with heuristic methods is that they yield solutions of varying quality, and it may not be known how far from optimal the solution is. A good illustration of this is a heuristic, discussed by Ansari and Heckel [1987], used by Hewlett–Packard for determining the frequency of consolidated Just–in–Time deliveries.

Ansari and Heckel comment that in most companies, delivery frequency is determined by attempting to minimize the sum of all incremental costs associated with

or influenced by these deliveries. Their heuristic considers only transportation and inventory–holding costs, and seeks the optimal tradeoff between these costs.

The model is illustrated through a mathematical example that yields a suggested delivery frequency of once per day (or 260 deliveries per business–year). In fact, the strategy that produces the lowest cost tradeoff (determined by the economic shipment quantity concept, discussed in the next section) is to have 21.8 deliveries per year, or a delivery about every 12 days.

Ansari and Heckel's heuristic produces erroneous results due to inconsistent units. For example, their "average daily shipping cost" actually is the average cost per shipment regardless of time. Their "average daily carrying cost" in fact calculates the annual average inventory–holding cost given the present delivery frequency. Correcting their formulae would be simple; the corrected result would be similar to the economic shipment quantity expression discussed in the next section.

5.3 Economic Shipment Quantity (ESQ) Concept

This section discusses the concept of an economic shipment quantity (ESQ). Like the economic order quantity (EOQ), the ESQ uses deterministic analysis to derive the load size that minimizes the sum of transportation, inventory–holding, and possibly other costs. This idea is not new, having been applied in many distribution studies, including Brennan [1981], Burns et al. [1985], Blumenfeld et al. [1985], Hall [1987], Abdelwahab and Sargious [1990], and others.

The objective of this section is to show how the minimum–cost shipment size can be calculated under various scenarios. This shipment size then can be used as

the target quantity in a quantity or a time–and–quantity shipment–release policy. It will be seen, however, that after considering such factors as vehicle capacity, this target quantity may not equal the economic shipment quantity. Nevertheless, even in situations where the ESQ assumptions do not hold, a deterministic optimal load size can provide useful insight to the general feasibility of a shipment consolidation program.

The economic shipment quantity expression can be derived in several ways. We will apply renewal theory, suggested by Brennan [1981], to the case where all orders are of equal weight. Thus, the accumulated quantity at any time is simply the number, N, of orders, rather than the total weight. The ESQ based on weight will be discussed later.

We define an order–holding cycle to begin when the first order arrives at time t=0, and to end with the dispatch of a consolidated load of N orders. Let N(t) be the number of orders waiting for shipment at time t, $t \geq 0$. If T_n, a non–negative random variable with all T_i independently and identically distributed, is the time between arrival of order n and order (n+1), then the counting process $\{N(t), t \geq 0\}$ is a renewal process. We then can determine the minimum–cost number, ESQ, of orders to be included in a consolidated load, ignoring for now any customer service restrictions imposed on the maximum value of time t.

Assume that orders arrive at a shipper's facility according to a renewal process with mean arrival rate â and mean interarrival time 1/â. A cost of r_i dollars per order per unit time is incurred for delaying each order, and that the only significant costs to vary with dispatch frequency are related to inventory–holding and transportation.

Given that T_i is the time between the arrival of orders i and i+1, i≥1, the expected inventory–holding cost, C_I, of a cycle of N orders is:

$$E[C_I] = E[(N-1)r_I T_1 + (N-2)r_I T_2 + \ldots + 2r_I T_{N-2} + r_I T_{N-1}]$$

$$= r_I (1/\hat{a}) (N/2) (N-1)$$

The transportation cost of shipping N orders is:

C_T = fixed cost of transportation + variable cost per unit shipped times quantity shipped

$$= F_L + f_C L$$

where F_L is the fixed cost of arranging and/or dispatching a transportation vehicle of a given capacity, f_C is the common carrier freight rate for a given distance, and L is the total weight of the load of N orders.

The expected total cost of a cycle of N orders is:

$$TC = (F_L + f_C L) + (r_I (1/\hat{a}) (N/2) (N-1))$$

$$= F_L + f_C N E[W] + r_I (1/\hat{a}) (N/2) (N-1)$$

where E[W] is the expected weight of a customer order. The average cost per order per cycle is:

$$TC/N = (F_L + f_C N E[W] + r_I (1/\hat{a}) (N/2) (N-1)) / N$$

Differentiating this expression gives the optimal number, ESQ, of orders to ship such that the average cost per cycle per order is minimized:

$$ESQ = \sqrt{\frac{2 \hat{a} F_L}{r_I}}$$

The most important conclusion from the ESQ concept is that it is not necessarily most cost effective to dispatch fully–loaded vehicles. Moreover, the total cost curve, like that in EOQ analysis, is fairly flat around the minimum–cost quantity.

Thus, small deviations in order–release scheduling due to shipping constraints (such as restrictions on the timing of vehicle dispatches) will not have a major detriment on total cost.

Moreover, application of the ESQ formula may yield a non–integer number of orders. In this case, per–unit cost for the integer quantities immediately above and below the ESQ value should be compared, with the lower cost quantity used as the ESQ.

In reality, the target quantity, or actual number of orders to ship per cycle, will not necessarily equal the economic shipment quantity. Rather, the target quantity N^* will be the lesser of the ESQ and vehicle capacity H (here, both N^* and H are expressed as number of orders); that is, N^* = minimum {ESQ, H}. The expected length of an order–holding cycle will be the time until the target number of orders has arrived (that is, $E[T] = (N^*-1)/\hat{a}$), subject to customer service constraints. Of course, if N^* is less than one, orders should be shipped individually without consolidation and without incurring inventory–holding costs.

We now extend the economic shipment quantity concept to include some practical considerations.

ESQ Under Private Carrier Rates

The economic shipment quantity for shipper–performed consolidation with private carriage (System 1a/P of Section 1.5) is:

$$ESQ = \sqrt{\frac{2\,\hat{a}\,(F_D + F_S + f_P M)}{r_I}}$$

where F_0 is the fixed cost of a vehicle dispatch, F_s is the fixed cost of making a customer stop, f_p is the private carrier transportation cost per unit distance, and M is the total supplier–to–customer round–trip distance. $\hat{a}$ and r_I are as defined in the previous section.

All variables being equal, private carrier will result in a larger economic shipment quantity than will common carrier. A shipper's linehaul transportation cost under common carriage is based mainly on quantity; with private carriage, it is based mainly on distance. Thus, for a given distance, shipping by private carrier typically yields a larger fixed transportation cost component per load than does common carriage. This illustrates the statement that common carrier often is more cost effective than private carrier for systems with small total throughput (Firth et al. [1988]).

ESQ Based On Shipment Weight

Our expression for ESQ states the economic shipment quantity in terms of number of orders, thus assuming that all orders are of equal weight. Vehicle capacity and common carrier freight rates, however, usually are based on weight or volume, and there is little relationship between these physical characteristics and the number of orders. We now present a simple modification of the ESQ formula to consider shipment weight under common carriage.

Denote the mean weight per order by E[W] and the total accumulated weight by TW. Thus, $r_I = r_W$ E[W], where r_I is the one–period inventory–holding cost per order and r_W is the one–period holding cost per pound. The economic shipment weight is obtained by replacing N with TW/E[W] and r_I with r_W E[W] in the expression for

average cost per cycle; â remains the mean arrival rate per order. Thus, the economic shipment weight, ESW, under common carriage is:

$$ESW = \sqrt{\frac{2\,\hat{a}\,F_s\,E[W]}{r_w}}$$

The target weight, W*, will be the lesser of the ESW and vehicle capacity H, where H is expressed as a weight; that is, W* = minimum {ESW, H}.

Because our formula for ESW does not require that all customer orders be of equal weight, this expression bridges the gap between simulation studies (many of which consolidate by weight without determining an economic shipment quantity) and analytic studies (which often calculate an ESQ but do not consolidate on the basis of weight).

ESQ Based On Volume

Determination of the optimal shipment quantity also must consider the volume capacity of the vehicle because a vehicle's volume capacity is often exhausted before its weight capacity is reached. Although we recognize the effect of volume capacity on the optimal shipment quantity, we do not pursue this constraint because charges for transporting freight normally are based on weight rather than on volume, and there is little relationship between shipment volume and weight. Also, it is not common practice to calculate inventory–holding cost based on volume. Lastly, a vehicle physically may be full before reaching its volume capacity if items are of irregular shapes. Thus, the upper bound on the number of orders to be shipped in a vehicle is not the volume capacity, but instead the optimal solution derived from a three–dimensional bin–packing algorithm.

Optimal Shipment Quantity With Transportation Weight Breaks

Section 2.4 noted that transportation cost benefits from shipment consolidation occur from spreading fixed transportation costs over larger load sizes and/or reductions in freight rate as total load weight increases. Our development of the common carrier ESQ formula considered only the former, thus assuming a fixed freight rate. We now consider the impact of common carrier rate weight breaks on the optimal order size.

As discussed in Section 2.4, shippers may apply the bumping clause by declaring heavier weights than actually exist to push the total shipment weight into a heavier weight bracket, thus qualifying for volume freight rates. The minimum weight ("WBT") at which this practice is cost–effective equals the minimum volume weight ("MWT") times the ratio of the volume freight rate f_V and the non–volume freight rate f_N, $f_V \leq f_N$; that is, WBT = MWT (f_V/f_N), WBT≤MWT. At this weight, WBT f_N = MWT f_V.

Example: Suppose that shipments weighing less than MWT=20000 lbs. are charged f_N=\$3.00 per cwt., and those above this minimum weight are charged f_V=\$2.25 per cwt. If a shipper ships 17500 lbs., the freight cost would be \$525, or \$3 per cwt. Under the bumping clause, the shipper would declare the weight of the shipment to be the minimum volume weight, 20000 lbs. Total freight cost would then be 20000 lbs. x \$2.25/cwt., or \$450 (\$2.57 per cwt. for the 17500 lbs. actually shipped). Over–declaring load weight would be cost–effective for weights at least WBT = 20000 (2.25/3.00) = 15000 lbs. Figures 5–1 and 5–2 illustrate the effect of over–declaring load weight on total and per–unit transportation cost without fixed cost. ■

Figure 5–1
Effect of Transportation Weight Breaks
On Transportation Cost Per Load

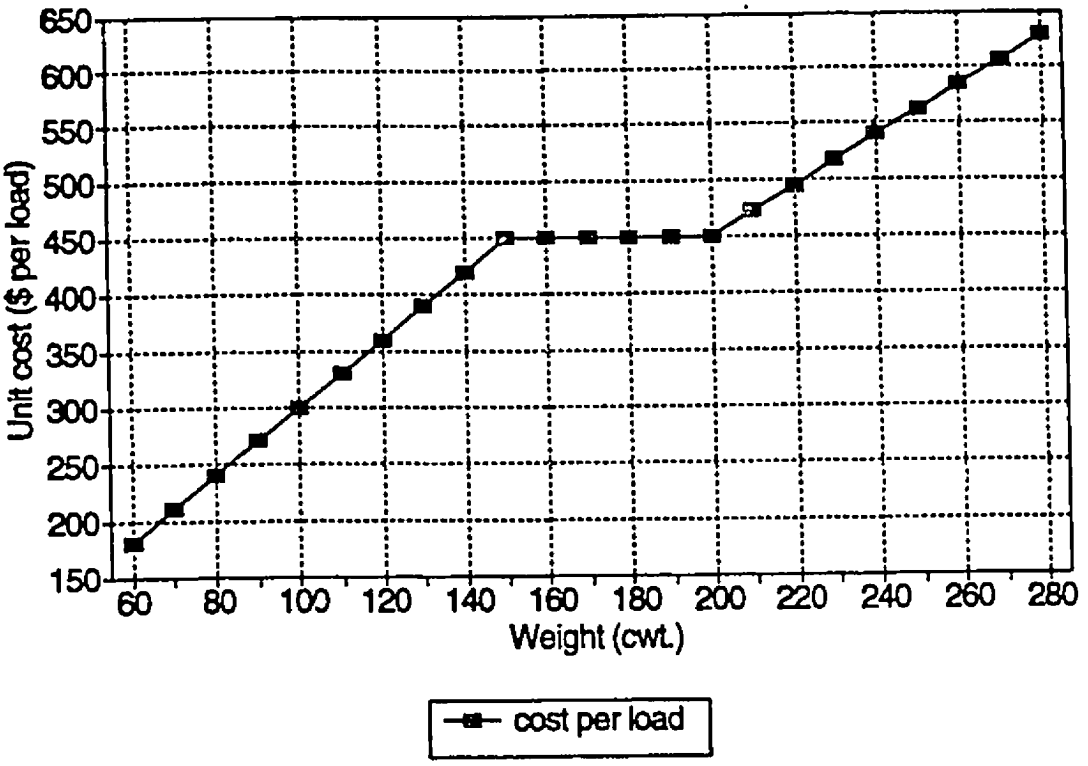


Figure 5–2
Effect of Transportation Weight Breaks
On Transportation Cost Per Cwt.

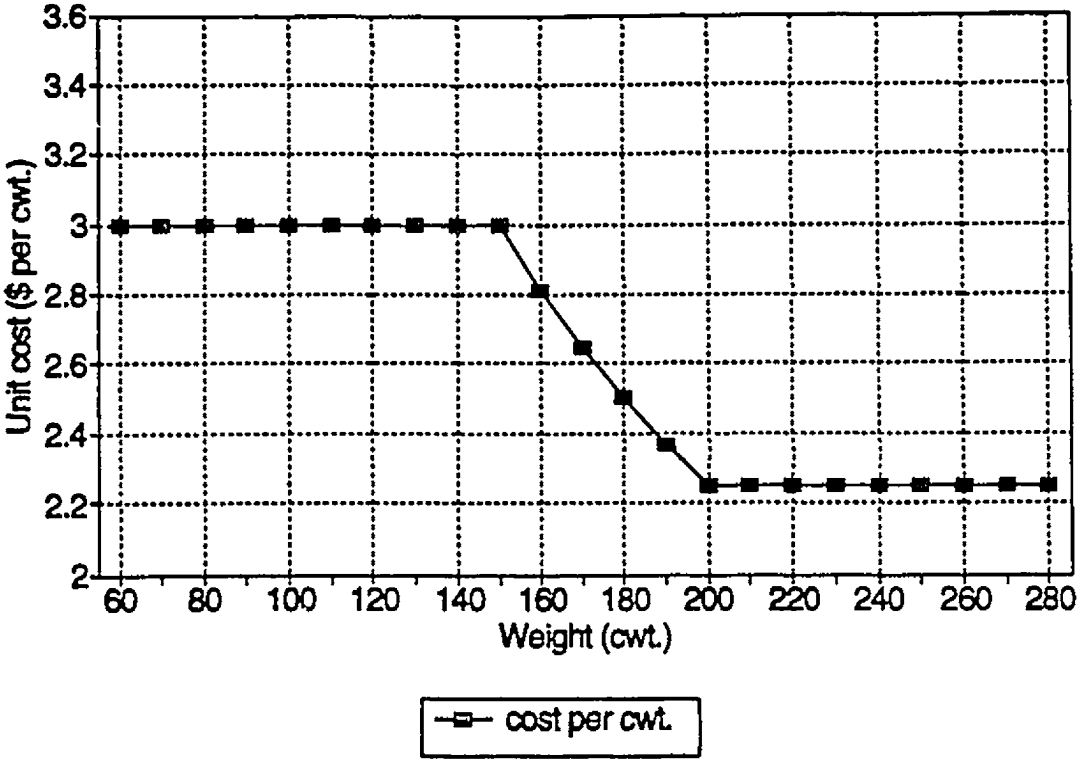


Figure 5–2 illustrates that, for weights between WBT and MWT (in the above example, 15000 and 20000 pounds respectively), over–declaring load weight decreases per–unit transportation cost. More important, we see that there is no transportation cost–advantage to shipping loads larger than the minimum volume weight (20000 pounds in Figure 5–2). Thus, under common carriage with no fixed transportation cost, orders should never be held for longer than is required to accumulate the minimum volume weight; doing so yields no transportation cost advantage, but incurs inventory–carrying costs and deteriorates customer service. Newbourne and Barrett [1972] agree: "as a general rule, the minimum truckload or carload weight will comprise the upper limit of the single–destination consolidation that will be economically desirable". An exception is if orderarrivals are such that small orders frequently are stranded and, as a result, later must move under non–volume rates.

Applying the economic shipment quantity concept when carrier weight breaks exist is similar to the use of the EOQ model with volume discounts (see, for example, Tersine, Larson, and Barman [1989]; Chase and Aquilano [1992]). In this EOQ model, the optimal order quantity occurs at either the EOQ or the smallest quantity of the volume discount category that yields minimum total cost.

Figures 5–1 and 5–2 show that the existence of one freight rate weight break results in three weight ranges: below the WBT weight, above MWT, and between WBT and MWT. When determining the most cost–efficient consolidated weight, however, not all weight ranges are relevant. We consider the cases with and without fixed transportation costs. It is important, however, that we first clarify our meaning of terms used in the next section and in the remainder of the thesis.

We will use the terms ESQ (economic shipment quantity) and ESW (economic shipment weight) to denote the number of orders or total weight that minimizes the sum of per–unit transportation and inventory–holding costs, *calculated without regard to vehicle capacity or carrier rate discounts*. This, in fact, is the way that the ESW was determined in Chapter 4 (see Section 4.4). Thus, the ESW is calculated by applying one of the formulae discussed earlier in this section.

We define the "minimum–cost quantity/weight" as the quantity/weight that minimizes total per–unit cost after considering carrier rate discounts, but before considering vehicle capacity. We will use the terms "minimum–cost quantity/weight" and "optimal shipment quantity/weight" interchangeably. The target quantity, N^*, or target weight, W^*, then will equal the smaller of the minimum–cost quantity/weight and vehicle capacity. Clearly, if vehicle capacity constraints are not binding, the target quantity/weight equals the minimum–cost quantity/weight.

Optimal Shipment Quantity With Transportation Weight Breaks and No Fixed Costs

Without fixed transportation costs, the ESW formula cannot be used. As noted previously, in this case, the optimal shipment weight under common carriage will never exceed the minimum volume weight. It may, however, be less depending on the order arrival rate and inventory–holding cost. It can be shown that weights between WBT and MWT will never yield a lower per–unit cost than both direct shipping or consolidating the minimum volume weight; this is seen in Figures 5–3 and 5–4. Thus, the weight range between WBT and MWT is not relevant to the dispatch decision when fixed costs do not exist. As a result, if the minimum per–unit cost does not

occur at the minimum volume weight MWT, it will result from not consolidating at all, yielding zero inventory–holding cost.

The minimum–cost weight will be the minium volume weight MWT if the expected per–unit cost of shipping MWT is less than the expected per–unit cost of shipping non–consolidated loads; that is, if:

$$\frac{r_W}{2\hat{a}} \frac{MWT - E[W]}{E[W]} + f_V < f_N$$

where f_V and f_N are the volume and non–volume rates per pound respectively. If this inequality does not hold, orders should be shipped directly without consolidation.

Example: Assume an arrival rate of â=3 orders per day, expected order weight of E[W]=2000 pounds, freight rates of f_V=\$2.25/cwt. and f_N=\$3.00/cwt., and minimum volume weight MWT=20000 pounds. Thus, WBT=15000 pounds.

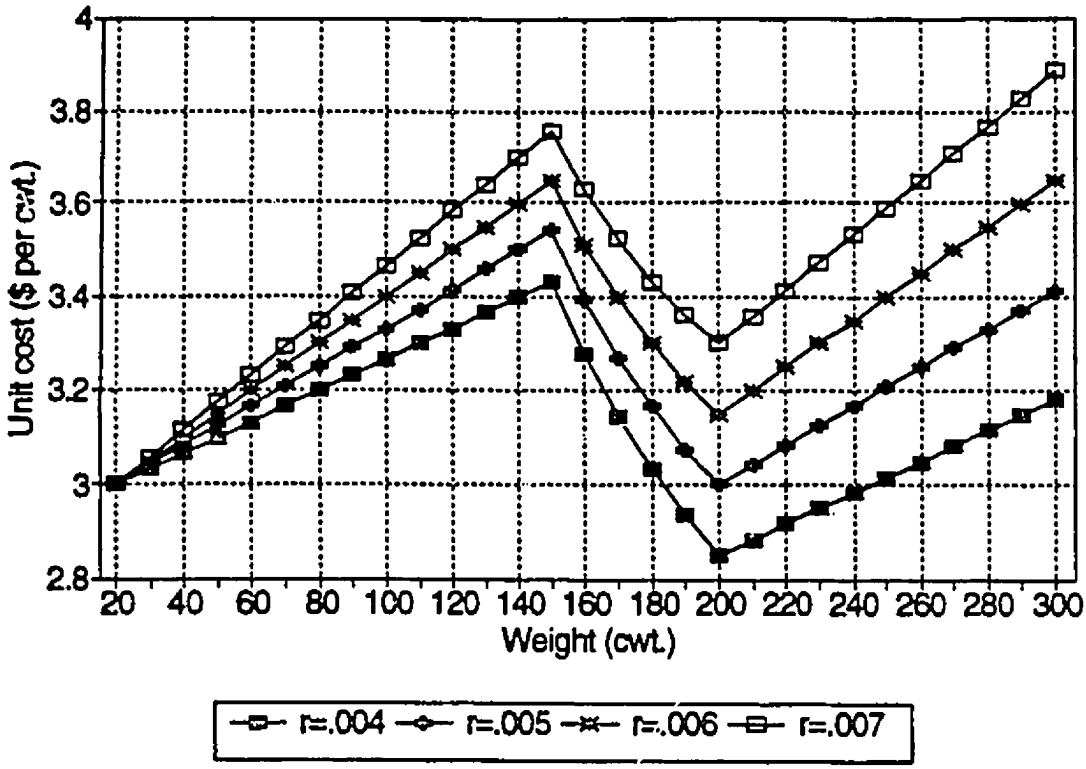
Rearranging the above inequality gives the break–even inventory–holding cost:

$r_W = 2\hat{a}\ E[W]\ (f_N - f_V)\ /\ (MWT - E[W])$

= \$ 0.005 per pound per day

For inventory cost r_W less than \$0.005 per pound per day, orders should be held until the minimum volume weight is reached; otherwise, all orders should be shipped directly. This is illustrated in Figure 5–3, which shows that, with r_W=0.004 per pound per day, the minimum per–cwt cost occurs at the minimum volume weight of 20,000 pounds. With r_W=0.006 and r_W=0.007, the minimum per–cwt. cost occurs at the expected order weight of 2000 pounds, implying that consolidation should not be performed. ■

Figure 5–3
Example: Optimal Shipment Weight With Transportation
Weight Breaks And No Fixed Costs
MWT=200 cwt.; WBT=150 cwt.



Optimal Shipment Quantity With Transportation Weight Breaks and Fixed Costs

Our algorithm for determining the minimum per–unit cost weight with one weight break and fixed transportation costs is as follows:

Step 1: Calculate the economic shipment weight (ESW), as discussed previously (thus ignoring transportation weight breaks).

Step 2: If the economic shipment weight ESW is equal to or greater than the minimum volume weight MWT, stop: the ESW will yield the lowest per–unit cost. Otherwise, go to Step 3.

Step 3: Calculate WBT, the smallest weight at which a shipment weighing less than the minimum volume weight can be transported under volume freight rates, as discussed previously.

Step 4: If ESW≥WBT, stop: the lowest per–unit cost occurs at the minimum volume weight MWT. Otherwise, go to Step 5.

Step 5: Calculate the expected per–unit cost for the minimum volume weight MWT and for the economic shipment weight ESW. The optimal shipment weight is the one that yields the lower per–unit cost.

Steps 3 and 4 are not necessary. However, they can be performed much quicker than can Step 5, thus speeding the algorithm when the economic shipment weight is between WBT and MWT.

For ESW≤MWT, MWT will yield a lower per–unit cost than ESW if:

$$\frac{r_W}{2\hat{a}}\frac{MWT - ESW}{E[W]} - F_L\,(ESW^{-1} - MWT^{-1}) \le f_N - f_V$$

The first term of the left–side is the inventory cost in dollars per pound per load, the second term is the per–load fixed transportation cost per pound, and the right–side is the variable transportation cost per pound.

If the above inequality does not hold, then from a cost standpoint, orders should be shipped individually rather than consolidated.

Example: As in the prior example, assume â=3 orders per day, E[W]=2000 pounds, f_V=\$2.25/cwt., f_N=\$3.00/cwt., MWT=20000 pounds, and WBT=15000 pounds. We now add a fixed transportation cost of \$30. Results for various values of the inventory–holding cost parameter r_w are illustrated in Figure 5–4. These are summarized below, where ESW is in pounds, and "LS" and "RS" refer to the values of the left–side and right–side of the inequality above.

r_w	ESW	LS	RS	shipping decision
.0007438	22,000	n/a*	n/a*	ship ESW
.0014062	16,000	.00009	.0075	ship MWT
.0036	10,000	.0015	.0075	ship MWT
.01	6,000	.008167	.0075	ship ESW

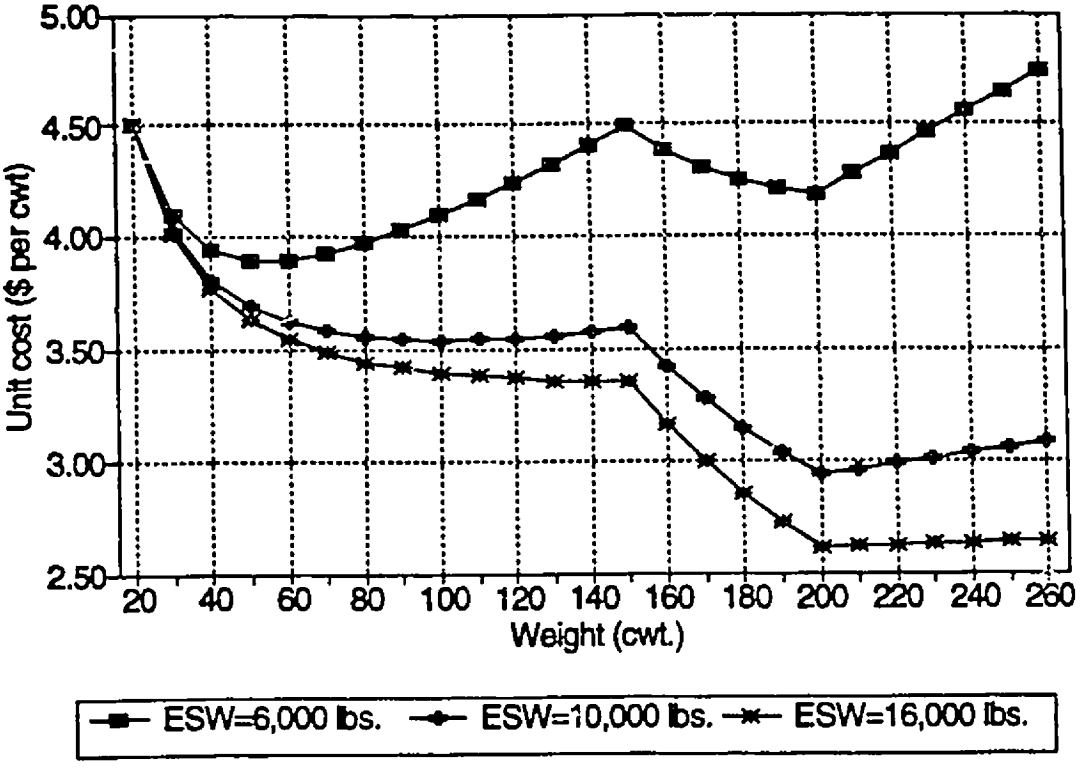
* "n/a" because ESW≥MWT; see Step 2 of above algorithm ■

Summary: Optimal Shipment Quantity With Transportation Weight Breaks

We now summarize the above conclusions regarding optimal shipment weight W* for the case with one weight break and no binding capacity constraints. First, when fixed costs do not apply, the economic shipment weight cannot be calculated. Then, depending on values of the cost parameters, the minimum per–unit cost will occur either at the minimum volume weight or by not consolidating at all.

If fixed costs exist, then, depending on values of the cost parameters, the optimal shipment weight will equal either: i) the economic shipment weight ESW such that 0<ESW≤WBT; ii) the minimum volume weight MWT; or iii) or the ESW such that ESW>MWT. The first two possibilities are seen in Figure 5–4 for ESW's of 6000 lbs. (where ESW<WBT) and 16000 lbs. (where WBT<ESW<MWT) respectively.

Figure 5–4
Example: Optimal Shipment Weight With Transportation
Weight Breaks and Fixed Costs
MWT=200 cwt.; WBT=150 cwt.



We add that Zahir and Sarker [1991] and Russell and Krajewski [1992] have proposed algorithms for calculating the economic shipment quantity with supplier purchase discounts and transportation weight breaks. Their minimum–cost shipment quantity occurs at either a purchase discount or a transportation weight break.

5.4 Conclusions

This chapter has introduced non–sequential methods for determining how long customer orders should held for consolidation. We showed that a deterministic economic shipment quantity can be calculated, and presented several variations of the basic ESQ formula. By considering this ESQ as well as practical considerations such as vehicle capacity and carrier weight breaks, a target shipment quantity can be determined. This target can then be used in a quantity or a time–and–quantity shipment–release policy.

The main problem with any target shipment quantity is that it may or may not be accumulated within an acceptable holding time. Many shippers attempt to include customer service considerations by incorporating both target quantity and elapsed waiting time in the shipment–release decision. This results in a time–and–quantity policy.

The arrival time of a customer order and its weight are, however, random variables. Thus, use of a maximum holding time forces consideration of the probability that an efficient consolidated load can be attained within the remaining consolidation cycle. As seen in Chapter 4, poor choice of a maximum time can lead to high transportation costs from unreasonably small loads or excessive inventory–holding

costs from long delays. Chapter 6 discusses stochastic non–sequential approaches for determining shipment–release timing, including methods for setting the value of the maximum delay ("oldest age") any order can incur for consolidation.

Chapter 6
NON-SEQUENTIAL APPROACHES FOR SETTING SHIPMENT-RELEASE PARAMETERS:STOCHASTIC MODELS

6.1 Introduction

Chapter 5 discussed deterministic methods for calculating the target quantity in quantity– and time–and–quantity shipment–release policies. We now extend our study of non–sequential approaches to include consideration of maximum holding time under stochastic conditions. Thus, the models developed in this chapter apply to all three shipment–release policies discussed in Chapter 4.

This chapter examines the following stochastic models for determining how long orders should be held for consolidation:

- probabilistic analysis of expected performance
- single–server bulk–service queue analysis
- stochastic clearing system analysis.

We begin by discussing simple probabilistic approaches to modeling shipment consolidation.

6.2 Probabilistic Analysis of Expected Performance

Probabilistic analysis of expected load size, expected total cost, and expected waiting time has two major benefits. First, shipment consolidation is examined from a more realistic stochastic standpoint. Second, customer service can be considered explicitly by including expected or maximum waiting times in the models.

Unfortunately, some disadvantages exist. Knowledge of the probability distributions of order arrival time and order size are required. As well, mathematical complexities may result when deriving or using these distributions.

We begin our examination of analysis of expected performance with the general case where consolidation is based on numbers of orders; that is, all orders are of an equal weight. These models are extended later in this section to consider accumulated weight.

Visual Inspection Of Cumulative Probability Plots

Assume that customer orders arrive at the shipper's facility according to a Poisson process with known arrival rate â, and that the economic shipment quantity N^* is less than vehicle capacity. We define $Pr\{N^*(t)\}$ to be the probability of accumulating at least N^* orders in time t.

Recall from Section 5.3 that: i) management is assumed to be aware of both the order arrival rate â and the optimal number of orders N^*; and ii) an order accumulation cycle begins with the arrival of the first order. Thus, we seek a value for T_{MAX} such that the probability $Pr\{N^*(T_{MAX})\}$ of accumulating at least the optimal number N^* of orders in time T_{MAX} is greater than or equal to some management–set parameter (we will ignore for now the value of this policy parameter). Because one order already is waiting, our analysis reduces to the simple calculation of cumulative Poisson probabilities of at least N^*-1 additional order arrivals in time T_{MAX}. Management consideration of these probabilities can provide insight to waiting times that yield reasonable tradeoffs between expected shipment size and maximum holding time.

Example: Assume that orders arrive at rate â=3 per day, and that management has calculated the optimal shipment quantity N^*. Figure 6–1 plots the cumulative Poisson probabilities of at least N^*-1 more orders arriving within the period $[0,T_{MAX}]$ for various values of T_{MAX}, given that the first order arrived at time T=0.

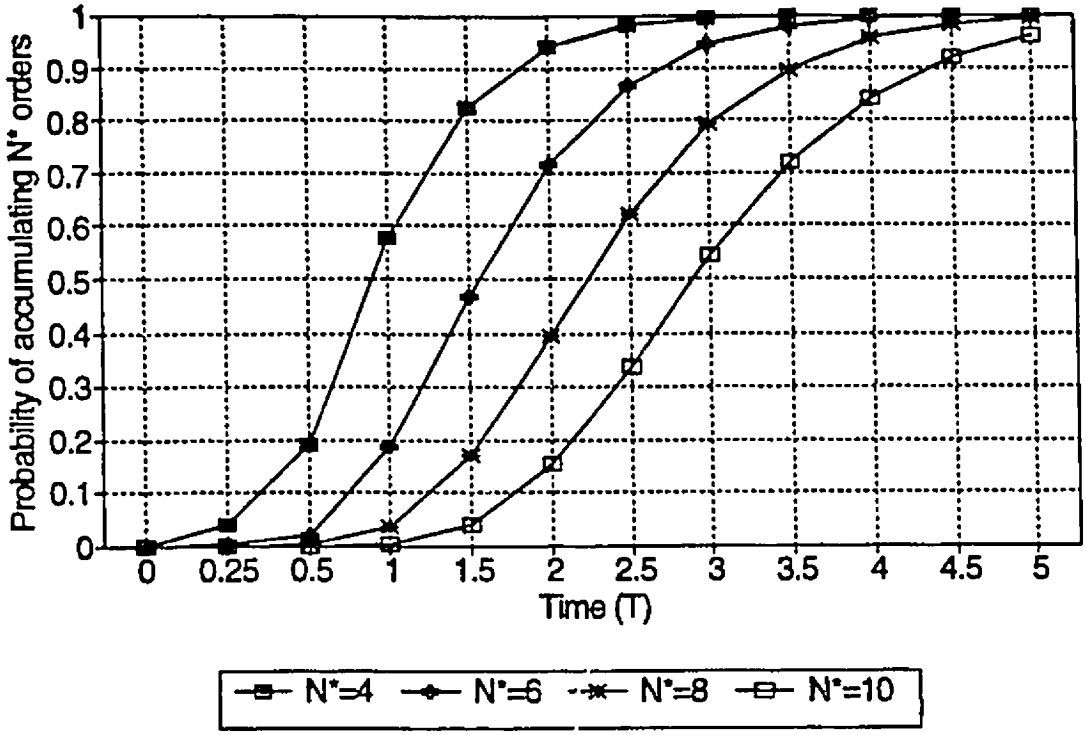
Visual inspection of Figure 6–1 shows that for $N^*=6$, the value of $Pr\{N^*(t)\}$ increases considerably from time T=1 to T=2, less from T=2 to T=2.5, and comparatively little from T=2.5 onwards. Thus, with an optimal shipment quantity of $N^*=6$, management should set T_{MAX} at a value not exceeding T=2.5. Moreover, for $N^*\leq 6$, if management seeks a $Pr\{N^*(t)\}$ value of, say, at least 85%, orders should not be held more than 2.5 days. Note that this does not mean that a vehicle load should be dispatched every 2.5 days; rather, it denotes the maximum age of the oldest order.■

Clearly, the decision following from the above analysis is subjective. Moreover, some values for Pr{N(t)} that are acceptable to management may result in high inventory–holding costs. However, this approach is fairly easy to understand and apply, and other criteria can be considered simultaneously. We will use cumulative probabilities of order arrivals and order weight in the following sections.

Analysis of Expected Order Cycle Length

If orders arrive according to a Poisson process with mean â, the inter–arrival times are exponentially–distributed with mean β=1/â. Therefore, for a consolidation cycle of N^* orders (where N^* is the optimal shipment quantity), the maximum time that any order must wait is (N^*-1)–Erlang distributed. Because N^*-1 is an integer, this $(N^*-$

Figure 6–1
Probability of Accumulating N* Orders in Time T
With Poisson(â=3) Order Arrivals



1)–Erlang distribution is identical to the gamma distribution with parameters (α=N*–1,β=1/â). The gamma distribution is denoted as Ga(α,β).

Important properties of the gamma distribution are given in Table 4–3. The gamma probability density function and cumulative distribution functions for the length T of a consolidation cycle are:

$$f_T(t) = \frac{\beta^{-\alpha} t^{\alpha-1} e^{-t/\beta}}{\Gamma(\alpha)} \qquad t \geq 0$$

$$F_T(t) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} \int_0^t x^{\alpha-1} e^{-x/\beta} dx \qquad t \geq 0$$

where the arrival of the first order occurs at time t=0.

If α is a positive integer, the gamma cumulative distribution function simplifies to:

$$F_T(t) = 1 - e^{-t/\beta} \sum_{k=0}^{\alpha} (t/\beta)^k / k! \qquad t \geq 0$$

We can use this distribution to calculate, for a specified maximum holding time T_{MAX}, the expected time that customer orders must wait. The expected transportation and inventory–holding costs per order also can be estimated.

Our analysis is simplified by the relationship between the gamma and Poisson distributions. If N, the number of events, is a Poisson random variable with mean â, the Poisson probability that N≥k in time T is equal to the gamma probability that time T≤t is required for the N events. Thus, T is Ga(α=k,β=1/â). For example, with â=3, the Poisson probability of N≥k=5 or more events in time t=1 is 0.1847. This equals the Ga(α=k,β=1/â) probability that the time T required for N events does not exceed t=1. This relationship is applied in the following example.

Example: Assume that the order arrival rate is $\hat{a}=3$ per day, and the target quantity is $N^*=6$ orders. Given that the first order arrives a time t=0, the expected time required to accumulate N^* is $(N^*-1)/\hat{a} = 1.667$ days.

Suppose that management now states that no order must wait more than $T_{MAX}=1$ day for consolidation. From the cumulative gamma distribution with $\alpha=N^*-1$ and $\beta=1/\hat{a}$, the probability that the consolidation cycle will last no more than one day is $Pr\{T \le T_{MAX}\}=0.1847$. As noted above, this is the same as the Poisson($\hat{a}$) probability that at least $N^*-1=5$ additional orders will arrive in one day. This Poisson probability can be read from the curves plotted in Figure 6–1 for $N^*=6$ and T=1.

There are two interpretations to this probability. With $T_{MAX}=1$ day, 18.47% of loads dispatched will contain the target quantity N^*. Also, if management seeks to dispatch only loads containing the economic shipment quantity, 18.47% of loads will be dispatched within one day; the other 81.53% will require more than one day to accumulate N^*.

Given a maximum delay of $T_{MAX}=1$ day and a target quantity of $N^*=6$ orders, the expected cycle length is:

$$
\begin{aligned}
E[T \mid T_{MAX}] &= E[T|T>T_{MAX}]\ Pr\{T>T_{MAX}\} + E[T|T\le T_{MAX}]\ Pr\{T\le T_{MAX}\} \\
&= T_{MAX}\ Pr\{T>T_{MAX}\} + E[T \mid T\le T_{MAX}]\ Pr\{T\le T_{MAX}\} \\
&= 1\ (0.8153) + \int_0^{Tmax} t\ \frac{\beta^{-\alpha}\ t^{\alpha-1}\ e^{-t/\beta}}{\Gamma(\alpha)}\ dt \\
&= 1\ (0.8153) + \int_0^{1} t\ \frac{3^5\ t^4\ e^{-3t}}{\Gamma(5)}\ dt \\
&= 0.9552 \text{ days}
\end{aligned}
$$

Similar analysis yields the expected number of orders accumulated during a consolidation cycle with a maximum holding time of one day. This simply is the expected number of arrivals within the holding time, limited to a maximum quantity of N^*-1:

$$\begin{aligned} E[N \mid T_{MAX}] &= 1 + \sum_{N=0}^{N^*-2} N \, Pr_{Poisson}\{N \text{ arrivals in } T_{MAX}\} \\ &\quad + (N^*-1) \, Pr_{Poisson}\{N \geq N^*-1 \text{ in } T_{MAX}\} \\ &= 1 + \sum_{N=0}^{N^*-2} N \, Pr_{Poisson}\{N \text{ arrivals in } T_{MAX}\} \\ &\quad + (N^*-1) \, Pr_{Gamma}\{T \leq T_{MAX}\} \\ &= 1 + 1.9417 + 5 \, (0.1847) \\ &= 3.8652 \text{ orders} \end{aligned}$$

The addition of 1 accounts for the first order to arrive, thereby starting the waiting time cycle. Note that deterministic analysis with $T_{MAX}=1$ would yield an expected accumulated size of $(1+\hat{a}T_{MAX})=4$ orders. This difference occurs because deterministic analysis assumes that all cycles are of length T_{MAX}, thus ignoring cycles that end at T_{MAX} without having attained the target weight. However, for large values of N^* relative to $\hat{a}T_{MAX}$, $Pr\{N \geq N^*-1 \text{ in } T_{MAX}\}$ approaches zero, thus the third component of the above expression does also, and $E[N|T_{MAX}]$ approaches the deterministic value of $(1+\hat{a}T_{MAX})$.■

Maximization of System "Gain"

We can extend the above analysis to find the value for T_{MAX} that yields the lowest cost per unit time. Consider shipment consolidation as an infinite–duration single–state process. As long as orders continue to be accumulated, we remain in the

single state. The dispatch of a consolidated load causes a transition outside of the state, followed by an instantaneous return to it.

The "gain" of the system is defined as the average reward per unit time if the process is allowed to operate indefinitely. We seek to minimize system "loss", or average cost per unit time. This objective is common in infinite–horizon Markov decision processes (discussed in Chapter 8 of this thesis). Howard [1971b] discusses gain criteria in detail.

The gain g(t) of this process is given by:

$$g(T_{MAX}) = E[TC \mid T_{MAX}] / E[T \mid T_{MAX}]$$

where $E[TC|T_{MAX}]$ is the expected cost of an order accumulation cycle, and $E[T|T_{MAX}]$ is the expected length of this cycle (i.e., expected time spent in the single state, or expected time between transitions), both given that the holding time is limited to a maximum length T_{MAX}. We seek the value of T_{MAX} such that $g(T_{MAX})$ is minimized.

The calculation of $E[T|T_{MAX}]$ was discussed and illustrated in the previous section. We now consider the calculation of expected cost per cycle. This cost calculation also will be used in a later section.

In Section 5.3, we noted that the total inventory–holding cost for a consolidation cycle of N^* orders was:

$$TC_R = r_I \, (1/\hat{a}) \, (N/2) \, (N-1)$$

If we dispatch N^* orders at time t, these orders arrived at an average rate of $(N^*-1)/t$ per unit–time. This can be seen by recalling that the expected length of a consolidation cycle was:

$$E[T] = (N^*-1) / \hat{a}$$

Thus, if the first order arrives at T=0, then t=E[T] denotes the time of a load dispatch, and $\hat{a}=(N^*-1) / E[T] = (N^*-1) / t$.

Substituting this value for $\hat{a}$ in the above expression for inventory–holding cost yields the inventory cost of holding N^* orders from time t=0 until a vehicle dispatch at time t>0:

$$TC_R = \frac{r_i\, t}{N^*-1} (N^*/2) (N^*-1)$$

$$= \frac{r_i\, t\, N^*}{2}$$

This expression is slightly easier to use than that for TC_R given earlier. Notice the similarity between it and that in the inventory–management literature.

Consider a time–and–quantity strategy: a consolidated load will be dispatched at the earlier of: i) the accumulation of a target number N^* of orders; or ii) the expiry of a maximum holding time T_{MAX}. A load will be dispatched at any time $t<T_{MAX}$ only if it consists of the target quantity. Thus, the total transportation and inventory cost for a cycle of length $t \leq T_{MAX}$ is:

$$TC = \text{transportation cost} + \text{inventory cost}$$

$$= F_L + (1/2)(N^*\, r_i\, t)$$

where F_T is the total transportation cost per load for a given distance.

A load dispatched at any time $t > T_{MAX}$ will contain N orders, $N<N^*$. The total cost for any cycle of length $t > T_{MAX}$ will be:

$$TC = \text{transportation cost} + \text{inventory cost}$$

$$= F_L + (1/2)(N\, r_i\, t)$$

Thus, the expected cost per cycle, given a maximum length of T_{MAX}, is:

$$E[TC \mid T_{MAX}]$$

$$= \int_0^{Tmax} f(t)\,[F_L + (1/2)(N\ r_i\ t)\ dt] + \int_{Tmax}^{\infty} f(t)\,[F_L + (1/2)(N^*\ r_i\ t)\ dt]$$

where f(t) is the density function of the cycle–time distribution. Consistent with previous assumptions, f(t) is the gamma density function, and $Pr\{t<T_{MAX}\}$ and $Pr\{t\geq T_{MAX}\}$ are found from the cumulative gamma distribution with $\alpha=N^*-1$ and $\beta=1/\hat{a}$.

The expected cycle length $E[T \mid T_{MAX}]$ was given in the previous section:

$$E[T \mid T_{MAX}] = E[T|T>T_{MAX}]\ Pr\{T>T_{MAX}\} + E[T|T\leq T_{MAX}]\ Pr\{T\leq T_{MAX}\}$$

Dividing the expected cost per cycle by the expected cycle length yields the gain of the process. This expression can be solved for various values of t to select the T_{MAX} value that yields the minimum average cost per unit time.

This analysis also could be used to determine the T_{MAX} value that yields the minimum–cost per unit–time when consolidation is based on weight, rather than on number of orders. Unfortunately, consolidation by weight makes the calculation of expected accumulation cycle length more difficult. When consolidation is based on orders, the arrival of orders forms a Poisson counting process. Without any limit on maximum holding time, the expected cycle length will equal the expected time at which order N^* order arrives, where N^* is the target number of orders and order N_1 arrives at time t=0. As seen in the previous example, imposing maximum holding time restriction does not complicate this calculation greatly.

When consolidation is based on weight, however, the expected cycle length equals the time at which the accumulated weight equals, *or exceeds for the first time*, the target weight. This expected time sometimes is referred to as the sojourn time,

discussed in Section 6.4 on stochastic clearing system models. In the next section, we continue the preceding analysis, but extend it to consider order weight.

Probabilistic Analysis of Consolidation By Order Weight

Shipment consolidation frequently is done based on weight rather than number of orders. We now seek a cumulative probability graph, similar to Figure 6–1, that gives the probability of attaining at least the economic shipment weight by a given time. This involves the convolution of two random variables, one representing the number of orders received by time t and the other representing the weight of each order. We assume that customer orders arrive according to a Poisson process, while, as in Chapter 4, order weight is distributed according to an unshifted gamma distribution. Basic properties of the gamma distribution are given in Table 4–3.

Let W denote the weight of a customer order. The unshifted gamma(α,β) probability density function of customer order weight W is:

$$f_W(w) = \frac{\beta^{-\alpha} w^{\alpha-1} e^{-w/\beta}}{\Gamma(\alpha)} \qquad w \geq 0$$

Total accumulated weight TW over an order cycle of N orders will be the sum of N gamma distributions, yielding a Ga(Nα, β) distribution.

N, the number of orders accumulated during one cycle, follows a Poisson distribution. Because an order cycle requires at least one order, this Poisson distribution must be truncated at zero, yielding the truncated Poisson density function:

$$f_N(n) = \frac{e^{-\hat{a}t} (\hat{a}t)^n}{(1-e^{-\hat{a}t})\, \Gamma(n+1)} \qquad n \geq 1$$

Conditioning on N, the probability density function for total weight TW accumulated in time t will be a conditional probability density composed of the truncated–Poisson and gamma distributions:

$$f_{TW}(w) = \sum_{n=1}^{\infty} \Pr\{W=w|N=n\} \Pr\{N=n\}$$

$$= \sum_{n=1}^{\infty} \frac{\beta^{-\alpha n} w^{\alpha n-1} e^{-w/\beta}}{\Gamma(\alpha n)} \frac{e^{-\hat{a}t}(\hat{a}t)^n}{\Gamma(n+1)(1-e^{-\hat{a}t})}$$

The probability that at least the optimal weight W^* has been accumulated by time t is:

$$\Pr\{TW \geq W^* \mid T=t\} = 1 - F_{TW}(W^*)$$

$$= 1 - \int_0^{W^*} \sum_{n=1}^{\infty} \frac{\beta^{-1} (w/\beta)^{\alpha n-1} e^{-w/\beta}}{\Gamma(n\alpha)} \frac{e^{-\hat{a}t}(\hat{a}t)^n}{\Gamma(n+1)(1-e^{-\hat{a}t})} dw$$

Simpson's rule or other numerical methods may be used to evaluate this expression. n is an integer by definition of the Poisson process, thus if α also is an integer, then $\Gamma(\alpha n)=(n\alpha-1)!$, and a closed form exists. Thus, for a given value of t, simplification yields:

$$\Pr\{TW \leq W^* \mid T=t\} = \sum_{n=1}^{\infty} \frac{e^{-\hat{a}t}(\hat{a}t)^n}{n!(1-e^{-\hat{a}t})} [1 - e^{-W^*/\beta} \sum_{k=0}^{\alpha n-1} (W^*/\beta)^k / k!]$$

This can be solved by summing successive evaluations over fixed values of n from 1 to ξ, where ξ is some value at which further evaluations do not reasonably affect the value of $\Pr\{TW \leq W^* \mid T=t\}$. Thus:

$$\Pr\{TW \geq W^* \mid T=t\} = 1 - \sum_{n=1}^{\xi} \Pr\{TW \leq W^* \mid T=t\}$$

However, because of the factorial terms, some computer languages will yield inaccurate results when $n\alpha$, and hence $(n\alpha-1)!$, is large. If $n\alpha$ is greater than 15, we can use the Normal distribution to approximate that part of the expression derived from the gamma distribution.

Example: Assume that customer orders arrive at Poisson rate â=3 per day, and that order weight follows an unshifted gamma distribution with α=2 and β=1000. This yields E[W]=2000 pounds.

Figure 6–2 plots the resulting compound gamma/truncated Poisson probability curves for various values of target weight W*. For example, for W*=12000, the probability that at least W* pounds is accumulated in a two–day waiting period is 0.592; this probability is 0.758 for two–and–one–half days. Reasoning similar to that at the beginning of this chapter is straightforward.

As with Figure 6–1, it is important to realize that Figure 6–2 does not show the cumulative probability of W* arrivals in time T. Because an order–holding cycle begins with the arrival of one order with an expected weight E[W], Figure 6–2 plots the cumulative probability Pr{W*–E[W] | T}; that is, the cumulative probability of accumulating, in time T, the additional weight required to reach W*. ■

Recall that the conditional density function for total weight per cycle of N orders is:

$$f_{TW}(w) = \Sigma^{\infty}_{n=1} \Pr\{W=w|N=n\} \Pr\{N=n\}$$

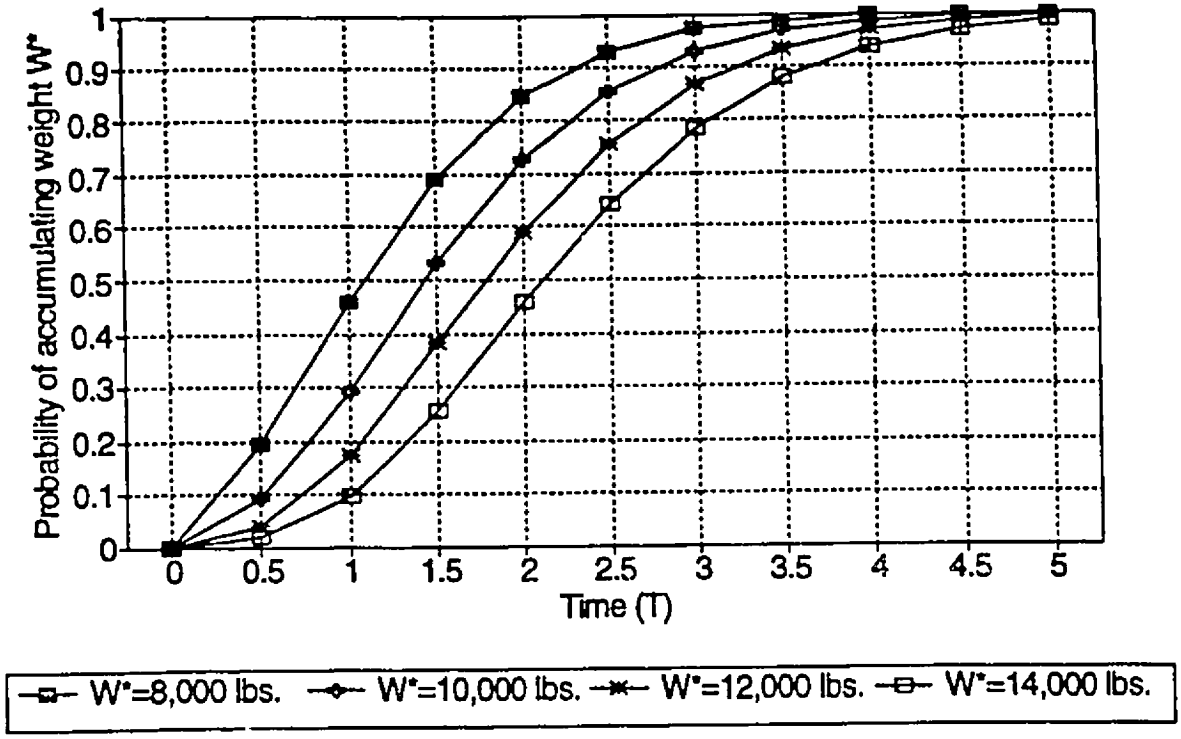
The expected total weight accumulated during the consolidation cycle is:

$$E[TW] = \int_0^{\infty} w\, f_{TW}(w)\, dw$$

We can simplify this expression by applying the maximum holding time (oldest order) rule of a time–and–quantity policy: hold all orders until the earliest of either attaining a minimum consolidated weight or reaching a pre–determined age.

Using logic similar to that of the section on system gain, we see that if T_{MAX} expires before weight W* is reached, the expected load size will be the sum of the

Figure 6–2
Probability of Accumulating Weight W* in Time T
With Poisson(â=3) Order Arrivals and
Gamma(α=2, β=1000) Order Weight



expected weight $E[W]=\alpha\beta$ of the first order and the expected weight of those orders that arrived after the first but before T_{MAX}:

$$E[TW \mid t > T_{MAX}] = E[W] + \int_0^{W^*-E[W]} w\, f_{TW}(w \mid t > T_{MAX})\, dw$$

Alternately, if weight W^* is attained before T_{MAX} is reached, the expected load weight will be W^*. Thus, the expression for expected total weight per cycle, given a maximum holding time of T_{MAX}, is:

$$E[TW \mid T_{MAX}] = E[W] + \int_0^{W^*-E[W]} w\, f_{TW}(w \mid t > T_{MAX})\, dw$$
$$+ (W^*-E[W])\, Pr\{TW \geq W^*-E[W] \mid t \leq T_{MAX}\}$$

$E[TW|T_{MAX}]$ can be used to estimate the per–unit inventory–holding and transportation costs for selected values of T_{MAX}. Management then can select the T_{MAX} value that gives a favourable tradeoff between expected cost and order delay. This method also can be used to determine the lowest–cost holding time, as illustrated in the following example.

Example: Consider a shipper sending consolidated loads between Toronto and London, Ontario, via private trucking. In Appendix C, we comment that Flood et al. [1984] derived a private–carrier transportation cost of \$0.879 per mile, or \$0.546 per kilometre. Applying this figure to the 190–kilometre travel distance yields a one–way transportation cost of \$103.74. We will take this cost as F_L=\$106.67 simply because doing so yields a value of ESW that is easier to work with. The cost of the back–haul movement also should have been considered, but was ignored to simplify calculations.

Assume order arrivals follow a Poisson distribution with $\hat{a}$=4.5 per day, and that order weight is gamma–distributed with α=2, β=1000, and $E[W]=\alpha\beta$=2000 pounds.

Letting the inventory–holding cost be \$0.0133 per pound per day, the ESW expression of Section 5.3 gives a target consolidated load size of 12,000 pounds.

Consider a T_{MAX} value of 1.6 days. Letting $W_D=W^*-E[W]$, the probability of accumulating at least W^* in T_{MAX} is $Pr\{TW \geq W_D \mid t \leq T_{MAX}\}=0.729$. This yields an expected load weight of:

$$E[TW \mid T_{MAX}] = E[W] + \int_0^{WD} w\, f_{TW}(w \mid t > T_{MAX})\, dw + W_D\, Pr\{TW \geq W_D \mid t \leq T_{MAX}\}$$

$$= 2000 + \int_0^{10000} w\, f_{TW}(w \mid t > T_{MAX}=1.6)\, dw + 10000\ (0.729)$$

where:

$$\int_0^{WD} w\, f_{TW}(w \mid T_{MAX}=t)\, dw$$

$$= \sum_{n=1}^{\infty} \frac{(\beta^{-\alpha}\, \hat{a}t)^n\, e^{-\hat{a}t}}{(1-e^{-\hat{a}t})(n\alpha-1)!n!} \qquad \int_0^{W/\beta} w\, e^{-w/\beta}\, (w/\beta)^{\alpha n-1}\, dw$$

$$= \sum_{n=1}^{\infty} \frac{e^{-\hat{a}t}(\hat{a}t)^n}{n!(1-e^{-\hat{a}t})}\, \alpha\beta\, [1 - e^{-W/\beta} \sum_{k=0}^{\alpha n} (W/\beta)^k / k!]$$

The last equality results because α is a positive integer.

We find that the expected load size with $W^*=12000$ lbs. and $T_{MAX}=1.6$ days is about $E[TW|T_{MAX}]=9728$ pounds, yielding a total inventory and transportation cost per 1.6–day cycle of about \$210, or about \$2.16 per cwt. This combination of cost and customer waiting time may or may not be acceptable to management.

Combining this approach with an appropriate search technique, the holding time yielding the minimum cost per hundredweight can be identified. This is summarized on page 122 for $W^*=12000$.

The minimum expected cost per cwt. occurs with a holding time of $T_{MAX}=1.35$ and an expected load size of 8678 pounds. For shorter holding times, savings from

spreading fixed transportation costs over larger loads outweigh the additional inventory–holding costs.

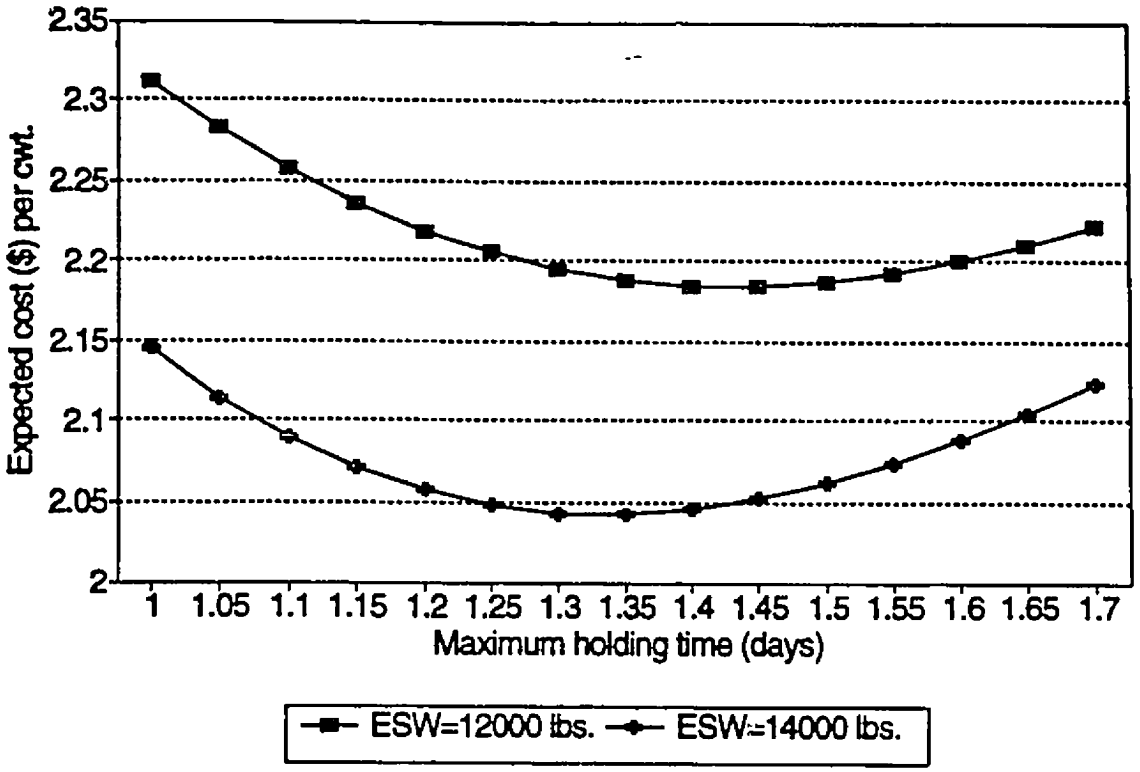
T_{MAX}	Pr{TW≥W*}	E[TW|T_{max}]	E[cost per cwt]
1.1	0.447	7427.469	$2.169
1.2	0.511	7944.489	2.143
1.3	0.572	8440.655	2.131
1.35	0.601	8677.986	2.129
1.4	0.629	8906.942	2.131
1.5	0.682	9337.125	2.142
1.6	0.729	9727.520	2.163
1.7	0.771	10076.628	2.192

Holding times exceeding T_{MAX} yield additional inventory–holding costs greater than transportation savings. This is illustrated in Figure 6–3, which also shows cost results for a target weight of 14000 pounds, derived by increasing the order arrival rate to 6.12 orders per day while keeping cost parameters constant. ■

Application of Probabilistic Analysis of Consolidation To Shipment Costing

Knowledge of expected consolidated load weight is useful for other management purposes, such as estimating the transportation cost of individual orders to be shipped as part of a consolidated load. Determining this cost is a common and important problem for reasons of accounting and cost control, and because the purchaser may want advance knowledge of freight charges as part of price negotiation. Shelley [1982] proposes a costing method for consolidated loads that requires estimates of the percentage of the total load that an individual order comprises and the expected number of delivery stops. Often, both these variables will be unknown until the consolidated load has been finalized. Our expression for

Figure 6–3
Comparison of Per–Cwt Cost
With Poisson(â=3) Order Arrivals and
Gamma(α=2, β=1000) Order Weight



$E[TW|T_{MAX}]$, the expected load size given T_{MAX}, provides a good estimate of the first of these unknowns.

The mean number of stops per pickup or delivery load will depend on several factors, including the arrival rate and size of customer orders from a region, shipper order–release policies, and the capacity of vehicles. Sometimes it is possible to estimate the number of stops per load from empirical data; Eilon et al. [1971] and Shelley [1982] give estimates for deliveries of oil and food products respectively.

Jaillet and Odoni [1988] suggest assigning to each customer or group of customers, a probability, independent of all other customers, that a vehicle stop will be required at that location. The expected number of stops is then the sum of these probabilities.

Burns et al. [1985] present a simple expression for the mean number of customer stops per delivery load. If d_i is the long–run average demand by customer i and $D=\sum_{i=1}^{\hat{C}} d_i$ is the total demand from $\hat{C}$ customers in a delivery region, the probability that any item in a load is destined for customer i is d_i/D. The probability that at least one of N items belongs to customer i is $[1-(1-d_i/D)^N]$. This is also the probability of making a stop at customer i given that the transportation vehicle contains N items, so the expected number of customer stops per load is:

$$\begin{aligned} E[S] &= \sum_{i=1}^{\hat{C}} (1 \text{ stop at customer } i) \Pr\{\text{making a stop at customer } i\} \\ &= \sum_{i=1}^{\hat{C}} [1 - (1 - d_i/D)^N] \\ &= \hat{C} - \sum_{i=1}^{\hat{C}} (1 - d_i/D)^N \end{aligned}$$

We will refer to this expression as the "Burns et al. formula".

Example: Figure 6–4 shows the results of applying this expression to a sample of 25 Ontario cities listed in Table 6–1. As with many analytical distribution models, demand from a city was assumed to be proportional to its population. The curve in Figure 6–4 increases relatively slowly and approaches $\hat{C}$ (=25) with large values of N, the number of items transported. For example, the expected number of stops is approximately E[S]=12 for N=30, and approximately E[S]=23 for N=300. Thus, if N is unknown, the formula is of limited guidance to management. ■

What is the value of N, the number of items carried by the vehicle? Two situations should be considered. First, if there are enough items to fill the vehicle, it should be dispatched full (see, for example, Blumenfeld et al. [1985], Burns et al. [1985], Daganzo and Newell [1985], Hall [1987]). Then, N equals vehicle capacity stated as number of items.

A different conclusion results from the scenario we have been examining, where customer orders arrive according to a Poisson process with rate â and management has set a maximum holding time T_{MAX}. Assuming that each order consists of one item and that vehicle capacity is not a binding constraint, the expected number of customer stops per delivery load is a compound distribution given by the Burns et al. formula and the Poisson probability density function truncated at one (because each load must have at least one order):

$$E[S] = \sum_{N=1}^{\infty} \{\sum_{i=1}^{\hat{C}} [1 - (1 - d/D)^N] [e^{-\hat{a}TMAX} (\hat{a}T_{MAX})^N / [N! (1-e^{-\hat{a}TMAX})]]\}$$

$$= \hat{C} + (1-e^{-\hat{a}TMAX})^{-1} [\hat{C}\, e^{-\hat{a}TMAX} - \sum_{i=1}^{\hat{C}} e^{-\hat{a}\,TMAX\,d/D}]$$

$$= (1-e^{-\hat{a}TMAX})^{-1} [\hat{C} - \sum_{i=1}^{\hat{C}} e^{-\hat{a}\,TMAX\,d/D}]$$

Figure 6–4
Expected Number of Vehicle Stops for Example of 25 Southern Ontario Cities
And Poisson(â=3) Order Arrivals

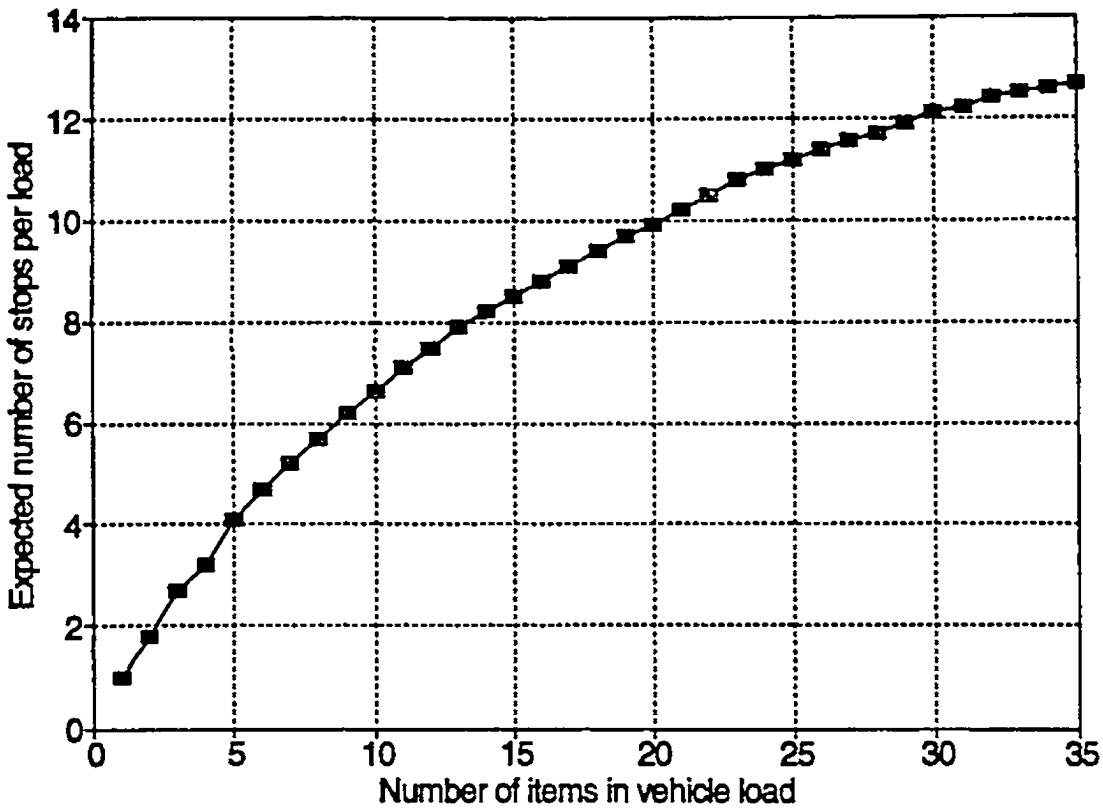


Table 6–1
Ontario Cities Used in E[S] Calculation Example

city	population	percentage population
Bancroft	2332	0.1
Barrie	34,389	1.5
Belleville	35,311	1.5
Brantford	66,950	2.9
Brockville	19,903	0.9
Chatham	38,685	1.7
Cornwall	46,121	2.0
Fort Erie	24,031	1.1
Goderich	7,385	0.3
Hamilton	312,003	13.6
Kingston	56,032	2.5
Kitchener	131,870	5.8
London	240,392	10.5
Niagara Falls	69,423	3.0
Ottawa	30,442	1.3
Owen Sound	19,525	0.9
Peterborough	59,683	2.6
Port Colborne	20,536	0.9
St. Catherines	123,351	5.4
St. Thomas	27,206	1.2
Sarnia	55,576	2.4
Smiths Falls	9,279	0.4
Stratford	25,657	1.1
Toronto (city)	633,318	27.8
Windsor	196,526	8.6
total	2,285,926	100.0

Applying this expression to our 25–city sample with an order arrival rate of $\hat{a}$=3 per day and T_{MAX}=2 days yields an expected number of vehicle stops of E[S]=4.5.

A more realistic formulation includes vehicle capacity restrictions. Let H denote the capacity of the vehicle expressed in number of items or orders. To determine the expected number of stops, we must consider two cases: i) vehicle capacity is reached before the maximum holding time expires, thus the load size equals vehicle capacity; and ii) the number of order arrivals within the maximum holding time is less than capacity. Again, the probability that a vehicle contains N items is given by a Poisson distribution truncated at 1. Given these two cases, the expected number of stops is:

$$\begin{aligned} E[S] &= \sum_{N=1}^{H-1} E[\text{stops}|N<H]\ \Pr\{N \text{ arrivals in } T_{MAX}|N<H\} \\ &\quad + E[\text{stops}|N=H]\ \Pr\{\geq H \text{ arrivals in } T_{MAX}\}) \\ &= \{ \sum_{N=1}^{H-1} [\hat{C} - \sum_{j=1}^{\hat{C}} (1 - d_j/D)^N] [(\hat{a}T_{MAX})^N/(N!(e^{\hat{a}TMAX}-1))\ ]\} \\ &\quad + \{[\ \hat{C} - \sum_{j=1}^{\hat{C}} (1 - d_j/D)^H] [(e^{\hat{a}TMAX}-1)^{-1} \sum_{k=H}^{\infty} (\hat{a}T_{MAX})^k/k!\ ]\} \end{aligned}$$

Repeating analysis from earlier in this section, the expected vehicle load E[L] (ie., number of orders) is:

$$\begin{aligned} E[L \mid T_{MAX}] &= 1 + E[L \mid N<H-1 \text{ arrivals in } T_{MAX}] \\ &\quad + (H-1)\ \Pr\{N\geq H-1 \text{ arrivals in } T_{MAX}\} \\ &= 1 + \sum_{N=0}^{H-2} N\, e^{-\hat{a}TMAX} (\hat{a}T_{MAX})^N/N! \\ &\quad + (H-1)\, e^{-\hat{a}TMAX} (\hat{a}T_{MAX})^{H-1}/(H-1)! \end{aligned}$$

As vehicle capacity H increases, the probability of accumulating more than H orders within the allowable time approaches zero. Thus, this expression may quickly reach a practical upper limit on the number of stops per tour.

Example: With â=3 orders per day and T_{MAX}=2 days, our 25–city example yielded an expected number of stops of E[S]=3.69 and expected vehicle load of E[L]=4.77 items if capacity was H=5, E[S]=4.47 and E[L]=6.84 if H=10, and approximately E[S]=4.5 and E[L]=7 for all H greater than 13. Note that, for large values of H* relative to ($âT_{MAX}$), E[L] approaches the deterministic value of ($1+âT_{MAX}$). ■

The values derived for E[S] and E[L | T_{MAX}] from the above expressions can be used in Shelley's model to estimate the cost of individual shipments comprising a consolidated load. One problem, however, is that the Burns et al. expression is stated in terms of items or orders, rather than weight. In the simplest case, this can be reconciled by setting, in the Burns et al. formula:

$$N = \lceil E[TW \mid T_{MAX}] / E[W] \rceil$$

where TW is the total weight carried by the vehicle, E[W] is the expected weight of an order, and $\lceil x \rceil$ is the smallest integer greater than or equal to x.

A more complicated approach is to calculate the expected number of customer stops per tour from a compound probability distribution derived from the Burns et al. expression and our probability density function of load weight. Both this, and modeling of the impact of shipment consolidation on the number of delivery stops and/or delivery tour length is suggested as further research.

Summary

The previous sections have illustrated a major advantage of probabilistic analysis of expected performance: it can be as simple or as complex as the situation dictates. For example, Table 6–2 summarizes the stochastic characteristics of the

three shipment–release policies of Chapter 4 when order arrival times are Poisson(â)–distributed.

Stochastic treatment of shipment consolidation is more realistic than deterministic analysis, and the general ideas are fairly easy to understand. It, however, requires knowledge of appropriate probability distributions, and mathematical complications may result when deriving or using these distributions. In the next section, we discuss a bulk–service queueing approach to determining T_{MAX}, the maximum length of time an order can be delayed for consolidation.

6.3 Single–Server Bulk–Service Queue Analysis

In the basic single–server bulk–service queue model, customers arrive singly and at random, forming a queue in sequence of arrival. After a time interval, they are served, usually on a first–come–first–served basis, in batches of no more than size H, H>0. Typically, arrivals are according to a Poisson process. New arrivals must wait for the next service epoch even if less than H currently are being served.

Most versions of the basic bulk–service queueing model assume that if the queue is empty when the server becomes idle, service restarts immediately when a minimum batch size is reached. Models by Bailey [1954] and Jaiswal [1961] allow the server to "serve nobody"; that is, service continues even when no customers are waiting, and new arrivals wait until the next service epoch. This model is sometimes called "bulk service with no control"; it might, for example, allow an empty vehicle to depart. Bloemena [1960] and Chaudhry and Templeton [1983] show that expressions

Table 6–2
Stochastic Characteristics of Basic Shipment–Release Policies
With Poisson(â) Order Arrivals

Quantity policy:

size of vehicle load (N_1):

N_1 is constant at a management–chosen value

time between load dispatches (T_1):

T_1 is N_1–Erlang distributed with parameter $1/\hat{a}$

$E[T_1] = N_1/\hat{a}$

expected waiting time per order (W_1):

$W_1 = (N_1-1)/2\hat{a}$

Time policy:

size of vehicle load (N_2):

N_2 is stochastic: vehicle load includes the first order to arrive plus those orders that arrive within the maximum holding time T_{MAX}; thus, the probability of a load consisting of N_2 orders equals the probability that exactly (N_2-1) more orders arrive in time T_{MAX}; this yields a shifted Poisson distribution:

$$f(N_2) = e^{-\hat{a}(Tmax)} (\hat{a}\, T_{MAX})^{N2-1} / (N_2-1)! \qquad N_2 \geq 1$$

$$E[N_2] = 1 + T_{MAX}\, \hat{a}$$

time between load dispatches (T_2):

T_2 is stochastic, consisting of a random time until the first order arrives plus the deterministic holding time T_{MAX}; this yields a shifted exponential distribution:

$$f(T_2) = \hat{a}\ e^{-\hat{a}(T2-Tmax)} \qquad T_2 \geq T_{MAX}$$

$$E[T_2] = T_{MAX} + 1/\hat{a}$$

Table 6–2 (continued)

expected waiting time per order (W_2):

$$W_2 = T_{MAX}/2\hat{a}$$

Time-and-quantity policy:

size of vehicle load (N_3):

N_3 is stochastic, and equals the minimum of N_1 (the deterministic maximum number of orders per load given by a volume strategy) and N_2 (the number from the shifted–Poisson distribution of a time strategy):

$$F_{N3}(n) = 1 - (1 - F_{N1}(n))\,(1 - F_{N2}(n))$$

$$= 1 - (1 - 0)\,(1 - F_{N2}(n)) \qquad \text{if } N_1 \geq n$$

$$= 1 - F_{N2}(n)$$

which yields a shifted–Poisson distribution right–truncated at N_1:

$$f(N_3) = \frac{e^{-\hat{a}(Tmax)}\,(\hat{a}\,T_{MAX})^{N2-1}}{(N2-1)!}\,(\Sigma^{N1}_{n=1}\,\hat{a}^{n-1}/(n-1!))^{-1} \qquad N_2 \geq 1$$

time between load dispatches (T_3):

T_3 is stochastic, and equals the minimum of T_1, which is N_1–Erlang–distributed from a volume strategy, and T_2, which follows the shifted–exponential distribution of a time strategy:

$$F_{T3}(t) = 1 - (1 - F_{T1}(t))\,(1 - F_{T2}(t))$$

$$= 1 - (1 - e^{-\hat{a}t}\,\Sigma^{N1-1}_{j=0}(\hat{a}t)^j/j!)(e^{\hat{a}\,TMAX} - e^{-\hat{a}(t-TMAX)}) \quad T_{MAX} \leq t \leq T_2$$

further simplification does yield any improvement.

derived by Bailey are valid even if service stops when the queue is empty and resumes immediately upon arrival of the next customer.

The basic model has been examined by many authors, including Bailey [1954], Downton [1955], Takács [1962], Giffin [1978], and Chaudhry and Templeton [1983]. The next section briefly discusses some variations to this model.

Variations of the Basic Bulk-Service Queue Model

The "general bulk–service rule" states that a minimum number of customers must be waiting before service can begin. Systems utilizing this policy were examined by Neuts [1967] for a general service distribution, and by Medhi [1975] for exponential service. A bulk–service queue model requiring a batch size of exactly H was discussed by Fabens [1961], Fabens and Perera [1963], Giffin [1978], Chaudhry and Templeton [1983], and Gross and Harris [1985]. Gross and Harris [1985] also examined the case where the server will take less than H, and arrivals during the service period immediately enter service if there are no more than H in service.

The basic model with general service distribution and finite capacity was discussed by Singh [1971, 1972]. Dick [1970] analyzed the case where the arrival rate changes when system capacity has been reached. A model where the service time is dependent on the size of the batch being served was presented by Nair and Neuts [1972]. Chaudhry and Templeton [1983] comment that "numerically manageable results do not appear possible" for this case.

Novaes and Frankel [1966] examined the Poisson arrival/general service bulk–service queue with balking, where the probability of balking is a function of the number

in the queue and the expected time between service epochs. Mercer [1968] discussed the case where bulk service is scheduled to occur at fixed intervals of time, and either occurs in those intervals or not at all, while Natarajan [1962] examined bulk–service queues where time is a discrete, rather than continuous, random variable.

Bulk–service with arrivals in batches has been discussed by several authors. Chaudhry and Templeton [1972] note that these systems are more complex than the simple bulk–service or batch–arrival cases because service batches rarely correspond to arrival batches. Powell [1986] presents methods for approximating the queue length and mean waiting time for such systems with Poisson and non–Poisson arrivals and various service strategies.

Analysis of Bulk-Service Queues

With a bulk–service queue, it is generally easier to obtain the distribution of the number in the queue rather than the number in the system. This leads to an imbedded Markov chain approach (Kendall [1951, 1955]), which analyses the state of the system at selected discrete points in time. For example, Bailey [1954] and Downton [1955] defined the states of the imbedded Markov chain as the number of customers waiting immediately before the beginning of a service epoch. This discrete–time solution can then be used as an approximation to the continuous–time case.

The process is ergodic if the average number of arrivals per service interval is less than the maximum batch size. Then, the long–run (limiting) probabilities exist. Solving for the steady state probabilities is complicated, involving derivation of the

difference equations for the infinite process, then reducing the infinite set to a single equation through geometric transformations (see, for example, Bailey [1954], Downton [1955], Giffin [1978]). It is necessary to determine the zeros z_i of a polynomial equation, where the number of zeros depends on the service time distribution. However, only the zeros inside the unit circle (that is, $|z_i|<1$) or outside the unit circle ($|z_i|>1$) are required to derive steady–state performance measures. For a given service time distribution, there will be a known number, usually (H–1), of these roots, where H is the maximum batch size.

The major hurdle to application of bulk–service queuing analysis is finding these zeros. If the maximum batch size is small, algebraic software, such as GAMS or MAPLE, may be used, or a quick approximation can be obtained by graphing the expression for arbitrary values of z using computer spreadsheet packages. However, if H is large or some zeros are complex, even the most powerful software may be unsuccessful. As a result, the imbedded Markov chain approach has been criticized as impractical.

Bagchi and Templeton [1972] and Neuts [1973, 1979] have discussed analytical methods that do not require finding the zeros. Powell [1985] proposed a method for obtaining approximate values of the zeros by solving H pairs of equations by simple search techniques.

Shipment Consolidation as a Bulk-Service Queue: Time-based Shipment Release Policy

Consider the use of a time–based shipment–release policy; that is, a consolidated load is dispatched when the oldest order reaches a maximum holding

time T_{MAX}. With Poisson–distributed order arrivals with rate â, one server, and infinite waiting space, a $M/D^H/1/\infty$ bulk–service queue results, where D^H denotes deterministic (constant) service in batches of maximum size H. In all our bulk–service queue models of shipment consolidation, the service time refers to the round–trip time required to transport a batch of orders. More generally, service time is the time between successive load dispatches.

Deterministic service implies that the time between shipment–releases is constant only if there are orders waiting after a load dispatch. If the system is empty, a random time passes until the first order of the next cycle arrives.

Performance measures for a $M/D^H/1$ bulk–service queue include:

mean number of orders waiting immediately before a consolidated load is dispatched:

$$N = \frac{H - (H-m)^2}{2(H-m)} + \sum_{i=1}^{H-1} (1 - z_i)^{-1}$$

mean waiting time per order:

$$Wq = \frac{H(m-H+1)}{2\hat{a}(H-m)} + (1/\hat{a}) \sum_{i=1}^{H-1} (1 - z_i)^{-1}$$

where m is the expected number of arrivals during a service interval (accumulation cycle): m=â E[length of service interval]. z_i are the H–1 simple zeros inside the unit circle ($|z_i|<1$) of the expression $[z^H e^{\hat{a}(1-z)} - 1]$.

A $M/D^H/1$ bulk–service queue provides a very simple, but computationally complex, model of shipment consolidation. Under a time–based shipment–release policy all orders waiting at time T_{MAX} are dispatched, implying that the maximum batch size H is infinity. This complicates the determination of zeros. Moreover, if the system

empties completely at each service, use of a clearing system model (discussed in the next section) is simpler and more intuitively–appealing.

For practical purposes, the $M/D^H/1$ bulk–service queue can be applied to a time–based consolidation policy if loads can only be dispatched at the end of the maximum waiting time and if load sizes are restricted to some practical limit, such as vehicle capacity. If loads can be dispatched before the holding time has elapsed because the maximum batch size has been reached, a time strategy is not being used; rather, a time–and–weight strategy is.

Even with a limit on maximum batch size, calculating performance measures frequently will be difficult, and approximations may be necessary. Bailey [1954], for example, presents a simple inequality for the mean waiting time in terms of maximum batch size and expected number of arrivals per service interval. Unfortunately, the resulting range is fairly wide if the batch size is large compared to the expected number of arrivals. Bailey [1952] and Downton [1955, 1956] give other approximations for expected queue length and mean waiting time.

A simple approach would be to first select values for T_{MAX}, then determine the expected queue length and mean waiting time from either exact or approximate methods. These results would be compared over all values of T_{MAX} that are acceptable to management. The preferred policy could be decided through an arbitrary rule, such as selecting the T_{MAX} that minimizes average cost per unit time (ie., system gain, as discussed previously).

Shipment Consolidation as a Bulk-Service Queue: Quantity-based Shipment Release Policy

Consider a quantity–based shipment–release policy: a consolidated load is dispatched when a target number N^* of orders has been accumulated. With Poisson–distributed order arrivals, the time between load–dispatches will be N^*–Erlang distributed with parameter $1/\hat{a}$.

This strategy could be modeled as a $M/E_{N^*}^{H,H}/1/\infty$ bulk–service queue: Poisson arrivals, N^*–Erlang service, service batches of exactly size H ($H=N^*$), single server, and infinite waiting space. It is easier just to treat this case as a Poisson point process. As a result, the time between load dispatches is $E[T]=N^*/\hat{a}$ and the expected waiting time per order is $W_q = (N^*-1)/2\hat{a}$.

We add that with a time–and–quantity policy, the resulting queue model is $M/D^H/1$ if the T_{MAX} is less than the expected time to accumulate the target weight, and $M/E_{N^*}^{H,H}/1$ if not.

Summary

The above examples consolidate by orders, not weight. Although a customer order could be viewed as a "batch of pounds", modeling shipment consolidation with such "batch arrivals" yields a process more complex than the usual batch–arrival, bulk–service queue. Research on the latter assumes that arrival batches can be broken for service, whereas restrictions exist as to the divisibility of customer orders. References pertaining to batch–arrival, bulk–service queues are given in Chaudhry and Templeton [1972].

The major benefit of viewing shipment consolidation as a bulk–service queue is the existence of a large body of queuing theory research. As a result, a wide variety of options and situations may be considered.

Unfortunately, there are several problems with this approach. The computational difficulties of determining zeros and developing expressions for performance measures were discussed previously. Modelers also must be careful of assumptions used in the literature when deriving or applying these measures.

There is a more fundamental problem with modeling shipment consolidation as a bulk–serve queue: use of queuing theory implies a possible dependence between the service time of batch n and the accumulation time of batch n+1. In shipment consolidation, this would mean that the time to accumulate batch n+1 could be dependent on the service time (ie., the transportation time) of batch n. This is unrealistic and not intuitively satisfying. Moreover, if the service time of batch n is unusually long, or arrivals to batch n+1 are unusually rapid, the maximum batch size (i.e., vehicle capacity) will be exceeded. Then, when batch n+1 is dispatched, some orders will have to be left behind. This implies that management did not act to reduce the accumulation of waiting orders.

One shipment consolidation scenario where the service time of batch n will affect the accumulation time of batch n+1 occurs if the frequency of load dispatch is limited by the number of delivery vehicles. The accumulation time then is a function of the time required for a vehicle to complete its route. This is an interesting and complex research topic in itself.

Other stochastic methods exist that are better suited to modeling shipment consolidation. The next section discusses one of these, stochastic clearing system analysis, and its use to determine T_{MAX}.

6.4 Stochastic Clearing System Analysis

In a stochastic clearing system, randomly–occurring inputs accumulate over time such that the cumulative amount is non–decreasing and right–continuous. At intermittent times, a "clearing" instantaneously restores the net quantity in the system to level M; M often is zero. This clearing occurs whenever the total input since the last clearing exceeds a critical level N^*. N^* typically is determined by optimizing some performance measure such as long–run average cost.

An important characteristic of stochastic clearing systems is the lack of interaction between the input and output processes. This allows simple closed–form expressions for important measures to be derived, a feature missing from many bulk–service queues.

The process of accumulating customer orders and dispatching consolidated shipments mirrors the characteristics of a stochastic clearing system. Our first section, based on work by Stidham [1974, 1977], summarizes the theory of such systems. We then discuss their application to shipment consolidation.

Theory of Stochastic Clearing Systems

Let N(t) be the cumulative input quantity at time t, N(0)=0, and V(t) be the net quantity in system at time t. A necessary assumption is that, if X_n is the time between

the (n–1)–th and n–th clearing, $(X_1, X_2, ...)$ is a renewal sequence and the behaviour of V(t) after a clearing is a probabilistic replica of the process beginning at time zero. Thus, V(t) is a regenerative process with respect to the renewal sequence $(X_1, X_2, ...)$.

Let T(n) denote the time t at which the cumulative quantity N(t) in the system first reaches level n, $n \geq 0$. T(n) is sometimes referred to as the "first entrance time", "first passage time", or "first hitting time". Because a clearing occurs when $V(t)=N^*$, $X_1=T(N^*)$.

Assume that the probability distribution of the cumulative quantity in the system is given by $F_{N(t)}(t,n) = Pr\{N(t) \leq n\}$, $t \geq 0$. W(n) is the expected length of time that the cumulative input N(t) does not exceed level n:

$$W(n) = \int_0^\infty F_{N(t)}(t,n)\, dt$$

$$= E[T(n)]$$

W(n) is referred to as the "sojourn measure" of $\{N(t), t \geq 0\}$ on $[0,\infty]$.

If T(0)=0, T(n) is the length of time t until N(t) first reaches level n. Then, the sojourn measure W(n) is the expected amount of time needed to accumulate n. W(n) is non–decreasing and right–continuous, $0<W(n)<\infty$, $W(0) \geq 0$, and $W(n) \to \infty$ as $n \to \infty$. Smith [1960] calls the sojourn measure the "infinitesimal renewal function".

The total cost incurred by the system in time [0,t] is:

$$c\, R(t) + \int_0^t g(V(s))\, ds = c\, R(t) + \int_0^t g(x)\, dx$$

where:

c = fixed positive cost per clearing
R(t) = number of clearings by time t
V(s) = net quantity in system at time s
g(x) = cost per unit time when the net quantity V(s) is x; g(x) is continuous and never negative

Stidham [1982] shows by elementary renewal theory that, with a fixed clearing level N, the long–run average cost of the system is:

$$(c + \int_0^N g(x)\, dW(x)) / W(N) = g(N) + (c - \int_0^N W(x)\, dg(x)) / W(N)$$

This expression is similar to the average cost per unit time seen in our discussion of system gain.

Assume that the cost function is linear: g(x)=rx, where r is a per–unit (variable) cost parameter and x is the quantity. The optimal clearing level N^* can be determined by solving for N^* in:

$$\int_0^{N^*} W(x)\, d(rx) = c$$

$$\Rightarrow \quad \int_0^{N^*} W(x)\, dx = c/r$$

$$\Rightarrow \quad \int_0^{N^*} \int_0^{\infty} F_X(t,x)\, dt\, dx = c/r$$

Depending on the sojourn function, an explicit solution for N^* may not be available. In that case, the optimal value N^* can be found by tabulating the integral $\int_0^{N^*} W(x)\, dx = c/r$. With deterministic inputs, where the time to accumulate n=W(n)=n/â (â>0; â=mean rate of growth), this expression yields $N^* = \sqrt{2âc/r}$, which is reminiscent of the EOQ formula.

Shipment Consolidation as a Stochastic Clearing System

By applying results given in Stidham [1977], Gupta and Bagchi [1987] use stochastic clearing system theory to model shipment consolidation. They assume that orders accumulate such that the quantity in the system at any time is gamma distributed. Costs relating to transportation, inventory–holding, and shipment–handling are considered. By treating the latter cost as a fixed amount per hundredweight per

day, shipment–handling cost and inventory–holding costs can be combined, thus simplifying calculations.

There are two problems with Gupta and Bagchi's example. First, modeling of shipment–handling cost as a dollar amount per physical unit per period is debatable. Second, as in ESQ analysis, this approach cannot be used if a fixed cost component does not exist. Although Bagchi and Gupta assume a fixed transportation cost, it appears that this cost was derived by calculating the common carrier transportation cost for a full vehicle, then incorrectly applying that same total cost to all load sizes (ie., the cost of a full load is used whether or not the vehicle actually is full).

In Section 6.2, we assumed that order weight followed a gamma(α,β) distribution. Thus, total accumulated weight given n orders had arrived was Ga($n\alpha$,β)–distributed. As well, because an order cycle requires at least one order, order arrivals were Poisson distributed truncated at zero. Thus, the probability that the total weight TW, accumulated by time t, is not more than Q is:

$$\Pr\{TW \le Q \text{ in } T=t\} = F_{TW}(t,Q)$$

$$= \int_0^Q \sum_{n=1}^{\infty} \frac{\beta^{-1}(w/\beta)^{\alpha n-1} e^{-w/\beta}}{\Gamma(\alpha n)} \cdot \frac{e^{-\hat{a}t}(\hat{a}t)^n}{\Gamma(n+1)(1-e^{-\hat{a}t})}\, dw$$

This leads to the sojourn function:

$$W(Q) = \int_0^{\infty} F_{TW}(t,Q)\, dt$$

$$= \int_0^{\infty} \int_0^Q \sum_{n=1}^{\infty} \frac{\beta^{-1}(w/\beta)^{\alpha n-1} e^{-w/\beta}}{\Gamma(\alpha n)} \cdot \frac{e^{-\hat{a}t}(\hat{a}t)^n}{\Gamma(n+1)(1-e^{-\hat{a}t})}\, dw\, dt$$

Assuming a linear cost function, the optimal consolidated weight W^* is found by solving for W^* in:

$$\int_0^{W^*} W(Q)\, dQ = c/r$$

$$\Rightarrow \quad \int_0^{W^*} \int_0^{\infty} \int_0^{Q} \sum_{n=1}^{\infty} \frac{\beta^{-\alpha n} w^{\alpha n-1} e^{-w/\beta}}{\Gamma(\alpha n)} \frac{e^{-\hat{a}t}(\hat{a}t)^n}{\Gamma(n+1)(1-e^{-\hat{a}t})}\, dw\, dt\, dQ$$

$$= F_L/r_W$$

where F_L is the sum of all per–load fixed costs, and r_w is the inventory–holding cost per–unit weight per–unit time.

This expression can be simplified. We have assumed that the input process is a compound process with Poisson arrivals and gamma order weights. Stidham [1974] shows that when input is from a compound process, the stationary distribution of V(t), the net quantity in the system, does not in any way depend on the particular interarrival–time distribution. Thus, we can ignore the Poisson arrival process and consider only the gamma "weight process".

Assuming a linear cost function, the optimal consolidated weight W^* is now found by solving:

$$\int_0^{W^*} W(Q)\, dQ = c/r$$

$$\Rightarrow \quad \int_0^{W^*} \int_0^{\infty} F_{TW}(t,Q)\, dt\, dQ = c/r$$

$$\Rightarrow \quad \int_0^{W^*} \int_0^{\infty} \int_0^{Q} \frac{\beta^{-\alpha n} w^{\alpha n-1} e^{-w/\beta}}{\Gamma(\alpha n)}\, dw\, d(\alpha n)\, dQ$$

$$= F_L/r_W$$

Stidham [1977] tabulates this integral for various values of F_L/r_W and provides corresponding W^* values.

Example: In Section 6.2, we considered a shipper sending consolidated loads between Toronto and London via private carrier. We assumed that order weight was Ga(α=2, β=1000)–distributed with an expected weight E[W] of 2000 pounds, and order

arrivals formed a Poisson process with rate â=4.5 per day. As well, we used cost parameters F_L=\$106.57 per trip and r_W=\$0.0133 per pound per day.

Using the ESW formula of Section 5.3, the economic shipment weight (ESW) equals 12,000 pounds. Section 6.2 applied stochastic analysis to determine the T_{MAX} value that minimized per-cwt. cost. This was found to be T_{MAX}=1.35 with an expected load size of 8677.99 pounds.

Let us now treat that example as a stochastic clearing system. With arrival rate â=4.5 orders per day and an expected order weight of 2000 pounds, the "build-up rate" is ρ=âE[W]=9000 pounds per day.

The optimal consolidated weight W* can be found by solving:

$$\int_0^{W^*} W(Q)\, dQ = F_L/r_W$$

Stidham [1977] provides a table of optimal clearing levels as a function of $F_L/\rho r_W$, thus eliminating the need to work directly with the sojourn function W(Q). Two conditions must be present before this table can be used. First, inputs to the system must be independent and gamma-distributed; that is, Pr{W(t)≤w}, t≥0, is gamma-distributed. Our assumption that order weights are i.i.d. gamma random variables satisfies this requirement.

Second, the total cost function must be linear: $TC = F_L + r_W Q/2$. Because we are dealing with private carrier, transportation cost for a given distance largely is fixed (see Appendix C), and a linear approximation to total cost is acceptable (such an assumption would not be true with common carrier). Thus, from Stidham's table, we find that the optimal clearing level for our values of ρ, F_L, and r_W is W*=7761.2513 pounds.

The expected time to accumulate W^* is given by the sojourn function:

$$W(W^*) = \int_0^{\infty} F_{TW}(t,W^*)\, dt$$

Stidham [1974] tabulates $W(W^*)$ for the standard process (ie., ρ=1). From this table, we find that the expected time to consolidate W^*=7761 pounds is 1.5174 days.

The following table compares the results of modeling this example by the different approaches.

	ESW concept section 5.3	stochastic analysis section 6.2	stochastic clearing system section 6.4
expected load	12,000 lbs.	8677.99 lbs.	7761.25 lbs.
expected length of consolidation cycle	1.11 days	1.35 days	1.52 days
total cost per cycle	$195.56	$184.77	$185.31
E[cost per cwt.]	$1.63	$2.13	$2.38
E[cost per day]	$176.04	$136.87	$121.91

The relative cost performances of the three approaches reflects the different objectives applied in their derivation. Both the deterministic ESW concept and the stochastic analysis of Section 6.2 sought the minimum cost per unit. The goal of the stochastic clearing system was to minimize long–run average cost, a time–based objective.

It also is interesting to compare how each of the above three methods considers the order arrival rate. The ESW concept, being deterministic, assumes that orders arrive at equal time intervals of length 1/â. The stochastic clearing system approach considers the expected arrival rate as part of the build–up rate, but ignores

its probability distribution. Only our stochastic analysis of section 6.2 explicitly treats order arrival time as a random variable. ■

Other insight into consolidation parameters can be gained from Stidham's [1974, 1977] research. For example, if $F_L/\rho r_w > 1.107$, then the optimal clearing weight W^* will exceed the average amount accumulated in one day. Second, if $W^*/\rho \geq 1$, the sojourn measure (the expected time to accumulate W^*) can be estimated by the asymptotic approximation $0.68+W^*/\rho$. Both these relationships provide useful and quick checks on the feasibility of shipment consolidation.

Summary

The application of stochastic clearing system modeling to shipment consolidation has several good features. Unlike the bulk–service queue approach, shipment–release in a stochastic clearing system is determined by the input process, not by the service process. This lack of interaction between the input and output processes is intuitively satisfying, and simplifies the derivation of performance measures because the output process can be ignored.

Unfortunately, for some probability distributions, it is difficult to derive closed form expressions for the sojourn function, and approximations must be used. The extent of research on this topic has been limited, especially that related to derivation of non–linear cost expressions that are mathematically tractable.

6.5 Conclusions

This chapter has examined stochastic non–sequential methods for determining how long customer orders should held for consolidation. We developed probabilistic models of consolidation system performance, and discussed shipment consolidation as a bulk–service queue and as a stochastic clearing system.

A non–sequential approach is most appropriate in a deterministic setting, or one of fairly regular randomness. Sequential methods, which make the shipment–release decision whenever an order arrives, may be preferred when there is high uncertainty (for example, if order arrivals or order weights are extremely variable or random, or if the range of order sizes is large) or where close monitoring of waiting orders is important (for example, if the items to be consolidated are of a perishable nature). Chapters 7 and 8 discuss sequential approaches to shipment release–timing.

Chapter 7
SEQUENTIAL APPROACHES FOR SETTING SHIPMENT-RELEASE PARAMETERS:HEURISTICS AND MARGINAL ANALYSIS

7.1 Introduction

Setting shipment–release parameters through non–sequential approaches provides management with general guidelines for timing the release of shipments. Although these decision aids could be reviewed at the beginning of each load cycle, typically they are set once and applied over a longer period. Non–sequential approaches also are useful when considering the feasibility of shipment consolidation.

There may, however, be insufficient certainty to develop or apply general guidelines to all situations. For example, items being consolidated may be perishable, or the weights or arrival times of customer orders may be highly random. In these cases, the shipment–release decision should be examined in greater detail throughout the order cycle, rather than just determining whether a minimum weight or maximum holding time has been reached.

The sequential approaches examined in this thesis make the shipment–release decision whenever an order arrives (sequential models do exist that allow shipping at times other than the arrival of an order; we will not study these in this thesis). Each time an order is received, the shipper must decide whether to dispatch the accumulated orders immediately or to delay them in hopes that they can be consolidated with the next order to arrive. Thus, given the uncertainties of arrival time and order weight (indeed, whether another order will arrive at all), the benefit from shipping potentially larger consolidated loads must be balanced against the risk that,

in the end, a larger load does not result and the waiting orders have been delayed needlessly. This concept is reminiscent of the newsvendor problem's "one–shot trade–off between the consequences of too much and too little" (Phillips, Ravindran, and Solberg [1976]). In the newsvendor problem, however, an action taken in one period has no effect on any other periods; this is not true in shipment consolidation.

Sequential approaches avoid the difficulty of setting a minimum weight, maximum holding time, or other target for an entire cycle. Because the sequential approaches studied here will make the dispatch decision only when a customer order arrives, lengthy delays could result if the time until the arrival of the next order is unusually long. Thus, combination sequential/non–sequential methods, which apply a sequential approach within maximum time and/or minimum quantity guidelines of a non–sequential approach, may be preferred. Alternatively, some sequential decision algorithms may be applied at any point during the order cycle to reflect revised information.

This chapter and the next discuss sequential approaches to shipment–release timing. Chapter 8 examines Markov decision processes. This chapter proposes two sequential decision heuristics based on marginal analysis. Our first model is applicable to private carriage, while the second model extends this logic to include maximum order holding time and characteristics of common carriage.

7.2 Private Carrier Sequential Decision Heuristic (PCSDH)

Assume that there are n–1 orders waiting to be shipped. Marginal analysis states that we should ship these orders now if the per–order cost of holding them until

the n–th order arrives exceeds the cost savings per order from a consolidated load of n orders. We define $TC_R(n-1)$ as the total inventory cost of holding n–1 orders until the n–th order arrives. $E[TC_R(n-1)]/(n-1)$ is the expected per–order inventory–holding cost of this decision, calculated as:

$$E[TC_R(n-1)]/(n-1) = [\, r_i/\hat{a} \sum_{i=1}^{n-1} i \,] / (n-1)$$

$$= (r_i/\hat{a})\ (n/2)\ (n-1) / (n-1)$$

$$= r_i\ n / 2\ \hat{a}$$

As discussed in Appendix C, private carrier transportation cost largely is fixed for a given distance. Let F_L be the sum of all transportation–related costs per load, so F_L/n is the transportation cost per order. Because F_L/n decreases as n increases, this quantity can be used as a measure of the transportation cost benefit of consolidating n orders. Note that this benefit is in per–unit, rather than per–load, terms.

If the (n–1) orders are delayed until the n–th order arrives, extra inventory–holding charges will be incurred with total certainty. However, additional transportation savings from consolidating n, rather than (n–1), orders will result only if the vehicle physically can transport them. Thus, whenever an order arrives, both the current consolidated weight TW and the difference W_D between vehicle capacity H and accumulated weight TW must be calculated: $TW=\sum_{i=1}^{n} W_i$ and $W_D=H-TW$, where W_i is the weight of order i. Our expression for consolidation benefit then must be adjusted for the probability that the total weight of n orders does not exceed the capacity of the vehicle.

The probability that the weight of the n–th order exceeds the remaining vehicle capacity W_D can be determined if the probability distribution of order weights is known. If this information is not available, Markov's inequality can be applied:

$$\text{Prob}\{W_n \le W_D\} = 1 - \text{Prob}\{W_n > W_D\}$$
$$\le 1 - (E[W] / W_D)$$

In summary, the expected per–unit benefit from consolidating n orders is:

$$E[B_n] = \Pr\{W_n \le W_D\}\ F_L / n$$

where W_n is the weight of the n–th order. The consolidation decision rule dictates to hold (n–1) orders until the n–th order arrives if:

$$E[TC_R(n-1)]/(n-1) \le E[B_n]$$

That is, continue to consolidate if:

$$r_i n / 2\hat{a} \le \Pr\{W_n \le W_D\}\ F_L / n$$
$$\le [1 - (E[W] / W_D)]\ [F_L / n] \quad \text{(applying Markov's inequality)}$$

subject to customer service constraints limiting shipment delays.

This decision rule can be restated as a function of order weight rather than of number of orders. We will delay (n–1) orders with total weight TW until the arrival of the n–th order if the expected per–unit cost of holding weight TW exceeds the expected per–unit benefit of shipping TW plus the expected weight E[W] of the next order. Thus, we delay shipping (n–1) orders until the arrival of the n–th order if:

$$E[TC_R(TW)] / TW \le E[\text{benefit from shipping } TW+E[W]]$$

$$\Rightarrow \quad \frac{(r_w/2\hat{a})(TW/E[W])\ (TW-E[W])}{TW} \le \frac{F_L}{TW + E[W]} \Pr\{W_n \le W_D\}$$

$$\Rightarrow \quad \frac{r_w}{2\hat{a}} \frac{TW - E[W]}{E[W]} \le \frac{F_L}{TW + E[W]} \Pr\{W_n \le W_D\}$$

Including consideration of $Pr\{W_n \leq W_0\}$ in the shipment–release decision is intended to reduce the chance of exceeding vehicle capacity. As a result, this model tends to dispatch loads slightly smaller than the minimum–cost quantity (ie., the ESW). The difference between load sizes of these two approaches is most evident when the ESW is close to or exceeds vehicle capacity.

Lastly, this decision rule (indeed, most decision rules) could suggest holding n–1 orders until order n arrives, only to find that order n is of a weight that causes the consolidated load to exceed vehicle capacity. The appropriate shipment–release action in this situation would depend on management objectives regarding customer service and cost.

Example: We tested our private carrier sequential decision heuristic (PCSDH) by simulating four load–dispatch strategies:

"Exactly ESW" policy: the consolidated load equals the economic shipment weight (ESW) and splitting of shipments between vehicle loads is allowed to attain exactly this shipment weight

"ESW + excess" policy: the consolidated load equals the economic shipment weight, but split shipments are prohibited; if an order causes the optimum weight to be exceeded, that entire order is included in the load as long as vehicle capacity is not exceeded

"Vehicle capacity" policy: a full vehicle is always dispatched, and split shipments may occur due to vehicle capacity restrictions; results for this policy are highly dependent on choice of vehicle capacity

"PCSDH policy": our private carrier sequential decision model with split shipments prohibited

In all cases, vehicle capacity constraints were strictly observed.

As in previous sections, order arrivals were modeled as a Poisson process, with order weights distributed according to an unshifted Gamma distribution with α=2,

β=1000, expected order weight E[W]=2000 pounds, and standard deviation of order weight of 1414 pounds. The economic shipment weight (ESW) was varied by changing the order arrival rate â. The following arrival rate/ESW combinations were simulated:

â (orders per day)	ESW (pounds)
1	12,000
3	13,856
6	16,971
7	18,330
8	19,560
9	20,785
10	21,909
12	24,000

Vehicle capacity was set at 24000 pounds so that runs with larger order arrival rates, hence larger economic shipment weights, would be affected by this constraint. Cost parameters were constant throughout, with inventory–holding cost r_w=\$0.25 per cwt. and fixed transportation cost F_L=\$30 per load.

Results are given in Figure 7–1 for mean cost per cwt. and Figure 7–2 for mean order delay. Significance testing of differences between policies was done using a paired–t confidence interval (Law and Kelton [1991]). All differences were statistically significant at the 90% confidence level, with the following exceptions:

- ESW+excess policy versus vehicle capacity policy for both cost and delay with arrival rate of â=12 orders per day: at this point, most shipments under the ESW+excess policy are equal to vehicle capacity;
- PCSDH versus ESW+excess policy for cost with arrival rate of â=9 per day: at this point, results for these two policies cross in Figure 7–1.

As expected, our private carrier sequential decision model accepts slightly smaller vehicle loads for improved customer service. As a result, it yields a lower

Figure 7–1
Private Carrier Sequential Decision Heuristic:
Mean Cost Per Cwt.

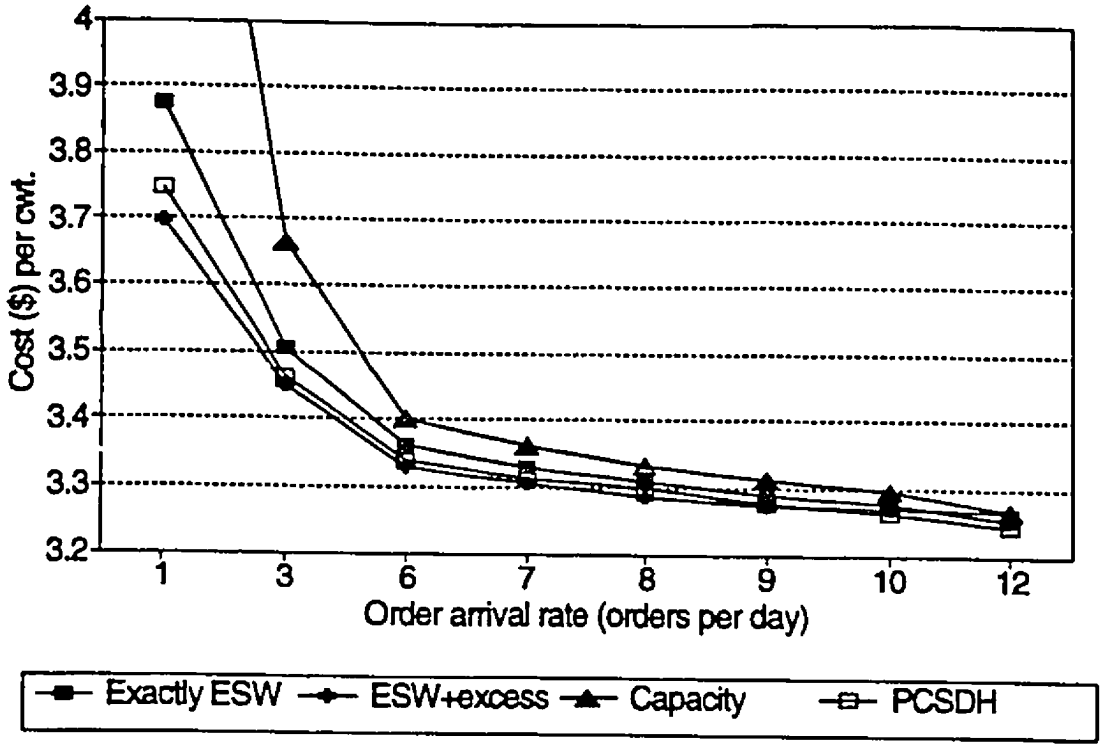
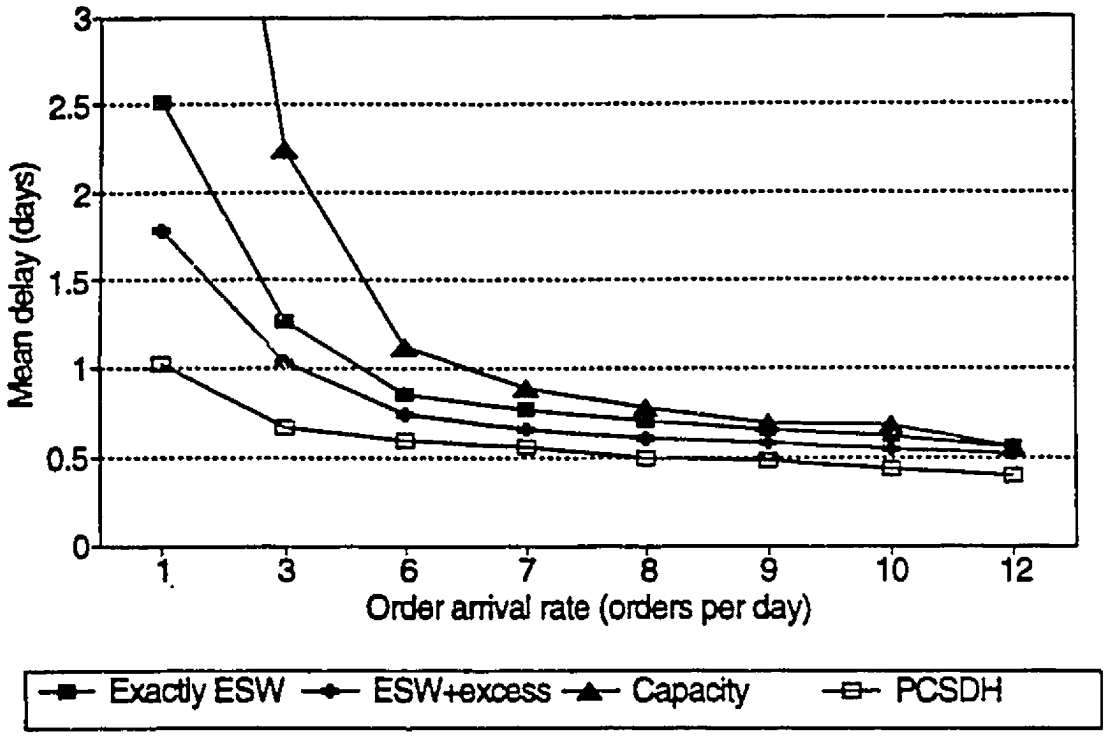


Figure 7–2
Private Carrier Sequential Decision Heuristic:
Mean Order Delay



mean order delay than do the other three strategies. Its per–cwt. cost performance is bettered only by the ESW+excess policy, which yields slightly larger loads and longer holding times than does our PCSDH model. Both the PCSDH and the ESW+excess policy have lower per–cwt. cost than does shipping exactly the economic shipment weight because splitting shipments under an exactly--ESW policy strands portions of orders for longer periods of times, thus incurring higher inventory–holding cost and mean delay than the PCSDH and ESW+excess strategies.

Figures 7–1 and 7–2 also show that, as the order arrival rate increases, the relative advantages of the various load–dispatch policies decrease. This occurs because vehicle capacity becomes an active constraint at large order arrival rates, leading to similar results from all strategies. For example, with an expected order weight of 2000 pounds, an arrival rate of 12 orders per day deterministically would attain vehicle capacity of 24000 pounds in one day. Simulation results showed that the exactly ESW, ESW+excess, and vehicle capacity policies tend to ship loads near or equal to vehicle capacity for arrival rates greater than or equal to â=12. The lower cost performance of the PCSDH seen in Figure 7–2 illustrates, however, that even when the ESW is close to vehicle capacity, a strategy other than dispatching full vehicles still can yield lower per–unit cost and per–order delay. ■

The major difference between our private carrier sequential decision heuristic and the ESW+excess policy is that the PCSDH considers vehicle capacity when deciding whether or not to dispatch an order. The ESW+excess strategy ignores vehicle capacity (unless capacity forces the dispatch), thus the PCSDH tends to dispatch slightly smaller loads than does the ESW+excess.

Our PCSDH resulted in lower per–unit cost than did the ESW+excess strategy for arrival rates of â=10 and â=12 orders per day, while the ESW+excess policy was cheaper than the PCSDH for â≤9. This leads to the question, "Under what conditions should vehicle capacity be considered (or ignored) when making the shipment–release decision?". The order arrival rate at which the per–unit cost performances of the two strategies is equal could be determined by further simulation runs. However, simple linear interpolation of cost results gives an indifference rate of approximately â=9.14 orders per day. This arrival rate yields an ESW of about 20949 pounds.

We can conclude that, with vehicle capacity of 24000 pounds, if ESW≤20949 pounds, capacity can be ignored in the dispatch decision; doing so does not have an adverse effect on average per–unit cost. Recalling that, in our simulation, the mean and standard deviation of order weight was 2000 pounds and 1414 pounds respectively, we see that the difference between vehicle capacity (24000 pounds) and this threshold ESW (20949 pounds) is, in fact, fairly small. Although further research is required to generalize these conclusions to other cases, it appears that the impact of vehicle capacity on the shipment–release decision may be much less than would be intuitively–thought.

The next section modifies our private carrier sequential decision heuristic to include maximum holding time restrictions and common carrier weight breaks.

7.3 Common Carrier Sequential Decision Heuristic (CCSDH)

The logic of our common carrier sequential decision model is similar to that of the private carrier heuristic. The major differences are: i) the existence of freight rate

weight breaks; ii) the consideration of customer service through inclusion of a maximum holding time; and iii) the irrelevance of vehicle capacity. Technically, the second difference changes this model from a sequential to a combination sequential/non–sequential approach. The third difference is based on the usual assumption that, with common carriage, another vehicle can always be found. This assumption ignores some realities of common carrier freight rates; we also will overlook these pricing characteristics.

As each customer order arrives, the accumulated weight TW is updated. If the maximum holding time is reached before the target weight W_{TAR} is attained, the consolidated shipment is dispatched immediately. Otherwise, the accumulated weight is subjected to the same test used in the private carrier model discussed previously; that is, the current accumulation of (n–1) orders is identified as a possible dispatch if:

$$E[TC_R(n-1)]/(n-1) > E[B_n]$$

$$r_l n / 2â > F_L/n$$

Note that because vehicle capacity constraints are not considered in this model, the mathematical definition of $E[B_n]$ is different than that of the private carrier sequential decision heuristic discussed in Section 7.2.

Whenever a potential load dispatch is flagged, a "benefit routine" is entered. This benefit routine first calculates the expected net marginal benefit (additional transportation savings minus additional inventory cost) of delaying vehicle dispatch until the minimum volume weight is reached. It then compares this benefit to the benefit of dispatching a consolidated load immediately. If the net benefit is positive, the load is delayed and consolidation continues.

If the result of this comparison is negative, the routine is repeated, now calculating the expected net benefit based on a target shipping weight of W_{TAR}=WBT (recall from Chapters 4 and 5 that WBT is the smallest weight at which over–declaring total load weight to reduce per–unit transportation cost is justified). This extra calculation is necessary because, although the remaining holding time may be too small to yield positive benefit for the minimum volume weight, there may be sufficient time to reach the WBT weight. Figures 5–3 and 5–4 show that weights between WBT and MWT may be cheaper per unit than weights below WBT. Thus, if the MWT cannot be accumulated within the maximum holding time, weights greater than WBT still might produce positive net benefit.

A flowchart of the sequential decision heuristic is given in Figure 7–3, while Table 7–1 outlines the benefit routine. Our algorithm considers only one weight break, however incorporating additional weight breaks would not be difficult.

Example: Chapter 4 discussed a simulation model for comparing time, quantity, and time–and–quantity shipment–release policies. Our common carrier sequential decision heuristic (CCSDH) was simulated using the same parameters and order data as the shipment–release policies examined in Chapter 4.

Results are given in Figures 7–4 to 7–7 for cost per cwt., and Figures 7–8 through 7–11 for mean delay per order. Results for the three policies simulated in Chapter 4 are identical to the results in Figures 7–4 to 7–11; for example, the cost performances shown in Figures 4–3 through 4–6 are reproduced in Figures 7–4 to 7–7. Figures 7–12 through 7–17 restate our simulation results in terms of order arrival rate.

Figure 7–3
Flowchart of Common Carrier Sequential Decision Heuristic

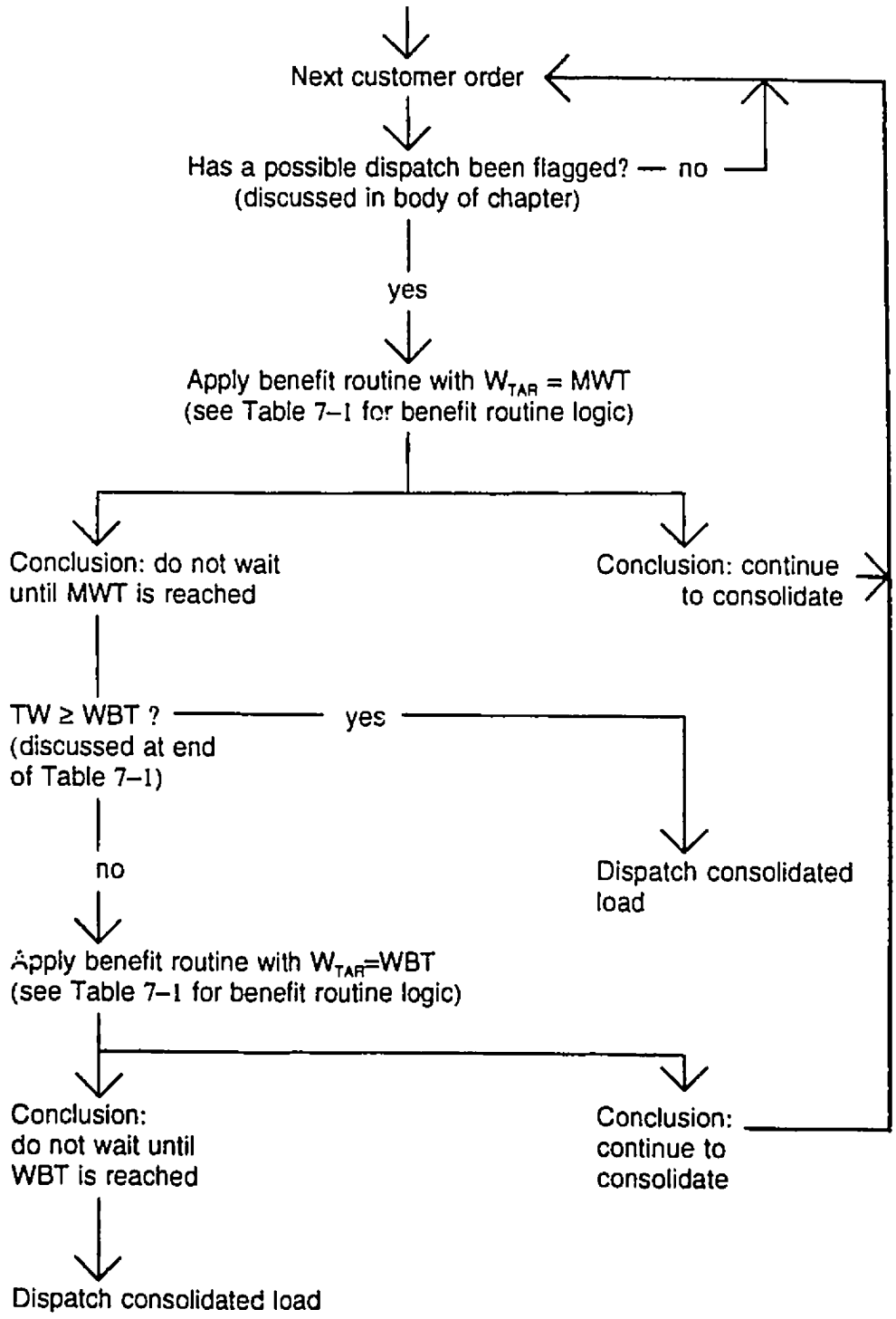


Table 7–1
Common Carrier Sequential Decision Heuristic
Benefit Calculation Algorithm

The following algorithm is used whenever the accumulated weight TW equals or exceeds the target shipment weight W_{TAR}.

Initialization:

1. Determine starting value for target weight W_{TAR}. Typically, this would be based on the economic shipment weight (ESW) formula. If ESW≥MWT, this algorithm is of no value; consolidate until the ESW is accumulated. The remainder of this algorithm assumes that ESW<MWT.

2. Set target weight W_{TAR} equal to the minimum volume weight MWT.

As orders arrive during a consolidation cycle:

3. Update accumulated weight TW as orders arrive.

4. If accumulated weight TW equals or exceeds target weight W_{TAR}, continued consolidation will increase per–unit cost. Thus, dispatch load and exit this routine.

5. If $TW<W_{TAR}$, calculate expected net marginal benefit from delaying load dispatch until the target weight W_{TAR} is reached (see comment at end of algorithm).

   a) Calculate $W_D = W_{TAR} - TW$.

   b) If a load is delayed for further consolidation, it will be held until the earliest of two events occurs: either the maximum holding time is reached, or the accumulated weight reaches W_{TAR}. Thus, calculate T_D:

      T_D = minimum of the time remaining until the maximum holding time T_{MAX} is reached, or the expected time required to accumulate W_D

      = min {T_{MAX} – elapsed time; W_D / arrival rate}

   c) Calculate the transportation savings from shipping weight W_{TAR} rather than TW. These marginal savings are derived from both the spreading of fixed transportation costs over a larger load size and the reduction in per–unit linehaul charges:

      $$B_W = (F_L/TW) - (F_L/W_{TAR}) + f_N - f_V$$

Table 7–1 (continued)

B_W is the net savings per pound from shipping W_{TAR} instead of TW; F_L is the sum of all per–load fixed transportation costs; f_N and f_V are the non–volume and volume freight rates per–pound respectively.

d) Calculate (B_W W_{TAR}), the net savings per load from shipping W_{TAR} instead of TW.

e) Adjust the per–load transportation savings for the probability that the target weight W_{TAR} actually can be attained within the remaining time:

$E[B_L] = B_W$ W_{TAR} Pr{accumulating at least W_D in T_D}

f) Estimate the additional inventory–holding cost from waiting until the earliest of either the maximum holding time or the target weight W_{TAR} is reached:

additional inventory–holding cost = T_D [r_W (TW + W_D/2)]

g) Test the following inequality:

$E[B_L] \geq$ additional inventory–holding cost

- If this inequality holds, do not dispatch a vehicle load. Exit this routine, and continue to consolidate.
- If this inequality does not hold and W_{TAR}=MWT, the cost of delaying shipment until the minimum volume weight is accumulated does not warrant waiting until MWT is reached. However, waiting until weight WBT is reached might yield net positive benefit. Thus, reset target weight W_{TAR} to equal WBT and return to Step 4.
- If this inequality does not hold and W_{TAR}=WBT, positive net benefit from transportation cost savings is unlikely to occur within the remaining allowable waiting time. Therefore, dispatch the load now.

Comment:

When TW≥WBT and further consolidation to reach the MWT is not justified, a possible second test might be to determine the probability of an order of *any* weight arriving in the remaining holding time, then using this probability to perform an expected net savings calculation similar to that above. This extra test was not included in our model.

Figures 7–4 through 7–17 are placed at the end of this chapter to avoid disrupting the text.

At the 90% level, the following differences between our CCSDH and the other policies were not statistically significant:

mean cost per cwt.:

- CCSDH and time policy for â=10.55 with maximum holding time T_{MAX}=1.0;
- CCSDH and quantity policy for â=10.55 with T_{MAX}=1.5 and T_{MAX}=2.0;
- CCSDH and time–and–quantity policy for â≥8.33 with T_{MAX}=2.0.

mean order delay:

- CCSDH and time–and–quantity policy for â≥8.33 with all T_{MAX} values tested;
- CCSDH and quantity policy for â≥8.33 with T_{MAX}=2.0.

As well, some differences were not statisticaliy significant when the plot of the performance of one policy crossed that of another; for example, cost per cwt. for CCSDH and quantity policy for both â=6.38 and â=10.55 for T_{MAX}=1.5.

Figures 7–4 through 7–7 illustrate that our CCSDH model consistently yields per–unit cost as low or lower than those for both a time policy and a time–and–quantity policy. As well, the CCSDH model always resulted in a smaller mean delay than a time policy. Thus, for all values of order arrival rate and maximum holding time, CCSDH dominates a time policy. It does not dominate, nor is it dominated by, a quantity policy nor a time–and–quantity policy.

In general, the CCSDH produces shorter mean delays than a quantity strategy, the exception being for small arrival rates with large holding times. CCSDH also is cheaper than a quantity policy for small arrival rate with large holding times. Under these conditions, the target weight of a quantity policy can be too small to be eligible for volume discounts, whereas the CCSDH, by considering remaining holding time, is

more likely to accumulate loads sufficiently large to qualify for these lower rates. For small holding times and small arrival rates, our CCSDH results in higher cost because larger loads, available under a quantity policy, cannot be accumulated in the relatively short holding time.

Our common carrier sequential decision heuristic produces the same mean delay as a time–and–quantity policy for large order arrival rates (â≥8.33), but yields longer delays when order arrival rates are small, especially for larger maximum holding times. When the holding time is large, the probability of reaching the minimum volume weight within that holding time is greater. Because the CCSDH considers this probability, it is more likely to hold orders until the minimum volume weight is reached. If the target weight under a time–and–quantity policy is less than the minimum volume weight, and the maximum holding time is large, a time--and–quantity policy will dispatch loads more frequently than would the CCSDH. This results in a shorter mean delay but larger per–unit cost. Of course, if management has decided upon a maximum holding time, there is little reason to condemn longer mean delay per order within that allowable holding time in return for lower cost per cwt.

Regardless of maximum order holding time, if the order arrival rate is small or moderate (â≤6.38), the preferred choice of a quantity, time–and quantity, or CCSDH policy is not clear. Monitoring of accumulated weight and remaining holding time then is vital. Our CCSDH could be recommended because it yields a reasonable tradeoff between per–unit cost and mean order delay.

Lastly, if both maximum order holding time and order arrival rate are large (in our simulation, two days and 8.333 to 10.55 respectively) so that the accumulating the

minimum volume weight within the holding time is fairly certain, any policy that monitors accumulated weight and ships when the minimum volume weight is reached is suitable. A time strategy, however, ignores accumulated weight, and with large arrival rates and holding times, frequently ships loads that are larger than necessary for volume rates, thus degrading customer service. ■

7.4 Conclusions

This chapter has introduced sequential approaches to determining shipment release. We have seen that sequential heuristics based on marginal analysis can perform as well or better than non–sequential methods that require setting of a maximum holding time or other time–based rules for an entire cycle. We recognize, however, that our models require some insight as to the order arrival rate and distribution of order weights. Further research is suggested to propose distribution–free methods.

The next chapter continues our discussion of sequential approaches to shipment–release timing by viewing shipment consolidation as a Markov decision process.

Figure 7–4
Common Carrier Sequential Decision Heuristic:
Mean Cost per Cwt.
Holding Time = 0.75 Days

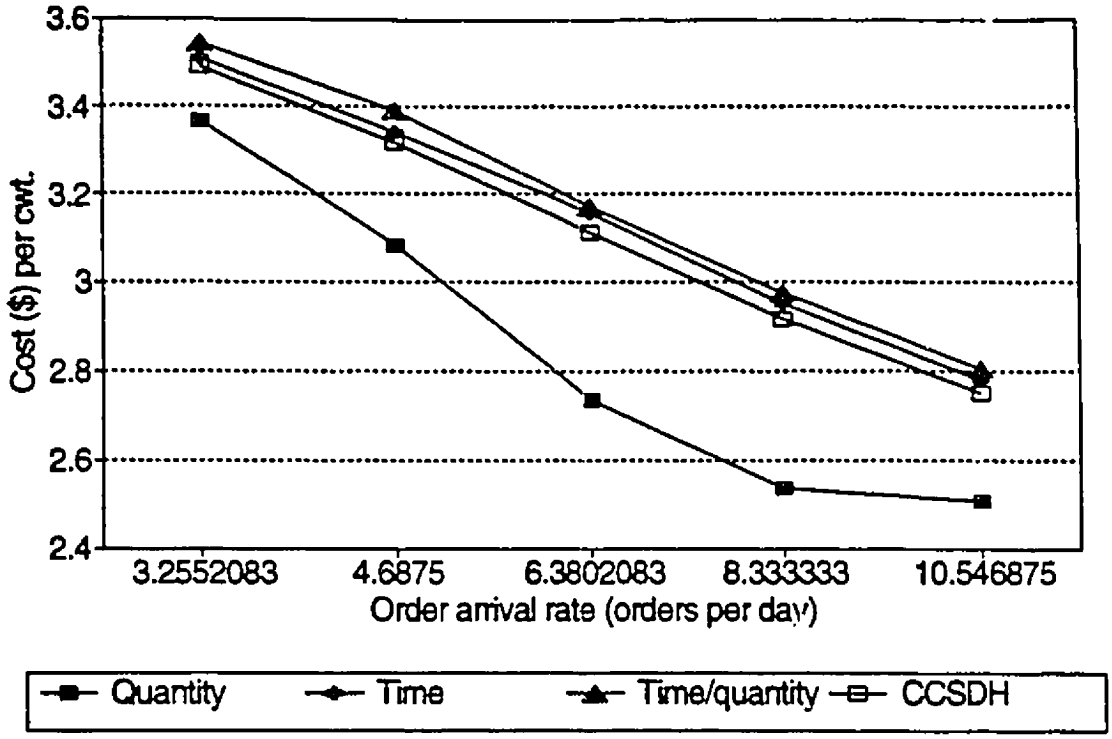


Figure 7–5
Common Carrier Sequential Decision Heuristic:
Mean Cost per Cwt.
Holding Time = 1.0 Days

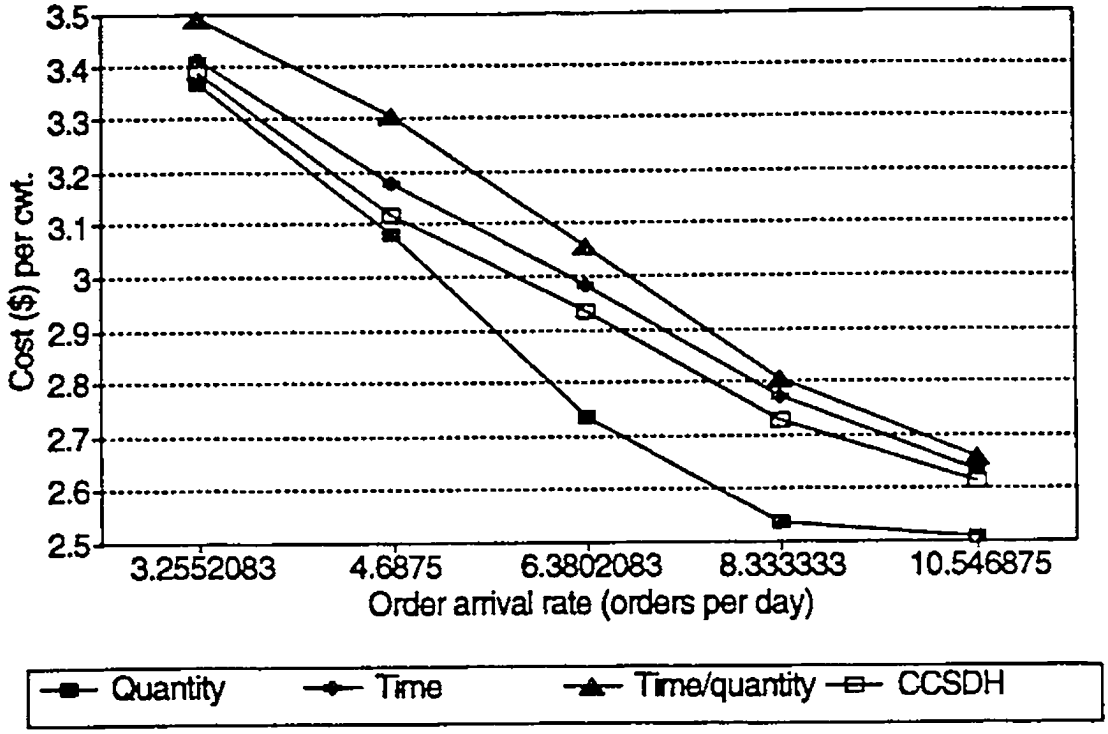


Figure 7–6
Common Carrier Sequential Decision Heuristic:
Mean Cost per Cwt
Holding Time = 1.5 Days

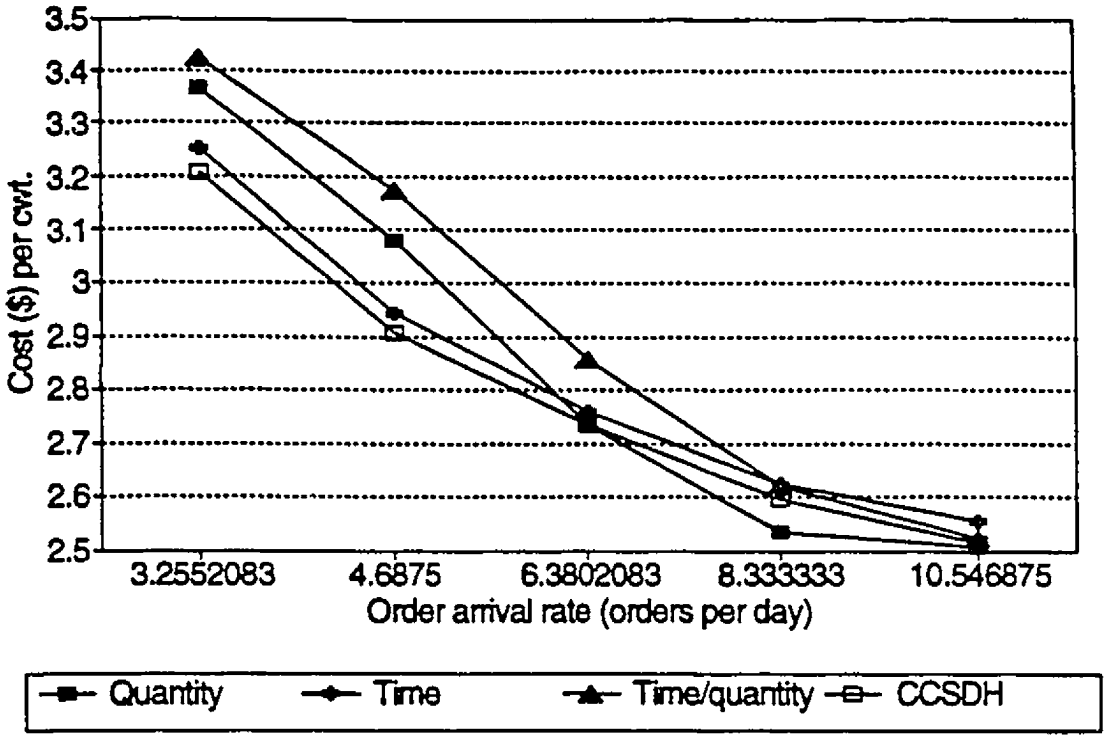


Figure 7–7
Common Carrier Sequential Decision Heuristic:
Mean Cost per Cwt.
Holding Time = 2.0 Days

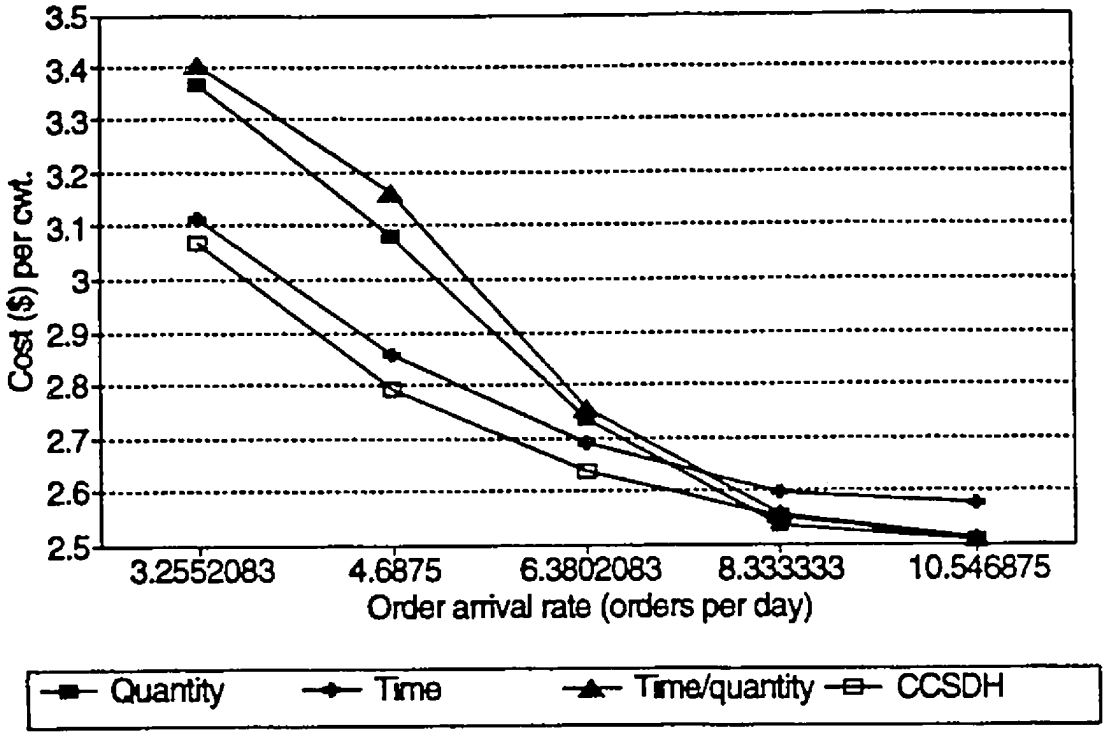


Figure 7–8
Common Carrier Sequential Decision Heuristic:
Mean Order Delay
Holding Time = 0.75 Days

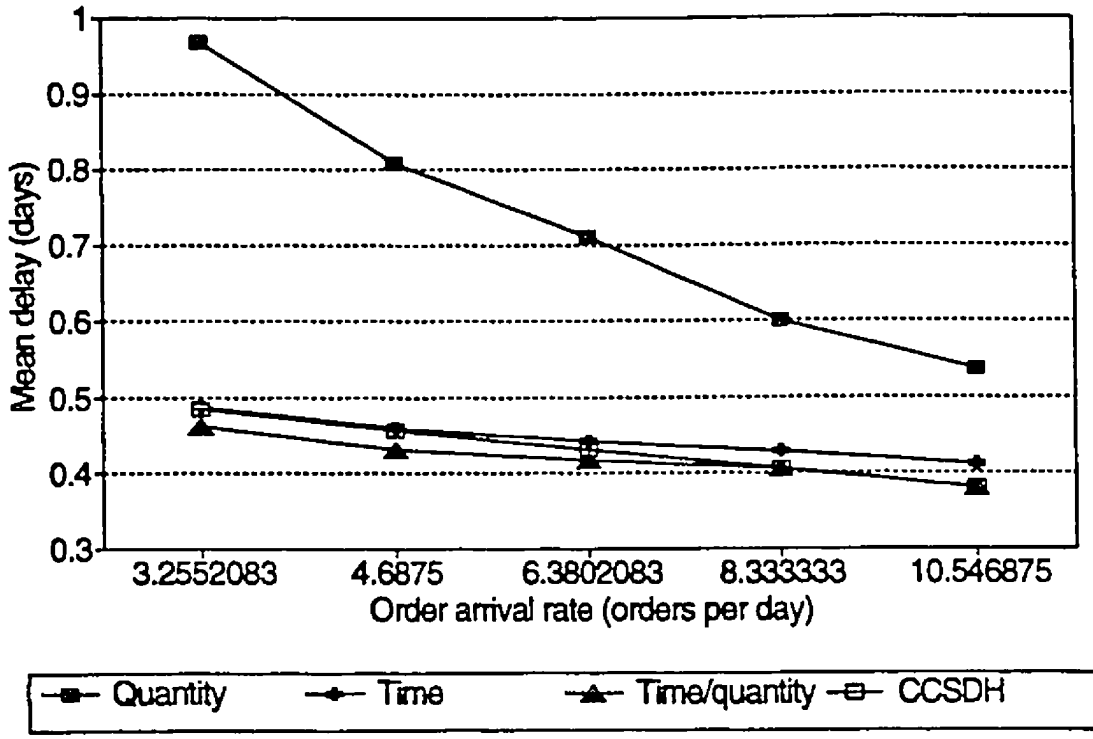


Figure 7–9
Common Carrier Sequential Decision Heuristic:
Mean Order Delay
Holding Time = 1.0 Days

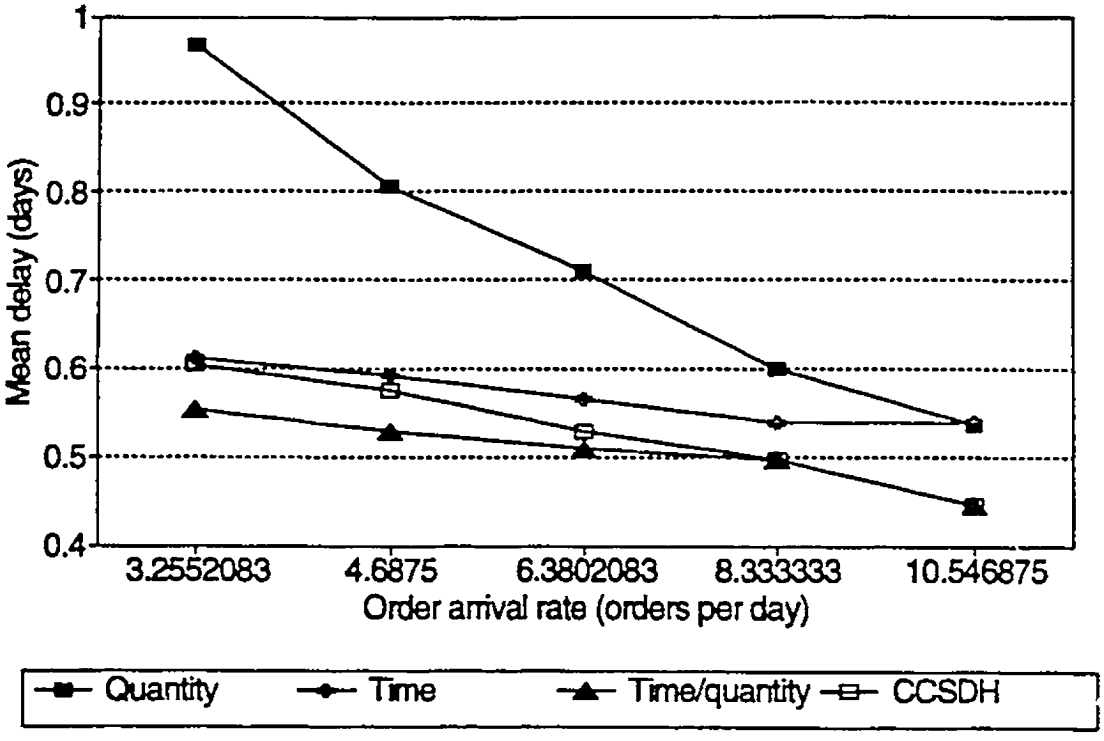


Figure 7–10
Common Carrier Sequential Decision Heuristic:
Mean Order Delay
Holding Time = 1.5 Days

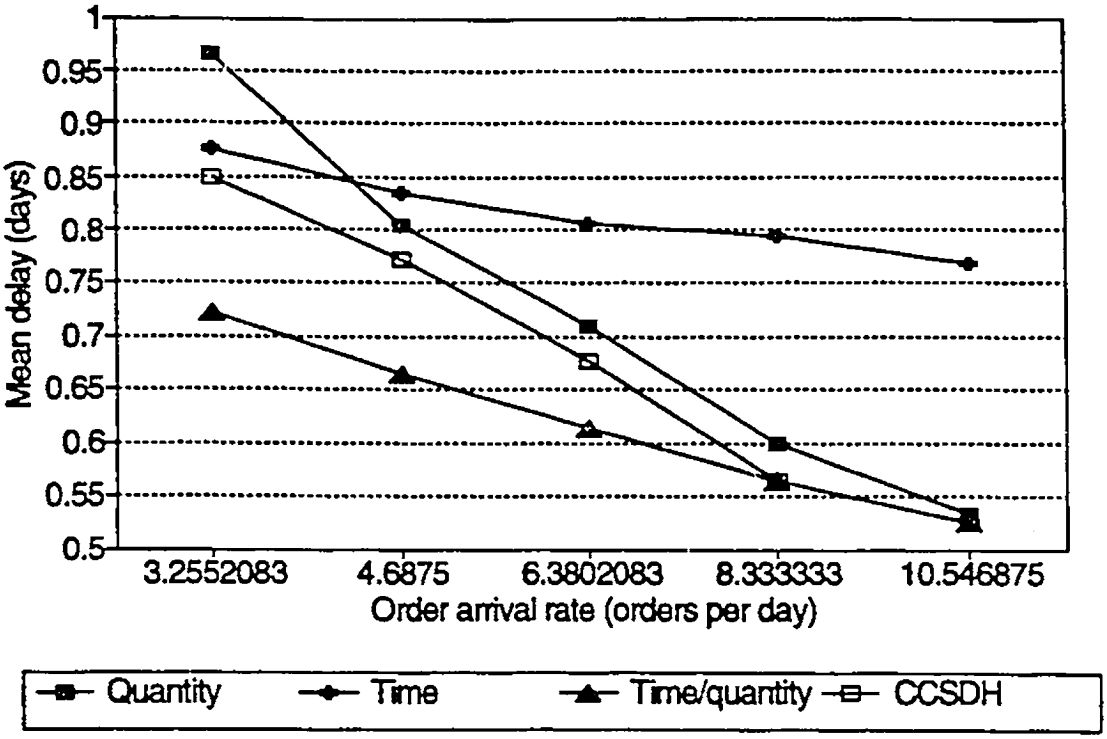


Figure 7–11
Common Carrier Sequential Decision Heuristic:
Mean Order Delay
Holding Time = 2.0 Days

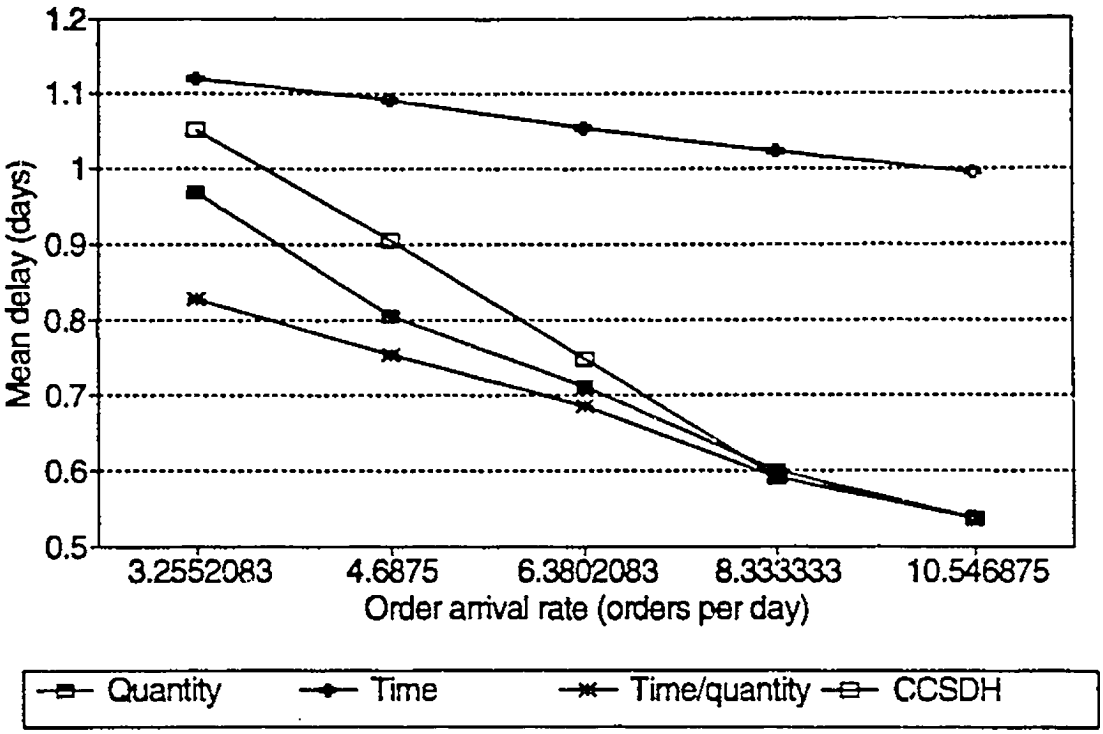


Figure 7–12
Common Carrier Sequential Decision Heuristic:
Mean Cost per Cwt.
Arrival Rate = 3.25 Orders Per Day

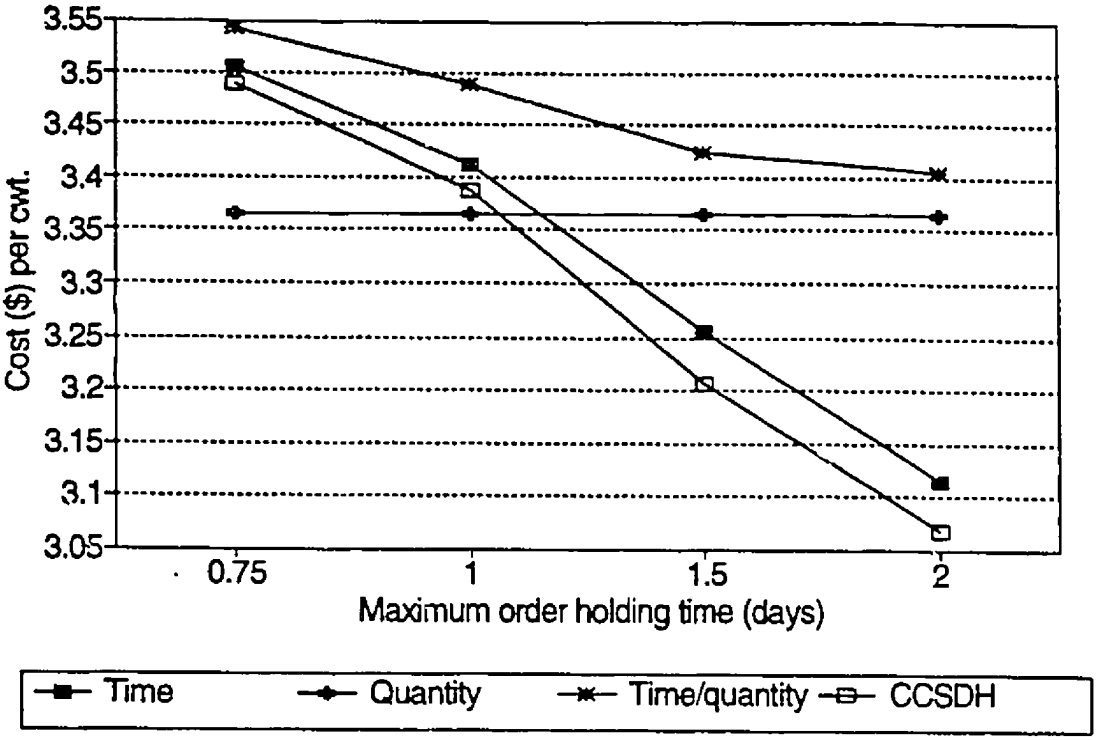


Figure 7–13
Common Carrier Sequential Decision Heuristic:
Mean Order Delay
Arrival Rate = 3.25 Orders Per Day

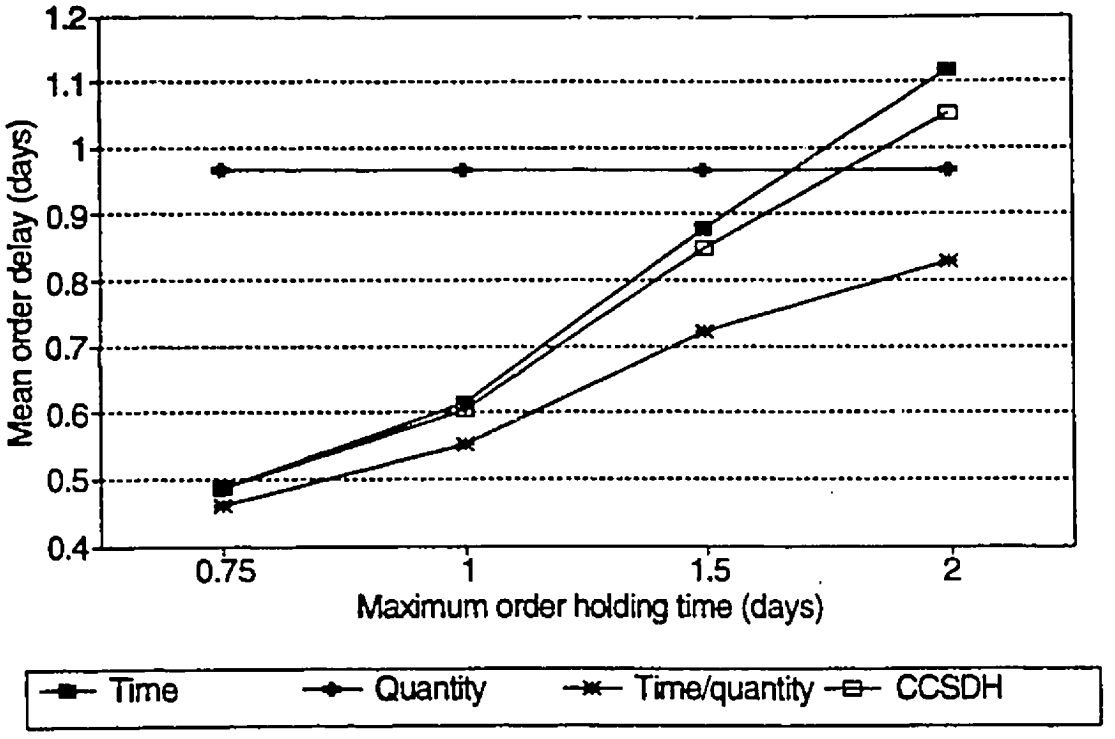


Figure 7–14
Common Carrier Sequential Decision Heuristic:
Mean Cost per Cwt.
Arrival Rate = 6.38 Orders Per Day

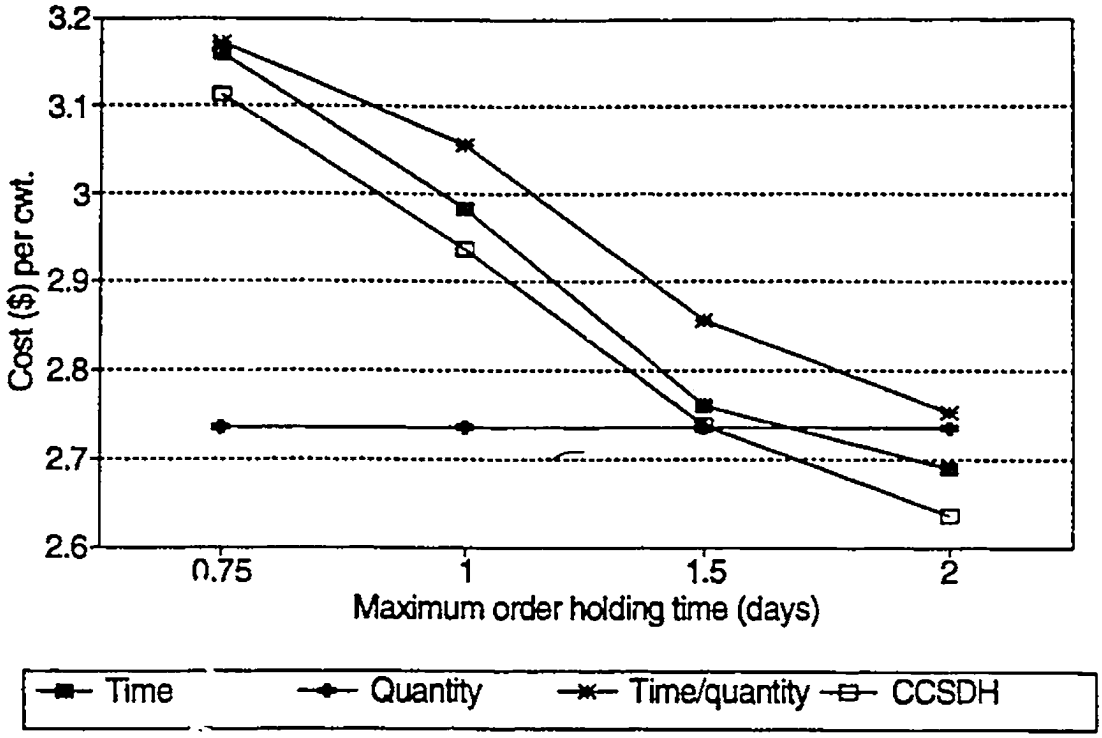


Figure 7–15
Common Carrier Sequential Decision Heuristic:
Mean Order Delay
Arrival Rate = 6.38 Orders Per Day

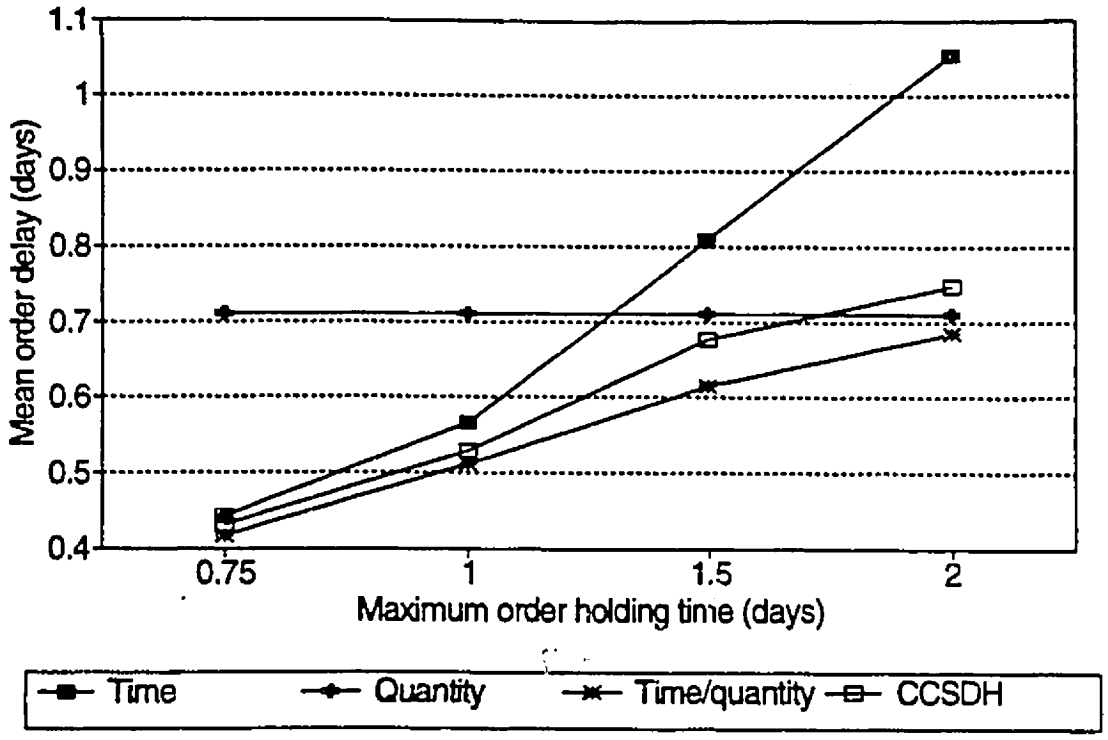


Figure 7--16
Common Carrier Sequential Decision Heuristic:
Mean Cost per Cwt.
Arrival Rate = 10.55 Orders Per Day

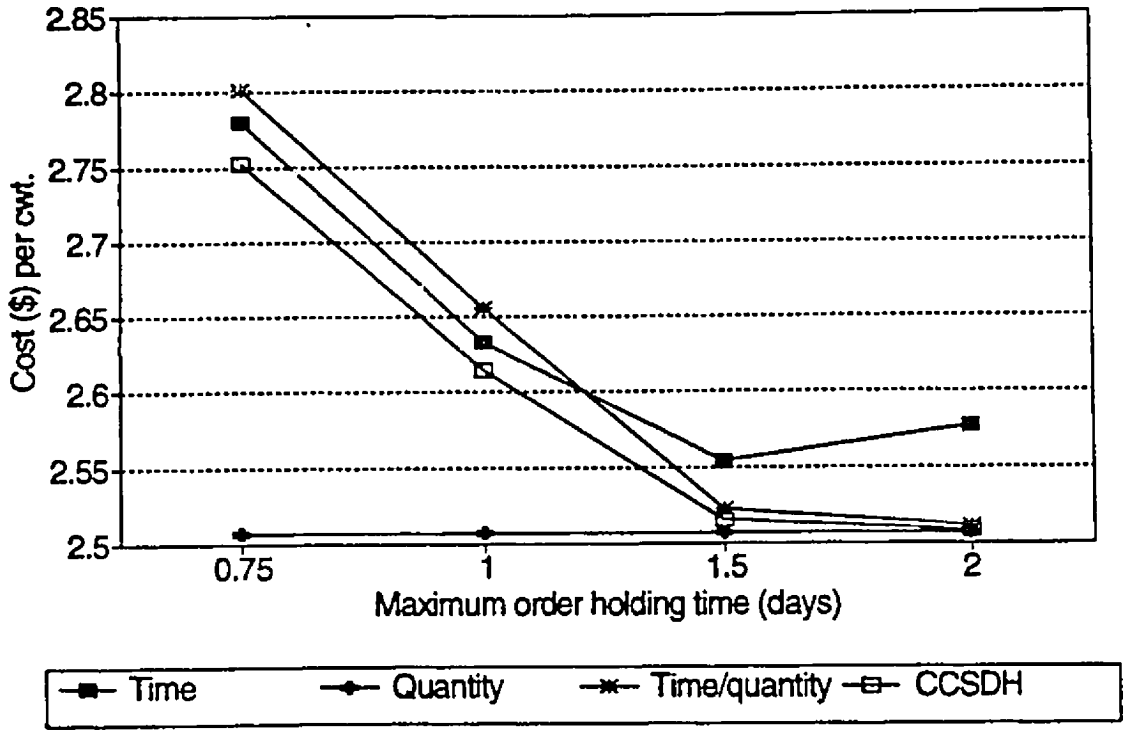
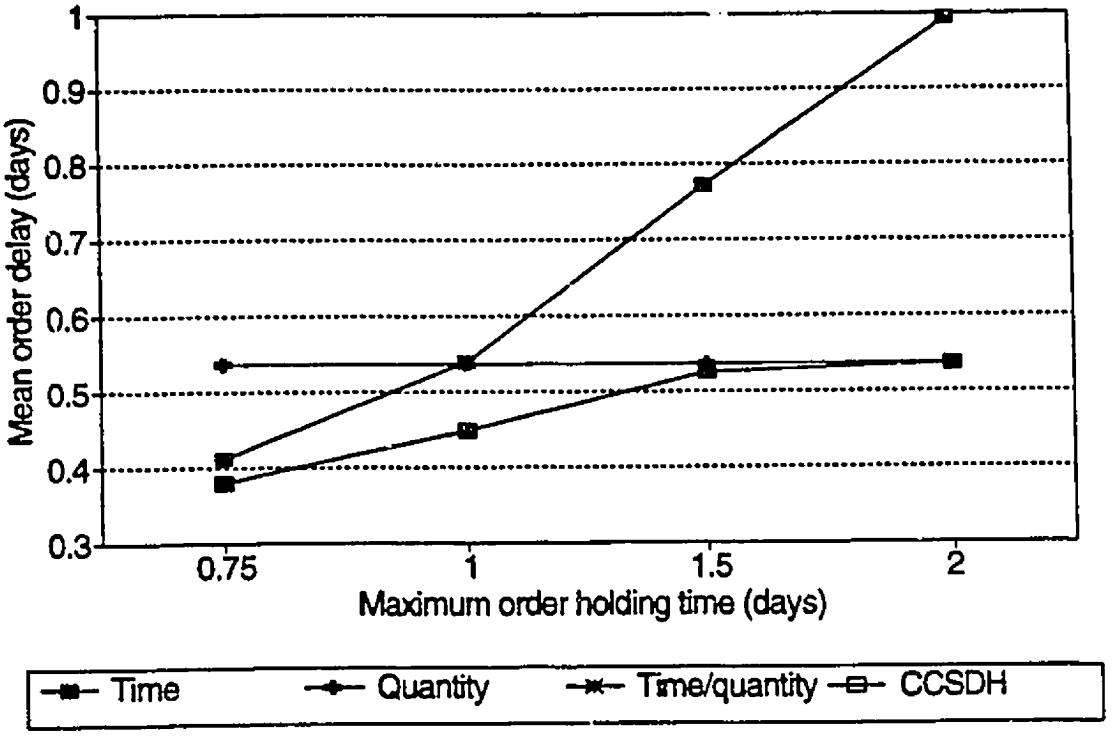


Figure 7–17
Common Carrier Sequential Decision Heuristic:
Mean Order Delay
Arrival Rate = 10.55 Orders Per Day



Chapter 8
SEQUENTIAL APPROACHES FOR SETTING SHIPMENT-RELEASE PARAMETERS:MARKOV DECISION PROCESSES

8.1 Introduction

In shipment consolidation, the random nature of successive order arrivals and order weights suggests a stochastic process where the state of the system is the total weight that has been accumulated. This was seen in Chapter 7, which introduced sequential approaches to shipment–release timing by presenting two probabilistic decision models based on marginal analysis.

This chapter continues our discussion of stochastic sequential methods by examining the use of Markov decision processes in determining whether or not to dispatch a consolidated load. We first outline the basic theory of MDP's. Following this, we present two discrete–time Markov decision models of shipment consolidation, then show how these models can be transformed to continuous–time processes. Lastly, we propose some extensions, including modifications to consider time–based objectives such as customer service.

8.2 Theory of Markov Decision Processes

Appending actions and rewards/costs to a Markov process yields a Markov decision process (MDP). A MDP consists of four basic components:

- a set S of states s that describes the condition of the system at any time;
- a set $A=\{a_s\}$ of actions a_s that may be chosen when in state s;
- a matrix of transition probabilities $p_{s,j}(a)$ denoting the probability of going directly from state s to state j if action a is selected; and

- an immediate reward or cost $r_s(a)$ conferred by a transition from state s by selection of action a.

Also, if the planning horizon is finite, a salvage–value function may exist.

At any time during the process, the system is in one of a set of states and the decision–maker must choose one of a set of possible actions. The action selected results in a transition to another state according to some known probability distribution, and confers an immediate reward or cost. The Markov assumption says that the subsequent state s_{n+1} and the reward/cost $r_s(a)$ resulting from a transition to that state depend only on the state s_n from which the transition was made and the action a_s chosen while in that state.

A Markov policy is any set of decisions specifying, for each state, the action to be taken when in that state, given that the Markov assumption holds. Our goal is to determine the optimal policy; that is, to find the appropriate action a_s to use when in state s so that a specific system measure (such as total reward or cost) is optimized.

There are several types of Markov decision processes. A **finite MDP** has a finite number of states and actions. A **discrete-time MDP** makes transitions between states only at specific, given points in time; often it is convenient to assume that the interval between transitions is constant. In a **continuous-time MDP**, transitions may occur at any moment, and the time spent in any state is exponentially–distributed. Hence, the transition probabilities must be redefined as transition rates.

A **finite-horizon MDP** features decisions over a specified number of periods or transitions, while an **infinite-horizon MDP** has no limit on the number of periods or transitions. The usual objective in a finite–horizon MDP is the optimization of

expected total reward or cost; in an infinite–horizon MDP, it is the optimization of long–run average reward or cost per unit time.

There are three common computational approaches for solving finite Markov decision processes: value–iteration, policy–iteration (also known as policy–improvement), and linear programming. Value–iteration (in effect, stochastic dynamic programming) is limited to the finite–horizon problem. Policy–iteration (which is similar, but not identical, to stochastic dynamic programming) and linear programming are used for solving infinite–horizon Markov decision processes.

Much of the literature on MDP's can be categorized as emphasizing either the theory of Markov decision processes or the solution techniques. Howard [1960] gives an excellent, easy–to–understand introduction to the theory, but focuses on the value–iteration and policy–improvement algorithms. Howard's work is heavily drawn upon by Bellman and Dreyfus [1962], who provide a concise summary of the topic. Ross [1983] covers the theory of Markov decision processes as part of his discussion of stochastic dynamic programming. Other introductory books include Derman [1970], Mine and Osaki [1970], and Thie [1983].

At a higher level, Heyman and Sobel [1984] provide encyclopedic treatment and in–depth examples. Advanced treatment of the subject also is given by Bertsekas [1976] and van der Wal [1981].

8.3 Shipment Consolidation as a Markov Decision Process

Let the state of the system at any time be the accumulated weight being held for consolidation. Whenever an order arrives, a decision must be made whether to

dispatch all waiting orders or to continue consolidation until at least the next order arrives. Thus, we define action 1 (ie., a_s=1, for all s) as "ship immediately", while action 0 (a_s=0, for all s) denotes "continue to consolidate". Selection of action 1 incurs a transportation cost, while action 0 incurs a charge for holding all waiting orders until the next one arrives. Our goal is to find the policy such that the cost per unit–weight averaged over the total consolidation cycle is minimized.

We will assume that order arrivals are independent and identically distributed Poisson random variables, while order weights are i.i.d. Gamma variables. Transition probabilities are stationary. Because the weight of a customer order is a random variable, the number of transitions (ie., customer orders) required to accumulate a consolidated load also is random. This implies that Markov decision models of shipment consolidation are infinite–horizon models, where the objective is to minimize average cost per unit time, rather than expected total cost.

Given our definition of state, the set of states is infinite. However, we can view this as a finite process by arbitrarily rounding all order weights to the nearest integer, then imposing some restriction on accumulated weight. Unfortunately, this still leaves a potentially huge state space. We can handle this (and also can ignore the need to treat order weights as integers) by aggregating states, discussed in the next section.

Aggregation of States

Many stochastic processes can be viewed as Markov decision processes with a suitably large number of states. This easily can yield a model too large to solve efficiently. By aggregating data into batches, the state space of the original Markov

decision process is replaced by a smaller set, thus trading some precision for reduced computational effort.

Aggregation techniques for Markov decision processes are discussed by Heyman and Sobel [1984]. Miller and Rice [1983] propose methods for expressing continuous probability distributions through discrete approximations. Some of their approaches may be useful for aggregation of transition probabilities in a Markov process.

We used the fixed–weight aggregation technique. This method sets a batch size b (b does not have to be constant throughout), then groups the states of the original model to create a new, smaller state space. Let s and k denote states in the original and aggregated state space respectively. For each old state s now included in new state k, we define a weighting constant ω_s, $\omega_s \geq 0$, such that:

$$\sum_{s \in k} \omega_s = 1$$

Let s and j represent states in the original unaggregated model, and k and m be states in the new aggregated model. The original transition probabilities $p_{s,j}(a)$ and single state rewards $r_s(a)$ are weighted by ω_s and restated as $p'_{k,m}(a)$ and $r'_k(a)$:

$$p'_{k,m}(a) = \sum_{s \in k} [\, \omega_s \sum_{j \in m} p_{s,j}(a) \,]$$

$$r'_k(a) = \sum_{s \in k} \omega_s \, r_s(a)$$

These calculations can be tedious, illustrating that sometimes, aggregating data to reduce the size of the model may require more effort than would use of the original data.

In our models, $\omega_s = 1/b$, where b is the batch size; thus, all original states are weighted equally. To reduce the number of calculations, the restatement of transition

probabilities was simplified by approximating the sum of individual probabilities $\sum_{j \in m} p_{s,j}(a)$ through cumulative probabilities. To illustrate, let new state k represent weights from c to c+b–1 pounds, and new state m represent weights from d to d=d+b–1 pounds, where (d–c)/b is a positive integer. We approximate:

$$p'_{k,m}(a) = \sum_{s \in k} \hat{\omega}_s \sum_{j \in m} p_{s,j}(a)$$

$$\approx (1/b) \sum_{i=0}^{b-1} \{ F[d+b-1-c-i] - F[d-1-c-i] \}$$

where F(.) is the cumulative distribution of order weight

Example: To physical distribution practitioners, a meaningful batch size would be 100 lbs. (ie., one hundredweight). In some cases, such as the models discussed here, this batch size still results in a fairly big model, thus a larger batch size is preferred.

Let the batch size of the aggregated model be b=1000 pounds. Thus, state k=4 represents weights from c=3000 to 3999 pounds and state m=6 represents weights from d=5000 to 5999. Using the above aggregation and approximation technique:

$$p'_{4,6}(a) \approx (1/b) \sum_{i=0}^{b-1} \{ F[d+b-1-c-i] - F[d-1-c-i] \}$$

$$= (1/1000) \sum_{i=0}^{999} \{ F[5999-3000-i] - F[4999-3000-i] \}$$

Assume that the system is in original state s=3999 pounds (which is included in aggregated state k=4). To move from aggregated state k=4 to aggregated state m=6, a transition from old state 3999 to any other original state that is contained in state m=6 is required; that is, the system must move from old state 3999 to any old state between 5000 and 5999 inclusive. The corresponding probability is the probability of an arrival (ie., a transition) of size between 1001 and 2000 pounds, which can be calculated by subtracting cumulative probabilities: [F(2000) – F(1000)]. This case occurs in the above mathematical expression when i=999.

If system is in original state s=3000, a transition to new state m=6 requires a transition to any original state between 5000 and 5999 pounds. This is given by the cumulative probability of a transition of 2999 pounds minus the cumulative probability of one of size 1999 pounds. This occurs in the above mathematical expression when i=0. ■

For many of the original states, the single–state rewards $r_s(a)$ are identical, thus aggregation of rewards is considerably easier. We further can reduce the size of our Markov decision models by imposing a finite number of states based on practical considerations, as discussed later.

Sections 8.4 and 8.5 present two discrete–time Markov decision process models of shipment consolidation. Section 8.6 discusses how these models can be transformed into continuous–time processes.

8.4 Common Carrier Discrete–Time Markov Decision Model

With a carrier–specified minimum volume weight of MWT pounds and batch size b pounds, our common carrier Markov decision model requires one state ("state 0") to denote an empty system, one state for weights of MWT and above, and $\lceil MWT/b \rceil$ states to model accumulated weights from one pound to (MWT–1) pounds ($\lceil x \rceil$ denotes the smallest integer greater than or equal to x). Thus, state 0 represents accumulated weight $W_c=0$, state 1 represents weights of $0<W_c\leq b$ pounds, state 2 represents weights of $b<W_c\leq 2b$, etc., and the final state ("state M") represents all $W_c\geq MWT$. Note that it is necessary to set both MWT and WBT (the minimum weight

at which over–declaring load weight is cost–justified, discussed in Chapters 4 and 5) as the lower limits of the states in which they are included.

Let W represent the actual accumulated weight, s and j represent states in the aggregated model, and F(.) denote the cumulative probability distribution of order weight. The transition probabilities $p'_{s,j}(a)$ are:

Action 0 (continue to consolidate):

$p'_{0,0}(a=0) = 0$

$p'_{0,j}(a=0) = F(m)–F(m–1)$ $\quad 1 \le j \le M$

$p'_{s,j}(a=0) \ge 0$ $\quad 1 \le s \le j \le M$

discussed in previous section on aggregating data and probabilities

$p'_{s,j}(a=0) = 0$ $\quad$ all $j<s$

$p'_{M,M}(a=0) = 0$

Action 1 (dispatch all waiting orders):

$p'_{s,j}(a=1) = 0$ $\quad 0 \le s \le j \le M$

$p'_{s,0}(a=1) = 1$ $\quad 1 \le s \le M$

Because this is a discrete–time model, We can assume that the times between state transitions are constant and equal to $1/\hat{a}$, where $\hat{a}$ is the mean order arrival rate. The immediate single–stage costs $r'_s(a)$ for transitions from state s to state j by action a are:

Action 0 (continue to consolidate):

$r'_s(a=0)$ = per–pound cost of holding W pounds for time $1/\hat{a} = r_W/\hat{a}$

Action 1 (dispatch shipment):

$r'_s(a=1) = f_N$ if W<WBT

$r'_s(a=1) = f_V$ if W≥MWT

if WBT≤W<MWT, $r'_s(a=1)$ equals the transportation cost per pound averaged over all original weights W aggregated in new state s; that is:

$$r'_s(a=1) = \Sigma_{Wcs}\ \omega_W\ f_V\ MWT\ /\ W$$

Note that all single–state costs are independent of the successor state. This is one property of a Markov decision process.

As well, for both actions, the resulting transition probability matrix satisfies the unichain assumption: the transition matrix defined by every policy induces a Markov chain with one communicating class, thus one chain of states. With a finite number of states, the unichain assumption implies that both the steady–state limiting probabilities and an optimal policy exist (Howard [1960]; Heyman and Sobel [1984].

Example: Assume the following parameters: arrival rate â=3 orders per day (thus, for this example, inter–arrival times are constant at 1/3 day), minimum volume weight MWT=20000 pounds, WBT=15000 pounds, and volume and non–volume freight rates f_V=$2.25 per cwt. and f_N=$3.00 per cwt. respectively.

Again, let order weights be gamma–distributed with parameters α=2 and β=1000 (so E[W]=2000 lbs.). We aggregate these weights into batches of b=1000 pounds using methods discussed in the previous section. The model is reduced to a finite number of states by recalling from Section 5.3 that no cost benefit occurs from continuing to consolidate past the minimum volume weight. Tests of a model that included states representing weights above the minimum volume weight produced optimal policies identical to the smaller model. Thus, 22 states are required: state 0

represents TW=0, state 1 represents accumulated 0<TW≤1000, state 2 represents accumulated 1000<TW≤2000, etc., and state 21 represents TW≥20000 pounds.

Our common carrier Markov decision process model is given in Table 8–1 for an inventory–holding cost of r_w=0.05 per pound per day. The optimal policy was found via the PROPS software package (Petersen and Taylor [1988]). PROPS applies a variation of the policy–iteration approach discussed in Howard [1960]; linear programming also could have been used.

The following table shows the optimal policy for various values of the inventory–holding cost parameter r_w. Note the gradual "fading in" effect as this parameter increases. This illustrates that higher inventory–holding costs make lengthy accumulation periods suggested by transportation savings less justifiable.

r_w	ship when in these states (otherwise, hold)
0.0465	20, 21
0.048	1 through 3, 20, 21
0.04875	1 through 6, 20, 21
0.0495	1 through 10, 20, 21
0.05	1 through 10, 19, 20, 21
0.0505	1 through 11, 19, 20, 21
0.05125	1 through 12, 19, 20, 21
0.052	1 through 13, 19, 20, 21
0.0535	1 through 14, 18 through 21
0.057	all states

These results show that the shipment action is preferred with very large or very small accumulated weights. In both cases, the expected per–unit cost is less by not continuing to consolidate. With small accumulated weights, loads sufficient to yield transportation savings would suffer from large inventory–holding costs. With large accumulated weights, further transportation savings would be surpassed by additional inventory–holding costs. ■

Table 8–1
Common Carrier Markov Decision Model

State i	Action k	Ending state 0	1	2	3	4	5	6	7	8	9	10	11
0	0	0	0.2642	0.3298	0.2069	0.1075	0.0512	0.023	0.0101	0.0043	0.0018	0.0007	0.0003
	1	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.001	0.0005
	1	1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.001
	1	1	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027
	1	1	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064
	1	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149
	1	1	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335
	1	1	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072
	1	1	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0.1516	0.3164	0.2557	0.1453
	1	1	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0.1516	0.3164	0.2557
	1	1	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0.1516	0.3164
	1	1	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0.1516
	1	1	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0
21	0	1	0	0	0	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0	0	0	0

8	9	10	11	12	13	14	15	16	17	18	19	20	21	Cost
0.0043	0.0018	0.0007	0.0003	0.0001	0.0001	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	10000
0.0064	0.0027	0.001	0.0005	0	0	0	0	0	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.0149	0.0064	0.0027	0.001	0.0005	0	0	0	0	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.0335	0.0149	0.0064	0.0027	0.001	0.0005	0	0	0	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.072	0.0335	0.0149	0.0064	0.0027	0.001	0.0005	0	0	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.001	0.0005	0	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.001	0.0005	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.001	0.0005	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.001	0.0005	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.001	0.0005	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.001	0.0005	0	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.001	0.0005	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0027	0.0015	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0064	0.0042	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0	0	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0149	0.0106	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0	0	0	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.0335	0.0255	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
0	0	0	0	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.072	0.059	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.029
0	0	0	0	0	0	0	0	0	0.1516	0.3164	0.2557	0.1453	0.131	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0273
0	0	0	0	0	0	0	0	0	0	0.1516	0.3164	0.2557	0.2763	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0257
0	0	0	0	0	0	0	0	0	0	0	0.1516	0.3164	0.532	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0243
0	0	0	0	0	0	0	0	0	0	0	0	0.1516	0.8484	0.0167
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0231
0	0	0	0	0	0	0	0	0	0	0	0	0	0	10000
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0225

The general linear programming formulation for solving Markov decision processes is:

maximize $\sum_s \sum_a x_{sa} r_s(a)$

subject to:

$\sum_a x_{ja} = \sum_s \sum_a x_{sa} p_{sj}(a)$ for all j

$\sum_s \sum_a x_{sa} = 1$

$x_{sa} \geq 0$ for all (s,a) pairs

The decision variable x_{sa} is the steady–state limiting probability that the process will be in state s and action a will be chosen when in that state. The objective function seeks to maximize the long–run average reward per unit–time.

The first constraint follows from the definition of x_{sa}, x_{ja}, and $p_{sj}(a)$. The second constraint considers the steady–state probability of a transition from state s via action a, and reflects that the long–run probabilities x_{sa} must sum to 1. Optimal values for x_{sa} are converted to an optimal policy through the expression: $D_{sa} = x_{sa} / \sum_j x_{ja}$, where $D_{sa}=\{0,1\}$. Because of the definition of x_{sa}, for each a, only one x_{sa} will not equal zero (that is, an optimal policy dictates that whenever in state s, we will always pick the same action a). Thus, D_{sa} will be an integer.

Example: Our common carrier Markov decision model has 24 state–action pairs (s,a), thus the linear programming formulation will have 24 x_{sa} variables. This can be reduced to 22 by ignoring action 1 when in state 0 and action 0 when in state 21. There will be 23 constraints (22 of the first constraint, plus one of the second).

The solution of most Markov decision processes via computer is fairly straightforward, with the main hurdle being the amount of input data. The PROPS

software package is fairly fast, easy to learn, and very user–friendly. MDP's also can be solved by simple computer programs built on Howard's [1960] policy–iteration algorithm, or by linear programming. Because of the amount of data input required, both methods are suitable only for models of a reasonable size. ■

An interesting aspect of our common carrier Markov decision model is that it exhibits some characteristics of a separable Markov decision process. Treating a Markov decision process as separable MDP can result in a reduction in the number of possible policies, a less dense coefficient matrix, and a considerably smaller model.

In a separable MDP, some or all the costs $r_s(a)$ can be split into two components, one associated with the state and the other with the selected action. For example, in our common–carrier model, for all states representing weights less than WBT, the per–pound cost of dispatching a shipment depends only on selecting the "ship" action, and not on the state from which it was chosen.

A second characteristic of a separable MDP is that some or all of the transition probabilities depend only on the action, and not on the originating state; that is, $p_{s,j}(a) = p_j(a)$. Because the arrival time of an order is independent of the accumulated weight, the probability of a transition from state s to state 0 is dependent only on whether the "ship" action is selected, and not on state s.

If the conditions for separability hold, one or more new states can be introduced, with "common transitions" (for example, transitions that always end in a certain state) dictated by a specified action allocated to the new state. When the specified action is selected, an instantaneous transition occurs to the new state, followed by a transition, with a probability dependent only on the selected action, to

one of the original states. Replacing several "common transitions" by one transition to the new state increases the number of states. However, it also reduces the number of policies to be considered because some policies that were originally defined by a state–action pair are now defined only by the action. This concept is analogous to shipment consolidation. If we replace several direct transportation links (transitions) between customers (states) by a terminal (new state) and re–route the shipments ("common transitions") through the terminal, the number of transportation routes (policies) between shipper and customer is reduced.

Exploiting the separable structure will not always be advantageous. The separable version will have more states than would the original problem. A smaller linear program will result only if the increase in the number of constraints (caused by adding more states) is overcome by a decrease in the number of variables (caused by reducing the number of state–action pairs). Loosely, the decrease in the number of variables from treating an MDP as separable is positively related to the number of actions available in each state (ie., state–action pairs). Thus, separability is useful only if there are sufficiently many actions available in a large number of states; in this case, the number of "common transitions" that can be replaced by a single transition to a new state may be large. However, in our model, there are at most two actions available in any state.

Viewing our common carrier MDP as a separable problem would aggregate all transitions dictated by "ship" as one state. Since the ship decision already causes an immediate transition to state zero regardless of the original state, formulating our model as a separable Markov decision process would provide no benefit. Discussions

of separable Markov decision processes are given by DeGhellinck and Eppen [1967], Denardo [1968], and Heyman and Sobel [1984].

Refinements to our common carrier MDP model are discussed later in this chapter.

8.5 Private Carrier Discrete–Time Markov Decision Model

The logic of the private carrier Markov decision process model is similar to that of the common carrier model. There are two main differences. First, transportation cost under private carrier largely is fixed per load, regardless of quantity carried. The second difference relates to the maximum quantity that can be shipped.

Under common carriage, unless the economic shipment quantity exceeds the minimum volume weight, continuing to consolidate after this weight has been reached yields no cost benefit. Thus, the minimum volume weight may form an upper bound on the quantity to be consolidated. However, loads exceeding the minimum volume weight still can be dispatched.

With private carrier, the minimum volume weight does not exist, and the maximum consolidated quantity is vehicle capacity. Because vehicle capacity cannot be exceeded, the possibility that some orders will remain after a consolidated load is dispatched forces changes to the model logic.

Let H be the capacity of the vehicle. Assume that s pounds have been accumulated, $s<H$, and that the next order weighs i pounds, with $s<H<s+i$. If orders must be shipped in first–in–first–out sequence, a vehicle carrying weight s would be

dispatched, leaving weight i for the next load. If, however, no restrictions exist on dispatch sequence, the load dispatched should be as large as possible.

A simple Markov decision process reports only the origin and destination states. This information is insufficient to determine which combination of orders will yield the largest load. Thus, we will apply the following arbitrary dispatch rule: if s pounds have been accumulated, and the next order of weight i ($s<H<S+i$) causes a load dispatch, ship weight s if $s \geq i$, otherwise ship weight i. Thus, our private carrier MDP retains the two actions of the previous model: continue to consolidate (action 0), and dispatch a load (action 1) of weight s or i, whichever is largest.

To handle our dispatch rule and to avoid violating the Markov principle, we must introduce new states, transitions to which are dependent on the state of the system *two* transitions ago. For example, suppose the arrival of an order of weight i ($i<H$) causes a transition from state s ($s<H$) to state j ($s,i<H<j \equiv s+i$). This is followed by an immediate (non–random) transition from state j to either state s or state i ($s,j<H$) under action 1. As will be seen, the addition of states representing accumulated weights greater than vehicle capacity results in a large state space.

Example: Consider a small private delivery service with vehicles of maximum capacity 4999 pounds. We ignore all orders of 5000 lbs. or more by assuming that they are shipped in a larger vehicle. This assumption is made only to keep our model small.

As in the common carrier model, we define the states as the accumulated weight pending consolidation, and treat order weight as gamma–distributed with $\alpha=2$, $\beta=1000$, $E[W]=2000$ pounds. Aggregating these weights into batches of 1000 lbs., we

start with six states: state 0 represents no orders waiting, state 1 represents accumulated weights between 1 and 999 pounds, etc., and state 5 represents accumulated weights of 4000 to 4999 pounds. Originally, the probabilities of transitions out of these states are identical to those for our common carrier model. However, these probabilities must be adjusted to eliminate orders exceeding 4999 pounds, so that order weight actually follows a truncated gamma distribution.

Whenever an order arrives that caused the accumulated weight to exceed vehicle capacity, a load is instantaneously dispatched. As discussed above, our dispatch rule is, ship the larger of either: i) the last order, or ii) the accumulated weight (excluding the last order) before the arrival of last order. This will leave the smaller of the two weights for the next load.

Table 8–2 summarizes the states that must be added to our model. The purpose of these extra states is twofold: to model accumulated weights exceeding vehicle capacity, and, for the dispatch rule, to "remember" which state was occupied prior to the arrival of the newest order. For example, in Table 8–2, state 6 can be reached only from state 1; $p_{1,6}(a=0)=0.02315$. Because all states of index greater than 5 represent accumulated weights above 4999 pounds, a transition from state 1 to state 6 implies that the newest order weighs more that the prior accumulated weight. Thus, the transition to state 6 is followed by an immediate transition back to state 1; $p_{6,1}(a=1)=1$, $r_6(a=1)=0.04463$.

When in some new states (ie., s>5), the dispatch of a load can cause a transition to one of two states. For example, a transition from state 4 (3000 to 3999 pounds) to state 12 (5000 to 5999 pounds) is triggered by the arrival of an order

Table 8–2
States Added to Private Carrier Markov Decision Model

		Action 0			Action 1	
from state s	to new state j	state j: weight	p(s,i)	return state i	p(j,i)	r(j)
1	6	5000-5999	0.02315	1	1	0.04463
2	7	5000-5999	0.07474	2	1	0.05108
2	8	6000-6999	0.02315	2	1	0.04463
3	9	5000-5999	0.15084	3	1	0.06931
3	10	6000-6999	0.07474	3	1	0.05108
3	11	7000-7999	0.02315	3	1	0.04463
4	12	5000-5999	0.26544	2	0.614174	0.05754
				3	0.385826	0.05754
4	13	6000-6999	0.15084	3	0.657852	0.05754
				4	0.342148	0.05754
4	14	7000-7999	0.07474	4	1	0.05108
4	15	8000-8999	0.02315	4	1	0.04463
5	16	5000-5999	0.33	1	0.444712	0.004463
				2	0.555288	0.004463
5	17	6000-6999	0.2069	2	0.614174	0.004463
				3	0.385826	0.004463
5	18	7000-7999	0.1076	3	0.657852	0.004463
				4	0.342148	0.004463
5	19	8000-8999	0.0512	4	0.677726	0.004463
				5	0.322274	0.004463
5	20	9000-9999	0.401	5	1	0.004463

weighing between 1001 and 2999 pounds. Since state 4 represents prior accumulated weights exceeding 2999 pounds, the accumulated weight given by state 4 is shipped. The remaining order will be that which just arrived, and will be of a weight modeled by either state 2 (1001 to 1999 pounds) or state 3 (2000–2999 pounds). Thus, entry to state 12 results in an immediate transition to either state 2 or state 3. The probability of this transition to state 2, for example, is just the probability that the most recent order is of weight 1000 to 1999 pounds, conditioned on the fact that a transition can be made only to states 2 or 3.

Table 8–2 shows that the basic six–state model has grown to 21 states. Four states can be eliminated by combining them with other states. For example, states 14 and 15 differ in the accumulated weight they represent, the size of the load dispatched, and the per–pound cost of a vehicle dispatch. However, both result from a transition from state 4, and, after applying our dispatch rule, both return to state 4. Thus, they can be combined as one state, with the per–pound transportation cost in the combined state being the expected value of the costs in the two original states. States 7 and 8 and states 9, 10, and 11 also can be combined, thus reducing the size of our model from 21 to 17 states.

Our private carrier Markov decision process model is given in Table 8–3. Here, the daily inventory–holding cost is r_W=\$0.05 per pound, the fixed transportation cost is F_L=\$200 per load, and the time between transitions is 1/â=1/3 days. ■

One difficulty with our dispatch rule is that it views a shipment as an aggregated weight, rather than as the individual orders that compose it. Thus, the dispatch rule could delay shipping one or more orders for long periods of time. Here,

it would be justified to trade higher total inventory–holding and transportation costs for improved customer service.

One refinement is to divide the "ship action" (action 1) into two actions; that is, if s pounds have been accumulated, s<H, and an order of weight i causes a load dispatch (ie., s+i≥H), ship weight s (action 1) or weight i (action 2). Because the orders composing weight s have been delayed longer, the (dispatch) cost of action 2 could be adjusted to encourage shipment of weight s. This adjustment would not be the same for all states, and would be based on both the mean time that weight s has been delayed, and the physical difference between weight s and weight i. For example, if weight s is small while weight i is large, the "encouragement factor" for weight s should be less than if weights s and i were similar.

The other major problem with our model is its size. The common carrier model had $N_C = (2 + \lceil MWT/b \rceil)$ states. After adding the additional states, but before combining some, our private carrier model has $N_P = 1 + \lceil H/b \rceil + \sum_{\ell=1}^{\lceil H/b \rceil} \ell = (1 + \lceil H/b \rceil)(1 + (1/2)\lceil H/b \rceil)$ states.

Example: A case study by Higginson [1991] discusses shipping cotton. The case notes that a box of cotton has dimensions of 34" by 24" by 14", a gross shipping weight of 110 pounds, and can be piled to a maximum height of 6 cartons. Logical batch sizes are b=990 or b=1100 pounds, so that each batch represents 9 or 10 boxes, respectively. Smaller batch sizes are possible, but will result in models with more states.

If shipped by rail (a common carrier), a typical 50–foot boxcar could hold 450 boxes. This maximum quantity is dictated by volume; the car's weight capacity is not

Table 8–3
Private Carrier Markov Decision Model

State i	Action k	Ending state 0	1	2	3	4	5	6	7	9
0	0	0	0.27532	0.34368	0.21561	0.11203	0.05336	0	0	0
	1	1	0	0	0	0	0	0	0	0
1	0	0	0.15738	0.32845	0.26544	0.15084	0.07474	0.02315	0	0
	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0.15738	0.32845	0.26544	0.15084	0	0.09789	0
	1	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0.15738	0.32845	0.26544	0	0	0.24873
	1	1	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0.15738	0.32845	0	0	0
	1	1	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0.2642	0	0	0
	1	1	0	0	0	0	0	0	0	0
6	0	1	0	0	0	0	0	0	0	0
	1	0	1	0	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0	0	0
	1	0	0	1	0	0	0	0	0	0
9	0	1	0	0	0	0	0	0	0	0
	1	0	0	0	1	0	0	0	0	0
12	0	1	0	0	0	0	0	0	0	0
	1	0	0	0.61417	0.38583	0	0	0	0	0
13	0	1	0	0	0	0	0	0	0	0
	1	0	0	0	0.65785	0.34215	0	0	0	0
14	0	1	0	0	0	0	0	0	0	0
	1	0	0	0	0	1	0	0	0	0
16	0	1	0	0	0	0	0	0	0	0
	1	0	0.44471	0.55529	0	0	0	0	0	0
17	0	1	0	0	0	0	0	0	0	0
	1	0	0	0.61417	0.38583	0	0	0	0	0
18	0	1	0	0	0	0	0	0	0	0
	1	0	0	0	0.65785	0.34215	0	0	0	0
19	0	1	0	0	0	0	0	0	0	0
	1	0	0	0	0	0.67773	0.32227	0	0	0
20	0	1	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0

7	9	12	13	14	16	17	18	19	20	Cost
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	1.49689
39	0	0	0	0	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0.13868
0	0.24873	0	0	0	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0.08111
0	0	0.26544	0.15084	0.09789	0	0	0	0	0	0.0167
0	0	0	0	0	0	0	0	0	0	0.05754
0	0	0	0	0	0.33	0.2069	0.1076	0.0512	0.0401	0.0167
0	0	0	0	0	0	0	0	0	0	0.04463
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.04463
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.04955
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.06154
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.05754
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.05754
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.04955
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.04463
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.04463
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.04463
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.04463
0	0	0	0	0	0	0	0	0	0	1000
0	0	0	0	0	0	0	0	0	0	0.04463

binding. 450 boxes would have a gross shipping weight of 49500 pounds, which exceeds the minimum carload weight MWT of 30000 pounds. Thus, our target is 30000 pounds rather than 450 boxes. Our common carrier MDP with batch size b=990 would have $N_c = 33$ states; with b=1100, 30 states.

From rail tariffs, we see that boxed cotton has a WBT of about 20455 pounds. Although we noted in Section 8.4 that MWT and WBT should form the lower limits of the states in which they are included, with a batch size of 1100 pounds, it would be easier to treat WBT as 20900 pounds, so that WBT would be the lower limit of the state representing accumulated weights of 20900 to 22000 pounds (with b=990 pounds, set WBT at 20790 pounds). The difference in cost would be negligible.

Alternatively, if shipped by private truck, a 45–foot trailer would be able to transport 396 boxes for a total shipping weight of 43,560 pounds. Applying a batch size of 1100 pounds would yield a private carrier Markov decision model with $N_p = 861$ states. The large difference in the number of states for the common carrier and private carrier MDP's results from the dispatch rule and from the existence of vehicle capacity limits. Of course, our dispatch rule is not the only one possible. This and other refinements to our private carrier Markov decision model are discussed later in this chapter. ■

8.6 Continuous–Time Markov Decision Models

The two previous discrete–time MDP models of shipment consolidation assumed that the time between state transitions was constant. In this section, we discuss how to transform these models to the more realistic case where orders arrive

at random times according to a Poisson process with rate â. This results in a continuous–time Markov chain where the time between state transitions is i.i.d. exponentially–distributed with mean 1/â. The "size" of each transition, being based on order weight, remains gamma–distributed.

Howard [1972] notes that there are two necessary conditions for a continuous–time Markov process. First, the length of time that the state has been occupied must have no effect on the determination of the destination state. This condition is met by the assumption of independence of arrival time and accumulated weight. Second, the time remaining until the next transition must be independent of how long the state has been occupied. This implies that time spent in a state is exponentially–distributed, as we have assumed through the Poisson arrival process.

A discrete–time Markov decision process differs from the continuous–time case in two main respects. First, the transition probability matrix of the discrete–time model is replaced by a continuous–time transition rate matrix. Second, rewards in the continuous–time case may be determined by time spent in a state, as well as being conferred whenever a transition is made.

Often, it is not difficult to convert a discrete–time Markov decision process to a continuous–time MDP. This section discusses this transformation, and outlines a simple solution method.

The Continuous-Time Transition Rate Matrix

The discrete–time transition probability matrix P for a Markov process can be transformed to a continuous–time transition rate matrix by the relationship:

$$\hat{A} = \Lambda (P - I)$$

where $\hat{A}$ is the matrix of transition rates $\hat{a}_{sj}$, Λ is a diagonal matrix of the reciprocal of the mean waiting time H_s in state s, P is the discrete–time transition probability matrix, and I is the identity matrix. $\hat{a}_{sj}$ is the rate at which a continuous–time Markov process makes a transition from state s to state j:

$$\hat{a}_{sj} = (1/H_s)\ p_{sj}$$

where H_s is the mean waiting time in state s, and p_{sj} is the transition probability of the discrete–time process.

The mean holding time H_{sj} is the expected length of time spent in state s before moving to state j. The mean waiting time H_s of state s is the unconditional mean holding time in state s when the successor state is unknown:

$$H_s = \Sigma^N_{j=1}\ p_{sj}\ H_{sj}$$

The reciprocal $1/H_s$ thus is the rate at which the process makes a transition when in state s.

Because orders arrive independently of the accumulated weight, our models assume that the exponential density functions $h_{sj}(t)$ of holding times are identical for all states s and j. Thus, the mean waiting time H_s is the reciprocal of λ_s, the parameter of the exponential probability density function. For our continuous–time Markov process, Λ is a diagonal matrix of the parameter λ_s of the exponential waiting–time density function of state s.

Rewards in Continuous-Time Markov Decision Processes

Rewards in continuous–time Markov processes are of two types. First, after entering state s, the system earns a reward r_{ss} per unit time spent in state s. Second, the process earns a reward r_{sj} each time it makes a transition from state s to state j ($s \neq j$). Thus, the two rewards are of differing units. The former, often called the "yield rate", is earned continuously for occupying state s, while the latter is a lump–sum amount earned at the time of a transition from state s to state j. This is reminiscent of a renewal–reward process: a reward is earned at time of shipment, with a renewal occurring when the system empties. The yield rates and lump–sum rewards must be real numbers, but can be positive or negative.

The "earning rate" of the system when in state s is:

$$q_s = r_{ss} + \sum_{j \neq s} \hat{a}_{sj} r_{sj}$$

This expression is similar in form and concept as that for expected earnings found in the discrete–time model.

Solving Continuous-Time Markov Decision Processes

Howard [1960] shows that, if all possible policies of the problem are completely ergodic, the discrete–time decision process and its corresponding continuous–time version are computationally equivalent, and the same computer program may be used for the solution of both after a simple data transformation. He presents modifications to the discrete–time policy–iteration method for solving continuous–time problems. The resulting method is "in all major respects completely analogous to the procedure used in the discrete–time process."

A continuous–time Markov decision process thus can be solved by a policy–iteration routine programmed for solving the discrete–time process by making two transformations. First, if the (computer) algorithm assumes that the discrete–time transition probabilities p_{sj} are $0 \le p_{sj} \le 1$, the diagonal elements $\hat{a}_{ss}$ of the continuous–time transition rate matrix A must be rescaled so that $-1 \le \hat{a}_{ss} \le 0$. This will result in input probabilities $0 \le p_{sj} \le 1$ for all values of s and j.

Next, the transition rate matrix of the continuous process must be transformed to "pseudo" transition probabilities according to the relation:

$$p_{sj} = \hat{a}_{sj} + \delta_{sj} \qquad \text{for all actions a}$$

where δ_{sj} is the Kronecker delta: $\delta_{sj}=1$ if $s=j$ and 0 if $s \ne j$.

The calculation of earnings rate, being similar to that for expected earnings of the discrete–time model, will be done by the policy–iteration algorithm. As a result, there is virtually no change in rewards beyond the ability to handle two types, as discussed above.

Example: The continuous–time version of our common carrier Markov decision process model is given in Table 8–4. The transition rate matrix was transformed from the transition probability matrix P of the discrete–time version given in Table 8–1 according to the relationship $A = \Lambda (P - I)$. As noted above, the matrix Λ of mean waiting time for state s has elements $\lambda_s=3$ on the diagonal and zero otherwise.

As with the discrete–time model, action a=0 denotes continued consolidation, where the only cost incurred is the inventory–holding cost. Thus, $r_s(a=0)$ = per–pound inventory–holding cost per unit time. Action a=1 denotes immediate shipment of the accumulated weight, and results in an instantaneous transition to state 0. Thus, we

Table 8–4
Common Carrier Continuous–Time Markov Decision Model

State i	Action k	Ending state 0	1	2	3	4	5	6	7	8	9	10	1
0	0	-3	0.7926	0.9894	0.6207	0.3225	0.1536	0.069	0.0303	0.0129	0.0054	0.0021	0.000
	1	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.0192	0.0081	0.003	0.001
	1	10	-10	0	0	0	0	0	0	0	0	0	
2	0	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.0192	0.0081	0.00
	1	10	0	-10	0	0	0	0	0	0	0	0	
3	0	0	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.0192	0.008
	1	10	0	0	-10	0	0	0	0	0	0	0	
4	0	0	0	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.019
	1	10	0	0	0	-10	0	0	0	0	0	0	
5	0	0	0	0	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.044
	1	10	0	0	0	0	-10	0	0	0	0	0	
6	0	0	0	0	0	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.100
	1	10	0	0	0	0	0	-10	0	0	0	0	
7	0	0	0	0	0	0	0	0	-2.545	0.9492	0.7671	0.4359	0.21
	1	10	0	0	0	0	0	0	-10	0	0	0	
8	0	0	0	0	0	0	0	0	0	-2.545	0.9492	0.7671	0.435
	1	10	0	0	0	0	0	0	0	-10	0	0	
9	0	0	0	0	0	0	0	0	0	0	-2.545	0.9492	0.767
	1	10	0	0	0	0	0	0	0	0	-10	0	
10	0	0	0	0	0	0	0	0	0	0	0	-2.545	0.949
	1	10	0	0	0	0	0	0	0	0	0	-10	
11	0	0	0	0	0	0	0	0	0	0	0	0	-2.54
	1	10	0	0	0	0	0	0	0	0	0	0	-1
12	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0	0	0	0	0	0	0	
13	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0	0	0	0	0	0	0	
14	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0-	0	0	0	0	0	0	
15	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0	0	0	0	0	0	0	
17	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0	0	0	0	0	0	0	
18	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0	0	0	0	0	0	0	
19	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0	0	0	0	0	0	0	
20	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0	0	0	0	0	0	0	
21	0	0	0	0	0	0	0	0	0	0	0	0	
	1	10	0	0	0	0	0	0	0	0	0	0	

10	11	12	13	14	15	16	17	18	19	20	21	Cost
0.0021	0.0009	0.0003	0.0003	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	10000
0.003	0.0015	0	0	0	0	0	0	0	0	0	0	0.0166
0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.0081	0.003	0.0015	0	0	0	0	0	0	0	0	0	0.0166
0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.0192	0.0081	0.003	0.0015	0	0	0	0	0	0	0	0	0.0166
0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.0447	0.0192	0.0081	0.003	0.0015	0	0	0	0	0	0	0	0.0166
0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.1005	0.0447	0.0192	0.0081	0.003	0.0015	0	0	0	0	0	0	0.0166
0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.216	0.1005	0.0447	0.0192	0.0081	0.003	0.0015	0	0	0	0	0	0.0166
0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.4359	0.216	0.1005	0.0447	0.0192	0.0081	0.003	0.0015	0	0	0	0	0.0166
0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.7671	0.4359	0.216	0.1005	0.0447	0.0192	0.0081	0.003	0.0015	0	0	0	0.0166
0	0	0	0	0	0	0	0	0	0	0	0	0.03
0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.0192	0.0081	0.003	0.0015	0	0	0.0166
0	0	0	0	0	0	0	0	0	0	0	0	0.03
-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.0192	0.0081	0.003	0.0015	0	0.0166
-10	0	0	0	0	0	0	0	0	0	0	0	0.03
0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.0192	0.0081	0.003	0.0015	0.0166
0	-10	0	0	0	0	0	0	0	0	0	0	0.03
0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.0192	0.0081	0.0045	0.0166
0	0	-10	0	0	0	0	0	0	0	0	0	0.03
0	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.0192	0.0126	0.0166
0	0	0	-10	0	0	0	0	0	0	0	0	0.03
0	0	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0447	0.0318	0.0166
0	0	0	0	-10	0	0	0	0	0	0	0	0.03
0	0	0	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.1005	0.0765	0.0166
0	0	0	0	0	-10	0	0	0	0	0	0	0.03
0	0	0	0	0	0	-2.545	0.9492	0.7671	0.4359	0.216	0.177	0.0166
0	0	0	0	0	0	-10	0	0	0	0	0	0.029
0	0	0	0	0	0	0	-2.545	0.9492	0.7671	0.4359	0.393	0.0166
0	0	0	0	0	0	0	-10	0	0	0	0	0.0273
0	0	0	0	0	0	0	0	-2.545	0.9492	0.7671	0.8289	0.0166
0	0	0	0	0	0	0	0	-10	0	0	0	0.0257
0	0	0	0	0	0	0	0	0	-2.545	0.9492	1.596	0.0166
0	0	0	0	0	0	0	0	0	-10	0	0	0.0243
0	0	0	0	0	0	0	0	0	0	-2.545	2.5452	0.0166
0	0	0	0	0	0	0	0	0	0	-10	0	0.0231
0	0	0	0	0	0	0	0	0	0	0	0	10000
0	0	0	0	0	0	0	0	0	0	0	-10	0.0225

assign a "large" value to the transition rate $\hat{a}_{s,0}(a=1)$; we have used "10" for this purpose in Table 8–4. That transition incurs a per–pound freight cost, so $r_s(a=1)=f_N$ or f_V, depending on the state from which the transition occurred. ■

8.7 Incorporating Customer Service in Markov Decision Models

Our Markov decision models seek to minimize the average cost of the process per unit time. This section briefly discusses methods for including customer service in our MDP models of shipment consolidation.

Adjust the Transition Probabilities to Consider Remaining Holding Time

Imposing a maximum order holding time adds a time component to the transition probabilities. This is contrary to the assumption of model stationarity, which states that these probabilities are independent of time. Moreover, because we have defined the state of the system to be the accumulated weight, transition probabilities are dependent on the distribution of order weight. Assuming that customers do not have knowledge of the state of the system at any time, this distribution should not be affected by elapsed waiting time.

Time–varying problems (ie., those where the transition probabilities change from one transition to another) can be solved as sequence of interrelated time–invariant problems. However, Howard [1971a] notes that since the transition probabilities can change at each transition, limiting state probabilities usually have little meaning. Thus, "because of the difficulties in specification and in computation, time–varying Markov models are not often used."

Adjust the Single-Stage Rewards

In our discussion of shipment dispatch rules for the private carrier Markov decision model, we suggested that shipment of older orders could be encouraged by adjusting the dispatch cost of shipping more recent arrivals. Similarly, lengthy consolidation cycles could be discouraged by applying a penalty cost to "continue to consolidate" action as the accumulated weight grows. This penalty could be based on expected elapsed time or more subjective measures, such perceived customer attitudes towards longer delays.

Apply a "Goodness Test" to the MDP Optimal Policy

Whenever our Common Carrier Sequential Decision heuristic (CCSDH) of Section 7.3 flagged a potential shipment release, a "benefit routine" was applied to decide whether the dispatch was justified. Similarly, whenever the optimal policy of a MDP suggests continued consolidation, a decision rule could be used to judge the "goodness" of this action in light of elapsed holding time or other criteria not explicitly included in the Markov model. Such a decision rule could reflect the marginal approach discussed in the previous chapter.

Treat Customer Service Outside the Model

The simplest way to include customer service in a Markov decision model is to apply a combination sequential/non–sequential approach. A maximum holding time would be set in light of customer service goals. The optimal MDP policy, being based on cost, would be followed until it dictates a shipment dispatch, or until the maximum

holding time had elapsed, which ever occurs first. Thus, this approach results in a time–and–quantity policy, discussed in Chapter 4.

8.8 Markov Decision Models of Shipment Consolidation: Other Extensions

Our Markov decision models considered only the costs of transportation and inventory–holding. As discussed in Appendix C, it is usual practice to allocate distribution costs to shipments on a weight basis. This allocation–base is consistent with our objective of minimizing average cost per unit weight per unit time. Extending our models to include other logistics costs, such as those of loading, unloading, and shipment–handling, would not be difficult.

A second refinement would include consideration of transportation mode in the ship action (our action a=1). For example, a MDP for consolidating small shipments could define actions such as "ship by mail", "ship by courier", and "ship by air–freight". This extension of the "ship" action would be most effective for weight ranges where many alternative transportation services are available, and for models with smaller weight aggregation batch sizes. It probably would not be useful in our examples, which have large order weights, weight range, and aggregation batch size, because the number of competing carrier services is limited.

8.9 Conclusions

Chapters 7 and 8 have discussed the use of sequential approaches to determine when to dispatch a consolidated load. By making the shipment–release decision whenever an order arrives, these methods avoid the difficulty of setting a

maximum holding time or other time–based rules to cover an entire cycle. There is the danger, however, that if order arrivals are infrequent, orders being held for the next arrival may incur excessive waiting times. Thus, the use of a sequential approach within the guidelines of a non–sequential method may be preferred.

Sequential approaches to shipment consolidation have received far less attention by researchers than have non–sequential methods. As such, both sequential and combination sequential/non–sequential approaches for determining dispatch–timing of consolidated loads appear to have good potential for additional research.

Some general conclusions and suggestions for further research related to shipment consolidation are given in the next chapter.

Chapter 9
CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

9.1 Conclusions

It was seen early in this thesis that shipment consolidation can be a valuable competitive tool. The potential benefits of reduced transportation cost, faster and more direct transportation, and improved equipment and employee utilization have led to increased use of this strategy, especially as means of handling spiralling distribution costs and customer demand for smaller order sizes, shorter leadtimes, and more frequent deliveries.

Although shipment consolidation has been recognized by both practitioners and academics, strategic and operational guidelines for a consolidation program were lacking before the present research was undertaken. This thesis examined shipment–release policies: approaches for determining how long customer orders should be held for consolidation. Such decision rules have been identified as a major factor in the success of a shipment consolidation program (see Table 3–1).

We first studied the impact on cost and order delay of three commonly–used shipment–release policies. We saw that the selection of a shipment–release policy can be very complex, requiring knowledge of the frequency and size of customer orders and the level of service required by customers.

Following this, we discussed non–sequential and sequential approaches to determining parameter values within a shipment–release policy. We noted that there are two types of analytical methods for determining how long customer orders should be held pending consolidation. Non–sequential approaches treat the shipment–

release question as a "one–time" decision, and result in set of guidelines or targets. Sequential approaches make the decision whenever an order arrives. Use of a sequential approach within the guidelines suggested by a non–sequential approach may be preferred, so as to encourage review of conditions throughout a consolidation cycle limited to a target weight or time.

We noted in Section 4.1 that decisions in a shipment consolidation program can be generalized as "What?", "When?", "Where?", "Who?", and "How?". The "When?" question studied in thesis is only one consideration that must be examined during the implementation and use of a shipment consolidation program; each of the other questions also require analysis. Clearly, shipment consolidation is a broad and important area for both logistics practitioners and researchers.

The next section discusses areas related to shipment consolidation that warrant further research. This is followed by some closing remarks.

9.2 Suggestions For Further Research

Extension of research on shipment-release policies and optimal order-holding time

Some models presented in this thesis can be extended. For example, Section 6.2 discussed setting T_{MAX} through visual inspection of cumulative probability plots. We sought a value for T_{MAX} such that the probability $Pr\{T_{MAX}|N^*\}$ of accumulating the optimal number N^* of orders in time T_{MAX} is greater than or equal to some management–set parameter. Determining an appropriate value of this management–

set probability was not discussed, and would require consideration of characteristics of items shipped and management attitudes toward cost and customer service.

Our research considered customers and shipments in aggregate, treating all consignees and all orders equal. Moreover, we have viewed consolidation from the standpoint of "orders", not customers. Taking a customer–based view of consolidation may require a more detailed analysis. For example, different types of items or customers may not have the same order arrival rates or service priorities. Instead of a probability distribution of order arrivals, we would consider the probability that an order will be placed by a specific consignee or for a specific product. The "spike–gap–tail distribution", which has been used to model demand from individual customers, may be of use here.

Similarly, some components of the physical distribution system may be treated separately, rather than in aggregate. For example, total order–cycle time may be viewed as the sum of the order–holding time and transportation time, each with a different probability distribution. Modeling the consolidation cycle in this manner may yield insights into the effect of these two components on the performance of a consolidation program.

Lastly, we note that many analytical studies of shipment–release timing have sought an optimal shipment size under varying conditions and distribution system designs. As mentioned in the conclusion to chapter 5, the economic shipment quantity concept (a non–sequential approach) has received considerable attention. Thus, further development of sequential approaches to shipment–release timing appears to hold promise of interesting research.

Investigation of methods to measure consolidation potential of specific customer orders

The ability to mix items of different types, due–dates, and destinations within the same vehicle can complicate the consolidation process. Methods are required to decide exactly which orders should be shipped immediately and which should be held for possible consolidation.

One approach to this problem is through a dynamic ABC–type analysis for classifying consolidation potential of orders. Major factors to be considered include the type of item, its weight and volume, and its consignee and destination. "A" items would be those that require immediate direct shipment; examples include shipments that require all or most of a vehicle, or those destined for an important Just–In–Time customer. "C" items would be those that could be delayed without impairing customer service and have a high probability of eventual consolidation. Of course, as the due date for a delayed "C" item nears, it would become an "A" item.

The major difficulty with such an ABC–analysis is the large degree of subjectivity when considering such factors as consignee priority. Indeed, this appears to be the chief difficulty in applying a multi–criteria ABC–analysis developed by Flores and Whybark [1986] for inventory management.

Combining the more subjective aspects of shipment consolidation with an objective bin–packing heuristic also may prove to be interesting research.

Examination of implications of "downstream" consolidation policies

Our research on shipment–release policies and timing has been limited to a single location and single level of the distribution system. Examination of shipment consolidation at other levels of the distribution system and/or with use of multiple facilities has been limited, and research on "downstream consolidation" and its impact on both suppliers and customers typically has been from a strategic, rather than operational, viewpoint (for example, Cooper [1984]).

Policies required for effective consolidation at a warehouse, for example, can be very different from those examined in this thesis. A warehouse may handle both single orders to be filled from stock and multiple orders arriving as part of a vehicle load for reloading and delivery. Delivery schedules both to and from suppliers and customers, choice of linehaul transportation strategy, backhaul planning, and utilization and location of other facilities in the distribution system also must be addressed. A good introductory discussion of some of these considerations is given in Bookbinder and Higginson [1990].

Integration of shipment consolidation and vehicle routing

Research on both vehicle routing and traveling salesman problems typically seeks to minimize total tour distance or total tour cost. Rarely, if ever, has such research considered the items being transported when determining tours.

The major question when integrating shipment consolidation and vehicle routing is, "should vehicle routes be determined first, then used to decide what shipments go on each vehicle, or should consolidated loads be assembled first, then used in setting

vehicle routes?" As well, research should recognize that, unlike much work on the traveling salesman problem, a vehicle may not be able or may not have to visit every possible destination on the tour (see, for example, Jaillet and Odoni [1988].

Interaction of shipment consolidation and facility location

The impact of shipment consolidation on facility location decisions has been outlined by Bookbinder and Higginson [1990]. This research can be expanded in several areas; for example:

- consideration of the effect of shipment consolidation on facility function;
- development of guidelines as to maximum and minimum number of consolidation terminals, and their function, required for effective consolidation;
- examination of impact of linehaul routing strategy on location of facilities;
- development of a facility location algorithm incorporating shipment consolidation.

Development of operational guidelines for inbound shipment consolidation

Inbound shipment consolidation is characterized by increased consignee control over demand for items in a shipment. This control, not usually present in outbound consolidation, gives the consignee the ability to alter the size, timing, and routing of inbound shipments to maximize service and cost benefits. This is especially valuable when the consignee operates under Just–In–Time conditions.

Although much study has been done in the area of joint replenishment of inventory items, only recently has research attempted to integrate inventory restocking

and transportation factors. Buffa [1986b, 1987, 1988], for example, has done considerable work on order grouping methods and order frequency in inbound logistics. Extension of this topic also should consider the impact of vehicle routing strategy on coordinated inventory control.

Research in integrating inbound and outbound consolidation also is limited. Such practice forces the scheduling of flows on inbound and outbound links to avoid exceeding facility or vehicle capacity constraints.

Examination of shipment consolidation by carrier

Quantitative analysis of shipment consolidation typically has focused on shipper–performed consolidation. Use of such analysis to assist a carrier in managing his consolidation program has been virtually ignored.

Common carrier consolidation differs from shipper–performed consolidation in several ways. Typically, a common carrier will have a wider variety of shipments, received from a greater number of origins and going to a larger set of destinations. There will be less control over arrival times, physical characteristics, and due dates, of items to be transported, though typically will have better access to break–bulk and make–bulk consolidation terminals than would a private carrier. Lastly, the cost structures will be different; for example, inventory–holding cost would be irrelevant to a common carrier, though not to the shipper whose goods are being transported.

Refinement of costing methods and cost estimation for consolidated shipments

When costs are incurred in providing service to more than one customer or product, a portion of the total cost must be allocated to each. This allocation is necessary for proper monitoring of costs, profitability analysis, and planning and budgeting.

Cost allocation is particularly important with shipment consolidation simply because some costs are incurred in aggregate for all items in the consolidated load. Complicating this is the fact that, frequently, the shipper does not know exactly what items will be shipped together until the load is dispatched. Thus, a simple method is required for both estimating and allocating the costs incurred by a consolidated load.

A discussion of distribution costing and accounting is given in Ernst and Whinney [1983, 1985]. Shelley [1982] outlines a method used by his company for allocating transportation costs to consolidated shipments. His approach draws heavily on company experience, thus is very company–specific. The application of standard costs (Lambert and Armitage [1979]) and activity–based costing (Lewis [1991]) appears well suited to this problem.

9.3 Closing Remarks

This thesis has shown that shipment consolidation offers:

to the shipper or manufacturer:

- significant savings in distribution costs (and possibly customer service) through careful analysis of sales and shipping patterns, customer service goals, and distribution alternatives;

to the carrier:

- improved equipment and employee utilization through increased vehicle load sizes and reduced shipment handling;

to the researcher:

- a large variety of interesting research questions in a topic that interacts with practically all areas of operations management within the firm.

Indeed, if availability of goods is increased through reduced costs, and productivity of employees, facilities, and transportation systems is improved by handling larger loads, then shipment consolidation benefits all of society.

Appendix A
DEFINITION OF LOGISTICS TERMS USED IN THIS THESIS

backhaul: i) the return movement of a transportation vehicle, whether empty or loaded, from the direction of its principal haul; ii) a shipment being transported over all or part of a transportation vehicle's return movement

break-bulk: to unload, dis–aggregate, and distribute all or part of the load carried by a transportation vehicle

break-bulk terminal: a facility whose primary function is to dis–aggregate loads consisting of two or more smaller shipments into separate lots for delivery

bumping clause: a legal option that allows the shipper to declare heavier weights than actually exist so as to push the weight of the load into a higher weight bracket, thus qualifying for a lower freight rate; also referred to as *phantom freight* or over–declared shipment

carload rate: the *freight rate* that applies to a *carload shipment*

carload shipment: a shipment, transported by rail, whose total weight equals or exceeds a stated *minimum volume weight*, and, as a result, is transported under a lower *freight rate* than that for a *less-than-carload shipment*; analogous to *truckload shipment*, except carried by rail rather than by motor vehicle

carrier: a person or organization which offers services for the transportation of property, passengers, or both

CL: carload; see *carload shipment*

common carrier: a for–hire *carrier* that is obligated to handle, at *published rates* between designated points, those commodities (property and passengers) that it holds itself out to the general public to haul, on a non–discriminatory basis as to shippers and receivers who request service (Flood et al. [1984])

consign: to send or address an item to an agent or *consignee* in another place to be cared for, sold, or used by such an agent (Merriam–Webster Inc. [1986])

consignee: the party to whom an item is consigned or shipped

consignor: the party that forwards or sends an item to a *consignee*

contract carrier: a for–hire *carrier* who provides a limited number of *shippers* with exclusive use of equipment under specific contractual agreement and pre–determined charges

contract rate: a *freight rate* that is pre–negotiated between a *carrier* and a *shipper*, thus allowing the carrier to tailor their services to shippers, to provide price incentives for large shipments, and to provide pricing exceptions to the *published rates*

demurrage: a charge by a railroad for the detention of rail–cars beyond the time allowed for loading, unloading, or other purposes

detention: a charge similar to railroad *demurrage* except that the detained piece of transport equipment is a motor vehicle, whether in motor carriage or rail TOFC ("trailer–on–flatcar") service

embargo: the refusal by a common carrier to handle certain types of freight or to provide service to or from specific shippers or locations

free on board (FOB): a term of sale used to indicate at which point title to goods sold passes to the buyer and to what point freight charges are paid for by the shipper; for example, under "F.O.B. origin" terms:
- buyer takes title to goods at origin;
- price is quoted as at the origin, thus does not include transportation charges past origin;
- buyer pays all freight charges past origin;
- buyer is responsible for risk of transportation and for filing any transportation damage and loss claims;
- buyer does not necessarily arrange transportation from origin

freight, all kinds rate: a *freight rate* which applies to shipments where i) the articles making up the shipment are not individually identified; and ii) the entire shipment moves under one *freight rate*

freight collect: a term of sale under which, at the time of delivery, the *consignee* pays all transportation charges directly to the *carrier*

freight rate: the common carrier *linehaul* charge, typically stated in cents per hundred pounds, for transporting a given quantity from its origin to its destination

in-transit arrangement: a *common carrier* service whereby a *shipper* can request that the transportation vehicle stop for partial loading or unloading before the final destination is reached

load: the items (property or passengers) carried by a transportation vehicle

LCL: less–than carload; see *less-than-carload shipment*

LCL rate: the *freight rate* that applies to a *less-than-carload shipment*

less-than-carload shipment: a shipment, transported by rail, whose total weight is less than a stated *minimum volume weight*, and, as a result, is transported under a higher *freight rate* than that for a *carload shipment*; analogous to *less-than-truckload shipment*, except carried by rail rather than by motor vehicle

linehaul: the portion of a transportation vehicle's route that is between *terminals*, as distinguished from the pickup and delivery portion of a route; may also be between the origin and destination of a shipment if the transportation is done directly between the two

logistics: all organizational activities related to the movement and storage of items from the time of raw material acquisition to the point of final consumption (Ballou [1992])

lot quantity costs: production– and acquisition–related costs that change as a result of a change in the logistics system

LTL: less–than truckload; see *less-than-truckload shipment*

LTL rate: the *freight rate* that applies to a *less-than-truckload shipment*

less-than-truckload shipment: a shipment, transported by motor vehicle, whose total weight is less than a stated *minimum volume weight*, and, as a result, is transported under a higher *freight rate* than that for a *truckload shipment*

make-bulk: to create larger transportation loads by consolidating smaller shipments in one larger load

make-bulk terminal: a facility whose primary function is to aggregate smaller shipments into larger consolidated loads

minimum carload weight: the minimum weight required for a shipment to be considered as a *carload shipment* and thus entitled to a *carload rate*

minimum truckload weight: the minimum weight required for a shipment to be considered as a *truckload shipment* and thus entitled to a *truckload rate*

non-volume shipment: a shipment whose weight is less than a specified *minimum volume weight*, thus is transported under a *non-volume rate*, such as a *less-than-carload rate* or a *less-than-truckload rate*

non-volume rate: the *freight rate* that applies to a *non-volume shipment*

order cycle time: the length of time from when an order is placed to when the ordered item(s) is received at its final destination

over-declared shipment: see *bumping clause*

private carrier: a not–for hire *carrier*, typically the consignor or consignee

published rate: a *freight rate* that has been filed with appropriate government and non–government agencies and has been listed in an authorized *tariff*

shipment consolidation: the active intervention by management to aggregate a number of small orders into larger transportation loads in order to reduce transportation costs and achieve operating efficiencies

shipment-release parameter: the specific parameters or parameter values for the decision variables of a shipment–release policy

shipment-release policy: a policy or rule used to determine when to release an order for shipment or when to dispatch a loaded transportation vehicle

shipper: the party who is responsible for preparing goods for transportation, but not necessarily to do the transportation function

tariff: a transportation industry publication that contains *freight rates*, rules and other information needed to purchase *common carrier* transportation services

terminal: a facility, generally operated by a *common carrier*, for the loading and unloading of transportation vehicles and the consolidating of shipments, but not for holding of items for any length of time

TL: truckload; see *truckload shipment*

truckload rate: the *freight rate* that applies to a *truckload shipment*

truckload shipment: a shipment, transported by motor vehicle, whose total weight equals or exceeds a stated *minimum volume weight*, and, as a result, is transported under a lower *freight rate* than that for a *less-than-truckload shipment*

volume shipment: a shipment whose weight equals or exceeds a specified *minimum volume weight*, thus is transported under a *volume rate*, such as a *carload rate* or a *truckload rate*

volume rate: the freight rate that applies to a *volume shipment*; it typically is lower than the *freight rate* that applies to smaller *non-volume shipments*

warehouse: a facility used for several logistical functions, including shipping, receiving, and storage of goods

Appendix B
LIST OF VARIABLES USED IN THIS THESIS

This list does not include some variables used only in Appendix C.

Subscripts:

I	"per item"
L	"per load"
W	"per unit–weight"

Variables:

α	shape parameter of Gamma distribution
β	scale parameter of Gamma distribution
ρ	order build–up rate
$\hat{a}$	mean order arrival rate
B_W	per unit–weight transportation savings from consolidating shipments
$\hat{C}$	number of customers in a delivery region
D	total average demand by all customers in a delivery region
d_i	average demand by customer i
$E[B_n]$	expected transportation savings, or benefit, from consolidating n orders
E[L]	expected vehicle load size (items, orders, pounds, etc., as specified)
E[N]	expected number of orders accumulated during a consolidation cycle
E[S]	expected number of customer stops per vehicle load
E[T]	expected length of consolidation cycle
E[TC]	expected total cost
E[TW]	expected total weight accumulated during a consolidation cycle
E[W]	expected weight of customer order
ESQ	economic shipment quantity: the shipment quantity that minimizes the sum of per–unit transportation and inventory–holding costs, calculated without regard to vehicle capacity or carrier rate discounts
ESW	economic shipment weight: the shipment weight that minimizes the sum of per–unit transportation and inventory–holding costs, calculated without regard to vehicle capacity or carrier rate discounts
F_D	per–load fixed cost of arranging transportation and/or dispatching transportation vehicle
F_L	sum of all fixed costs related to transporting a load a given distance
F_S	fixed cost of making a customer stop
f_C	common carrier freight rate (cost per pound) for a given distance
f_N	common carrier non–volume freight rate (cost per pound) per pound for a given distance
f_V	common carrier volume freight rate (cost per pound) for a given distance
f_P	private carrier transportation cost per unit distance

g(t)	gain of system: average reward/cost per unit time if system was to operate for an infinite length of time
H	capacity of transportation vehicle
L	actual load size carried by transportation vehicle
M	total transportation distance, including pickup, delivery, and backhaul if applicable, per vehicle load
MWT	common carrier minimum volume weight
N	actual number of orders or items waiting for consolidation or being carried by transportation vehicle
N^*	minimum–cost quantity: the consolidated number of orders, determined with regard to the economic shipment quantity, carrier rate discounts, and other economic considerations, that minimizes total per–unit cost
Q	shipment quantity, expressed as items, orders, weight, etc., as stated when used
r_I	inventory–holding cost per item or order per time period
r_W	inventory–holding cost per pound per time period
T	length of a consolidation cycle
t	elapsed time, starting at t=0
T_D	time remaining in consolidation cycle
T_{MAX}	a management–set parameter representing the maximum length of time that a customer order can be held for consolidation
TC	total cost
TC_R	total inventory–holding cost
TC_T	total transportation cost per vehicle load
TW	total accumulated weight being held for consolidation
W	weight of a customer order
W(n)	sojourn measure: expected length of time that the quantity in the system does not exceed level n
W^*	minimum–cost weight: the consolidated weight, determined with regard to the economic shipment weight, carrier rate discounts, and other economic considerations, that minimizes total per–unit cost
W_D	*in Chapter 6:* difference between total weight and expected weight of one order
W_D	*in Chapter 7:* difference between target weight W_{TAR} and accumulated weight TW
W_{TAR}	any weight designated by the shipper as a "target weight"
WBT	minium weight at which over–declaring of load weight is cost–effective

Markov decision process variables:

A	set of actions a_s
a_s	action that may be selected when in state s
$\hat{A}$	continuous–time transition rate matrix
$\hat{a}_{sj}$	(continuous–time) rate at which transitions are made from state s to state j
b	aggregation batch size

H_s	mean waiting time in state s: unconditional mean holding time in state s when the successor state is unknown
H_{sj}	mean holding time in state s: expected length of time spent in state s before moving to state j
N	total number of states
P	discrete–time transition probability matrix
$p_{sj}(a)$	(discrete–time) transition probability of making a transition from state s to state j by choosing action a
$p'_{sj}(a)$	(discrete–time) transition probability $p_{sj}(a)$ restated for aggregated data
$r_s(a)$	single–state reward/cost resulting from a transition from state s by selection of action a
$r'_s(a)$	single–state reward/cost $r_s(a)$ restated for aggregated data
S	set of states s
s	state that represents the condition of the system at any time
ω_s	weighting constant for original state s in aggregated model
Λ	diagonal matrix of the reciprocal of the mean waiting time in state s
λ	parameter of exponential distribution

Appendix C
MODELING SHIPPER COSTS IN PHYSICAL DISTRIBUTION ANALYSIS

C.1 Introduction

Analysis of any physical distribution activity must include some consideration of cost. This typically is done through algebraic cost expressions. In this appendix, we discuss the modeling of shipper costs through an examination of some such expressions and their relevance to actual costs.

We first present some general comments about modeling distribution costs, then outline cost classifications relevant to a shipper. The components of each classification are examined by reviewing cost expressions from the literature and discussing practical aspects of these formulations.

Examples of cost expressions found in the distribution literature are classified and summarized in Tables C–1, C–2, and C–3. Table C–4 provides a general decision table for developing expressions for total distribution cost. As well, in several places, approximate values for actual costs or regression coefficients given in the literature have been included for illustrative purposes. These values will be different for every situation, and users should refer to the original work to determine their applicability to any situation being modeled.

C.2 Costs and Cost Modeling in Physical Distribution

The starting point of any discussion of physical distribution costs is the "total cost concept". This view states that all costs of the logistics system are interrelated, and any analysis must examine the effect on all costs rather than on just a few.

Models in the distribution literature, however, typically have focused on one activity, such as vehicle routing, facility location, or load planning, and have ignored the impact of the analysis on other activities of the distribution system. Moreover, models have almost exclusively focused on direct costs (those that can be easily traced to a shipment or order) and have omitted those that are less visible. Thus, while the total cost concept has been recognized for many years, it has not received adequate application by theoretical models. The total cost concept in physical distribution has been discussed by several authors, including LeKashman and Stolle [1965].

Expressions for the total cost of a shipper's logistics system have been proposed by various authors, including Willis [1977], Lalonde and Lambert [1977], Stock and Lambert [1987], and Tyworth, Cavinato, and Langley [1987]. Practical approaches for costing shipments also have been developed (for example, Temple, Barker, and Sloane [1982]). While these models differ as to the detail in which certain activities are considered, the costs included can be classified as those relating to transportation, inventory–holding, shipment handling and warehousing, and indirect activities. These classifications form the basis for our examination of distribution cost modeling, and are introduced next.

Transportation Costs

Transportation cost refers to the round–trip cost incurred by the shipper to physically move goods from origin to destination. This movement may consist of three phases: pickup, linehaul (from one terminal or region to another), and delivery. The return trip ("backhaul") of the vehicle also may be relevant.

All phases of transportation can be performed by two general types of carrier: common (for–hire) carrier or private carrier. The modeling aspects of each can be very different, thus it is important that researchers clearly state which form is being assumed. Although other types of carrier exist, our discussion will consider only these two, and unless otherwise noted, will be restricted to motor carriage.

Inventory-Holding Costs

Inventory–holding cost refers to the cost of holding an item in inventory for a specified length of time. It includes both real costs, such as insurance, interest, and breakage, and opportunity costs, such as the cost of capital that could be used elsewhere. It should not be confused with "storage cost", the cost charged by public warehouses for storage of goods. The theory and calculation of inventory–holding costs have been well covered in the literature.

Shipment-Handling and Warehousing Costs

Shipment–handling refers to all activities, other than actual transportation, involved in the movement of items. This includes order selection, consolidation, packing, staging, loading, and unloading. If shipments are made to a warehouse instead of directly to customers, costs for receiving, inspection, and moving to storage also must be considered.

Warehousing costs can be classified as those related to handling, storage, and indirect activities. Warehousing may be done by public or private facilities. When using public warehouses, the shipper's cost will be the sum of charges for storage and

handling. Private warehousing requires consideration of indirect costs beyond shipment storage and handling, such as administration, security, maintenance, insurance, and taxes.

Indirect Costs

Indirect costs (or "overhead") result from activities that support the distribution function but are not easily traced to a particular shipment or load. Examples include costs of clerical support, information and order processing, claims handling, marketing, and general administration.

Not all these costs will be relevant to every analysis. For example, some activities may not be affected significantly by proposed system changes. However, researchers must carefully consider both the short–term and long–run relevance of each cost classification to their analysis.

We now discuss the modeling of specific components of each cost category.

C.3 Modeling Shipper Transportation Costs

Most transportation cost expressions in the literature take one of four general forms:

Form (a): transportation cost is a function of quantity (typically number of items or total weight) and freight cost per unit–quantity, without explicit consideration of distance traveled;

Form (b): transportation cost is a function of distance and freight cost per unit–distance, without explicit consideration of quantity shipped;

Form (c): transportation cost is a function of distance, quantity, and freight cost per unit–quantity per unit–distance; or

Form (d): transportation cost is a stated cost per load, without explicit consideration of quantity shipped and distance traveled.

Differences between transportation cost models largely relate to the individual costs included and the mathematical approaches used to determine the values of parameters. Examples of these forms in the literature are listed in Table C–1.

Form (a) assumes a constant transportation cost per unit shipped, regardless of distance traveled. Distance may be considered implicitly in setting the per–unit–quantity freight cost. The dependence of total transportation cost on quantity rather than on distance implies that this form is more applicable to common carriage than to private carriage.

Form (b) models transportation cost as the product of freight cost per unit–distance and total distance, where the freight cost parameter may be a common carrier rate per unit–distance or the cost per unit–distance of operating a private vehicle. The usual assumption that total transportation cost is linearly related to distance is questionable, and an estimate of the total travel distance is required. This formulation, as with form (d), is better suited to modeling private carrier than common carriage.

Most distribution models consider either quantity shipped or distance traveled as part of the freight cost parameter, as in forms (a) and (b). Form (c), which occurs almost exclusively in the facility location literature, treats these parameters explicitly.

The fourth form, found in many vehicle routing models, assumes that distances between all origins and destinations are known. As a result, distance is not included

TABLE C-1: EXAMPLES OF DISTRIBUTION COST MODELS: TRANSPORTATION COST

Type of Cost	Modeled as form (a): function of quantity and cost per unit-quantity	Modeled as form (b): function of distance and cost per unit-distance	Modeled as form (c): function of quantity, distance, & cost per unit-quantity per unit-distance	Modeled as form (d): stated (fixed) cost per occurrence
Transportation costs:				
Linehaul/ pickup/ delivery costs	Abdelwahab & Sargious; Arcelus & Rowcroft; Aucamp; Buffa; Baumol & Vinod; Das; Constable & Whybark; Hall [1984]; Hall [1985]; Perl & Daskin; Perl & Sirisonsip; Powell & Sheffi; Samuelson	Burns et al.; Daganzo [1978]; Daganzo & Newell; Hall [1985]; Perl & Daskin	McFadden et al.; many facility location models, eg, Eilon et al. and median-weight method	Ball et al.; Blumenfeld et al.; Hall [1984]; many vehicle routing models
Dispatch cost	n/a	n/a	n/a	Burns et al.
Stop cost	n/a	n/a	n/a	Burns et al.; Hall [1985]; Shelley

n/a No examples are known.

explicitly in the formulation, and transportation cost per load is expressed as a stated amount for each origin–destination pair regardless of quantity shipped. Some models expand this parameter to include costs relating not only to travel but also to other activities. Ball, Golden, Assad, and Bodin [1983], for example, include crew and vehicle waiting costs at each origin and destination.

The costs of loading and unloading a vehicle, although more accurately classified as shipment handling, are treated by many models as part of the transportation cost, usually as a fixed cost per load. Most models also distinguish between the costs of linehaul and pickup/delivery transportation, and may model the two stages differently. For example, the linehaul cost in Perl and Daskin [1985] is a function of quantity (Form (a)), while the delivery cost is a function of distance (Form (b)). We retain this distinction in our discussions.

C.4 Modeling Transportation Costs Under Common Carriage

Common carrier transportation services may involve pickup, linehaul transportation, delivery, and handling at intermediate points. As noted by Davis and Dillard [1983] and Tyworth, Cavinato, and Langley [1987], terminal services and pickup and delivery on both truckload and less–than–truckload shipments usually are included in the carrier's linehaul freight rate . Loading and unloading at origin and destination frequently are not, and will depend on the product being moved and the mode of transportation. For example, for bulk items moving by truck, unloading typically is performed by the carrier and included in their rate. However, as one shipper of

palletized items told us, "*We* load and unload; common carrier provides transportation only. Drivers *may* assist in these operations."

In the simplest form, transportation cost incurred by the shipper under common carriage can be modeled as the fixed cost of arranging transportation plus the carrier's linehaul freight charges:

$$TC_T = F_L + f_c\, Q$$

where TC_T is the transportation cost per vehicle load, F_L is the fixed cost of a load of Q items or weight, and f_c is the common carrier linehaul freight rate per item or per unit–weight. These costs are discussed in the next sections.

Fixed Costs of Common Carriage

In theory, it seems reasonable to assume that whenever a shipment is dispatched, the shipper incurs some costs regardless of the size of the load. These may relate to time spent arranging transportation and preparing shipping documents, or to common carrier fixed charges. Aucamp [1982], for example, includes a fixed amount per vehicle used.

In practice, however, it is doubtful that a fixed cost is incurred by a shipper when using common carriage. Our discussions with shippers reported that the cost of preparing documentation tends to be dependent on shipment size. Moreover, carrier freight charges include the cost of delivering a truck or trailer to the shipper's premises and picking up empty trailers and containers. As one shipper commented, "fixed costs are negligible".

As noted in Chapter 5, many authors have derived expressions for the optimal shipment size similar to the economic shipment quantity concept. These formulae require a fixed cost component, which typically is assumed to be related to transportation. Our conclusion regarding common carrier fixed charges implies that such analysis cannot be applied to this type of carriage.

Common Carrier Linehaul Charges

Common carrier linehaul charges typically are derived by multiplying the weight of the shipment by a carrier freight rate (usually cents per hundred pounds) for a stated distance. Regression analysis frequently has been used to investigate the relationship between these three factors (weight, distance, and carrier rate) and a shipper's linehaul transportation cost under common carriage. For example, using rates from Ballou [1992] and other sources, we found that the linehaul charge for rail shipments could be approximated by the expression:

linehaul charge per hundred pounds = $c\ (a + M_L^b)$

where a and b are regression coefficients, c is a factor to adjust for the classification rating of the items shipped, and M_L is the linehaul distance; a≈38.54, b≈0.74, c≈1.67.

Samuelson [1977] derived the expression:

linehaul charge per unit shipped = $\exp[a + b \ln(Q)]$ $\quad b<0$

where Q is the shipment quantity, and a and b are regression coefficients. The a coefficient includes factors not related to shipment size, such as product value, density, and length of haul. A variation of this model for total, rather than unit, freight cost is given by Abdelwahab and Sargious [1990]:

total linehaul charge = exp[a + b ln(Q)] b>0

McFadden, Winston, and Boerch–Supan [1986] proposed the expression:

total linehaul charge = $(a_1 + b_1 M_L) + (a_2 + b_2 M_L)(Q)$

where M_L is the linehaul distance. Although they state that "this specification was used to capture the nonlinear nature of the rate schedule", the distance component is not explicit in the final version of this expression. For example, for truck shipments of agricultural produce, their formula is: total linehaul charge = 776.8 + 0.474 (pounds). Ballou [1991] examines errors resulting from using linear approximations of trucking rates based on distance, and discusses ways to improve this estimate.

The existence of discounted "volume rates" for shipments exceeding specified weights makes modeling common carrier linehaul charges more complex. The simplest case, where volume rates apply above only one weight break, can be handled as:

total linehaul charge = $\delta f_V W$

where f_V is the volume freight rate per unit shipped, W is the weight shipped, and δ is the ratio of non–volume rate f_N ($f_N \leq f_V$; f_N applies to shipments weighing less than the weight break) to the volume rate f_V. δ equals 1 if Q equals or exceeds the weight break. A second formulation is:

total linehaul charge = $[\alpha f_N + (1-\alpha) f_V] W$

where α equals 1 if the weight W shipped is less than the weight break, and 0 otherwise. Unfortunately, these formulations are of limited use when a continuous mathematical cost expression is required. They also ignore many subtleties of common carrier pricing, such as "phantom freight", the ability to declare larger

quantities than actually exist in order to push the shipment into a lower rate category, thus reducing total transportation charge. As a result, consideration of carrier weight quantity discounts largely has been ignored in the literature, typically by assuming the quantity shipped falls under a single freight rate.

The impact of deregulation on transportation cost modeling is not clear. Results from regression modeling of linehaul charges now are less universal, often L .g applicable only to a few parties. As well, the ability of shippers to negotiate transportation charges has introduced game–theoretic aspects to the analysis; see, for example, Bookbinder and Fraser [1990]. Buffa and Munn [1989], however, comment that negotiation has resulted in freight rates that are more closely related to the cost of transportation, and thus are easier to model. Common carrier freight rates, which have provided cost data for many studies in the literature, are still published, and many carriers quote this rate when transportation services are first solicited.

Common Carrier Small Package Charges

Small packages moving by common carrier usually are charged a flat amount, regardless of weight. On a per–pound basis, this can be extremely expensive, thus lighter shipments (for example, no more than 75 pounds) may move under "small package rates".

Arcelus and Rowcroft [1991] examined small package rates for shipments from Moncton to Montreal (653 miles), Toronto (1030 miles), Winnipeg (2051 miles), and Vancouver (3577 miles). Using regression analysis, they developed the expression:

small package freight cost = $a\ e^{b\ \ln(W)}$

where a and b are regression coefficients (a was between 3.02 and 3.93; b between 0.3862 and 0.3804) and W is the weight of the shipment. Except for shipments to Vancouver, values for the coefficients did not vary appreciably among the destinations examined, implying that a single cost expression may be adequate for modeling small package transportation cost for a large range of distances.

Hall [1985a] modeled the cost of shipping by U.P.S. (United Parcel Service) by a simple linear function of weight with a minimum charge per shipment. We found that the form suggested by Arcelus and Rowcroft yielded a much better fit (with addition of a constant term) to prices obtained from Purolator, a Canadian courier, than did a linear expression. [In fact, an even better approximation was given by the model: small package rate = $a + e^{b1} + W^{b2}$, with a=1.26, b1=0.33, and b2=2.03.] Modeling of courier charges is further simplified because most use "zone pricing": in our analysis, all packages of a given weight originating in Ontario and Quebec with destinations anywhere in those provinces were charged the same amount. Such pricing reduces considerably the number of cost expressions that must be considered when modeling small package shipments.

C.5 Modeling Transportation Costs Under Private Carriage

Costs of private carrier transportation may be classified as those related to:

- dispatching vehicle from load origin (fixed cost per dispatch);
- round–trip linehaul transportation from origin to breakbulk centre or directly to destination (variable cost per load);
- dispatching local delivery of individual shipments from warehouse or other breakbulk facility (fixed cost per delivery load);
- transporting delivery shipments from breakbulk facility to customers (variable cost per delivery load); and
- making delivery stops (fixed cost per stop).

The cost of pickups, if relevant, would be similar to that of deliveries. As well, any revenues or savings resulting from performing backhauls should be deducted from the cost of the forward haul. Most models in the literature avoid this complication by assuming vehicles return empty.

The total cost of transporting a load via private carrier will be the sum of the above five costs:

$$TC_T = F_V + f_P M_L + F_D + f_D M_D + F_S G_S$$

where:

F_V = fixed cost per dispatch of linehaul vehicle
f_P = private carrier linehaul transportation cost per unit distance
M_L = total linehaul distance (including backhaul distance if applicable)
F_D = fixed cost per dispatch of delivery vehicle
f_D = private carrier delivery transportation cost per unit distance
M_D = total local pickup or delivery distance
F_S = fixed cost per customer stop
G_S = mean number of customer stops per route or load

This expression is similar to that in Blumenfeld et al. [1985] and Burns et al. [1985]. We now discuss variable and fixed costs of private carrier transportation.

Private Carrier Variable Transportation Costs

Variable costs of private linehaul and pickup/delivery transportation can be estimated from historical data. For example, Flood, Jablonski, and Callson [1984] give an example of the derivation of the per–mile cost for a motor carrier; the result was $0.879 per mile. Their analysis includes direct costs such as driver salaries, fuel, and maintenance, and indirect charges, such as depreciation, insurance, taxes, and licences. Costs related to both tractors and trailers are included, based on the assumption that there are three trailers per tractor (we have seen advertisements for

motor carriers that give ratios of seven–to–one). Total distance is that travelled in one year by both loaded and empty vehicles. Thus, a simple and practical way of modeling private carrier linehaul transportation cost is by deriving a per–unit–distance parameter based on all transportation–related costs for a representative length of time, then multiplying this parameter by the distances being analyzed.

Research on estimation of theoretical linehaul distances is discussed by Love, Morris, and Wesolowsky [1988]. Commercial distance data–bases using actual distances also are readily available. Approaches for estimating the total distance travelled in a pickup or delivery tour are discussed by: i) Eilon et al. [1971] and Love et al. [1988] for direct distance approaches; ii) Eilon et al. [1971], Banks, Driscoll, and Stanford [1982] for modified direct distance approaches; and iii) Beardwood, Halton, and Hammersley [1959], Eilon et al. [1971], Stein [1978], Larson and Odoni [1981], and Lawler, Lenstra, Rinnooy Kan, and Shymoys [1985] for traveling–salesman–problem approaches.

The calculation of linehaul cost illustrates an important difference between transportation cost under private and common carriers. Because a common carrier's freight rate includes consideration of shipment origin and destination (hence distance), the shipper's linehaul cost under common carriage is a function of quantity shipped. However, based on a per–mile allocation of costs as discussed above, the linehaul transportation cost for private carriage is a function of distance. Thus, for a given distance, the linehaul transportation cost for private carriage is largely fixed regardless of quantity shipped; this statement has been echoed by several authors, including

Webb [1968]. This is one reason common carrier is often more cost effective than private carrier for systems with small total throughput (Firth et al. [1988]).

The importance of distance on private carrier transportation cost is recognized by Daganzo and Newell [1985], who include the speed of a vehicle and portion of time it spends in maintenance, non–working, and waiting activities. From this they derive the minimum size of vehicle fleet required, and ultimately, total transportation cost based on fleet size, vehicle operating cost per unit–distance, and distance travelled.

Some models have treated loaded and empty vehicle movements separately. Ball et al. [1983] combine transportation and loading/unloading costs, then apply different cost parameters to loaded and empty vehicles; with the latter, vehicle crews are assumed not to wait while shipments are loaded or unloaded. Powell and Sheffi [1989] also consider the cost of moving empty trailers to locations where they are required.

Private Carrier Fixed Transportation Costs

The major fixed transportation costs of using private carriage are those of dispatching linehaul and pickup/delivery vehicles, and making pickup/delivery stops. These costs can be derived from industry experience and company records. Shelley [1982], for example, estimated a fixed cost of $60 per delivery stop for his company. Ernst and Whinney [1983] recommend allocation of dispatch costs as a fixed charge per load based on the total number of loads handled in one year. Thus, a fixed transportation cost per load is more relevant under private carriage than under common carriage.

The number of stops per pickup or delivery load will depend on several factors, including the arrival rate and size of customer orders, shipper order–release policies, and the capacity of vehicles. Analytical methods for estimating the mean number of stops per load are discussed by Burns et al. [1985] and Jaillet and Odoni [1988]; Shelley [1982] uses a figure derived from empirical data.

Models including a fixed cost of dispatching a linehaul vehicle and/or fixed cost per customer stop are listed in Table C–1. Daganzo [1987], however, ignores the fixed cost per customer stop by assuming vehicles stop for relatively short periods of time.

C.6 Inventory–Holding Costs

Practically all models of shipper distribution activities include inventory–holding costs; consistent exceptions are those dealing with vehicle routing or facility location. Treatment of inventory–holding costs varies in two main aspects: the position of inventory in the system and the level of uncertainty with regard to time. These are discussed in the next sections.

Position of Inventory in the Distribution System

Inventory in a distribution channel may be held in three places: awaiting shipment by the shipper or consignor, in–transit, or awaiting usage by the purchaser or consignee. Practically all models include the first, typically treated as the product of inventory–holding cost parameter, total value of waiting items, and time between shipments. Burns et al. [1985], Blumenfeld et al. [1985], and Perl and Sirisoponslip [1988], for example, consider a manufacturer with a constant production rate, thus the

average inventory waiting for shipment is one–half the shipment size. As seen in Table C–2, some models also consider inventory awaiting usage by the purchaser or consignee.

The holding cost of in–transit inventory is included in many models; examples are given in Table C–2. Blumenfeld et al. [1985] and Buffa [1986b] consider in–transit inventory–holding cost to be the cost of inventory actually being transported plus that waiting for delivery at a terminal. Buffa [1986b] includes the cost of transportation as part of the item value; this treatment is technically correct from an accounting standpoint. Daganzo [1987], however, assumes that inventory cost is negligible because items are "cheap". A shipper's in–transit inventory–holding costs also may be ignored by assuming goods are shipped *FOB origin*, thus being the property of the buyer rather than the shipper.

The holding cost of in–transit inventory usually is calculated as the product of inventory–holding cost parameter, value of items shipped, and transportation time. Transportation time typically is assumed to be a known and constant average, however analytical expressions treating it as a function of other factors have been developed; these are discussed briefly in the next paragraphs.

Larson and Odoni [1981] divide the travel of a vehicle into an acceleration stage, a cruising stage, and a deceleration stage. From these, they derive formulae for expected transportation time given the probability density function of travel distance or its expected value. These expressions have been supported by empirical studies of emergency vehicle response time within urban regions. We found, however, that when applying these formulae to travel times between actual urban areas, because

Table C–2:
Examples of Distribution Cost Models: Inventory–Holding Costs

Type of Cost	Modeled As A: (Fixed) cost per occurrence per occurrence	Function of quantity and and cost per unit–quantity	Function of quantity, time, & cost per unit–quantity per unit–time
Shipper inventory–holding cost	n/a: would be a facility cost	n/a: see next column	practically all models
In–transit inventory–holding cost	n/a	n/a	Baumol & Vinod; Burns et al.; Das; Constable & Whybark; Hall [1985a]; Hall [1985b]; *Blumenfeld et al.; *Buffa [1986b]; *Perl & Sirisonslip
Recipient inventory–holding cost	n/a	n/a	Baumol & Vinod; Blumenfeld et al.; Burns et al.; Hall [1985b]; Daganzo [1987]
Safety stock/ backorder holding cost	n/a	n/a	Baumol & Vinod; Buffa; Das; Constable & Whybark; Perl & Sirisonslip

* These models include the cost of holding inventory at terminal or distribution centre as part of the holding cost of in–transit inventory or inventory waiting for shipment.

of the differing quality and directness of non–urban roads, it frequently was difficult to determine values for the acceleration/deceleration rate and the cruising speed constant which could be applied to all vehicle tours.

The effect on transportation time of such factors as distance, load weight, number of pieces, and number of stops has been recognized by some computerized routing and control packages (such as Statistical Supervisor, discussed in Flood et al. [1984]). Daganzo and Newell [1985] give an expression for transportation time based on time per customer stop, distance per tour, speed of vehicle, and number of stops. Actual transportation times were analyzed by DeHayes [1968], Piercy [1977], and Chiang and Roberts [1980]. Both DeHayes and Chiang and Roberts concluded that a Gamma probability distribution provided the best fit to their empirical data. Chiang and Roberts' model states that, for a given distance, transportation time can be modeled by a shifted gamma distribution with an expected time of [a M^b + c + M tan d], where a, b, c, and d are regression coefficients, and M is the distance for a given haul. They suggest three models; the basic model has a=0.21, b=0.0003, c=0.13, and d=.003.

Applying these results, the in–transit inventory–holding cost expression would become:

(item value) (inventory–holding rate) (number of days)

= (item value) (rate) [a M^b + c + M tan d]

Chiang and Roberts limited their analysis to regular–route common carrier. However, their expression is particularly useful for private carriage because it allows us to express the costs of in–transit inventory–holding and linehaul and delivery

transportation as a function of the distance travelled. Optimization of such a distance–based expression would yield the distance that results in the minimum–cost combination of transportation and inventory–holding. This may be of use in planning vehicle delivery tours.

Load size frequently affects transportation time, especially under common carriage. Thus, in many analyses, the expression for in–transit inventory–holding cost should include a term relating to load size. This is considered by Buffa [1987], who modeled transportation time as a linear function of distance. The longer in–transit times resulting from non–volume loads are adjusted for by multiplying expected transportation time by a factor exceeding 1 for non–volume shipments.

The separation of inventory–holding cost between inventory waiting for shipment and that in–transit is important if the analysis will have an effect on the length of time that goods are in–transit. In many other cases, this separation may be unnecessary or unrealistic. Unless the transit time is very long, it is unlikely that the inventory–holding cost parameter (cost per unit per time) will be significantly different for pipeline inventory and standing inventory. It also seems impractical to compute a separate inventory–holding cost parameter for in–transit shipments with short transportation times.

Level of Time or Demand Uncertainty

Total inventory–holding cost is dependent on the length of time that items are held. Models in the distribution literature differ as to the level of uncertainty related to

length of order lead–time (a major component being transportation time) or demand within this time.

As mentioned previously, most models treat transportation time as deterministic by assuming total lead–time can be approximated by a known and constant average; see, for example, Baumol and Vinod [1972], Constable and Whybark [1978], Ball et al. [1983], Blumenfeld et al. [1985], Burns et al. [1985], Hall [1985b]. This treatment is adequate when shipments move between known origin–destination pairs or along pre–defined vehicle tours. Some models (for example, Baumol and Vinod [1970], Das [1972], Constable and Whybark [1978], Buffa [1986b]) also claim that if transportation time is uncertain, safety stock must be carried to meet unexpected customer demand during this time. A term based on stochastic assumptions is then included to calculate the cost of holding safety stocks or backordering to allow for additional demand due to variability in lead–time. However, these models do not recognize this variability when calculating in–transit inventory–holding cost, instead treating transportation time as a constant average. This implies that time is deterministic for inventory in transit, yet stochastic for demand during the entire lead–time.

Research, such as that by DeHayes [1968] and Chiang and Roberts [1980], which suggests a probability distribution for transportation time allows estimation of time variance. For example, the variance of the gamma–distributed model hypothesized by Chiang and Roberts is: var(days) = a M^b, where a and b are regression coefficients as discussed previously and M is transportation distance. Use of this expression, along with knowledge of representative distances, load sizes, and per–unit inventory–holding costs, allows the modeler to estimate the impact of time

variability on total in–transit inventory–holding cost, and decide to what extent such variability should be considered in the model.

C.7 Shipment–Handling and Warehousing Costs

Shipment–handling includes activities such as order selection, consolidation, packing, staging, loading, unloading, and inspection. Rather than treat each activity separately, models in the literature usually aggregate them into three general classifications: loading of vehicles at the origin, unloading vehicles at the destination, and handling in–transit shipments at a terminal. Vehicle loading and unloading costs often are treated as part of a fixed cost of dispatching a vehicle or making a delivery stop. However, in–transit shipment–handling costs and costs other than for loading and unloading have generally been ignored by distribution models. Models that include cost expressions for shipment–handling and facility activities are listed in Table C–3.

Analyzes including shipment–handling usually consider only the variable costs of handling a shipment and ignore those facility costs that do not vary with volume. For example, shipment–handling costs at a private facility should include an allocation on a per–shipment or per–weight basis of the fixed costs of operating the facility. Blumenfeld et al. [1985] note that an allocation of facility overhead and operating costs may be factored into the inventory–holding cost.

Blumenfeld et al. [1985] treat shipment–handling time as the sum of average handling time at terminal and the waiting time at the origin and destination. Buffa [1986b] calculates shipment–handling cost as the product of a per–hour

Table C–3:
Examples of Distribution Cost Models:
Shipment–Handling, Facility Costs, and Indirect Costs

Type of Cost	Modeled As A: (Fixed) cost per occurrence	Function of quantity and cost per unit–quantity	Function of quantity, time, & cost per unit–quantity per unit–time
Shipment-handling costs	n/a: would be facility or indirect cost	Deming; Powell & Sheffi	Buffa [1986b]; Ball et al. (time only); Gupta & Bagchi
Terminal/ warehouse/ facility costs	Eilon et al.; Perl & Sirisonslip; Willis	Buffa [1986b]; Willis; Eilon et al.; Perl & Daskin; Perl & Sirisonslip	n/a: see inventory holding cost (Table C–2)
Indirect costs:			
Administr'n, clerical, marketing	Temple, Barker, Sloane	n/a	n/a
Claims	n/a; may be a useful simplification	Temple, Barker, Sloane	n/a

n/a No examples are known and/or this treatment seems unusual.

loading/unloading cost and the time required to load/unload a shipment. Powell and Sheffi [1989] model the costs of shipment–handling at origin and destination as the number of truckloads handled at a terminal times the handling cost per truckload. Gupta and Bagchi [1987] include a fixed charge per hundredweight (cwt.) per day, thus allowing the costs of shipment–handling and inventory–holding to be combined as a cost per cwt. per time period.

In their study of maritime cargo ships, Jansson and Shneerson [1978] expressed the tonnage that could be loaded or unloaded per unit time as a function of ship's freight–holding capacity: tonnage handled per unit time = a H^b, where H is the ship's capacity, and a and b are regression coefficients (a reflects products handled, labour productivity, etc.; b reflects the elasticity of handling speed). Adapting their expression gives the expression:

shipment–handling time = load weight / a H^b

Jansson and Shneerson include the capacity of the ship because it influences the number of holds and hatches, hence the handling time. With the exception of bulk products, this is less true with motor and rail freight, where actual load, rather than capacity, is much more important in determining handling time. This is reflected in the model by Deming [1978].

Deming's relationship between handling time and shipment weight takes the form:

shipment–handling time = a + b W^c

where a, b, and c are regression coefficients and W is the weight of the shipment. The values of the coefficients depend on the method (manual, dragline, or forklift)

used to move the freight and the type of movement (truck–to–platform, platform–to–truck, and truck–to–truck); separate regression equations were derived for eight cases. With forklift, for example, b=0.66 for loading and 1.0 for unloading, while a=0.6 and c=0.54 for both activities. These values imply that, under identical conditions, loading takes less time than unloading, an observation not usually included in modeling of shipment–handling time.

The shipment–handling time derived from Deming's expression can be multiplied by a cost per unit time to yield a handling cost per shipment. The cost per unit time parameter can be estimated from accounting records by dividing the relevant total costs by the total shipment handling time for a representative period. Such a simple estimate, however, negates some of the depth of study done by Deming, and cost parameters developed from more activity–specific tasks may be justified or preferred. Lastly, use of Deming's model requires knowledge of the specific shipment–handling techniques used and how changes being analyzed will affect them. This knowledge may be beyond the scope of many general analytic studies.

The annual cost of operating a facility such as a warehouse or terminal can be modeled as:

TC_W = fixed costs + shipment–handling costs + storage and inventory–holding costs

$$= F_W + r_H W_V + r_W (W_I/2 + W_B)$$

where:

TC_W = annual cost of operating warehouse or terminal
F_W = warehouse/terminal fixed cost
r_H = warehouse/terminal shipment handling cost per unit weight
W_V = annual warehouse/terminal throughput weight
r_W = warehouse storage cost per unit weight per year
$W_I/2$ = mean warehouse inventory level

W_B = mean warehouse safety stock level

Fixed costs include those related to taxes, security, clerical and administrative tasks, etc. As discussed earlier, shipment–handling costs relate to the labour and equipment costs of receiving, shipping, and those operations that are not a function of receiving or shipping, such as inspection, repacking, consolidation, restocking, repackaging, rotation, and establishment of order picking. Storage and inventory costs will include, in addition to the usual inventory–holding components, an expected return on warehouse space and storage equipment. Clearly, storage and inventory–carrying costs will not be relevant if the function of the facility is only to transfer shipments between vehicles.

This expression is often simplified, as in Eilon et al. [1971] and Willis [1977], to:

$$TC_W = F_W + r_H W_V + b \sqrt{W_V}$$

where b is a constant to reflect the cost of carrying stock. Economies of scale from larger throughputs are considered through the square root, which is derived from the relationship of the economic order quantity to demand. While this simplification is reasonable for warehouse inventory level W_I, it overlooks the dependence of safety stock level W_B on the variance of throughput W_V.

The values for F_W, r_H, and b will depend on many factors, such as facility size and function, labour and equipment cost and productivity, type, value, volume, weight, and density of products handled, and whether the facility is operated for private or public use. For example, we have seen cases in the literature with handling costs ranging from \$4.50 to \$12 per pallet and administrative costs from \$6 to \$50 per pallet.

In practice, however, most private warehousing costs are fixed and should be allocated to shipments based on weight, size, or other suitable measure.

Public warehouse rates are typically stated as the sum of a handling cost, a storage cost, and a profit component. Handling charges may be based on weight, cost, size, volume, or hazard of service. Storage charges are usually based on volume or weight, or may be set as a percentage of the stated value of the goods. Of course, the total cost to the shipper of using a public facility will be quoted as a single rate, typically based on volume. Thus, including in the model a term as simple as "quantity multiplied by warehouseman's rate" may be sufficient; a time component should also be included if the analysis causes items to be held for varying lengths of time. We add that, as Buffa [1986b] has noted, because of deregulation, shipment handling and short–term storage costs are often negotiable if performed by a common carrier.

C.8 Administration and Other Indirect Costs

Indirect costs include those costs relating to activities, such as administration and clerical support, information and order processing, and claims handling, which support the shipper's logistics system but are not directly associated with or traceable to individual shipments or orders. These costs typically are charged to shipments as a flat amount per shipment, although Temple, Barker, and Sloane [1982] proposed that claims cost be allocated based on weight.

Some models in the literature include an allocation, usually for clerical and shipment processing, as a fixed–cost–per–shipment term. In general, however,

models rarely consider administration and other indirect costs. This implies an incremental approach to shipment costing; that is, any changes resulting from the analysis are assumed not to affect indirect activities. This approach would also, for example, state that if a vehicle is empty on the return portion of a haul, the cost of performing a backhaul is negligible because travel costs must be incurred anyway.

The incremental approach should only be used for small changes to the distribution system. It should not be used for long–term planning because it leads to poor decision making and unrealistic application of costs.

C.9 Conclusions and Discussion

Comments regarding the relevance and appropriateness of specific approaches to modeling physical distribution costs have been presented throughout this appendix. We close with some general conclusions and discussion of issues regarding these statements.

The first and most important step in any analysis of distribution activities is the definition of the system and its components. Important questions here relate to the mode and form of transportation, the existence and function of facilities, and the type of product being handled. The second step, following from this definition, is the identification of those costs relevant to the analysis.

Table C–4 at the end of this Appendix provides a decision flowchart to assist the modeler in these two steps. This table provides only general guidance, and certainly does not cover all possible situations.

The major shortcoming of many distribution models is that they have failed to reflect the total cost concept, instead taking an incremental approach. This has meant that the long–term impact of recommended system changes have been ignored.

For example, costs included in many analytical models of physical distribution activities have been limited to transportation and inventory–holding, both of which can be modeled fairly easily. Shipment–handling activities, which affect almost all aspects of physical distribution and are impacted by most changes to the system, have not been included in most models. The cost and impact of these activities should be recognized at least qualitatively, if not explicitly.

The major problem in implementing a total cost approach to distribution system analysis is that it requires a much more exact definition of the activities involved. The resulting system may be much larger than many modelers desire, and include costs that are unknown or less–well defined. This may result in a less generalizable analysis, and may be one reason that some distribution costs have not received the attention they deserve.

Researchers frequently have not stated clearly whether their analysis assumes common or private carriage, or whether facilities are for public or private use. For the shipper, each of these alternatives will result in different costs, thus requiring different modeling approaches.

The major advantage to modeling common carriage is that many costs required for accurate modeling of private carriage distribution are summarized in one or a few common carrier rates. These rates often can be obtained from carriers or published tariffs. However, the modeler must determine which services are included in the

freight rate and which must be paid (and modeled) separately. Moreover, unlike private carriage, the size of the load must be considered unless it can be assumed to be fixed throughout the analysis.

The impact on transportation cost of volume discounts and other common carrier adjustments deserves greater attention by modellers. We have seen models of common carriage which incorrectly apply the same total transportation cost to all scenarios, even though load size varies throughout the analysis.

Including quantity discounts introduces discontinuities to the total cost function, making some methods, such as continuous optimization, difficult. Moreover, the weight breaks must be known; this requires assumption of mode of transportation and possibly product. At the least, the modeler must be aware of their impact, and should interpret results accordingly.

The effect of transportation time variability resulting from such factors as different load sizes should be considered when calculating in-transit inventory holding cost. An estimate of the magnitude of this variability and the range of in–transit inventory costs should be made. If this range is large, some recognition of the fact should be given.

In most cases, a probability distribution of transportation time will not exist. This is especially true in theoretical analyzes such as vehicle routing, where the travel time (if being considered at all) of a route changes with each iteration. Empirical analysis may find that the collar–impact of transportation time variability is negligible.

Some allocation to shipments of some fixed costs, such as those relating to clerical support, information processing, and operating terminals should be considered when developing cost models. If it can be accurately assumed that changes to the distribution system will not affect indirect costs, omitting this cost allocation can be justified. However, system changes often will have a long–term effect, and the changes to administrative costs that result must be considered.

The major hurdle in including indirect costs is obtaining them. Ernst and Whinney [1983] reported that "most companies are not capturing data at a level desirable for effective transportation accounting and control". We do note that several studies have used cost data from actual companies, so estimates of indirect costs also may be possible.

The impact of costs on the performance and profitability of a physical distribution system cannot be underestimated. As a result, accurate and realistic modeling of these costs must be an important part of any analysis of distribution activities.

Table C–4
Decision Table for Modeling Shipper Distribution Costs

Transportation Cost

1–1 Are both quantity shipped and distance travelled constant throughout the analysis? If so, transportation cost can be modeled as a stated constant amount regardless of form of carriage; otherwise, the form of carriage must be considered.

1–2. Is the transportation done by common (for–hire) carrier or by private carrier? If done by common carrier, go to Question 1–3; if private carrier, go to Question 1–8.

Common Carrier

1–3. Does the analysis involve varying distances? If so, the transportation cost expression should include distance explicitly, or the freight rates used should have been derived with some consideration of distance. Otherwise, transportation cost usually can be modeled as function of quantity only.

1–4. Are the shipment weights fairly small (for example, under 75 pounds)? If so, consider applying small package rates. Under such an assumption, Question 1–7 below is irrelevant.

1–5. Is the cost of loading and unloading included in the common carrier freight rate? If yes, an explicit loading/unloading cost is not required; go to Question 1–7. If no, an explicit loading/unloading cost should be included; go to Question 1–6.

1–6. Is the quantity being shipped highly variable? Does the analysis focus on different or changing load sizes? If no to both questions, a fixed cost per dispatch or delivery may be appropriate to cover loading, clerical, and other activities. Otherwise, a cost based on quantity should be applied. This cost may be a separate component of the cost expression or may be part of shipment–handling costs.

1–7. Do discounts exist for various quantities or weights shipped? If yes, the mode of transport should be stated and the weight– breaks identified. The relevant range of shipment quantity should be examined to see if it covers more than one discount category. If so, analysis and results must be to include consideration of discounts.

Table C–4 (continued)

Private Carriage

1–8. Is the distance of shipment known and constant throughout the analysis? If yes, a straight fixed transportation cost per trip may be appropriate. Otherwise, use an expression that includes distance, such as "distance times transportation cost per–unit distance".

1–9. Does the analysis involve or affect backhauls? If so, any incremental revenues and costs of the backhaul should be considered.

1–10 Do vehicles make delivery stops at facilities other than those run by the shipper? If so, a cost per stop should be included. The cost of dispatching the vehicle and stopping at shipper facilities also should be considered, but may be treated as part of shipment–handling costs.

Inventory–Holding Cost

2–1. Is the transportation time very long? If not, the values of the inventory–cost parameter (cost per unit per time) for inventory waiting for shipment and for inventory in transit probably will be sufficiently similar. Then, a separate component for in–transit inventory cost will not be necessary.

2–2. Is the transportation time highly variable? If yes, an estimate of the magnitude of this variability and the range of in–transit inventory costs should be made. If both this range and the inventory–holding cost parameter are large, some recognition of the impact of transportation time variability on inventory cost should be made.

Shipment–Handling Costs

3–1. Does the analysis examine activities that may have an effect on shipment–handling in the distribution system? For example, will the analysis affect the number of shipments being transported or handled at a facility, or the method of handling? If yes, shipment–handling costs should be considered. Moreover, in many cases, a change in shipment–handling activities will cause changes to administrative and other indirect costs.

3–2. Is the transportation by common carrier? If so, some shipment–handling costs, such as those at a transloading terminal, are included in the carrier's freight charges.

Table C–4 (continued)

3–3. Do shipment–handling activities go beyond loading and unloading of vehicles? If not, a simple expression based on quantity per vehicle dispatch may be sufficient. Otherwise, a clear definition of activities involved is required; this may be expressed as part of facility costs.

Facility, Warehouse, and Terminal Costs

4–1 Does the analysis include facilities, excluding the point of origin, for which the shipper is responsible for costs? These facilities may include warehouses, terminals, and distribution centres. If yes, the impact of changes to operating costs of these facilities should be considered.

4–2 What is the function of the facilities? A warehouse will include costs of longer–term storage, and may require a separate inventory–carrying cost parameter; a terminal will not. If the facility is a transportation terminal and transportation is done by common carrier, costs related to shipment–handling (and possibly short–term storage) are included in the common carrier rate.

4–3. Are the facilities owned by the shipper or by an outside party (usually a public facility)?

If the facility is owned by an outside party, costs of handling and storage will be quoted to the shipper, typically based on volume. Then, it may be sufficient to include in the model a term such as "quantity multiplied by rate", with a time component if holding times vary in the analysis.

If facilities are privately–owned, total facility cost per shipment should include components for storage and inventory costs, handling costs, and an allocation of administrative and fixed costs.

Administrative and Other In–direct Costs

5–1. Will the analysis result in major or long–term changes to the distribution system or to activities such as order and information processing, administration, clerical support, etc.?

If yes, the analysis should include some allocation, usually as a cost per shipment, of administrative and other indirect costs.

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.

Ackerman, K. [1990]. *The Practical Handbook of Warehousing.* New York: Van Nostrand Reinhold.

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.

Allen, W.B., M.M. Mahmoud, and D. McNeil [1985], "The importance of time in transit and reliability of transit time for shippers, receivers, and carriers". *Transportation Research*, vol. 19B, no. 5, pp. 447–456.

Ansari, A., and J. Heckel [1987], "JIT purchasing: impact of freight and inventory costs". *Journal of Purchasing and Materials Management*, vol. 23, no. 2, pp. 24–28.

Arcelus, F.J., and J.E. Rowcroft [1991], "Small order transportation costs in inventory control". *The Logistics and Transportation Review*, vol. 27, no. 1, pp. 3–13.

Aucamp, D.C. [1982], "Nonlinear freight costs in the EOQ problem". *European Journal of Operational Research*, vol. 9, no. 1, pp. 61–63.

Bagchi, P.K. [1988], "Management of materials under Just–in–Time inventory system: a new look". *Journal of Business Logistics*, vol. 9, no. 2, pp. 89–101.

Bagchi, P.K., and F.W. Davis [1988], "Some insights into inbound freight consolidation". *International Journal of Physical Distribution and Materials Management*, vol. 18, no. 6, pp. 27–33.

Bagchi, T.P., and J.G.C. Templeton [1972]. *Numerical Methods in Markov Chains and Bulk Queues.* New York: Springer–Verlag.

Bagchi, U., J.C. Hayya, and J.K. Ord [1984], "Modeling demand during lead time". *Decision Sciences*, vol. 15, no. 2, pp. 157–176.

Bailey, N.T.J. [1954], "On queuing processes with bulk service". *Journal of the Royal Statistical Society*, vol. 16B, no. 1, pp. 80–87.

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

Ballou, R.H. [1976], "Computer methods in transportation–distribution". *Transportation Journal*, vol. 15, no. 2, pp. 72–85.

Ballou, R.H. [1991], "The accuracy in estimating truck class rates for logistical planning". *Transportation Research*, forthcoming.

Ballou, R.H. [1992]. *Business Logistics Management, 3rd edition.* Englewood Cliffs, New Jersey: Prentice–Hall Inc.

Banks, J., W. Driscoll, and R. Standford [1982], "Design methodology for an airport limousine service". *Transportation Science*, vol. 16, no. 2, pp. 127–148.

Baumol, W.J. and H. D. Vinod [1970], "An inventory theoretic model of freight transportation choice". *Management Science*, vol. 16, no. 7, pp. 413–421.

Beardwood, J., J.H. Halton and J.M. Hammersley [1959], "The shortest path through many points". *Proceedings of the Cambridge Philosophical Society*, vol. 55, pp. 299–327.

Beckmann, M., C.B. McGuire, and C.B. Winsten [1956]. *Studies in the Economics of Transportation*, New Have: Yale University Press.

Bellman, R.E., and S.E. Dreyfus [1962]. *Applied Dynamic Programming.* Princeton, New Jersey: Princeton University Press.

Bertsekas, D.P. [1976]. *Dynamic Programming and Stochastic Control.* New York: Academic Press.

Bhat, U.N. [1964], "Imbedded Markov chain analysis of a single server bulk queue". *Journal of the Australian Mathematical Society*, vol. 4, pp. 244–263.

Bloemena, A.R. [1960], "On queuing processes with a certain type of bulk service". *Bulletin Institut International de Statistique*, vol. 37, pp. 219–226.

Blumenfeld, D.E., and M.J. Beckmann [1985], "Use of continuous space modeling to estimate freight distribution costs". *Transportation Research*, vol. 19A, no. 2, pp. 173–187.

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*, vol. 19B, no. 5, pp. 361–380.

Bookbinder, J.H., and C.I. Barkhouse [1991], "An information system for simultaneous consolidation of inbound and outbound shipments". Working paper, University of Waterloo.

Bookbinder, J.H., and N.M. Fraser [1990], "The role of game theory in shipper–carrier negotiations". *Journal of the Transportation Research Forum*, vol. 30, no. 2, pp. 495–505.

Bookbinder, J.H., and J.K. Higginson [1990], "Shipment consolidation in location–distribution models". Presented at the May TIMS/ORSA meeting, Las Vegas, Nevada.

Bowersox, D.J., D.J. Closs, and O.K. Helferich [1986]. *Logistical Management, 3rd edition.* New York: Macmillan Publishing Co. Inc.

Brennan, J.J. [1981]. *Models and Analysis of Temporal Consolidation.* Doctoral dissertation, University of California at Los Angeles.

Buffa, F.P. [1986a], "Restocking inventory in groups: a transport–inventory case". *International Journal of Physical Distribution and Materials Management*, vol. 16, no. 3, pp. 29–44.

Buffa, F.P. [1986b], "Inbound logistics: analysing inbound consolidation opportunities". *International Journal of Physical Distribution & Materials Management*, vol. 16, no. 4, pp. 3–32.

Buffa, F.P. [1987], "Transit time and cost factors: their effect on inbound consolidation". *Transportation Journal*, vol. 27, no. 1, pp. 50–63.

Buffa, F.P. [1988], "An empirical study of inbound consolidation opportunities". *Decision Sciences*, vol. 19, no. 3, pp. 635–653.

Buffa, F.P., and J. Munn [1989], "A recursive algorithm for order cycle–time that minimizes logistics cost". *Journal of the Operational Research Society*, vol. 40, no. 4, pp. 367–377.

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

Canadian Permanent Committee on Geographical Names [1988]. *Gazetteer of Canada: Ontario.* Ottawa.

Chakravarty, A.K. [1981], "Multi–item inventory aggregation into groups". *Journal of the Operational Research Society*, vol. 32, no. 1, pp. 19–26.

Chase, R.B., and N.J. Aquilano [1992]. *Production and Operations Management, 6th edition.* Homewood, Illinois: Richard D. Irwin Inc.

Chaudhry, M.L., and J.G.C. Templeton [1972], "The theory of bulk–arrivals and bulk–service queues". *Opsearch*, vol. 9, pp. 103–121.

Chaudhry, M.L., and J.G.C. Templeton [1983]. *A First Course in Bulk Queues.* New York: John Wiley and Sons.

Chentnik, C.G. [1977], "Fixed facility location techniques", in M. Christopher and P. Schary, eds., *European Insights in Distribution.* Bradford, England: MCB Books.

Chiang, Y.S., and P.O. Roberts [1980], "A note on transit time and reliability for regular–route trucking". *Transportation Research*, vol. 14B, no. 1/2, pp. 59–65.

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.

Constable, G.K. and D.C. Whybark [1978], "The interaction of transportation and inventory decisions". *Decision Sciences*, vol. 3, no. 4, pp. 688–699.

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 strategies". *International Journal of Physical Distribution and Materials Management*, vol. 14, no. 6, pp. 47–67.

Daganzo, C.F. [1984a], "The length of tours in zones of different shapes". *Transportation Research*, vol. 18B, no. 2, pp. 135–145.

Daganzo, C.F. [1984b], "The distance traveled to visit N points with a minimum of C stops per vehicle: an analytic model and an application". *Transportation Science*, vol. 18, no. 4, pp. 331–350.

Daganzo, C.F. [1985], "Supplying a single location from heterogenous sources". *Transportation Research*, vol. 19B, no. 5, pp. 409–418.

Daganzo, C.F. [1987], "The break–bulk role of terminals in many–to–many logistic networks". *Operations Research*, vol. 35, no. 4, pp. 543–555.

Daganzo, C.F. [1988a], "Shipment consolidation enhancement at a consolidation center". *Transportation Research*, vol. 22B, no. 2, pp. 103–124.

Daganzo, C.F. [1988b], "A comparison of in–vehicle and out–of–vehicle freight consolidation". *Transportation Research*, vol. 22B, no. 3, pp. 173–180.

Daganzo, C.F. [1991]. *Logistics System Analysis*. Berlin: Springer–Verlag.

Daganzo, C.F., and G.F. Newell [1985], "Physical distribution from a warehouse: vehicle coverage and inventory levels". *Transportation Research*, vol. 19B, no. 5, pp. 397–405.

Das, C. [1972], "Choice of transport service: an inventory theoretic approach", *The Logistics and Transportation Review*, vol. 10, no. 2, pp. 181–187.

Davis, G.M., and J.E. Dillard, Jr. [1983]. *Physical Logistics Management*. Lanham, Maryland: University Press of America.

DeGhellinck, G.T., and G.D. Eppen [1967], "Linear programming solutions for separable Markovian decision problems". *Management Sciences*, vol. 13, no. 5, pp. 371–394.

DeHayes, D.W. [1968]. *The General Nature of Transit Time Performance For Selected Modes in the Movement of Freight*. Doctoral dissertation, Ohio State University.

Deming, W.E. [1978], "On a rational relationship for certain costs of handling motor freight: over the platform". *Transportation Journal*, vol. 17, no. 4, pp. 5–11.

Denardo, E.V. [1968], "Separable Markovian decision problems". *Management Sciences*, vol. 14, no. 7, pp. 451–462.

Denardo, E.V., and L.G. Mitten [1967], "Elements of sequential decision processes". *Journal of Industrial Engineering*, vol. 18, no. 1, pp. 106–112.

Derman, C. [1970]. *Finite State Markov Decisions Processes*. New York: Academic Press.

Dick, R.S. [1970], "Some theorems of a single server queue with balking". *Operations Research*, vol. 18, no. 6, pp. 1193–1205.

Dirick, Y.M., and L.P. Jennergren [1975], "On the optimality of myopic policies in sequential decision problems". *Management Science*, vol. 21, no. 5, pp. 550–556.

Downton, F. [1955], "Waiting time in bulk service queues". *Journal of the Royal Statistical Society*, vol. 13B, no. 2, pp. 256–261.

Eilon, S., C.D.T. Watson–Gandy, and N. Christofides [1971]. *Distribution Management: Mathematical Modelling and Practical Analysis*. London: Charles Griffin & Company.

Ernst and Whinney [1983]. *Transportation Accounting and Control: Guidelines for Distribution and Financial Management*. Oak Brook, Illinois: National Council of Physical Distribution Management.

Ernst and Whinney [1985]. *Warehouse Accounting and Control: Guidelines for Distribution and Financial Management*. Oak Brook, Illinois: National Council of Physical Distribution Management.

Fabens, A.J. [1961]. "The solution of queuing and inventory models by Semi–Markov processes". *Journal of the Royal Statistical Society*, vol. 23B, no. 1, pp. 113–127.

Fabens, A.J., and A.G.A.D. Perera [1963]. "A correction to 'The solution of queuing and inventory models by Semi–Markov processes'". *Journal of the Royal Statistical Society*, vol. 25B, no. 2, pp. 455–456.

Ferguson, W., and L.W. Glorfeld [1981]. "Modeling the present motor carrier rate structure as a benchmark for pricing in the new competitive environment". *Transportation Journal*, vol. 21, no. 2, pp. 59–66.

Firth, D., J. Apple, F.R. Denham, J. Hall, P. Inglis, and A.I. Saipe [1988]. *Profitable Logistics Management, 2nd edition*. Toronto: McGraw–Hill Ryerson.

Flood, K.U., O.G. Callson, and S.J. Jablonski [1984]. *Transportation Management, 4th edition*. Dubuque, Iowa: Wm. C. Brown Company Publishers.

Flores, B.E., and D.C. Whybark [1986]. "Multiple criteria ABC analysis". *International Journal of Operations and Production Management*, vol. 6, no. 3, pp. 46.

Foster, F.G. [1953]. "On the stochastic matrices associates with certain queuing processes". *Annuals of Mathematical Statistics*, vol. 24, pp. 355–360.

Giffin, W.C. [1978]. *Queuing: Basic Theory and Applications*. Columbus, Ohio: Grid Inc.

Griffith, T.F., N.E. Daniel, D.L. Shrock, and M.T. Farris [1983]. "Inbound freight and deregulation: a management opportunity". *Journal of Purchasing and Materials Management*, vol. 19, no. 3, pp. 16–21.

Gross, D., and C.M. Harris [1985]. *Fundamentals of Queuing Theory, 2nd edition*. New York: John Wiley and Sons.

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. [1985a], "Dependence between shipment size and mode in freight transportation". *Transportation Science*, vol. 19, no. 4, pp. 436–444.

Hall, R.W. [1985b], "Determining vehicle dispatch frequency when shipping frequency differs among suppliers". *Transportation Research*, vol. 19B, no. 5, pp. 421–431.

Hall, R.W. [1987], "Consolidation strategy: inventory, vehicles and terminals". *Journal of Business Logistics*, vol. 8, no. 2, pp. 57–73.

Heyman, D.P., and M.J. Sobel [1984a]. *Stochastic Models in Operations Research, volume 1: Stochastic Processes and Operating Characteristics*. New York: McGraw–Hill Book Company.

Heyman, D.P., and M.J. Sobel [1984b]. *Stochastic Models in Operations Research, volume 2: Stochastic Optimization*. New York: McGraw–Hill Book Company.

Higginson, J.K. [1991], "Canadian Cotton Distributors Inc." Unpublished case study, Waterloo: Department of Management Sciences, University of Waterloo.

Howard, R.A. [1960]. *Dynamic Programming and Markov Processes*. Cambridge, Massachusetts: The M.I.T. Press.

Howard, R.A. [1971a]. *Dynamic Probabilistic Systems, volume 1: Markov Models*. New York: John Wiley & Sons Inc.

Howard, R.A. [1971b]. *Dynamic Probabilistic Systems, volume 2: Semi-Markov and Decision Processes*. New York: John Wiley & Sons Inc.

Jackson, G.C. [1980], "Order consolidation research to 1980 and beyond", in D.J. Lambert, ed. [1980], *A Decade of Distribution Research*. Columbus, Ohio: The Ohio State University.

Jackson, G.C. [1981], "Evaluating order consolidation strategies using simulation". *Journal of Business Logistics*, vol. 2, no. 2, pp. 110–138.

Jackson, G.C. [1985], "A survey of freight consolidation practices". *Journal of Business Logistics*, vol. 6, no. 1, pp. 13–34.

Jaillet, P., and A.R. Odoni [1988], "The probabilistic vehicle routing problem", in B.L. Golden and A.A. Assad, eds., *Vehicle Routing: Methods and Studies*. Amsterdam: North–Holland.

Jaiswal, N.K. [1960a], "Bulk–service queuing problem". *Operations Research*, vol. 8, no. 1, pp. 139–143.

Jaiswal, N.K. [1960b], "Time–dependent solution of the bulk–service queuing problem". *Operations Research*, vol. 8, no. 6, pp. 773–781.

Jaiswal, N.K. [1961], "A bulk–service queuing problem with variable capacity". *Journal of the Royal Statistical Society*, vol. 23B, no. 1, pp. 143–148.

Jansson, J.O., and D. Shneerson [1978], "Economies of scale of general cargo ships". *The Review of Economics and Statistics*, vol. 60, no. 2, pp. 287–293.

Kendall, D.G. [1951], "Some problems in the theory of queues". *Journal of the Royal Statistical Society*, vol. 13B, no. 2, pp. 151–185.

Kendall, D.G. [1953], "Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chain". *Annals of Mathematical Statistics*, vol. 24, pp. 338–354.

Lalonde, B.J., and D.M. Lambert [1977], "Inventory carrying costs: significance, components, means, functions", in M. Christopher and P. Schary, eds., *European Insights in Distribution*. Bradford, West Yorkshire, England: MCB Books.

Lambert, D.M. [1976]. *The Development of an Inventory Costing Methodology: A Study of the Costs Associated with Holding Inventory*. Chicago, Illinois: National Council of Physical Distribution Management.

Lambert, D.M., and H.M. Armitage [1979], "Distribution costs: the challenge". *Management Accounting*, vol. 60, no. 11, pp. 33–45.

Lambert, D.M., M.L, Bennion, Jr., and J.C. Taylor [1987], "Solving the small order problem". *International Journal of Physical Distribution and Materials Management*, vol. 17, no. 2, pp. 7–19.

Larson, R.C. and A.R. Odoni [1981]. *Urban Operations Research*. Englewood Cliffs, New Jersey: Prentice–Hall, Inc.

Law, A.M, and W.D. Kelton [1991]. *Simulation Modeling and Analysis*. New York: McGraw–Hill Book Company.

Jaillet, P., and A.R. Odoni [1988], "The probabilistic vehicle routing problem", in B.L. Golden and A.A. Assad, eds., *Vehicle Routing: Methods and Studies*. Amsterdam: North–Holland.

Jaiswal, N.K. [1960a], "Bulk–service queuing problem". *Operations Research*, vol. 8, no. 1, pp. 139–143.

Jaiswal, N.K. [1960b], "Time–dependent solution of the bulk–service queuing problem". *Operations Research*, vol. 8, no. 6, pp. 773–781.

Jaiswal, N.K. [1961], "A bulk–service queuing problem with variable capacity". *Journal of the Royal Statistical Society*, vol. 23B, no. 1, pp. 143–148.

Jansson, J.O., and D. Shneerson [1978], "Economies of scale of general cargo ships". *The Review of Economics and Statistics*, vol. 60, no. 2, pp. 287–293.

Kendall, D.G. [1951], "Some problems in the theory of queues". *Journal of the Royal Statistical Society*, vol. 13B, no. 2, pp. 151–185.

Kendall, D.G. [1953], "Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chain". *Annals of Mathematical Statistics*, vol. 24, pp. 338–354.

Lalonde, B.J., and D.M. Lambert [1977], "Inventory carrying costs: significance, components, means, functions", in M. Christopher and P. Schary, eds., *European Insights in Distribution*. Bradford, West Yorkshire, England: MCB Books.

Lambert, D.M. [1976]. *The Development of an Inventory Costing Methodology: A Study of the Costs Associated with Holding Inventory*. Chicago, Illinois: National Council of Physical Distribution Management.

Lambert, D.M., and H.M. Armitage [1979], "Distribution costs: the challenge". *Management Accounting*, vol. 60, no. 11, pp. 33–45.

Lambert, D.M., M.L, Bennion, Jr., and J.C. Taylor [1987], "Solving the small order problem". *International Journal of Physical Distribution and Materials Management*, vol. 17, no. 2, pp. 7–19.

Larson, R.C. and A.R. Odoni [1981]. *Urban Operations Research*. Englewood Cliffs, New Jersey: Prentice–Hall, Inc.

Law, A.M, and W.D. Kelton [1991]. *Simulation Modeling and Analysis*. New York: McGraw–Hill Book Company.

Lawler, E.L., J.K Lenstra, A.H.G. Rinnooy Kan and D.B. Shmoys, eds. [1985]. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization.* Chichester, England: John Wiley and Sons.

LeKashman, R., and J.F. Stolle [1965], "The total cost approach to distribution". *Business Horizons*, vol. 8, pp. 33–46.

Lewis, R.J. [1991], "Activity–based costing for marketing". *Management Accounting*, vol. 73, no. 5, pp. 33–38.

Liberatore, M.J. [1979], "A model of freight transport selection", *Transportation Journal*, vol. 18, no. 4, pp. 92–100.

Love, R.F., J.D. Morris, and G.O. Wesolowsky [1988]. *Facilities Location: Models and Methods.* New York: North–Holland.

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

Maranzana, F.E. [1964], "On the location of supply points to minimize transportation costs". *Operational Research Quarterly*, vol. 15, no. 3, pp. 261–270.

Marglin, S.A. [1963]. *Approaches to Dynamic Investment Planning.* Amsterdam: North–Holland Publishing Company.

Masters, J.M. [1980], "The effects of freight consolidation on customer service". *Journal of Business Logistics*, vol. 2, no. 1, pp. 55–74.

McFadden, D., C. Winston and A. Boerch–Supan [1985], "Joint estimation of freight transportation decisions under non–random sampling", in A.F. Daugherty, ed., *Analytical Studies in Transport Economics.* Cambridge, England: Cambridge University Press.

McKinnon, A.C. [1989]. *Physical Distribution Systems.* London: Routledge.

McKenzie, D.R., M.C. North, and D.S. Smith [1989]. *Intermodal Transportation.* Omaha, Nebraska: Simmons–Boardman Books Inc.

Merriam–Webster Inc. [1986]. *Webster's Third New International Dictionary of the English Language, unabridged edition.* Springfield, Massachusetts: Merriam–Webster Inc. Publishers.

Min, H., and M.C. Cooper [1990], "A comparative review of analytical studies on freight consolidation and backhauling". *The Logistics and Transportation Review*, vol. 26, no. 2, pp. 149–170.

Medhi, J. [1975], "Waiting time distribution in a Poisson queue with a general bulk–service rule". *Management Science*, vol. 21, no. 7, pp. 777–782.

Mercer, A. [1968], "A queue with random arrivals and scheduled bulk departures". *Journal of the Royal Statistical Society*, vol. 30B, no. 1, pp. 185–189.

Meyers, C.F., J.L. Fanelli, and D.C. Boger [1987]. "Assessing the impact of consolidation on inbound vendor traffic". *Journal of the Transportation Research Forum*, vol. 27, pp. 230–236.

Miller III, A.C., and T.R. Rice [1983], "Discrete approximations of probability distributions", *Management Science*, vol. 29, no. 3, pp. 352–362.

Mine, H., and S. Osaki [1970], *Markov Decision Processes.* New York: Elsevier Publishing Company.

Moulton, G.W. [1990]. "Book review: The practical warehousing book". *Canadian Transportation*, vol. 93, no. 11, p. 50.

Nair, S.S., and M.F. Neuts [1972], "Distribution of occupation time and virtual waiting time of a general class of bulk queues". *Sankhä Series A*, vol. 34, no. 1, pp. 17–22.

Natarajan, R. [1962], "Discrete–time bulk service queuing processes". *Defence Science Journal*, vol. 12, pp. 318–326.

Newbourne, M.J. [1976]. *A Guide to Freight Consolidation for Shippers.* Washington, D.C.: Traffic Service Corp.

Newbourne, M.J., and C. Barrett [1972], "Freight consolidation and the shipper," Parts 1 to 5. *Transportation and Distribution Management*, vol. 12, nos. 2 to 6.

Neuts, M.F. [1967], "A general class of bulk queues with Poisson input". *Annals of Mathematical Statistics*, vol. 38, pp. 759–770.

Neuts, M.F. [1973], "The single server queue in discrete time, numerical analysis I". *Naval Research Logistics Quarterly*, vol. 20, no. 2, pp. 297–304.

Neuts, M.F. [1979], "Queues solvable without Rouché's theorem". *Operations Research*, vol. 27, no. 4, pp. 767–781.

Novaes, A, and E. Frankel [1966], "A queuing model for unitized cargo generation". *Operations Research*, vol. 14, no. 1, pp. 100–132.

Perl, J. and M.S. Daskin [1985], "A warehouse location–routing problem". *Transportation Research*, vol. 19B, no. 5, pp. 381–396.

Perl, J. and S. Sirisoponslip [1988], "Distribution networks: facility location, transportation, and inventory". *International Journal of Physical Distribution and Materials Management*, vol. 18, no. 6, pp. 18–26.

Petersen, E.R., and A.J. Taylor [1988]. *PROPS: Probabilistic Optimization Spreadsheets*. Kingston, Ontario: Alwington Press.

Piercy, J.E. [1977]. *A Performance Profile of Several Transportation Freight Services*. Doctoral dissertation, Case Western Reserve University.

Pollock, T. [1978], "A management guide to LTL consolidation." *Traffic World*, April 3, pp. 29–35.

Powell, W.B. [1985], "Analysis of vehicle holding and cancellation strategies in bulk arrival, bulk service queues". *Transportation Science*, vol. 19, no. 4, pp. 352–377.

Powell, W.B. [1986], "Iterative algorithms for bulk arrival bulk service queues with Poisson and non–Poisson arrivals". *Transportation Science*, vol. 20, no. 2, pp. 65–79.

Powell, W.B., and P. Humblet [1986]. "The bulk service queue with a general control strategy: theoretical analysis and a new computational procedure". *Operations Research*, vol. 34, no. 2, pp. 267–275.

Powell, W.B., and Y. Sheffi [1983], "The load planning problem of motor carriers: problem description and proposed solution approach". *Transportation Research*, vol. 17A, no. 6, pp. 471–480.

Powell, W.B., and Y. Sheffi [1989], "Design and implementation of an interactive optimization system for network design in the motor carrier industry". *Operations Research*, vol. 37, no. 1, pp. 12–29.

Ross, S.M. [1983]. *Introduction to Stochastic Dynamic Programming*. Orlando, Florida: Academic Press Inc.

Ross, S.M. [1990]. *An Introduction to Probability Models, 4th edition*. Orlando, Florida: Academic Press Inc.

Russell, R.M., and L. Krajewski [1991], "Inbound freight consolidation policies with transportation economies and price–quantity discounts". Working paper, Columbia: The University of South Carolina.

Samuelson, R.D. [1977], "Modeling the freight rate structure". *Report #77-7*, Cambridge, Massachusetts: Centre for Transportation Studies.

Schuldenfrei, R.L., and J.M. Shapiro [1980], "Inbound collection of goods: the reverse distribution problem". *Interfaces*, vol. 10, no. 4, pp. 30–33.

Schwarz, L.B. [1989], "A model for assessing the value of warehouse risk–pooling: risk–pooling over outside–supplier leadtimes". *Management Science*, vol. 35, no. 7, pp. 828–842.

Shapiro, R.D., and J.L. Heskett [1985]. *Logistics Strategy: Cases and Concepts.* St. Paul, Minnesota: West Publishing Co.

Sheahan, D. [1982], "Know thy carrier: All kinds of cost–saving opportunities exist in LTL consolidation". *Handling and Shipping Management*, vol. 23, pp. 44–46.

Sheffi, Y. [1986], "Carrier/shipper interactions in the transportation market: an analytical framework". *Journal of Business Logistics*, vol. 17, no. 1, pp. 1–27.

Shelley, D.F. [1982], "Costing pool shipments: A simple formula for determining the base–delivered prices on goods moving in mixed–load consolidated shipments". *Handling and Shipping Management*, vol. 23, pp. 67–68.

Shuster, A.D. [1979], "The economics of shipment consolidation". *Journal of Business Logistics*, vol. 1, no. 2, pp. 22–35.

Singh, V.P. [1971], "Finite waiting space bulk service system". *Journal of Engineering Mathematics*, vol. 5, no. 4, pp. 241–248.

Singh, V.P. [1972], addendum to "Finite waiting space bulk service system". *Journal of Engineering Mathematics*, vol. 6, no. 1, pp. 85–88.

Stein, D.M. [1978], "Scheduling dial–a–ride transportation systems". *Transportation Science*, vol. 12, no. 3, pp. 232–249.

Stidham, S., Jr. [1974], "Stochastic clearing systems", *Stochastic Processes and Their Applications*, vol. 2, no. 1, pp. 85–113.

Stidham, S., Jr. [1977], "Cost models for stochastic clearing systems", *Operations Research*, vol. 25, no. 1, pp. 100–127.

Stock, J.R., and D.M. Lambert [1987]. *Strategic Logistics Management, 2nd edition.* Homewood, Illinois: Richard D. Irwin Inc.

Sutton, R.M. [1966], "A checklist for reducing transportation costs". *Transportation and Distribution Management*, August, pp. 30–31.

Takács, L. [1962]. *Introduction to the Theory of Queues.* New York: Oxford University Press.

Taff, C.A. [1984]. *Management Of Physical Distribution and Transportation, 7th edition.* Homewood, Illinois: Richard D. Irwin Inc.

Temple, Barker and Sloane, Inc. [1982]. *A Carrier's Handbook for Costing Individual Less-Than-Truckload Shipments.* Prepared for the Regular Route Carrier Conference, Washington, D.C.

Tersine, R.J., P.D. Larson, and S. Barman [1989], "An economic inventory/transport model with freight rate discounts". *The Logistics and Transportation Review*, vol. 25, no. 4, pp. 291–306.

Thie, P.R. [1983]. *Markov Decision Processes.* Lexington, Massachusetts: Consortium for Mathematics and Its Applications.

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

van der Wal, J. [1981]. *Stochastic Dynamic Programming.* Amsterdam: Mathematisch Centrum.

Weart, W.L. [1984], "The techniques of freight consolidation: an area of increased interest aided by computers". *Traffic World*, March 14, pp. 46–60.

Webb, M.H.J. [1968], "Cost functions in the location of depots". *OR Quarterly*, vol. 19, no. 3, pp. 311–320.

Willis, R. [1977]. *An Analytical Approach to Physical Distribution Management.* London, England: Kogan Page.

Wishart, D.M.G. [1956], "A queuing system with chi–square service–time distribution". *Annals of Mathematical Statistics*, vol. 27, pp. 768–779.

Zahir, S., and R. Sarker [1991], "Joint economic ordering policies of multiple wholesalers and a single manufacturer with price–dependent demand functions". *Journal of the Operational Research Society*, vol. 42, no. 2, pp.157–164.