2013

Lead-Time Quotation When Customers are Sensitive to Reputation

Susan A. Slotnick

Cleveland State University, s.slotnick@csuohio.edu

Lead-time quotation when customers are sensitive to reputation

Susan A. Slotnick*

Department of Operations and Supply Chain Management, Monte Ahuja College of Business Administration, Cleveland State University, Cleveland, OH, USA

Firms consider a variety of factors when making lead-time promises, including current shop status and the size of the incoming order. The profit-maximising model presented in this paper is the first to include reputation effects explicitly in a lead-time optimisation model. Reputation is considered to be the lasting effect on the market of a firm's delivery performance over time, and so it affects the future as well as the current profits. The model is complicated, and a counter-example demonstrates that qualitative monotonicity results are not obtainable. A computational study explores the relationships between shop status, order size, reputation, market characteristics and the lead-time decision. Regression analysis sheds light on these relationships and suggests three heuristics, which provide near-optimal solutions with relatively short running times.

Keywords: scheduling; due-date assignment; dynamic programming

1. Introduction

Firms that provide services or manufacture to order must take a variety of factors into consideration when they promise customers that goods or services will be completed at a specific time. In addition to current shop backlog and processing time for the incoming order (which may be known or estimated), the firm may also want to consider how its present lead-time promise will affect the current congestion as well as the future business. For example, a lead-time promise that is too optimistic, in order to lure the customer into placing a firm order, will increase congestion in the facility. If the overly optimistic promise results in late delivery, this may affect the firm's reputation for on-time delivery and deter future customers. However, including the concern for reputation complicates the trade-off that the firm faces when making lead-time decisions. In addition to balancing the pressure to quote a short lead-time (in order to retain impatient customers) with the costs of delivering late if the lead-time is too optimistic, the firm must also consider the possibility of losing potential future customers, if its history of delivery performance is below expectations.

The importance of on-time delivery to a company's reputation is well documented in the practical literature as well as in research articles. When viewed as one aspect of quality (Gjerde and Slotnick 1997; Hua, Wang, and Cheng 2010), timely delivery is likely to affect future business: '[the] dependence of patronage on the quality of service is widely taken for granted by managements of companies in service industries' (Sobel 1973). For example, a study of health-care providers found that 'patient satisfaction dropped significantly with each five minutes of waiting time' (Beck 2010). Studies have found that a medical facility's reputation for treating patients on time is a major determinant of whether they will return and/or recommend the provider to others (Hill and Joonas 2005), and there are similar results for other service industries (Keaveney 1995). In the airline industry, delays are a leading cause of traveler complaints; the US Department of Transportation keeps records which are available online (McCartney 2005), and airlines are likely to lose customers because of repeated delays (Anonymous 2000).

Reputation for on-time delivery is also important to manufacturing firms. Boeing has incurred costs in the billions, as well as potential and actual loss of customers, because of its considerable delay in the delivery of the Dreamliner (Lunsford 2007, 2008a,b; Sanders 2009, 2010a,b, 2011). Its competitor Airbus 'has seen its reputation tarnished and other business fly off amid delays in the superjumbo program', including the potential loss of UPS as customer and loss of other customers to Boeing (Michaels 2006; Annett 2007a,b). In its competition to win customers from Dell, Hewlett Packard responded to customer complaints about late deliveries by improving its delivery performance by 30%, as well as informing customers about availability of products on order (Lawton 2007).

Given these real-world examples of the relationships between on-time delivery, reputation and customer attraction and retention, the present paper seeks to answer the following questions:

*Email: s.slotnick@csuohio.edu

(1) How should the firm's reputation for on-time delivery influence its lead-time for an order?

(2) How should the firm balance the attractiveness of relatively short lead-times with the possible degradation of its reputation due to an increased likelihood of tardiness?

(3) How are the above decisions influenced by market characteristics such as customer "memory" (the persistence of reputation), customer impatience (preference for shorter lead-times) and the relative importance of reputation for on-time delivery in the market?

(4) Does the reputation of the firm make a difference in balancing these factors?

While there is an extensive literature on modelling lead-time quotation, and a variety of studies incorporating decisions related to reputation, *the present paper is the first to include reputation effects explicitly in a lead-time optimisation model.* Analysis of this more complicated model confirms previous results about the importance of considering shop status and order processing time when setting lead-times, suggests how a firm's reputation should be included, and employs near-monotone properties to develop high-quality heuristics.

The rest of this paper is structured as follows. Section 2 discusses related research on lead-time decisions and reputation. The details of the model are presented in Section 3, the analysis is in Section 4 and Section 5 describes the computational study and its results. Section 6 presents a summary and conclusions.

2. Related work

The literature on lead-time decisions spans a number of research areas, including scheduling, inventory policy and production planning. The most recent comprehensive survey of lead-time research is by Keskinocak and Tayur (2004). Zorzini, Corti, and Pozzetti (2008) discuss lead-time and due-date setting papers from the perspective of capacity planning, and Upasani and Uzsoy (2008) review lead-time analysis in joint production-marketing models. See also the literature review sections in some of the papers discussed below (in particular, Slotnick and Sobel 2005, Slotnick 2011, Pekgun, Griffin, and Keskinocak 2008). The following discussion focuses on those papers most relevant to the present work, in which lead-time is a decision variable.

A series of related papers present profit-maximising models of lead-time quotation with tardiness penalties. Chatterjee, Slotnick, and Sobel (2002) develop an optimal lead-time policy for a sales department that does not have exact information about current shop status. Slotnick and Sobel (2005) compare that policy with the case when the firm does have complete information about shop backlogs and delays, and find conditions under which it is worthwhile for the sales department to expend resources in order to secure that knowledge. In a paper motivated by the delivery improvement task force of a steel mill, Slotnick (2011) develops a lead-time model for a bottleneck process with minimum batch size requirements. The present paper extends this previous work by adding the factor of reputation to what the firm should consider when setting its lead-times.

A number of papers use lead-times to set delivery promises when lateness will not be tolerated. Keskinocak, Ravi, and Tayur (2001) develop online and offline algorithms when orders have release times and lead-time sensitive revenue. Kaminsky and Lee (2008) present online algorithms that minimise the sum of quoted due dates for a single processor. Two classes of customers with different revenues and sensitivities to delivery times characterise the model of Kapuscinski and Tayur (2007). Plambeck (2004) also considers two customer classes, and includes order sequencing in the decision problem; a simulation study shows the accuracy of diffusion approximations. Ata and Olsen (2009) develop a model of capacity, lead-time and sequencing decisions, and investigate the optimal policies for different delay cost functions.

While most research on the effect of reputation has been fairly recent, there are a few related papers that were written decades ago. In a paper on inventory theory, Schwartz (1966) models how the disappointment resulting from unavailable product (a stockout) changes the future actions of a customer, that is, demand is perturbed. Since reputation-related costs (i.e. goodwill and lost sales) are experienced in the future, the customer's memory of disappointment is modelled using exponential smoothing. Almost a decade later, Sobel (1973) presents a queueing model in which arrival rates are affected by past waiting times. In a study of a service facility (a commercial bank), customers will switch providers if waits become too long. Here as well, the customer's memory of waiting times is modelled by an exponentially weighted average.

More recent work includes studies of how the expenditure of resources on quality efforts may depend on the effects of reputation, as well on as the cost and revenue structure of the manufacturing or service organisation. From the perspective of economics, reputation is considered an asset, about which customers have imperfect information (e.g. Shapiro 1983). In marketing studies, reputation is one factor in 'brand success' (e.g. Herbig and Milewicz 1995). For comparative discussions of the research on reputation, quality and related learning effects in the fields of operations, economics and marketing, see Ernst and Powell (1995), Hall and Porteus (2000), Gans (2002) and Gjerde and Slotnick (2004).

A number of papers focus on the effect of consumer response to uncertain levels of service. Gans (2002) presents a model in which customers are not well informed about quality levels, but behave in a Bayesian fashion as they collect information over time: "the customer myopically chooses the supplier that has the highest probability of being good". This response, and ability of customers to switch suppliers, drives competing firms to establish quality standards for the industry, which increase with the number of suppliers. Gaur and Park (2007) model the situation in which the customer chooses a retailer based on a history of service (the impact of which diminishes over time). In their study, consumers react differently to good and bad experiences, and the authors find that retailers do better when they take into account asymmetries in customers' sensitivity to service, and also consider the interaction between marketing and operations.

Hall and Porteus (2000) develop a multi-period dynamic model of two firms which must decide on capacity levels, and whose customers are sensitive to the availability of service or product (service denial or stockout). Current service failures have future consequences, since a customer who is denied in the present period will switch to the competitor in the next period. Results include a measure of the value of a customer over multiple periods (time), a measure of customer defection and the optimal capacity level. Aflaki and Popescu (2011) employ a dynamic programming model to investigate how a firm can manage customer retention over time by using responsive service strategies. Adelman and Mersereau (2013) consider how a firm should allocate its capacity to serve a diverse collection of customers who remember past fill rates; using approximate dynamic programming, they discover conditions when it is optimal for the firm to act myopically, and when it is beneficial to employ a policy that considers tradeoffs between short- and long-term value.

Building on an earlier paper that considers on-time delivery to be one component of quality (Gjerde and Slotnick 1997), Gjerde and Slotnick (2004) model the interaction of price, production quantity, product quality and the persistence of reputation in the marketplace. Reputation is said to have a "half life", which is represented by an exponential smoothing formula. Taking into account the persistence of reputation and the length of time that quality efforts make a difference, along with the firm's cost and market structure, the authors find conditions under which it is advisable for a firm to raise or lower its expenditures on quality, change its production quantity or both.

The present research extends the literature by developing a model that unifies reputation (specifically, the firm's history of on-time delivery) with other factors involved in lead-time quotation. For the firm, the trade-off inherent in the lead-time decision involves the desire to keep the customer by quoting a relatively short lead-time, which may in turn result in increased congestion in the shop, tardiness costs *and* loss of future business, based on reputation for on-time delivery in the future. Including reputation increases the complexity of the model, but provides insights into the effects of a combination of internal (backlog, on-time delivery history) and external (order characteristics, customer sensitivities and reputation) factors on lead-time decisions.

3. The model

Prospective customers arrive individually at the firm and request lead-time quotations at an exogenous sequence of times comprising a Poisson process with intensity $\lambda > 0$. The firm quotes a lead-time to each prospective customer who responds either by placing a firm order or by balking. The firm's objective is to maximise its expected total profit from N prospective customers.

Let L_n denote the lead-time quotation to prospective customer n ($n = 1, 2, \ldots, N$) and let $\delta_n = 0$ if the customer balks; otherwise $\delta_n = 1$. The balking probability depends on L_n and on a scalar summary T_n of the firm's reputation for tardy delivery. At the time when it makes the lead-time quotation, the firm knows the size a_n of the prospective customer's order. It is assumed that the order sizes $a_1, a_2, \ldots, a_N$ are independent and identically distributed random variables which do not depend on the lead-time quotations and with a_n revealed to the firm at the time that prospective customer arrives.

The customer places a firm order with probability $e^{-(\xi_n L_n + \gamma_n T_n)}$ and balks with probability $1 - e^{-(\xi_n L_n + \gamma_n T_n)}$. These two expressions represent the probabilities of transfer to a state where the order is accepted and joins the queue for processing, and transfer to a state where the customer balks, and the queue is not augmented, respectively.

That is,

$$P\{\delta_n = 1\} = e^{-(\xi_n L_n + \gamma_n T_n)} \text{ and } P\{\delta_n = 0\} = 1 - e^{-(\xi_n L_n + \gamma_n T_n)} \tag{1}$$

The parameters ξ_n and γ_n reflect the customer's sensitivity to lead-time and tardiness reputation, respectively. For a given lead-time L_n, the higher the value of ξ_n, the less likely that the customer will submit the order. Similarly, for a given value of T_n, the higher the value of γ_n, the less likely that the customer will submit the order. To determine values for these parameters in actual practice, a firm would employ its knowledge of the market, and its past data on customer behaviour. For example, a steel manufacturer knows that important customers such as automobile fabrication plants insist on short-lead-times *and* on-time delivery, and so would be more sensitive to both parameters than other customers (Slotnick 2011).

If the order is placed, the firm's profit consists of revenue minus proportional tardiness penalties. The revenue credited is $w^r a_n$, which is the size of the order (a_n) times the per-unit revenue rate w^r. A tardy order results in the reduction of revenue by per-unit penalty rate w^p multiplied by the amount of time that an order is delivered after its promise date. Let B_n denote the processing-time backlog in the shop. Under the assumption that the shop processes orders on a first-in first-out basis, if prospective customer n does not balk, its tardiness will be $(a_n + B_n - L_n)^+$.

Reputation in this model consists of a smoothed value of the history of tardy deliveries. Smoothing effects have been used extensively in previous studies to model reputation (see the references in Section 2 above, and particularly page 6 of Aflaki and Popescu (2011)). Here, reputation is reflected in the history of tardy deliveries. If customer n balks, i.e. if $\delta_n = 0$, then $T_{n+1} = T_n$. Otherwise, T_n is updated as

$$T_{n+1} = \alpha(a_n + B_n - L_n)^+ + (1 - \alpha)T_n \tag{2}$$

where $0 \le \alpha \le 1$ is a smoothing constant that reflects the impact of any tardiness incurred by the arriving job on the firm's reputation. If the current job is not tardy, this index diminishes; otherwise it increases. This factor T will be referred to as "reputation for on-time delivery" or "tardiness index". Note that a higher value of T indicates a *worse* reputation. The values for the parameter α that might be used in practice would be derived from the firm's knowledge of the nature of information about reputation, and customer memory, in the marketplace. For example, Adelman and Mersereau (2013) describe customers with varying degrees of "memory" about their service experiences, and Gans (2002) describes a situation where customers are myopic about reputation (which would correspond to a higher value of α). Gaur and Park (2007) also model reputation which diminishes over time.

It is convenient to define the unit of time as the length of time it takes the shop to process one unit of goods. That is, the unit of time is defined so that it would take the shop a_n units of time to process customer n's order. In these units, let τ_n denote the interval of time which separates the arrivals of the nth and $n + 1^{st}$ prospective customers. The Poisson arrival process assumption implies that $\tau_1, \tau_2, \ldots, \tau_N$ are independent and identically distributed exponential random variables with rate λ. Let τ denote a generic random variable with the same exponential distribution. The first-in first-out assumption implies that the shop backlogs satisfy the recursion

$$B_{n+1} = (B_n + \delta_n a_n - \tau_n)^+ \tag{3}$$

for the following reasons. If customer n balks, then $\delta_n = 0$ and the backlog when customer $n+1$ arrives will be $max\{B_n - \tau_n, 0\}$ because the shop requires B_n units of time to process the backlog. If customer n does not balk, then $\delta_n = 1$, the backlog increases by the amount a_n, and the shop needs $B_n + a_n$ units of time to process the larger backlog. So $B_{n+1} = max\{B_n + a_n - \tau_n, 0\}$.

Let $z(a, B, L) = (a + B - L)^+$. Then the firm's expected single-period reward is

$$E\left(e^{-(\xi_n L_n + \gamma_n T_n)} \left[w^r a_n - w^p z(a_n, B_n, L_n) \right] \right). \tag{4}$$

Let $\eta \ge 0$ be a continuous-time discount factor and let β denote the expected present value (expected value of the present value), evaluated at the time when customer n arrives, of a unit profit τ_n units of time later. Using a standard derivation,

$$\beta = E(e^{-\eta \tau})$$
$$= \int_0^\infty e^{-\eta x} \lambda e^{-\lambda x} dx$$
$$= \frac{\lambda}{\lambda + \eta}.$$

The present value of the profit during the planning horizon is

$$\sum_{n=1}^N \left(e^{-\eta \sum_{j=1}^{n-1} \tau_j} \right) \delta_n \left[w^r a_n - w^p z(a_n, B_n, L_n) \right], \tag{5}$$

and using a standard derivation, the expected present value is

$$E\left(\sum_{n=1}^{N}\left(e^{-\eta\sum_{j=1}^{n-1}\tau_j}\right)e^{-(\xi_n L_n+\gamma_n T_n)}[w^r a_n - w^p z(a_n, B_n, L_n)]\right)$$

$$= E\left(\sum_{n=1}^{N}\beta^{n-1}e^{-(\xi_n L_n+\gamma_n T_n)}[w^r a_n - w^p z(a_n, B_n, L_n)]\right). \qquad (6)$$

The optimisation problem is to maximise the expected present value of profit (6) by choosing lead-time quotations $L_1 \geq 0, L_2 \geq 0, \ldots, L_N \geq 0$ non-anticipatively.

The sources of randomness in this model are balking and the times between arriving jobs. The exponential inter-arrival times of the exogenous Poisson process imply that the model is a continuous-time MDP. However, the discounted criterion and the specification of the planning horizon in terms of the number of jobs permit a standard reduction of the continuous-time model to an equivalent discrete-time MDP.

The MDP model corresponds to the following dynamic program with $f_{K+1}(\cdot, \cdot, \cdot, \cdot) \equiv 0$:

$$f_k(a, B, T) = \qquad (7)$$
$$max_{L>0}\left\{e^{-\xi L-\gamma T}[w^r a - w^p z(a, B, L)] + \beta e^{-\xi L-\gamma T}\left[E[f_{k+1}(a', B' - \tau, T')]\right]\right.$$
$$\left. +\beta(1 - e^{-\xi L-\gamma T})E[f_{k+1}(a', B - \tau, T)]\right\}$$

where:

K	planning horizon	k	k current period (customer)
	(total number of arriving customers)	a	processing time of arriving order (RV)
B	current shop backlog	T	current tardiness index (reputation)
L	lead-time quotation	ξ	customer impatience
γ	customer sensitivity to tardiness	w^r	per-unit reward
w^p	per-unit tardiness penalty	$z(a, B, L)$	tardiness function
a'	size of next arriving order	B'	updated shop backlog $(B + a)$
T'	updated tardiness level (reputation)	α	smoothing constant, $0 < \alpha < 1$
τ	elapsed time since last arrival	β	discount factor

The maximisation in (7) refers to a customer who has just arrived, and to whom a due-date must be quoted. The first term is the expected immediate reward: revenue minus any tardiness penalties (cf. (5) and (6)). The second term is the expected future reward if the customer places the order, and the third term is the expected future reward if the customer balks. The dynamics of the model (transition equations) are (2) and (3), the recursions for reputation T and backlog B, respectively. The transition probability is defined by (1).

4. Analysis of the model

4.1 *Introduction*

How can the MDP described in Section 3 contribute to an understanding of lead-time decisions, when the customers hold the firm accountable for its past delivery performance? Intuition and previous work (e.g. Chatterjee, Slotnick, and Sobel 2002; Slotnick and Sobel 2005; Slotnick 2011) suggest that shop backlog and order size should be taken into account. How should the firm's reputation for on-time delivery influence the delivery promise for the current order? On one hand, it seems reasonable that a firm with a less than stellar reputation should strive to improve its record by quoting a conservative lead-time, that is, one that can be easily met. On the other hand, a conservative lead-time is a longer one, which may drive away the current customer.

What role is played by the impact of tardiness on the firm's reputation? If the customers have a long memory, that is, the effect of *previous* tardiness is relatively important (smaller α), is that an argument for longer lead-times that will not make that history worse? Or should that argument apply to cases where the tardiness that may be incurred by the *present* order is more important (larger α)? Does the situation change depending on the level of the firm's reputation?

Customer and market characteristics also inform the lead-time decision. It seems obvious that impatient customers (higher ξ) demand shorter lead-times, but what about sensitivity to reputation (γ)? How should the firm balance the need to retain

the current customer's order by promising a relatively quick delivery, with the desire to preserve its reputation which might mean a longer lead-time promise (with higher risk of balking)?

The combination of present and future factors, as formulated in (7), reflects the complicated situation of a firm that must consider these tradeoffs in order to maximise its expected profit over time. However, the very complexity of this model precludes closed-form solutions or monotonicity results. The next subsection presents a counterexample that confirms the intractability of the model in this regard.

4.2 *Counterexample to monotonicity of state and decision variables*

It turns out that lead-time is not always an increasing function of order size in this model. Consider an example where the maximum order size, backlog, tardiness index and lead-time all equal 20, $w^r = 2$, $w^p = 1$, $\alpha = 0.5$, $\xi = 0.05$ and $\gamma = 0.5$. It is sufficient to show that there is a counterexample in the one-period problem. Looking at the expected immediate reward (4) in the state where $a = 11$, $B = 0$ and $T = 0$, the optimal lead-time for this order is $L = 9$. The optimal lead-time for the previous state where $a = 10$ was $L = 10$, so if lead-time were a non-decreasing function of order size, the optimal L would be greater than or equal to 10.

The numerical details of this example are as follows:
When $L = 9$, the expected immediate reward (4) is

$$e^{(-0.05)(9)}e^{(-0.05)(0)}\left[2(11) - 1(0 + 11 - 9)^+\right] = (0.637628151)(1)[20] = 12.75256303$$

When $L = 10$, the expected immediate reward (4) is

$$e^{(-0.05)(10)}e^{(-0.05)(0)}\left[2(11) - 1(0 + 11 - 10)^+\right] = (0.606530659)(1)[21] = 12.73714385$$

Since the decision variable is not monotonic in the state variables, this model lacks the property of supermodularity. However, the computational study described below discovers and exploits *near*-monotone properties of the model, and provides insights into how managers should balance the trade-offs among cost, revenue and reputation.

5. Computational study

The computational study consists of two parts. Section 5.1 describes the implementation of the MDP, the design of the study and the analysis of relationships among the decision variable and three state variables. Section 5.2 details the development of three heuristics, and the computational study that demonstrates their performance.

5.1 *Investigation of relationships among state and decision variables*

The first part of the computational study examines the relationship between the decision variable L, the three state variables a, B and T and the model parameters α, γ and ξ. It seems reasonable that lead-time should be increasing in order size, since it would take longer to process a longer order; however, the counterexample in Section 4.2 shows that monotonicity is not a general property. Similarly, a larger backlog B should result in a longer lead-time, since more processing to be done before the present job, all else equal, will result in that job finishing later. The intuition behind the relationship between lead-time

Table 1. Parameter settings for the computational study.

Parameter	Test Number								
	1	2	3	4	5	6	7	8	9
α	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
γ	0.5	0.5	0.5	0.3	0.3	0.3	0.8	0.8	0.8
ξ	0.05	0.5	0.005	0.05	0.5	0.005	0.05	0.5	0.005
	10	11	12	13	14	15	16	17	18
α	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
γ	0.5	0.5	0.5	0.3	0.3	0.3	0.8	0.8	0.8
ξ	0.05	0.5	0.005	0.05	0.5	0.005	0.05	0.5	0.005
	19	20	21	22	23	24	25	26	27
α	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
γ	0.5	0.5	0.5	0.3	0.3	0.3	0.8	0.8	0.8
ξ	0.05	0.5	0.005	0.05	0.5	0.005	0.05	0.5	0.005

and reputation (tardiness index) is more complicated: a firm might quote a shorter lead-time if T is larger, in an attempt to keep the current customer by reducing the probability of balking because of a long lead-time; or it might quote a longer lead-time in order to improve its reputation for on-time delivery in the future. As discussed in Section 3, these parameters correspond to customer and market characteristics with which a firm would be familiar.

This part of the computational study was designed to examine these three relationships, in a variety of situations. Three values, each of the parameters α, γ and ξ were chosen after pilot studies revealed the range of values that generated a range of lead-times over all states, rather than just boundary values. A full-factorial design resulted in 27 tests (see Table 1). For all tests, the per-unit reward w^r is two and the per-unit tardiness penalty w^p is one. Order size takes values from 1 to 20, backlog takes values from 1 to 20, and the tardiness index takes values from 0 to 200. This results in $20 \times 21 \times 201 = 84,420$ states per test, or 2,279,340 states for all 27 tests. Lead-times were bounded at 20. The discount factor β was set at 0.95, and each test was run for 500 periods (corresponding to the total number of stages or arriving customers). To generate optimal values, the MDP was coded as a dynamic program in Fortran 95 and run on a Macintosh with 3.06 GHz Intel Core 2 Duo processor under the OS X operating system (10.6.8). Convergence of the value function was verified by checking that $\|f_{n+1} - f_n\| \leqslant [\|f_1 - f_0\|/(1 - \beta)]\beta^n$, where $\|x\|$ is the largest absolute value in the vector x, β is the discount factor and $n + 1$ is the number of iterations (see Heyman and Sobel 1984, p. 160, Theorem 4.8). Here the largest value of the bound $[\|f_1 - f_0\|/(1 - \beta)]\beta^n$ is $2.1038e^{-08}$.

The MDP was implemented as follows:

Procedure MDP

For each order size;
For each backlog level;
For each level of tardiness index;
For each possible lead-time quotation;
{ Calculate the expected profit
If the current profit is the highest, save this lead-time quotation and profit }

Please see the Appendix for more details about this implementation.

Regression analysis was used to analyse the relationship between the dependent variable L, six independent variables and three interaction effects involving the tardiness index (INT1, INT2 and INT3 are the product of the independent variables T and α, γ and ξ, respectively). The regression model is:

$$L = \beta_0 + \beta_1 a + \beta_2 B + \beta_3 T + \beta_4 \alpha + \beta_5 \gamma + \beta_6 \xi + \beta_7 INT1 + \beta_8 INT2 + \beta_9 INT3 + \epsilon$$

The regression was run using the "Fit Model" procedure of JMP (version 9.0.2). The combination of relatively high R^2 (0.74) and low p-values suggest that the model is a good fit. All variance inflation factors equal one, so multicollinearity is not a problem here (see Table 2).

These results provide insights into the effects of the state variables, the model parameters and their interactions, on lead-time quotation. The main effects are straightforward and intuitive. First, lead-time quotations should be *longer* when order size and shop backlog are longer, that is, the lead-times reflect the work to be done on this order as well as the orders already

Table 2. Results of regression analysis.

Variable	β	t-value	p-value
Intercept	12.556811	1062.5	< 0.0001
Order size (a)	0.1139498	273.36	< 0.0001
Backlog (B)	0.540971	1362.8	< 0.0001
Tardiness index (T)	−0.014589	−352.2	< 0.0001
α	0.5995489	51.25	< 0.0001
γ	−1.463842	−125.1	< 0.0001
ξ	−22.46366	−2089	< 0.0001
INT1	0.0032045	16.14	< 0.0001
INT2	0.0058725	29.57	< 0.0001
INT3	−0.047702	−261.2	< 0.0001

Table 3. Performance of heuristics (all numbers are percentages).

Measure	ODP	BDP	TDP
Average deviation (heuristic value/optimal value)	4.04	0.16	0.28
Worst case deviation	27.32	25.65	10.09
Average worst case deviation	11.49	2.01	1.37
Average percent time (heuristic time/optimal time)	34	43	66

Table 4. Comparison of MDP and ODP (running time in seconds).

Test	1	2	3	4	5	6	7	8	9
Avg deviation (%)	8.18	3.07	0.34	10.06	2.00	0.35	5.79	5.36	0.38
Max deviation (%)	18.93	12.39	1.15	19.51	16.50	1.02	18.75	14.65	1.71
Optimal time	934.23	930.26	931.47	938.19	933.40	935.07	938.30	937.55	944.83
Heuristic Time	223.13	511.86	215.75	218.79	545.14	210.77	234.06	485.94	226.52
% Optimal time	23.88	55.02	23.16	23.32	58.40	22.54	24.95	51.83	23.97

Test	10	11	12	13	14	15	16	17	18
Avg deviation (%)	4.77	4.17	0.21	6.44	3.95	0.32	3.48	5.14	0.17
Max deviation (%)	19.21	12.75	1.14	17.18	27.32	1.07	18.68	13.40	1.30
Optimal time	944.27	938.17	942.94	937.14	936.40	937.49	939.21	935.07	935.47
Heuristic time	235.03	491.55	220.78	221.57	510.70	214.31	246.46	483.50	224.71
% Optimal time	24.89	52.39	23.41	23.64	54.54	22.86	26.24	51.71	24.02

Test	19	20	21	22	23	24	25	26	27
Avg deviation (%)	11.338	2.697	0.669	12.368	0.622	0.294	10.371	6.012	0.507
Max deviation (%)	18.518	10.791	1.660	19.347	3.702	0.933	18.105	18.874	1.635
Optimal time	935.88	933.59	936.25	935.06	936.29	933.94	932.03	931.75	932.00
Heuristic time	223.30	549.42	226.44	229.10	587.81	203.34	226.31	500.11	231.27
% Optimal time	23.86	58.85	24.19	24.50	62.78	21.77	24.28	53.67	24.81

Table 5. Comparison of MDP and BDP (running time in seconds).

Test	1	2	3	4	5	6	7	8	9
Avg deviation (%)	0.0023	0.0000000004	0.0117	0.0002	0.00001	0.0001	0.0102	0.0000000002	0.57
Max deviation (%)	0.3356	0.0000354900	0.2438	0.0162	0.00542	0.0049	0.6483	0.0000155089	11.15
Optimal time	934.23	930.26	931.47	938.19	933.40	935.07	938.30	937.55	944.83
Heuristic time	245.33	725.95	225.95	230.04	702.34	222.14	281.69	746.29	228.73
% Optimal time	26.26	78.04	24.26	24.52	75.25	23.76	30.02	79.60	24.21

Test	10	11	12	13	14	15	16	17	18
Avg deviation (%)	0.0002	0.000003	0.00002	0.0001	0.00001	0.0002	0.0008	0.000	0.0078
Max deviation (%)	0.1563	0.003592	0.00230	0.0185	0.00532	0.0094	0.3239	0.000	0.5172
Optimal time	944.27	938.17	942.94	937.14	936.40	937.49	939.21	935.07	935.47
Heuristic time	282.35	735.06	227.60	244.41	717.09	227.72	307.38	745.49	227.24
% Optimal time	29.90	78.35	24.14	26.08	76.58	24.29	32.73	79.73	24.29

Test	19	20	21	22	23	24	25	26	27
Avg deviation (%)	0.02	0.00	1.47	0.00001	0.00	0.0003	0.14	0.00	2.18
Max deviation (%)	0.67	0.00	10.65	0.00398	0.00	0.0487	3.70	0.00	25.65
Optimal time	935.88	933.59	936.25	935.06	936.29	933.94	932.03	931.75	932.00
Heuristic time	228.54	725.01	229.62	236.43	696.91	217.58	231.48	743	232.02
% Optimal time	24.42	77.66	24.53	25.29	74.43	23.30	24.84	79.70	24.89

in queue. Second, a higher tardiness index T (i.e. worse reputation for tardiness) means that lead-times should be *shorter*; if customers are more likely to balk because of $e^{-\gamma T}$, where T is relatively high, then try to raise the probability that they will stay by lowering L in the second factor $e^{-\xi L}$. Third, more weight on the current tardiness value for reputation (higher

Table 6. Comparison of MDP and TDP (running time in seconds).

Test	1	2	3	4	5	6	7	8	9
Avg deviation (%)	0.0132284	0.015	0.456	0.000	0.005	0.000	0.053	0.008	3.301
Max deviation (%)	0.0000002	0.422	1.910	0.000	0.246	0.000	0.739	1.235	10.085
Optimal time	934.23	930.26	931.47	938.19	933.40	935.07	938.30	937.55	944.83
Heuristic time	779.44	289.39	803.14	803.26	316.13	809.67	739.38	269.82	771.99
% Optimal time	83.43	31.11	86.22	85.62	33.87	86.59	78.80	28.78	81.71

Test	10	11	12	13	14	15	16	17	18
Avg deviation (%)	0.0000003	0.010	0.000	0.001	0.007	0.000	0.0000001	0.011	0.790
Max deviation (%)	0.0016113	0.225	0.000	0.048	0.257	0.000	0.0006688	1.389	3.603
Optimal time	944.27	938.17	942.94	937.14	936.40	937.49	939.21	935.07	935.47
Heuristic time	753.55	286.69	806.75	792.17	304.97	807.88	724.88	267.55	798.06
% Optimal time	79.80	30.56	85.56	84.53	32.57	86.17	77.18	28.61	85.31

Test	19	20	21	22	23	24	25	26	27
Avg deviation (%)	0.0174	0.0008	0.9081	0.0124	0.0002	0.0300	0.1423	0.0241	1.7797
Max deviation (%)	0.3786	0.0471	4.7704	0.1907	0.0426	0.1699	1.6035	1.2195	8.4146
Optimal time	935.88	933.59	936.25	935.06	936.29	933.94	932.03	931.75	932.00
Heuristic time	798.80	290.76	774.45	790.71	323.21	809.94	772.30	270	774.96
% Optimal time	85.35	31.14	82.72	84.56	34.52	86.72	82.86	28.96	83.15

α) suggests longer lead-times; the firm seeks to avoid tardiness costs now, and also to keep its reputation from degrading in the future. Finally, both sensitivity parameters (ξ and γ) are negatively correlated with lead-time, so the firm should attempt to keep the current customer by quoting a lower lead-time when the customer is more sensitive to either lead-time quotation or reputation.

The interaction effects suggest that the lead-time decision in the light of the firm's reputation can be moderated by market characteristics. Given a certain reputation (tardiness index T), a stronger recency effect on reputation (higher α) or customers more sensitive to reputation (higher γ) results in relatively longer lead-times (attempt to improve reputation in the future). On the other hand, a market with relatively more impatient customers (higher ξ) suggests a shorter lead-time, in order to keep those customers.

The regression analysis serves two purposes. First, it confirms that the model has predictive value. In a practical sense, a firm could improve its delivery performance by running the regression model and using the results to inform its lead-time decisions, or simply by basing its lead-time quotations on the relative magnitude of the independent variables. Second, these results suggest heuristics, in particular, a further investigation of the effects of the state variables. This will be discussed in the next section.

5.2 Near-monotonic heuristics

As discussed in Section 4.2, counterexamples can be found to the preposed monotonicity of one state variable (order size a), and so the MDP is not supermodular. Nevertheless, the regression model suggests three heuristics, based on near-monotone properties of order size, backlog and tardiness index. That is, the optimal lead-time L is *generally* increasing in a and B and decreasing in T. The three heuristics were implemented as follows:

Procedure ODP
For each backlog level;
For each level of tardiness index;
For each order size;
For each possible lead-time quotation, *where possible lead-times are bounded below*
by the best lead-time value for the previous order size at this backlog and tardiness level;
{ Calculate the expected profit
If the current profit is the highest, save this lead-time quotation and profit }

```
    Procedure BDP
For each level of tardiness index;
For each order size;
For each backlog level;
For each possible lead-time quotation, where possible lead-times are bounded below
by the best lead-time value for the previous backlog at this order size and tardiness level;
{ Calculate the expected profit
If the current profit is the highest, save this lead-time quotation and profit }
```

```
    Procedure TDP
For each order size;
For each backlog level;
For each level of tardiness index;
For each possible lead-time quotation, where possible lead-times are bounded above
by the best lead-time value for the previous tardiness index at this order size and backlog ;
{ Calculate the expected profit
If the current profit is the highest, save this lead-time quotation and profit }
```
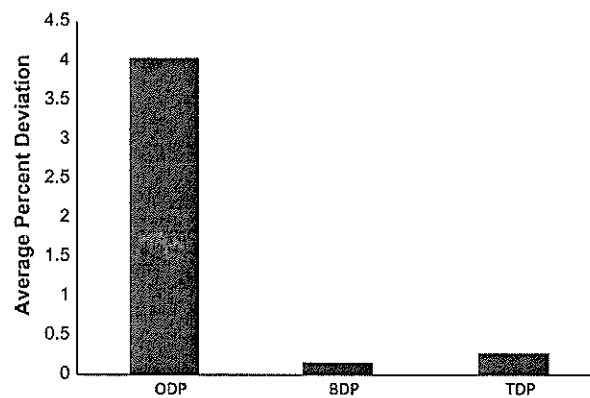

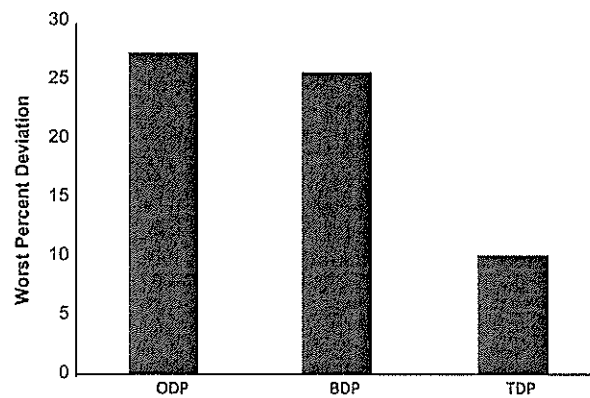Figure 1. Average percent deviation of three heuristics.

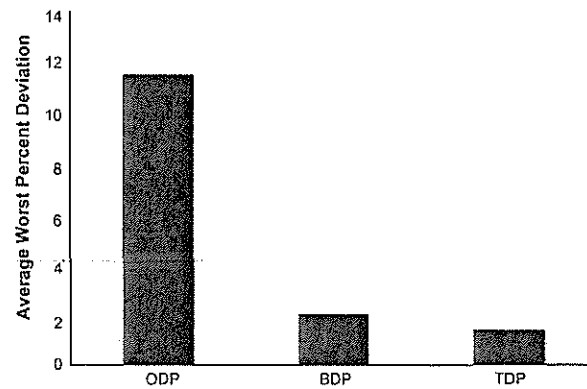Figure 2. Worst percent deviation of three heuristics.

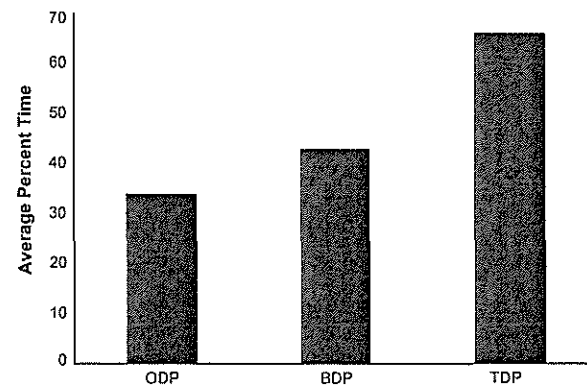Figure 3. Average worst percent deviation of three heuristics.

Figure 4. Average percent time of three heuristics.

Please see the Appendix for more details about these implementations.

For the second part of the computational study, all three heuristics were run on the original full-factorial data, and the results were compared to those of the optimal procedure. See Table 3 for a summary, and Tables 4–6 for more details. Note that the average performance of all three heuristics in terms of deviation from the optimal value was always less than 5%, for BDP and TDP, it was less than 1%, and that all three ran in considerably less time than the MDP. The best-performing heuristic was BDP, with overall average deviation of 0.16% and average maximum deviation of 2.01%. Next best was TDP with overall average deviation of 0.28% and average maximum deviation of 1.37%. ODP had an overall average deviation of 4.04% and average maximum deviation of 11.5%. All three heuristics were faster than the optimal procedure. ODP was consistently the fastest, with the average percent of running time 34% of the optimal running time. BDP came second at 43% and TDP was the slowest at 66%. See Figures 1–4 for a comparison of the three heuristics on four performance measures.

6. Conclusions

This research extends the study of lead-time models to include the consequences of the firm's past actions, that is, its reputation for on-time delivery. Consistent with previous work, the present study suggests that when the completion time of an order will be relatively later, because of shop backlog and/or the processing time associated with this order, then lead-time quotations should also be longer. This conclusion is intuitive but not always what is done in practice (e.g. the steel mill reported in Slotnick 2011 was using a "standard lead-time" for all orders). If the firm has a relatively worse reputation for late delivery, it should compensate by quoting a shorter lead-time, all else equal, so that the customer is more likely to stay.

Market conditions also have a bearing on lead-time decisions. If the present action (lead-time decision) has a relatively strong influence on reputation, the quotation should be longer, in order to sustain the reputation in the future as well as

incur lower tardiness costs in the present. When customers are relatively more sensitive to reputation or to lead-times, the present lead-time should be shorter, all else equal, to stop customers from leaving. However, at a given level of reputation, the lead-time should be longer if the impact of the tardiness of *this* order on that reputation is stronger. Similarly, at a given level of reputation, the lead-time should be longer if the market is more sensitive to reputation. In these two situations, the firm needs to protect its reputation for the future by lowering tardiness in the present. So, managers should consider market characteristics, as well as the firm's past history for on-time delivery, when making lead-time decisions.

The regression results also suggest that the model possesses near-monotone properties, that is, the decision variable is generally increasing or decreasing in each of the state variables. Three heuristics exploit this property, and give very good performance with considerable savings in computational expense. The idea of using near-monotonicity to provide faster results with minimal degradation of performance (profit) (Slotnick 2011) is extended here to all of the state variables.

The results described here have practical implications for managers who decide on delivery-time promises for their products and services. First, both the regression model and heuristics strongly suggest which internal and external factors should be taken into consideration when making these decisions. Second, the regression model might serve as a tool for generating lead-time quotations. For input to the model, it is reasonable to assume that a manager would have a good idea of the status of processing in the facility and the time to process the current order, and could characterise the effects of reputation and the market. Finally, the three heuristics provide answers that are close to optimal, and are fast enough to be used for "what-if" and sensitivity analysis.

References

Adelman, D., and A. Mersereau. 2013. "Dynamic Capacity Allocation to Customers Who Remember Past Service." *Management Science* 59 (3): 592–612.

Aflaki, S., and I. Popescu. 2011. *Managing Retention in Service Relationships*. Insead Working Paper No. 2011/80/DS, Boulevard de Constance 77305 Fontainebleau.

Annett, T. February 23, 2007a. "The Afternoon Report: Nothing but Static: Online Edition." *Wall Street Journal*, Online edition.

Annett, T. January 31, 2007b. "The Evening Wrap: Price Patrol." *Wall Street Journal*, Online edition.

Anonymous. 2000. "Flight Cancellations Decrease at United." *New York Times*, August 10, p. 6.

Ata, B., and T. L. Olsen. 2009. "Near-Optimal Dynamic Lead-time Quotation and Scheduling Under Convex-Concave Customer Delay Costs." *Operations Research* 57 (3): 753–768.

Beck, M. October 19, 2010. "The Doctor will See You Eventually." *Wall Street Journal*, p. D1.

Chatterjee, S., S. A. Slotnick, and M. J. Sobel. 2002. "Delivery Guarantees and the Interdependence of Marketing and Operations." *Production and Operations Management* 11 (3): 393–410.

Ernst, R., and S. Powell. 1995. "Optimal Inventory Policies Under Service-Sensitive Demand." *European Journal of Operational Research* 87 : 316–327.

Gans, N. 2002. "Customer Loyalty and Supplier Quality Competition." *Management Science* 48 (2): 207–221.

Gaur, V., and Y. H. Park. 2007. "Asymmetric Consumer Learning and Inventory Competition." *Management Science* 53 (2): 227–240.

Gjerde, K. P., and S. Slotnick. 1997. "A Multidimensional Approach to Manufacturing Quality." *Computers and Industrial Engineering* 32 (4): 879–890.

Gjerde, K. P., and S. Slotnick. 2004. "Quality and Reputation: The Effects of External and Internal Factors Over Time." *International Journal of Production Economics* 89 (1): 1–20.

Hall, J., and E. Porteus. 2000. "Customer Service Competition in Capacitated Systems." *Manufacturing and Service Operations Management* 2 (2): 144–165.

Herbig, P., and J. Milewicz. 1995. "The Relationship of Reputation: And Credibility to Brand Success." *Journal of Consumer Marketing* 12 (4): 5–10.

Heyman, D. P., and M. J. Sobel. 1984. *Stochastic Models in Operations Research, Volume II: Stochastic Optimization*. New York: McGraw-Hill.

Hill, C. J., and K. Joonas. 2005. "The Impact of Unacceptable Wait Time on Health Care Patients Attitudes and Actions." *Health Marketing Quarterly* 23 (2): 69–87.

Hua, G., S. Wang, and T. Cheng. 2010. "Price and Lead Time Decisions in Dual-channel Supply Chains." *European Journal of Operational Research* 205 : 113–126.

Kaminsky, P., and Z. H. Lee. 2008. "Effective On-line Algorithms for Reliable Due Date Quotation and Large-Scale Scheduling." *Journal of Scheduling* 11 (3): 187–205.

Kapuscinski, R., and S. Tayur. 2007. "Reliable Due-date Setting in a Capacitated MTO System with Two Customer Classes." *Operations Research* 55 (1): 56–74.

Keaveney, S. 1995. "Customer Switching Behavior in Service Industries: An Exploratory Study." *Journal of Marketing* 59 : 71–82.

Keskinocak, P., R. Ravi, and S. Tayur. 2001. "Scheduling and Reliable Lead-Time Quotation for Orders with Availability Intervals and Lead-time Sensitive Revenues." *Management Science* 47 : 264–279.

Keskinocak, P., and S. Tayur. 2004. "Due Date Management Policies." In *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era*, edited by D. Simchi-Levi, S. Wu, and Z. J. Shen, 485–556. Boston, MA: Kluwer.

Lawton, C. June 4, 2007. "Hard Drive: How H-P Reclaimed Its PC Lead Over Dell." *Wall Street Journal*, p. A.1.

Lunsford, J. L., April 10, 2008a. "Boeing Delays Dreamliner Delivery Again Six more Months Needed Despite Making Progress on Jet 'Start-up' Issues'." *Wall Street Journal*, p. B3.

Lunsford, J. L. October 23, 2008b. "Boeing Strikers Dig in Heels Even as Economy Turns Sour." *Wall Street Journal*, p. A.1.

Lunsford, J. L. September 17, 2007. "Boeing's Tall Order: On-Time 787." *Wall Street Journal*, p. A8.

McCartney, S. December 6, 2005. "The Middle Seat: Airline Passenger Complaints Rise Sharply; Total Number Jumps 29%, With Flight Cancellations And Delays Topping List." *Wall Street Journal*, p. D1.

Michaels, D. July 3, 2006. "Airbus Problems Lead to Ouster of Key Executives." *Wall Street Journal*, p. A.1.

Pekgun, P., P. Griffin, and P. Keskinocak. 2008. "Coordination of Marketing and Production for Price and Leadtime Decisions." *IIE Transactions* 40 : 12–30.

Plambeck, E. L. 2004. "Optimal Leadtime Differentiation via Diffusion Approximations." *Operations Research* 52 (2): 213–228.

Sanders, P. August 27, 2010b. "Boeing has New Delay for Dreamliner." *Wall Street Journal*, p. B.6.

Sanders, P. January 19, 2011. "Corporate News: Boeing Again Delays 787 Delivery." *Wall Street Journal*, p. B3.

Sanders, P. November 26, 2010a. "Boeing 787 is Set Back as Blaze Forces Fix." *Wall Street Journal*, p. A1.

Sanders, P. October 7, 2009. "Boeing Settles in for a Bumpy Ride." *Wall Street Journal*, p. B.1.

Schwartz, B. 1966. "A New Approach to Stockout Penalties." *Management Science* 12 (12): B538–B544.

Shapiro, C. 1983. "Premiums for High Quality Products as Returns to Reputations." *The Quarterly Journal of Economics* 98 (4): 659–679.

Slotnick, S. A. 2011. "Optimal and Heuristic Lead-time Quotation for an Integrated Steel Mill with a Minimum Batch Size." *European Journal of Operational Research* 210 : 527–536.

Slotnick, S. A., and M. J. Sobel. 2005. "Manufacturing Lead-time Rules: Customer Retention vs. Tardiness Costs." *European Journal of Operational Research* 163 (3): 825–856.

Sobel, M. 1973. "Queuing Processes at Competing Service Facilities." *Management Science* 19 (9): 985–1000.

Upasani, A., and R. Uzsoy. 2008. "Incorporating Manufacturing Lead Times in Joint Production-marketing Models: A Review and Some Future Directions." *Annals of Operations Research* 161 : 171–188.

Zorzini, M., D. Corti, and A. Pozzetti. 2008. "Due Date (DD) Quotation and Capacity Planning in Make-to-order Companies: Results from an Empirical Analysis." *International Journal of Production Economics* 112 (2): 919–933.

Appendix 1. Details of optimal and heuristic procedures

Procedure MDP

Initialise counters, parameters (α, γ, ξ), arrays
Begin timing

For each stage
For each state (order size, backlog level, tardiness index)
For each possible lead-time quotation:

Calculate the expected immediate reward:
Calculate the probability of not balking PNB: $e^{-(\xi L + \gamma T)}$
Multiply order size a by per-unit reward w^r
Calculate the probability of tardiness $z(a, B, L)$
Multiply expected tardiness by per-unit penalty w^p
Subtract expected tardiness cost to get net revenue
Multiply by PNB to get expected value

Calculate the expected future reward if customer stays:
Update backlog B: add processing time a, subtract interarrival time τ as in (3)
Update tardiness index T with exponential smoothing formula as in (2)
Calculate expected future reward (weighted sum) as described below.

Calculate the expected future reward if customer balks:
Calculate probability of balking (1-PNB)
Update backlog B by subtracting interarrival time τ
Calculate expected future reward (weighted sum) as described below.

Calculate the expected reward for this state:
expected immediate reward plus discounted expected future reward

If the expected reward for this lead-time is the best found so far for this state,
save this expected reward and lead-time

Continue through all states and lead-time values
Print the optimal lead-time and expected reward for each state
End timing

The expected future reward is the weighted sum of the value of each possible future state, times the probability of being in that state. The procedure loops through all possible values of processing time for the incoming order a' and interarrival times τ:

$$\sum_{j}^{B+j-i\leq U}\sum_{i}^{B+j-i\leq U} P\{a'=j\}P\{\tau=i\}f_{k+1}(j, B+j-i, \tau')$$
$$+ \sum_{j}^{B+j-i>U}\sum_{i}^{B+j-i>U} P\{a'=j\}P\{\tau=i\}f_{k+1}(0, B-i, T)$$

The first expression enumerates the states where the updated backlog is less than a maximum (the capacity of the queue), and the second expression enumerates the states where the updated backlog would exceed that capacity, and so the order is rejected (note 0 instead of a' as the first state variable in the second expression). The probabilities $P\{a'=j\}$ and $P\{\tau=i\}$ are calculated as a geometric distribution with an empirically determined parameter (0.145). The geometric distribution is the unique non-negative discrete random variable that has the memorylessness property of the exponentially distributed interarrival times of the Poisson arrival process. The maximum order size was set at 20, and the maximum interarrival time was set at 5.

$$P\{a'=j\} = \frac{(0.145)(1-0.145)^j}{\sum_{m}^{20}(0.145)(1-0.145)^m}$$

$$P\{\tau=i\} = \frac{(0.145)(1-0.145)^i}{\sum_{m}^{5}(0.145)(1-0.145)^m}$$

The three heuristic procedures are the same as **MDP** *except* that the lead-times considered are restricted, according to the near-monotone properties described in Section 5.2. In particular:

Procedure ODP: instead of 'For each possible lead-time quotation', only consider lead-times for this order size that are at least as large as the best lead-time value for the previous (smaller) order size, with backlog and tardiness index the same. The idea is that lead-time increases with order size. In other words, the larger the processing time of this order, the longer it will take to get through the shop, and so the lead-time quotation should be relatively longer.

Procedure BDP: instead of 'For each possible lead-time quotation', only consider lead-times for this backlog that are at least as large as the best lead-time value for the previous (smaller) backlog, with order size and tardiness index the same. The idea is that lead-time increases with backlog. In other words, the larger the backlog on the shop floor, the longer this order will take to get through the shop, and so the lead-time quotation should be relatively longer.

Procedure TDP: instead of 'For each possible lead-time quotation', only consider lead-times for this tardiness index that are no larger than the best lead-time value for the previous (smaller) tardiness index, with order size and backlog the same. The idea is that lead-time decreases with tardiness index. In other words, the worse a firm's reputation for tardiness, the more it needs to try to keep customer orders by quoting a shorter (more attractive) lead-time.