How to Determine the Order-up-to Level When Demand is Gamma Distributed with Unknown Parameters

Janssen, E.; Strijbosch, L.W.G.; Brekelmans, R.C.M.

Publication date:
2007

CentER

Discussion Paper

No. 2007–71

HOW TO DETERMINE THE ORDER-UP-TO LEVEL WHEN DEMAND IS GAMMA DISTRIBUTED WITH UNKNOWN PARAMETERS

By E. Janssen, L.W.G. Strijbosch, R.C.M. Brekelmans

September 2007

ISSN 0924-7815

How to determine the order-up-to level when demand is gamma distributed with unknown parameters

E. Janssen, L. W. G. Strijbosch, R. C. M. Brekelmans
Department of Econometrics and OR, Tilburg University, Warandelaan 2, 5037AB Tilburg, the Netherlands.
{E.Janssen_1@UvT.nl, L.W.G.Strijbosch@UvT.nl, R.C.M.Brekelmans@UvT.nl}

Inventory models need information about the demand distribution. In practice, this information is not known with certainty and has to be estimated with often relatively few historical demand observations. Using these estimates leads to underperformance. This paper focuses on gamma distributed demand and a periodic review, order-up-to inventory control policy, where the order-up-to level satisfies a service equation. Under this policy the underperformance is quantified analytically under strong assumptions and with help of simulation if these assumptions are relaxed. The analytical results can be used to improve the attained service level, such that it approaches the desired service level more closely, even if the assumptions are not met. With help of simulation we show that in some cases this improvement results in reaching the desired service level. For the remaining cases, i.e., the cases in which the desired service level is not reached, the underperformance decreases; improvements range from almost 17% up to over 90%. Moreover, with help of simulation and linear regression further improvements can be obtained. The desired service level is reached in more cases and the underperformance in the other cases is decreased even more compared to using only the first improvement. These improvements range from 57% up to 99% compared to the base case (i.e., do not use analytical results) and from 35% up to over 90% compared to using the analytical results, except for a few cases in which the service hardly improved, but in those cases the attained service level was already very close to the desired one. Finally, the method developed in this paper is applied to real demand data using simulation. The total improvements in this case study range from 53% up to 96%.

Keywords: Unknown Demand Parameters, Inventory Control, Gamma Distribution, Service Level Criterion, Case Study

JEL-classification: C13, C53.

1. Introduction

Inventory control models need information about the demand distribution. These models are developed assuming that all the information they need (e.g., moments, family of distribution, parameters) are known with certainty. However, in practice often relatively few historical demand observations are known and these are used to estimate the demand distribution characterization that is needed. Although this problem is already known for a long time, see, e.g., Scarf (1958) and Hayes (1969), most literature still considers either forecasting or inventory control and not the effect of forecasting on inventory control. Recent papers in which this effect is studied, include e.g. Silver and Rahnama (1986, 1987), Watson (1987), Strijbosch and Heuts (1992), Snyder et al. (2002), Bertsimas and Thiele (2006), Lu et al. (2006), Syntetos and Boylan (2006), and Janssen et al. (2006); the last mentioned study provides a short overview of literature on how to deal with uncertainty about the demand distribution. Strijbosch and Heuts (1992) use simulation to show the trade-off between attained service and expected average costs while estimating the lead time demand in four different ways, including a distribution-free approach. Snyder et al. (2002) consider both forecasting and inventory control and use simulation to show the effects of using their forecasting model, which incorporates the possibility of having nonconstant variance. Syntetos and Boylan (2006) show the effect of using four different estimators under a periodic review, order-up-to inventory control policy with help of simulation. The other studies will be considered in the remainder of this section.

In this paper the family of distributions to which the demand belongs, is assumed to be known, but its parameters are not. Hence, estimates are needed to use the inventory model and the effect of using these estimates is studied. Scarf (1958), Bertsimas and Thiele (2006) and Lu et al. (2006) also consider demand uncertainty explicitly, but differently from the approach here. All three papers consider a cost criterion and furthermore Scarf (1958) and Bertsimas and Thiele (2006) assume that the mean and variance of the demand are known, while the family to which it belongs, is not. Lu et al. (2006) focus on the way the demand forecasts evolve as more information becomes available over time and use that to find solution bounds and cost error bounds for general dynamic inventory models with possibly nonstationary and autocorrelated demands.

Silver and Rahnama (1986, 1987) and Janssen et al. (2006) studied the effects of using estimates with normally distributed demand; the former have considered a cost criterion while the latter considered service criteria. Although the normal distribution is commonly assumed in inventory control, it certainly has some problems. According to Burgin (1975) demand distributions generally only exist for nonnegative values of demand and the shape of the density function changes from monotonic decreasing (low mean demand) via a unimodal distribution that is skewed to the right to a normal type distribution that is truncated at zero (high mean demand). A normal distribution does not fit all these criteria: the probability that a normally distributed variable can be negative, is nonnegligible (more than 1%) if the coefficient of variation (standard deviation divided by mean) is larger than 0.43; further, the normal distribution is symmetric. The gamma distribution does fit the criteria of Burgin (1975), since it is nonnegative and the value of the shape parameter can be adjusted to get all three forms described. The gamma distribution also has some nice properties which makes it relatively easy to work with, although maybe not as easy as the normal distribution and that is probably why that distribution is used so often in literature and in practice. But the gamma distribution has proven its worth. Watson (1987) considers an Erlang distribution (i.e., a special case of the gamma distribution) and studies the effect forecasting has on attaining the desired service level using simulation. Note that the Erlang assumption implies that demand during lead time has a relatively small coefficient of variation, hence demand cannot be highly variable, which limits the applicability of Watson (1987). Furthermore, he does restrict his research to intermittent demand. Segerstedt (1994) develops another inventory control policy, which also uses gamma distributions. Yeh (1997) slightly adapts this policy to implement it in an electronics industrial company. Both mention that parameters in the model need to be estimated, but they do not show the effects of doing this. The consultancy firm Involvation (www.involvation.com) uses the gamma distribution in their stock control software and its customers are satisfied with the achieved improvements. This consultancy firm generously provided demand data of one of their customers, the Dutch Ministry of Defence, to us which will be used to test the method developed in this paper.

In this paper demand is assumed to follow a gamma process, i.e., the demand during a period of length ℓ, denoted by X_ℓ, has a gamma distribution with shape parameter $\ell\rho$ and scale parameter $1/\lambda$, or $X_\ell \sim \Gamma(\ell\rho, \lambda)$ for short. If $\ell = 1$ demand will be denoted by X. Also, demands during disjoint time intervals are independent. Furthermore, an (R, S) inventory control policy is used with $R = 1$ (without loss of generality). This policy states that every R periods the inventory level is reviewed and replenished up to S. The order is then delivered after a fixed and deterministic lead time L. Demand is assumed to be stationary for $t + 1 + L$ periods, which means that the actual demand during the first t periods can be used to estimate the demand during the last $(L+1)$ periods. This method of forecasting is known as moving average and it is used in practice because either one wants to account for nonstationarity or one has only few observations. The order-up-to level S is chosen such that the required service level is reached. The two service level criteria considered in this paper are P_1 (cycle service) and P_2 (fill rate). The demand that cannot be satisfied immediately, is backlogged.

The remainder of the paper consists of three parts. Section 2 considers the P_1 service criterion. The theoretically correct order-up-to level is provided and the impact of using estimates for the unknown parameters is investigated. We prove that underperformance always occurs under the assumption that demand is exponentially distributed and that it occurs when the desired service level is over 50% in case of Erlang demand with a known shape parameter. Relaxing these assumptions leads to intractable results and at this point simulation is used to show that underperformance exists for values of the desired service level that are commonly used in practice. With help of simulation a correction to the order-up-to level is found such that the desired service level is reached and regression techniques are applied to estimate the relation between the correction needed and the parameters of the model. Using the regression equation to correct the order-up-to level results in reaching the desired service level more closely; improvements range from 75% up to over 99%. Section 3 has approximately the same layout, but focuses on the P_2 service criterion. It provides the theoretically correct order-up-to level under the assumption that the complete demand distribution is known and the effects of using estimates. We prove that the order-up-to levels under the P_1 and the P_2 criterion are equal in case of exponentially distributed demand and hence the results of the P_1 criterion can be adopted. For the remainder of this section simulation is used to show that underperformance exists if the desired service level is relatively high. Further, a correction to the estimated order-up-to level is provided and using this correction decreases the underperformance significantly; improvements range from 57% to almost 96%. Section 4 will use the order-up-to levels developed in Sections 2 and 3 on actual demand data. In case of the P_1 criterion the total improvement ranges from 63% to 96%; in case of the P_2 criterion it ranges from 53% to 95%. Section 5 concludes this paper with a short summary of the results and ideas for further research.

2. P_1 service level criterion

This section focuses on the P_1 service level criterion, which states that the fraction of replenishment cycles without backlogged demand should be at least α, or mathematically $P(X_{1+L} \le S_1) \ge \alpha$. Since keeping inventory as low as possible is desirable, equality will hold. Furthermore, we know that $X_{1+L} \sim \Gamma((1+L)\rho, \lambda)$. Define $\widetilde{\rho} := (1+L)\rho$. If the parameters of a gamma distribution are known, the order-up-to level is easily determined:

$$\alpha = F_{\widetilde{\rho},\lambda}(S_1) \qquad \Leftrightarrow \qquad S_1 = F^{-1}_{\widetilde{\rho},\lambda}(\alpha) \qquad \Leftrightarrow \qquad S_1 = \tfrac{1}{\lambda} F^{-1}_{\widetilde{\rho},1}(\alpha). \tag{1}$$

The function $F_{\rho,\lambda}$ ($F^{-1}_{\rho,\lambda}$) is the distribution function (inverse distribution function) of a gamma distribution with parameters ρ and λ. In the next section also $f_{\rho,\lambda}$ is used, which denotes the corresponding density function.

2.1. Using estimates for the determination of the order-up-to level

The order-up-to level determined in (1) is correct, given that the parameters ρ and λ are known. This is of course not true in practice. So this section will consider the effect of estimating the parameters. First, only λ is considered to be unknown, so ρ and thus the coefficient of variation ($\nu = \rho^{-1/2}$) are assumed to be known; the last part of this subsection considers the situation that ρ is unknown too. One possible estimator for λ is derived from the relation $E[X] = \rho/\lambda$, leading to $\hat{\lambda} = \rho/\overline{x} = (\nu^2\overline{x})^{-1}$, where $\overline{x} = \frac{1}{t}\sum_{i=1}^{t} x_i$ is the sample mean. The estimated order-up-to level in this case is then $\widehat{S}_1 = F^{-1}_{\widetilde{\rho},1}(\alpha)\nu^2\overline{x}$. Define $g_\alpha = F^{-1}_{\widetilde{\rho},1}(\alpha)\nu^2$; note that g_α consists of non-random terms and that $\overline{x} \sim \Gamma(t\rho, t\lambda)$. This results in $\widehat{S}_1 \sim \Gamma(t\rho, \frac{t}{g_\alpha}\lambda)$. Now let us consider the fraction of replenishment cycles with backlogged demand when using the order-up-to level $\widehat{S}_1$:

$$P\left(X_{1+L} > \widehat{S}_1\right) = P\left(\lambda X_{1+L} > \lambda \widehat{S}_1\right) = P\left(X^*_{1+L} > \widehat{S}^*_1\right).$$

Note that $X^*_{1+L} = \lambda X_{1+L} \sim \Gamma(\widetilde{\rho}, 1)$ and $\widehat{S}^*_1 = \lambda \widehat{S}_1 \sim \Gamma(t\rho, \frac{t}{g_\alpha})$, so λ does not play a role in the derivation. Moreover, assuming $\widetilde{\rho} \in \mathbb{N}$ and $t\rho \in \mathbb{N}$, $f_{\widetilde{\rho},\cdot}(\cdot)$ and $f_{t\rho,\cdot}(\cdot)$ are density functions of an Erlang distributed variable, which is used in the derivation below, after (2).

$$P\left(X^*_{1+L} > \widehat{S}^*_1\right) = \int_0^\infty P\left(X^*_{1+L} > s\right) f_{t\rho,\frac{t}{g_\alpha}}(s)ds = \int_0^\infty (1 - F_{\widetilde{\rho},1}(s)) f_{t\rho,\frac{t}{g_\alpha}}(s)ds \tag{2}$$

$$= \int_0^\infty \left(e^{-s}\sum_{i=0}^{\widetilde{\rho}-1}\frac{(s)^i}{i!}\right)\left(\frac{\left(s\frac{t}{g_\alpha}\right)^{t\rho-1} e^{-s\frac{t}{g_\alpha}}\frac{t}{g_\alpha}}{(t\rho-1)!}\right)ds$$

$$= \sum_{i=0}^{\widetilde{\rho}-1}\int_0^\infty e^{-s}e^{-s\frac{t}{g_\alpha}}\frac{s^i s^{t\rho-1}\left(\frac{t}{g_\alpha}\right)^{t\rho}}{i!(t\rho-1)!}ds$$

$$= \sum_{i=0}^{\widetilde{\rho}-1}\frac{(t\rho-1+i)!}{\left(\frac{t}{g_\alpha}\right)^i (t\rho-1)!i!}\int_0^\infty e^{-s} f_{t\rho+i,\frac{t}{g_\alpha}}(s)ds \tag{3}$$

$$= \sum_{i=0}^{\widetilde{\rho}-1}\frac{(t\rho-1+i)!}{\left(\frac{t}{g_\alpha}\right)^i (t\rho-1)!i!}\left(\frac{\frac{t}{g_\alpha}}{\frac{t}{g_\alpha}+1}\right)^{t\rho+i}$$

$$= \left(\frac{t}{t+g_\alpha}\right)^{t\rho}\sum_{i=0}^{\widetilde{\rho}-1}\binom{t\rho-1+i}{i}\left(\frac{g_\alpha}{t+g_\alpha}\right)^i \tag{4}$$

Table 1 Values of α for which the non-stockout probability using order-up-to level $\widehat{S}_1$ equals α (ρ known).

	$L=0$				$L=1$				$L=3$			
ρ	$t=2$	$t=10$	$t=20$	$t=50$	$t=2$	$t=10$	$t=20$	$t=50$	$t=2$	$t=10$	$t=20$	$t=50$
2	0.2499	0.2612	0.2627	0.2636	0.3288	0.3474	0.3500	0.3517	0.3688	0.3932	0.3971	0.3996
6	0.3729	0.3817	0.3828	0.3836	0.4055	0.4172	0.4189	0.4200	0.4253	0.4400	0.4423	0.4438
10	0.4038	0.4107	0.4116	0.4122	0.4274	0.4366	0.4380	0.4388	0.4423	0.4537	0.4556	0.4567
20	0.4329	0.4380	0.4387	0.4391	0.4489	0.4556	0.4565	0.4571	0.4593	0.4674	0.4687	0.4695
50	0.4579	0.4612	0.4616	0.4619	0.4678	0.4720	0.4727	0.4730	0.4743	0.4794	0.4803	0.4808

Note that the integral part of (3) is the Laplace-Stieltjes transform of an Erlang-$(t\rho+i)$ distribution. Now consider the special case of exponentially distributed demand ($\rho=1$) and zero lead time. Then $g_\alpha=-\ln(1-\alpha)$ and (4) simplifies to the following equation.

$$P\left(X^*_{1+L}>\widehat{S}^*_1\right)=\left(\frac{t}{t+g_\alpha}\right)^t=\left(\frac{t}{t-\ln(1-\alpha)}\right)^t$$

Now we prove that the desired service level is not reached, in other words that $(\frac{t}{t-\ln(1-\alpha)})^t>1-\alpha$. This can be rewritten as $(\frac{t+\ln((1-\alpha)^{-1})}{t})^t<(1-\alpha)^{-1}$; note that the left hand side is increasing in t. We finish our derivation as follows.

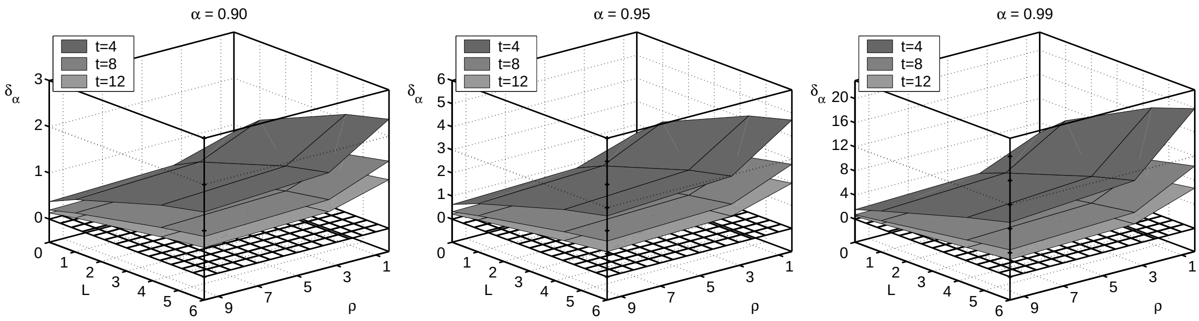
$$\left(1+\frac{\ln\left(\frac{1}{1-\alpha}\right)}{t}\right)^t<\lim_{t\to\infty}\left(1+\frac{\ln\left(\frac{1}{1-\alpha}\right)}{t}\right)^t=e^{\ln\left(\frac{1}{1-\alpha}\right)}=\frac{1}{1-\alpha}$$

So in case of exponentially distributed demand and zero lead time, the desired service level will not be attained and the underperformance is larger if t is smaller. If we would replace α by $\alpha'=1-\exp(t(1-(1-\alpha)^{-1/t}))$, where obviously $\alpha'>\alpha$, the desired service would be attained again. Unfortunately, we cannot find such analytical results if $L>0$ or if $\rho\in\mathbb{N}/\{1\}$.

Now let us assume that $\widetilde{\rho}\in\mathbb{N}$ and $t\rho\in\mathbb{N}$, with $\rho\geq 2$. In this case we can use (4) to investigate the attained service level numerically. These calculations, not shown here, indicate that at low desired service levels the attained service level exceeds the desired one, while this reverses at higher values of α. If α approximates either 0 or 1, the attained service level approximates the desired one. This is easily explained: if $\alpha=0$, $g_\alpha=0$ and thus $\widehat{S}_1=0$, so demand is not satisfied in any period. On the other hand, if $\alpha=1$, $g_\alpha\to\infty$ and thus $\widehat{S}_1\to\infty$. In that case we can always satisfy demand. Further, there is one $\alpha\in(0,1)$ such that the attained service equals the desired service level. These breakeven points are shown in Table 1 for several combinations of ρ, t and L. Table 1 clearly shows that the breakeven point gets higher if ρ is larger, if t is larger or if L is larger. It also appears to converge to some value if ρ increases. In that case the coefficient of variation gets smaller and the gamma distribution looks more and more like a normal distribution. For this distribution it can easily be shown that the breakeven point is at $\alpha=0.50$ (see Janssen et al. 2006).

Now let us consider $\widetilde{\rho}>0$ and $t\rho>0$, not necessarily integer. In this case it is no longer possible to derive a closed-form expression, but we can numerically evaluate the integral in (2). The results of these evaluations are shown in Figure 1 for three non-integer values of ρ ($\rho\in\{\frac{1}{3},\frac{44}{13},\frac{165}{17}\}$). This figure shows the relative deviation from the stockout probability, which is defined as $\delta_\alpha(\hat{\alpha}_0)=\frac{(1-\hat{\alpha}_0)-(1-\alpha)}{1-\alpha}$, where $\hat{\alpha}_0$ is the service level determined by evaluating (2). We have chosen for this measure, since we think that the *perceived* customer service is mainly determined by stock out occurrences, so we choose to

Figure 1 Relative deviation ($\delta_\alpha(\hat{\alpha}_0)$) with non-integer but known values of ρ.



measure the performance relative to the probability of having a stockout $(1-\alpha)$. If $\delta_\alpha(\cdot)$ is positive, the attained service level is lower than the desired one and if $\delta_\alpha(\cdot)$ is negative, it is higher. Note that the z-axes of different desired service levels have different scales. Further, we have chosen to show the results only graphically; the corresponding numerical results listed in tables are available upon request. Only high values of α are considered, since these are used in practice. The results in Figure 1 show that the earlier findings for $t\rho$ and $\widetilde{\rho}$ both integer also hold when these assumptions are relaxed: the desired service level is not reached. Furthermore we see that the underperformance is larger if ρ is smaller (ceteris paribus; c.p. for short), if t is smaller (c.p.), if L is larger (c.p.) and if α is larger (c.p.). If ρ is small, the coefficient of variation is large and hence demand will be more variable, which implies that $\widehat{S}_1$ will be more variable as well (cf. Janssen et al. 2006). Also if t is small, the estimator $\overline{x}$ will be more variable and this also implies that $\widehat{S}_1$ will be more variable. If L is large, we have to estimate demand for a longer period of time using the estimate of $E[X]$ for one period. However, this estimate is multiplied by g_α and this factor is larger when L is larger. So the error made by estimating $E[X]$ is enlarged if L is larger and hence $\widehat{S}_1$ will be more variable. The same line of reasoning applies to α being larger: in that case g_α is larger and the error made by estimating $E[X]$ is enlarged, hence $\widehat{S}_1$ will be more variable. So intuitively the relative deviation of the desired service level will be larger in these cases and the numerical results confirm this intuition.

We assumed that ρ is known. In the remainder of this subsection this assumption is relaxed, hence, an estimate of ρ is needed. The sample mean and sample variance $\left(s^2 = \frac{1}{t-1}\sum_{i=1}^{t}(x_i-\overline{x})^2\right)$ are used to estimate ρ and λ: $\hat{\rho} = \overline{x}^2/s^2$ and $\hat{\lambda} = \overline{x}/s^2$. Using these in the estimate of the order-up-to level results in $\widehat{S}_1 = F^{-1}_{(1+L)\hat{\rho},1}(\alpha)/\hat{\lambda}$ and this is no longer gamma distributed. So derivations become intractable and hence simulation is used to determine the attained service level $\hat{\alpha}_1$ (see Appendix A). This simulation is restricted to high values of α, since values lower than 0.90 will not often be used in practice. The simulation has $n = 100,000$ replicates for each combination of ρ, t, α and L and its results are displayed in Figure 2. Note that the z-axes of the graphs equal the z-axes of Figure 1 for easy comparison. The same holds for the z-axes of Figures 3 and 5. Figure 2 clearly shows that in all cases considered the desired service level is not reached. The underperformance again is larger if ρ is smaller (c.p.), if t is smaller (c.p.), if L is larger (c.p.) and if α is larger (c.p.).

In the case of an exponential distribution ($\rho = 1$) and zero lead time, using α' instead of α ascertains that the desired service level is met. Using α' while ρ is unknown and $L \geq 0$ will probably not lead to meeting the desired service level, but since $\alpha' > \alpha$ it will certainly increase the attained service level, denoted by $\hat{\alpha}_2$. The results of replacing

Figure 2 Relative deviation ($\delta_\alpha(\hat{\alpha}_1)$) when α is used (ρ unknown).

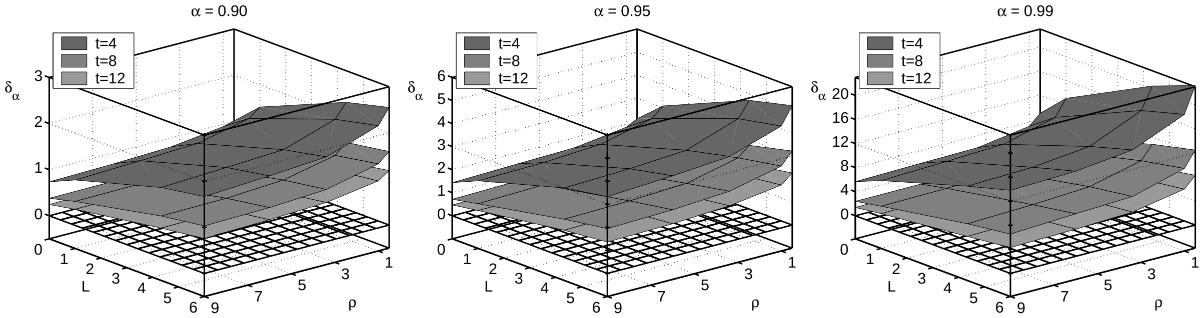
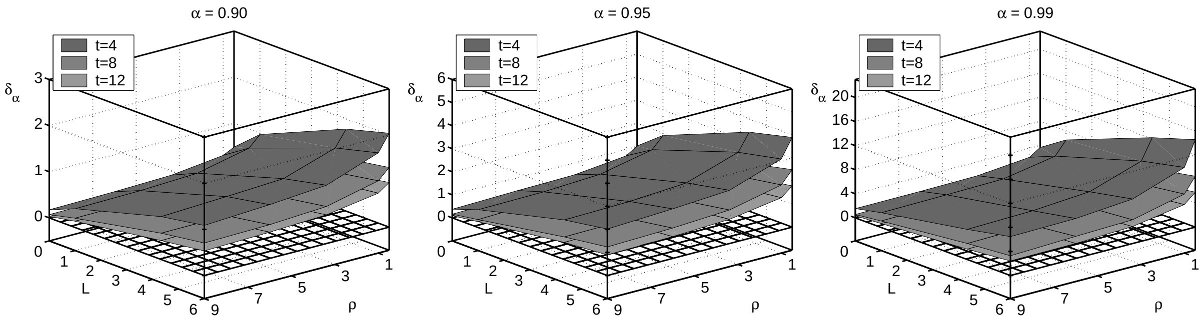


Figure 3 Relative deviation ($\delta_\alpha(\hat{\alpha}_2)$) when α' is used (ρ unknown).



α by α' are shown in Figure 3 and these show that indeed the performance improves significantly compared to using α; improvements range from almost 18% up to almost 80%. These relative improvements, denoted by $\mathcal{I}_\alpha(\hat{\alpha}_i, \hat{\alpha}_j)$, are measured by considering the percentage change in the attained stockout probability. $\hat{\alpha}_i$ is the attained service level using the order-up-to level determined by method i, here $\hat{\alpha}_1$, while $\hat{\alpha}_j$ is the attained service level using method j, $\hat{\alpha}_2$. The improvements are calculated using

$$\mathcal{I}_\alpha(\hat{\alpha}_i, \hat{\alpha}_j) = \frac{\delta_\alpha(\hat{\alpha}_i) - \delta_\alpha(\hat{\alpha}_j)}{\delta_\alpha(\hat{\alpha}_i)} \cdot 100\% = \frac{(\alpha - \hat{\alpha}_i) - (\alpha - \hat{\alpha}_j)}{\alpha - \hat{\alpha}_i} \cdot 100\%.$$

So if $\mathcal{I}_\alpha(\hat{\alpha}_i, \hat{\alpha}_j)$ is between 0 and 100%, $\hat{\alpha}_j$ is closer to the desired service level than $\hat{\alpha}_i$. If it is negative, this reverses, i.e., the $\hat{\alpha}_i$ is closer to the desired service level than $\hat{\alpha}_j$. If $\mathcal{I}_\alpha(\hat{\alpha}_i, \hat{\alpha}_j)$ is larger than 100%, using method j instead of i results in overperformance, if under i there was underperformance and vice versa. If it is smaller than 200%, $\hat{\alpha}_j$ is closer to the desired service level compared to $\hat{\alpha}_i$ and this reverses if the improvement is larger than 200%.

Four cases are considered in detail; see Table 2. In the first case the attained service level increases from 0.8016 ($\delta_\alpha(\hat{\alpha}_1) = 0.984$) to 0.8350 ($\delta_\alpha(\hat{\alpha}_2) = 0.650$), which is an improvement of $(0.8350 - 0.8016)/(0.90 - 0.8016) \cdot 100\% = 33.94\%$. Using the relative deviations ($\delta_\alpha(\cdot)$), the same improvement is found: $(0.984 - 0.650)/0.984 \cdot 100\% = 33.94\%$. In the second case one can see that, although the attained service level is already pretty close to the desired service level (compared to the first and fourth case), a large improvement is possible using α' instead of α. This leads to almost reaching the desired service level. In the third case the desired service level is reached even closer. In the fourth case the attained service level is a lot closer to the desired service level, but there is still a large underperformance. This

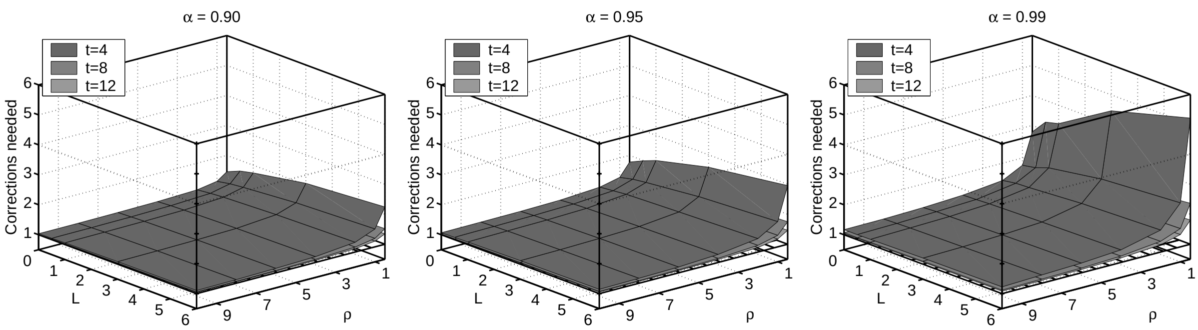
Table 2 Examples of improvement of attained service using α' instead of α.

α	ρ	t	L	$\hat{\alpha}_1$ $(\delta_\alpha(\hat{\alpha}_1))$	$\hat{\alpha}_2$ $(\delta_\alpha(\hat{\alpha}_2))$	$\mathcal{I}_\alpha(\hat{\alpha}_1, \hat{\alpha}_2)$
0.90	$\frac{44}{13}$	8	$4\frac{1}{3}$	0.8016 (0.984)	0.8350 (0.650)	33.94%
0.95	9	12	1	0.9178 (0.644)	0.9375 (0.250)	61.18%
0.95	6	12	0	0.9262 (0.464)	0.9449 (0.102)	78.57%
0.99	$\frac{1}{2}$	4	6	0.7579 (23.210)	0.8445 (14.550)	37.31%

Table 3 Parameter values used for finding correction values.

Parameter	Values used in simulation
ρ	0.5, 1, 2, 4, 6, 8, 10
t	4, 6, 8, 10, 12, 20
α	0.90, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99
L	0, 0.5, 1, 3, 6

Figure 4 Corrections needed while using a P_1 service criterion.



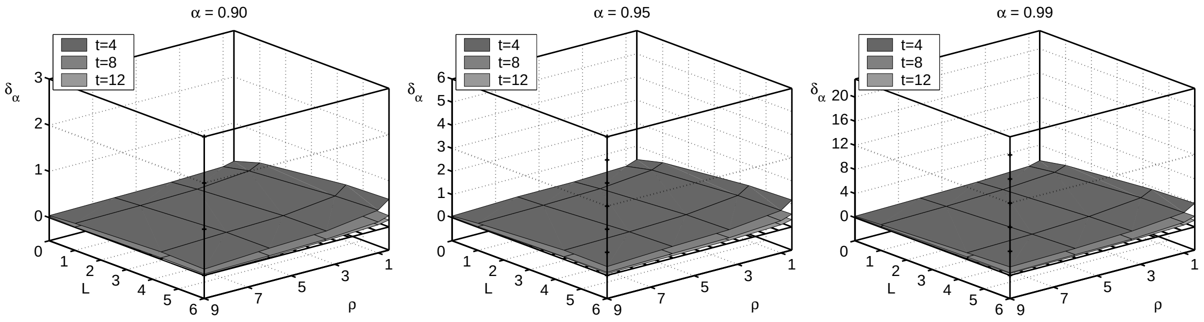
case has the most difficult parameter setting: both ρ and t are small while both α and L are large.

2.2. Determine the correction

As seen in Subsection 2.1 the desired service level is not met when using estimates in the determination of the order-up-to level. The attained service level can be improved by using α' instead of α, but still the desired service level is not reached. Another idea could be to use (an estimate of) the variance of the forecast error instead of the variance of demand (cf. Janssen et al. (2006)). We have tried using this correction, but unfortunately the initial simulation results did not show consistent improvement and hence we decided to consider other methods to improve the attained service level.

In this section it is shown that the attained service level is further improved (compared to only using α' instead of α) by using a multiplicative correction. That is, the estimated order-up-to level is multiplied by a certain factor that depends on the value of ρ, t, α and L. First, simulation is used to find the value this factor should have for different values of ρ, t, α and L; the used parameter values are listed in Table 3. The values of the corrections are found by first determining the order-up-to levels one would get while using $\hat{\rho}$, $\hat{\lambda}$ and α'. Then the factor by which this order-up-to level should be multiplied in order to reach the desired service level is determined numerically. Figure 4 shows the corrections needed for different values of ρ, t, α and L. These values could be tabulated

Figure 5 Relative deviation ($\delta_\alpha(\hat{\alpha}_3)$) using $\widehat{S}_1 \cdot e^{\hat{k}_1(\hat{\rho},t,\alpha,L)}$.



and then used to correct the order-up-to level, but a formula for the correction is easier to use. A regression technique is applied to find such a function. The natural logarithm of the correction is used as the dependent variable, since this transformation ensures that the correction increases more smoothly. The resulting function, $\hat{k}_1(\rho,t,\alpha,L)$, is shown in (5). Note that at the right hand side a is used, where $a = \ln((1-\alpha)^{-1})$. The idea of using a instead of α originates from the fact that $g_\alpha = a$ in case of exponentially distributed demand and zero lead time, hence a influences the order-up-to level directly.

$$\begin{aligned}\hat{k}_1(\rho,t,\alpha,L) = &-0.0014 - 0.0988t^{-1.10} + (0.0005 + 0.0860t^{-1.80})a^{1.90} \\ &+ [0.0613 - 0.3845t^{-0.45} + (-0.0043 + 0.5375t^{-0.85})a^{0.85}]\rho^{-1.00} \\ &+ \big(-0.0282 + 0.0518t^{-0.15} + (0.0000 - 0.0231t^{-3.00})a^{2.75} \\ &+ [0.0703 - 0.0225t^{0.35} + (0.0044 + 0.1840t^{-1.45})a^{0.90}]\rho^{-0.75}\big)L^{0.55}\end{aligned} \qquad (5)$$

In short, (5) is found as follows: we first choose only one dependent variable (L in case of (5)) and regress that on the logarithm of the correction needed, with different values for the power. We choose the power that results in the lowest sum of squared errors. Next we choose a second variable (ρ) and that is regressed on the coefficients found in the first regression. Also the third (a) and fourth (t) variable are treated in this manner; see Janssen et al. (2006) for a detailed description of the method. We choose to use this method, instead of, e.g., stepwise linear regression, because of the way the values of the powers are determined.

Using (5) on the parameter values listed in Table 3 results in an R^2 (determination coefficient) of 0.9988 (adjusted $R^2 = 0.9987$), which is very high. However, using (5) implies that ρ is known, which is obviously not true in practice. This could be solved by using the estimate for ρ, but that will be at the expense of a lower attained service. This is checked with help of simulation ($n = 100{,}000$) and in this simulation the coefficients are rounded to 10^{-4}, which means that the term $a^{2.75}L^{0.55}$ is not included, since its coefficient is zero, if rounded. The order-up-to level in this simulation is thus determined by $\widehat{S}_1 \cdot \exp(\hat{k}_1(\hat{\rho},t,\alpha,L))$ and the attained service level is denoted by $\hat{\alpha}_3$. Figure 5 shows that indeed the desired service level is reached more closely (in case of ρ large, $\alpha = 0.90$ and $t = 12$ the desired service is reached completely); additional improvements for the remainder of the cases range from 60% to 99%. Total improvements for the cases in which the desired service level is not met range from 76% up to 99%. See Table 4 for the four cases that were also considered in Table 2. In the first case the attained service level is improved a lot and the desired service level is almost reached. In the second and third case we can state that it actually is reached and in the fourth case we again see a large improvement upon the situation without using a correction, but unfortunately the underperformance

Table 4 Examples of improvement of attained service using the correction $\hat{k}_1(\hat{\rho}, t, \alpha, L)$ instead of only α'.

α	ρ	t	L	$\hat{\alpha}_2$ ($\delta_\alpha(\hat{\alpha}_2)$)	$\hat{\alpha}_3$ ($\delta_\alpha(\hat{\alpha}_3)$)	$\mathcal{I}_\alpha(\hat{\alpha}_2, \hat{\alpha}_3)$	$\mathcal{I}_\alpha(\hat{\alpha}_1, \hat{\alpha}_3)$
0.90	$\frac{44}{13}$	8	$4\frac{1}{3}$	0.8350 (0.650)	0.8911 (0.089)	86.31%	90.96%
0.95	9	12	1	0.9375 (0.250)	0.9498 (0.004)	98.40%	99.38%
0.95	6	12	0	0.9449 (0.102)	0.9493 (0.014)	86.27%	97.06%
0.99	$\frac{1}{2}$	4	6	0.8445 (14.550)	0.9508 (3.920)	73.06%	83.11%

Table 5 Extreme deviations for $\alpha \in \{0.90, 0.95, 0.99\}$ when using $\widehat{S}_1 \cdot e^{\hat{k}_1(\rho,t,\alpha,L)}$.

Desired service level	Minimum attained service ($\delta_\alpha(\hat{\alpha}_4)$)	Maximum attained service ($\delta_\alpha(\hat{\alpha}_4)$)
$\alpha = 0.90$	0.8972 (0.028)	0.9047 (-0.047)
$\alpha = 0.95$	0.9473 (0.054)	0.9522 (-0.044)
$\alpha = 0.99$	0.9889 (0.110)	0.9909 (-0.090)

is still quite large in this case. The fact that the desired service level is not reached completely in almost all cases (see Figure 5), is due to using $\hat{\rho}$ instead of ρ. If this true value could be used, simulation shows (resulting in attained service levels $\hat{\alpha}_4$) that the desired service level would be reached; the extreme deviations are denoted in Table 5.

3. P_2 service level criterion

This section considers the P_2 service level criterion, which states that at least a fraction β of demand has to be satisfied immediately. Mathematically this is denoted by $E[(X_{1+L} - S_2)^+] - E[(X_L - S_2)^+] \leq (1-\beta)E[X]$, where $x^+ = \max(0, x)$. Note that shortages at the start of a replenishment cycle are included and if $L = 0$, this term (the second expectation) vanishes. When assuming that both parameters are known, it is not difficult to find the order-up-to level that satisfies this criterion. We have to find the value of S_2 that satisfies the equality version of the inequality above. First note that the left hand side can be rewritten.

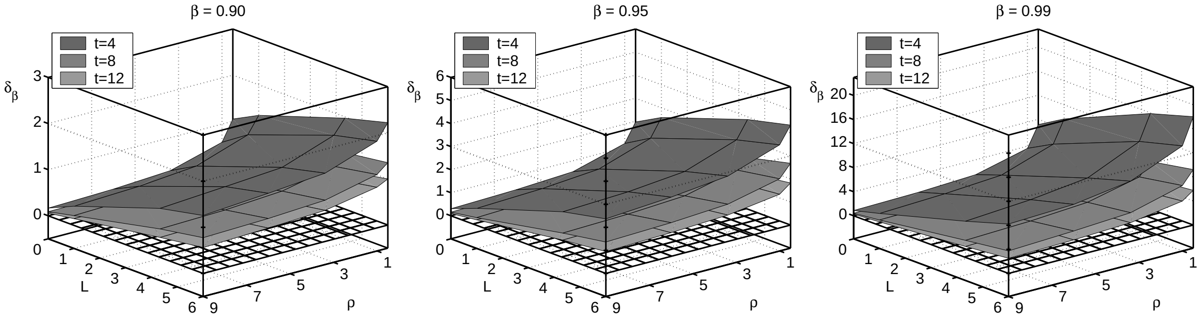
$$E\left[(X_{1+L} - S_2)^+\right] - E\left[(X_L - S_2)^+\right] \\ = \int_{S_2}^{\infty} (x - S_2) f_{\tilde{\rho},\lambda}(x)dx - \int_{S_2}^{\infty} (x - S_2) f_{L\rho,\lambda}(x)dx =: \mathcal{L}_{\tilde{\rho},\lambda}(S_2) - \mathcal{L}_{L\rho,\lambda}(S_2) \quad (6)$$

It is well known that $\mathcal{L}_{\rho,\lambda}(y) = \int_y^\infty (x-y) f_{\rho,\lambda}(x)dx = \frac{\rho}{\lambda}[1 - F_{\rho+1,\lambda}(y)] - y[1 - F_{\rho,\lambda}(y)]$. Note that there is no closed-form expression for the order-up-to level in general, hence we need to solve it numerically. However, if an exponential distribution is assumed, the order-up-to level using the P_2 criterion equals the order-up-to level using the P_1 criterion when $\alpha = \beta$ (see Appendix B). Hence, if $\rho = 1$, $S_2 = \frac{1}{\lambda} F_{1+L,1}^{-1}(\beta)$.

3.1. Using estimates for the determination of the order-up-to level

Let us first consider the case that only λ is unknown and $\rho = 1$, hence we have exponentially distributed demand during the review period. Since in the case of exponentially distributed demand the theoretically correct order-up-to levels of the P_1 and the P_2 criterion are equal, the order-up-to level in case of the P_2 criterion can be estimated by $\widehat{S}_2 = g_\beta \overline{x}$, where $g_\beta = F_{1+L,1}^{-1}(\beta)$. Then it is known that $\widehat{S}_2 \sim \Gamma(t, \frac{t}{g_\beta}\lambda)$. The attained service, using again

Figure 6 Relative deviation from the desired service levels ($\delta_\beta(\hat{\beta}_0)$) ($\rho$ known).



the fact that $1/\lambda$ is a scale parameter and $\widehat{S}_2^* = \lambda\widehat{S}_2$, will be as follows (using relation (10) in Appendix B).

$$E\left[(X_{1+L}^* - \widehat{S}_2^*)^+\right] - E\left[(X_L^* - \widehat{S}_2^*)^+\right] = E\left[E\left[(X_{1+L}^* - \widehat{S}_2^*)^+ \middle| \widehat{S}_2^*\right] - E\left[(X_L^* - \widehat{S}_2^*)^+ \middle| \widehat{S}_2^*\right]\right]$$

$$\overset{\text{App.}}{=} E\left[P\left(X_{1+L}^* > \widehat{S}_2^* \middle| \widehat{S}_2^*\right)\right] = P\left(X_{1+L}^* > \widehat{S}_2^*\right)$$

$$= \left(\frac{t}{t+g_\beta}\right)^t \sum_{i=0}^{L} \binom{t-1+i}{i} \left(\frac{g_\beta}{t+g_\beta}\right)^i \overset{L=0}{=} \left(\frac{t}{t+g_\beta}\right)^t > 1-\beta$$

Hence, in case of zero lead time the attained service will always fall short of the desired one. In fact, as long as we are considering exponentially distributed demand during the review period, the results of the P_1 criterion also hold for the P_2 criterion. Unfortunately, if $\rho \neq 1$, no tractable results can be derived, due to the non-existence of a closed-form expression for $\widehat{S}_2$. Hence, simulation is used in this case (again with $n = 100,000$ replicates) to obtain the performance of using an estimate for λ ($\hat{\lambda} = \rho/\overline{x}$) in determining the order-up-to level. Note that ρ is still assumed to be known. The relative deviation from the desired fraction of backlogged demand ($\delta_\beta(\hat{\beta}_0) = \frac{(1-\hat{\beta}_0)-(1-\beta)}{1-\beta}$, where $\hat{\beta}_0$ is the service attained in simulation; see Appendix A) is shown in Figure 6. Note that the scales of the z-axes vary across the different desired service levels. Figure 6 shows that the desired service level is not met when we have to use estimates instead of the true value of one of the parameters. Also the underperformance is larger if ρ is smaller (c.p.), if t is smaller (c.p.), if L is larger (c.p.) and if β is larger (c.p.).

If also ρ is assumed to be unknown, we will have to estimate that parameter too. The estimates for the parameters ρ and λ are $\hat{\rho} = \overline{x}^2/s^2$ and $\hat{\lambda} = \overline{x}/s^2$, where $\overline{x}$ is the sample mean and s^2 the sample variance. The order-up-to level based on these estimates ($\widehat{S}_2$) is determined by numerically solving (7) to get $\widehat{S}_2^*$ and then $\widehat{S}_2 = \widehat{S}_2^*/\hat{\lambda}$.

$$\mathcal{L}_{(1+L)\hat{\rho},1}(\widehat{S}_2^*) - \mathcal{L}_{L\hat{\rho},1}(\widehat{S}_2^*) = (1-\beta)\hat{\rho} \tag{7}$$

Next, simulation is used to estimate the performance of the order-up-to level determined in this way; $\hat{\beta}_1$ denotes the attained service. The results of this simulation, again based on $n = 100,000$ replicates, are shown in Figure 7. The z-axes in this figure equal the z-axes in Figure 6 for easy comparison, just as the z-axes in Figures 8 and 10. Comparing this figure to Figure 6 clearly shows that the underperformance is larger if both ρ and λ are assumed to be unknown. Furthermore, we again see that the underperformance is larger when ρ is smaller (c.p.), when t is smaller (c.p.), when L is larger (c.p.) and when β is

Figure 7 Relative deviation ($\delta_\beta(\hat{\beta}_1)$) when β is used (ρ unknown).

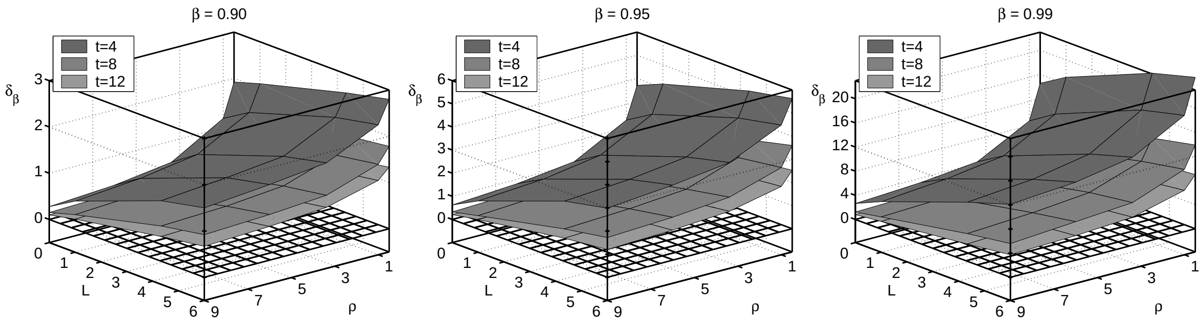


Figure 8 Relative deviation ($\delta_\beta(\hat{\beta}_2)$) when β' is used (ρ unknown).

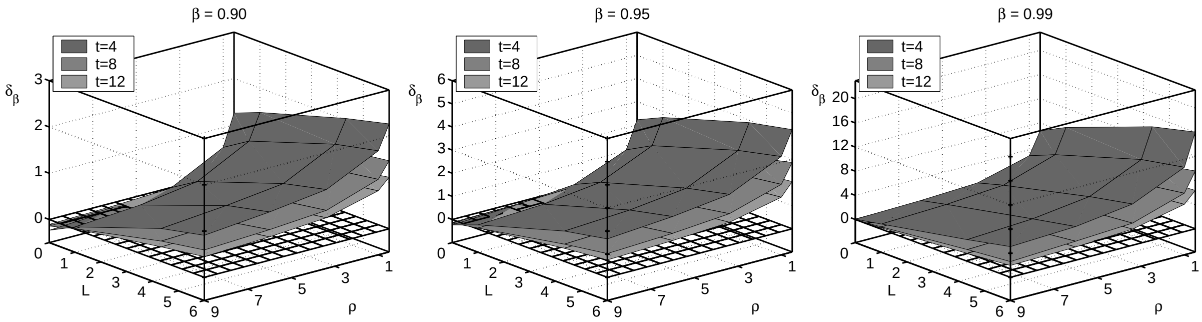


Table 6 Examples of improvement of attained service using β' instead of β.

β	ρ	t	L	$\hat{\beta}_1$ ($\delta_\beta(\hat{\beta}_1)$)	$\hat{\beta}_2$ ($\delta_\beta(\hat{\beta}_2)$)	$\mathcal{I}_\beta(\hat{\beta}_1, \hat{\beta}_2)$
0.90	$\frac{44}{13}$	8	$4\frac{1}{3}$	0.8065 (0.935)	0.8391 (0.609)	34.87%
0.95	9	12	1	0.9294 (0.412)	0.9477 (0.046)	88.83%
0.95	6	12	0	0.9366 (0.268)	0.9538 (-0.076)	128.36%
0.99	$\frac{1}{2}$	4	6	0.7390 (25.100)	0.8295 (16.050)	36.06%

larger (c.p.). In case of exponential demand and zero lead time under the P_1 criterion the desired service level could be attained by using α' instead of α. Since the order-up-to levels under the P_1 and P_2 criterion are equal in case of exponential demand, the same holds here; i.e., if β is replaced by $\beta' = 1 - \exp(t(1-(1-\beta)^{-1/t}))$, the desired service level is met again, when lead time is zero. Of course, when demand is not exponentially distributed, this will no longer hold, but we can use this correction as a first improvement. The results of using β' instead of β are shown in Figure 8; $\hat{\beta}_2$ denotes the attained service level. This figure clearly shows that, although the desired service is still not met (except for some cases where ρ large, $L = 0$ and t large), the attained service is significantly improved. In the cases where the desired service level is not met, improvements range from 16% to 99%; see Table 6 for the four cases of the previous examples. The improvements shown in this table are actually quite similar compared to the improvements for the P_1 criterion, except for the third case. In this case the desired service level is met, which leads to an improvement of more than 100%, i.e., overperformance. In the second case the desired service level is almost reached (closer compared to the P_1 criterion) and in the fourth case the attained service levels are a little below those of the P_1 criterion.

Figure 9 Corrections needed while using a P_2 service criterion.

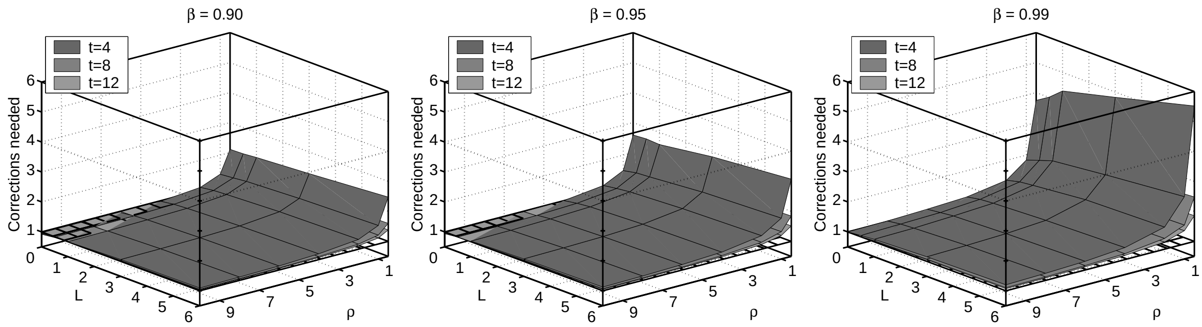
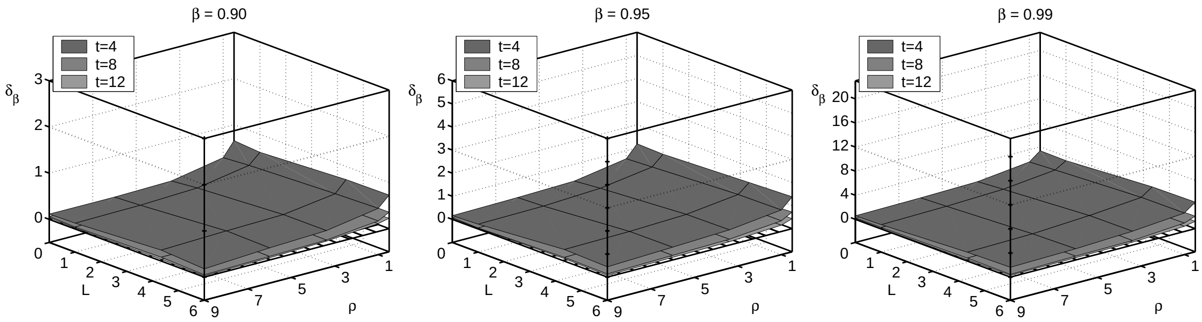


Figure 10 Relative deviation ($\delta_\beta(\hat{\beta}_3)$) using $\widehat{S}_2 \cdot e^{\hat{k}_2(\hat{\rho},t,\beta,L)}$.



3.2. Determine the correction

Analogously to Section 2.2 we try to find a multiplicative correction in case a P_2 criterion is used. First the sizes of these corrections are determined with help of simulation: the order-up-to levels based on $\hat{\rho}$, $\hat{\lambda}$ and β' are calculated for different values of ρ, t, β and L (see Table 3, with $\alpha = \beta$). The corrections needed to attain the service level are then determined numerically and these corrections for different combinations of ρ, t, β and L are shown in Figure 9. The same linear regression technique as outlined in Section 2.2 is used to find a function that estimates the logarithms of the corrections needed and the resulting function is shown in Equation (8), where $b = \ln((1-\beta)^{-1})$.

$$\begin{aligned}\hat{k}_2(\rho,t,\beta,L) = &-0.0154 - 1.0112t^{-1.25} + (-0.1363 + 0.2797t^{-0.20})\rho^{-1.45} \\ &+ [0.0034 + 0.4644t^{-1.15} + (0.0082 - 0.2634t^{-0.75})\rho^{-1.15}]L^{0.35} \\ &+ \big(-0.0014 + 1.2026t^{-2.90} + (0.0230 + 0.7037t^{-1.05})\rho^{-0.85} \\ &+ [0.0029 - 17.2361t^{-5.85} + (-0.0034 + 0.1449t^{-1.00})\rho^{-0.80}]L^{0.55}\big)b^{0.85}\end{aligned} \tag{8}$$

Using this function to estimate the corrections needed results in an R^2 of 0.9987 (adjusted $R^2 = 0.9987$), which is again very high. However, this function implies that one needs to know the true value of ρ, which is not known in practice. Hence, $\hat{\rho}$ is used to calculate the correction needed and simulation ($n = 100,000$) is performed to determine the effect of using $\widehat{S}_2 \cdot \exp(\hat{k}_2(\hat{\rho},t,\beta,L))$ to estimate the order-up-to level; the attained service level is denoted by $\hat{\beta}_3$. Like in the case of the P_1 criterion, the coefficients are rounded to 10^{-4}. The results are shown in Figure 10. This figure shows that the desired service level is reached more closely, but not reached completely yet; also in the few cases it was reached before (see Figure 8) the service is not reached, since the correction needed is smaller than

Table 7 Examples of improvement of attained service using the correction $\hat{k}_2(\hat{\rho}, t, \beta, L)$ instead of only β'.

β	ρ	t	L	$\hat{\beta}_2$ $(\delta_\beta(\hat{\beta}_2))$	$\hat{\beta}_3$ $(\delta_\beta(\hat{\beta}_3))$	$\mathcal{I}_\beta(\hat{\beta}_2, \hat{\beta}_3)$	$\mathcal{I}_\beta(\hat{\beta}_1, \hat{\beta}_3)$
0.90	$\frac{44}{13}$	8	$4\frac{1}{3}$	0.8391 (0.609)	0.8881 (0.119)	80.46%	87.27%
0.95	9	12	1	0.9477 (0.046)	0.9484 (0.032)	30.43%	92.23%
0.95	6	12	0	0.9538 (-0.076)	0.9486 (0.028)	136.84%	89.55%
0.99	$\frac{1}{2}$	4	6	0.8295 (16.050)	0.9459 (4.410)	72.52%	82.43%

Table 8 Extreme deviations for $\beta \in \{0.90, 0.95, 0.99\}$ using $\widehat{S}_2 \cdot e^{\hat{k}_2(\rho,t,\beta,L)}$.

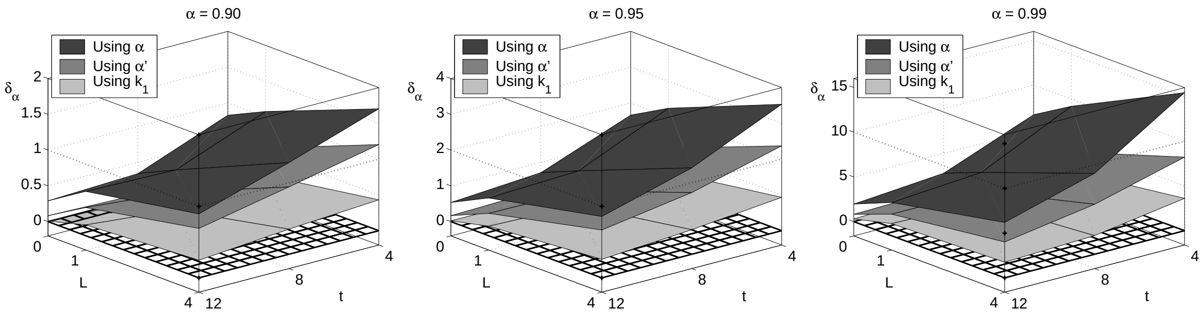
Desired service level	Minimum attained service $(\delta_\beta(\hat{\beta}_4))$	Maximum attained service $(\delta_\beta(\hat{\beta}_4))$
$\beta = 0.90$	0.8944 (0.056)	0.9031 (-0.031)
$\beta = 0.95$	0.9475 (0.050)	0.9519 (-0.038)
$\beta = 0.99$	0.9887 (0.130)	0.9911 (-0.110)

1 for those cases. In general, additional improvements range from only a few percent or even a little decline, in the cases were the desired service level was (almost) met, to 94%. The total improvements range from 66% to 94%. Table 7 shows the four cases we have considered earlier. In the first case the attained service level is improved upon considerably and the desired service level is almost reached. In the second case we only see a minor improvement, due to the fact that the attained service level was already very close to the desired one. The third case shows what is mentioned above, i.e., due to the fact that the correction needed is smaller than one, the attained service level declines a little. It is again below, but very close to the desired service level. Finally, in the fourth case the underperformance is still substantial, however, using the correction improves the attained service considerably. Again, not reaching the desired service level completely is due to the fact that $\hat{\rho}$ is used instead of ρ. If the true value of ρ is used, simulation (resulting in attained service levels $\hat{\beta}_4$) shows that the desired service level is reached; the extreme deviations are denoted in Table 8.

4. Case study

The consultancy firm Involvation provided daily demand data for a period of 9 years and 4 months of the Dutch Ministry of Defence, which is one of their customers. The data is from a department that manages the inventories of all kinds of not perishable products, ranging from screws and bolts to first aid equipment and spare parts for vehicles. We use these data to test the method for determining the order-up-to level developed in Sections 2 and 3. Since gamma demand is assumed, the demand should be continuous. However, with daily demand this will be difficult to accomplish, so we aggregated the daily demand data to monthly demand data (resulting in 112 months). The review period R is set to 1 month and the lead time L is expressed in months. Still a lot of articles have intermittent demand and therefore we selected articles that faced demand in every month for at least 17 consecutive months; 2462 articles satisfied this requirement. Of those 2462 articles 602 had two or more periods with monthly demand occurrence for at least 17 consecutive months, resulting in 3153 demand streams to work with. The length of the demand stream ranges from 17 to 112 months, with most (84%) between 17 and 50 months. Next, we used a two-tailed Anderson-Darling test to test whether each of the 3153 demand streams are gamma distributed or not at a significance level of 5%. In 401 cases there is evidence to support that the demand is *not* gamma distributed; in the remaining 2752 cases there is not enough evidence. However, we do not discard those 401 cases from our simulation.

Figure 11 Relative deviation (δ_α) using the P_1 criterion in the case study simulation.

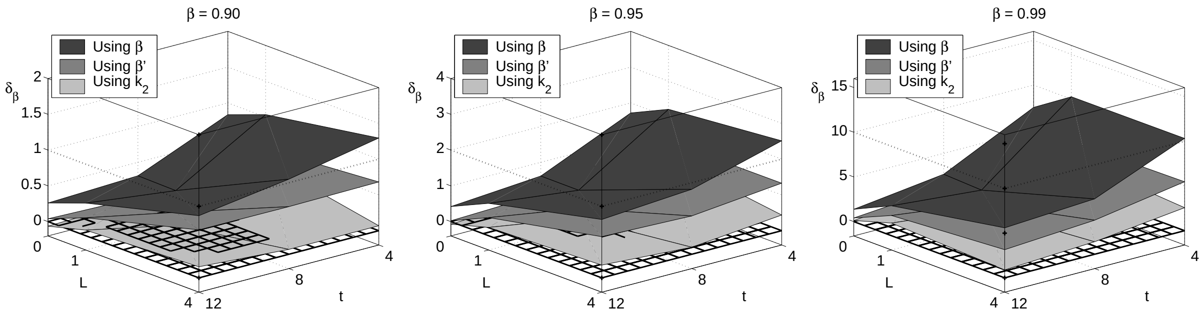


Since the shape parameter ρ is now determined by the demand data available, we only have 3 parameters left: the number of observations used ($t \in \{4, 8, 12\}$), the desired service level ($\alpha(\beta) \in \{0.90, 0.95, 0.99\}$) and the lead time ($L \in \{0, 1, 4\}$). For every demand stream we determined the order-up-to levels under the $P_1(P_2)$ criterion using $\alpha(\beta)$, $\alpha'(\beta')$ and the correction $\hat{k}_1(\hat{k}_2)$ for every combination of t, $\alpha(\beta)$ and L as follows. Since we need independent observations of the stockout occurrences (P_1) and the demand backlogged (P_2), the data streams were split up into parts, each containing $t + L + 1$ observations (therefore we selected articles with at least 17 consecutive periods with demand: $12 + 4 + 1 = 17$). The first t are used to estimate ρ and λ and those estimates are used to find the order-up-to levels. Next the demand during the lead time (the $(t{+}1)$th up till the $(t + L)$th observation) is subtracted, if $L \neq 0$, to get the net inventory at the start of the replenishment cycle. Finally the demand in the replenishment cycle (the $(t + L + 1)$th observation) is subtracted, which results in the net inventory at the end of the replenishment cycle. These can then be used to determine the attained service level; see Figures 11 and 12 for the resulting underperformance.

One can clearly see in Figure 11 that using α' instead of α results in less underperformance; improvements range from 22% to 85%. It can be verified (results not shown here), using a t-test with a significance level of 5%, that the improvement is significant, except for two cases ($t = 12$, $L = 0$ and $\alpha = 0.90, 0.95$). Using the correction instead of just α' again results in less underperformance; additional improvements range from almost 50% to 94%. In one case ($t = 8$, $L = 0$ and $\alpha = 0.90$) an improvement of over 100% is achieved, so here underperformance changes to overperformance. All the additional improvements turn out to be significant. Also, all the total improvements, ranging from 63% to 96% (plus one case with 113%), are significant as well. Hence, using the correction function given in (5) results in a significantly better performance in this case study.

Also in case of the P_2 criterion using the corrections improves the attained service level; see Figure 12. Using β' instead of β results in improvements ranging from almost 23% to 98%, plus one case ($t = 8$, $L = 1$, $\beta = 0.90$) in which the desired service level is attained; the attained service level is 0.9004. Unfortunately, not all of these improvements are significant according to a t-test with a significance level of 5%; 10 out of the 27 cases are not: all the cases at which $t = 12$ plus the case at which $t = 8$, $L = 4$ and $\beta = 0.90$. If not only β' is used, but also the correction function given in (8), additional improvements, ranging from 2% up to 90% can be achieved in case the desired service level is not reached completely. In four more cases the desired service level is reached: $\alpha = 0.90, 0.95$, $t = 8, 12$ and $L = 0$. Only if $L = 4$ and $t = 4, 8$ these improvements turn out to be significant (for all the desired service levels). The total improvements range from 53% to 95% for the 22 cases in which the desired service level is not reached completely, and from 110% to

Figure 12 Relative deviation (δ_β) using the P_2 criterion in the case study simulation.



161% for the five cases in which the underperformance changes to overperformance. If we consider the total improvements all of them, except for $t = 12$, $L = 1$ and $\beta = 0.90, 0.99$, are significant. Hence we can state that using the correction function and β' together will improve the attained service level.

5. Conclusions and further research

In this paper we considered the case of an (R, S) inventory control model with gamma demand and a service criterion. It is shown that using estimates in the determination of the order-up-to levels derived under the assumption that all parameters are known, leads to underperformance. If demand is exponentially distributed and the lead time is zero, it is shown that the desired service level is never reached. For the case of the P_1 criterion, Erlang demand, integer lead time and known shape parameter, we derived closed-form expressions for the attained service and use these to show that the desired service is not met for higher values of α, i.e., $\alpha \geq 0.50$. For the most realistic situation treated in this paper (demand is truly gamma distributed with unknown parameters) simulation is used to show that indeed underperformance exists for both the P_1 and P_2 criterion and the desired service level set at 0.90 or higher; these values are used in practice. Part of this underperformance could be solved by using α' instead of α for the P_1 criterion (improvements range from 18% to 80%) and β' instead of β for the P_2 criterion (improvements range from 16% to 99% in case the desired service level is not met; it is met in 17 out of 180 cases).

Further improvements are obtained by applying a multiplicative correction to the estimated order-up-to level. This correction is found using simulation and with help of linear regression a function is constructed. Using this correction function, given in (5) for the P_1 case and in (8) for P_2, causes the attained service to reach the desired service even more closely. However, the desired service level is not reached completely, due to the fact that the correction functions are determined using the true value of ρ while in practice only $\hat{\rho}$ can be used. The additional (total) improvements range from 60% to 99% (76% to 99%) for the P_1 criterion. The results for the additional improvements of the P_2 criterion are a bit more complicated, since in some cases the desired service level is (almost) met without using the multiplicative correction. So in some cases the attained service level declines a little, but in general the additional improvements are between a few percent up to 94%. Total improvements range from 66% to 94%.

We also applied the corrections developed in this paper to real demand data, that was provided by Involvation and the Dutch Ministry of Defence. In case of the P_1 criterion the total improvements range from 63% to 96% and in one case the desired service level is met. For the P_2 criterion the total improvements range from 53% to 95% in the 22 cases in which the desired service level is not met; it is met in the remaining 5 cases.

In this paper only the (R,S) policy is considered. One direction for further research is to consider other policies, like (s,Q), (R,s,S) or (R,s,Q). Also the lead time is assumed to be deterministic, while in practice it can very well be random to a greater or lesser degree, so this randomness could be taken into account in a follow-up study. A third idea is to consider other forecasting methods. In this paper one of the most simple forecasting methods is used: the moving average. More complex forecasting methods might lead to less underperformance, but it will probably be impossible to find analytical results using those methods. Finally, we consider the gamma distribution, which implies that demand occurs in every period. However, not having demand in each period is very common for a lot of product types; consider, e.g., spare parts. So the fourth direction for further research is to include this possibility, i.e., consider intermittent demand.

Appendix A: The attained service level in simulation

i	: Simulation run $(i=1,\dots,n)$;
d_{iR}	: Demand during review in ith run;
d_{iL}	: Demand during lead time in ith run;
s_i	: Estimated order-up-to level in ith run;
$\mathrm{I}(c)$	: Indicator function: 1 if c is true and 0 otherwise.

In case of a P_1 criterion the attained service level ($\hat{\alpha}_j,\ j\in\{1,2,3,4\}$) is determined as follows.

$$\hat{\alpha}=\frac{\sum_{i=1}^{n}\mathrm{I}(s_i>d_{iR}+d_{iL})}{n}$$

In case of a P_2 criterion the attained service level ($\hat{\beta}_j,\ j\in\{0,1,2,3,4\}$) is determined as follows.

$$\hat{\beta}=1-\frac{\sum_{i=1}^{n}(d_{iR}+d_{iL}-s_i)^+-(d_{iL}-s_i)^+}{\sum_{i=1}^{n}d_{iR}}$$

Appendix B: Equality order-up-to levels under P_1 and P_2 if demand is exponential

Remember that in case of the P_1 criterion the order-up-to level S_1 satisfied the equality $P\left(X_{1+L}>S_1\right)=1-\alpha$ or $P\left(X^*_{1+L}>S^*_1\right)=1-\alpha$, where $S^*_1=\lambda S_1$. Let us now consider the service equality in case of a P_2 criterion. Choose S_2 such that it satisfies the following. Note that, analogously to the P_1 case, $X^*_\ell=\lambda X_\ell\sim\Gamma(\ell,1)$ and $S^*_2=\lambda S_2$.

$$\begin{array}{ll} & E\left[(X_{1+L}-S_2)^+\right]-E\left[(X_L-S_2)^+\right]=(1-\beta)E\left[X\right] \\ \Leftrightarrow & \frac{1}{\lambda}\left(E\left[(X^*_{1+L}-S^*_2)^+\right]-E\left[(X^*_L-S^*_2)^+\right]\right)=(1-\beta)\frac{1}{\lambda} \\ \Leftrightarrow & E\left[(X^*_{1+L}-S^*_2)^+\right]-E\left[(X^*_L-S^*_2)^+\right]=1-\beta \end{array} \tag{9}$$

Since $\alpha=\beta$, the right hand sides of both service equations are equal, so if the left hand sides of the service equations are equal as well, then $S^*_2=S^*_1$ or $S_2=S_1$. So consider the left hand side of (9). Note that $rf_{1+r,1}(x)=xf_{r,1}(x)$ and $F_{1+r,1}(x)=F_{r,1}(x)-f_{1+r,1}(x)$.

$$\begin{array}{l} E\left[(X^*_{1+L}-S^*_2)^+\right]-E\left[(X^*_L-S^*_2)^+\right]=\mathcal{L}_{1+L,1}(S^*_2)-\mathcal{L}_{L,1}(S^*_2) \\ =(1+L)[1-F_{2+L,1}(S^*_2)]-S^*_2[1-F_{1+L}(S^*_2)]-\mathcal{L}_{L,1}(S^*_2) \\ =(1+L)[1-F_{1+L,1}(S^*_2)+f_{2+L}(S^*_2)]-S^*_2[1-F_L(S^*_2)+f_{1+L}(S^*_2)]-\mathcal{L}_{L,1}(S^*_2) \\ =(1+L)[1-F_{1+L,1}(S^*_2)]-S^*_2[1-F_L(S^*_2)]+(1+L)f_{2+L}(S^*_2)-S^*_2f_{1+L}(S^*_2)-\mathcal{L}_{L,1}(S^*_2) \\ =1-F_{1+L,1}(S^*_2)+\mathcal{L}_{L,1}(S^*_2)+S^*_2f_{1+L}(S^*_2)-S^*_2f_{1+L}(S^*_2)-\mathcal{L}_{L,1}(S^*_2) \\ =1-F_{1+L,1}(S^*_2)=P\left(X^*_{1+L}>S^*_2\right) \end{array} \tag{10}$$

Thus, the left hand sides of the service equations are equal as well and therefore the order-up-to levels under the P_1 and P_2 criterion will be equal if demand is exponentially distributed.

Acknowledgments
We would like to thank Involvation and the Dutch Ministry of Defence for providing the data used in the case study, and Dick den Hertog for his suggestions on improving this paper.

References

Bertsimas, D., A. Thiele. 2006. A robust optimization approach to inventory theory. *Oper. Res.* **54**(1) 150–168.

Burgin, T. A. 1975. The gamma distribution and inventory control. *Operational Research Quarterly* **26**(3) 507–525.

Hayes, R. H. 1969. Statistical estimation problems in inventory control. *Management Sci.* **15**(11) 686–701.

Janssen, E., L. W. G. Strijbosch, R. C. M. Brekelmans. 2006. Assessing the effects of using demand parameters estimates in inventory control. CentER Discussion Paper 90, Tilburg University, The Netherlands. URL: http://greywww.uvt.nl:2080/greyfiles/center/2006/90.html.

Lu, X., J.-S. Song, A. Regan. 2006. Inventory planning with forecast updates: Approximate solutions and cost error bounds. *Oper. Res.* **54**(6) 1079–1097.

Scarf, H. 1958. A min-max solution of an inventory problem. K. J. Arrow, S. Karlin, H. E. Scarf, eds., *Studies in the Mathematical Theory of Inventory and Production*. Stanford University Press, Stanford, CA, 201–209.

Segerstedt, A. 1994. Inventory control with variation in lead times, especially when demand is intermittent. *Int. J. Prod. Econ.* **35** 365–372.

Silver, E. A., M. R. Rahnama. 1986. The cost effects of statistical sampling in selecting the reorder point in a common inventory model. *J. Oper. Res. Soc.* **37**(7) 705–713.

Silver, E. A., M. R. Rahnama. 1987. Biased selection of the inventory reorder point when demand parameters are statistically estimated. *Eng. Cost. Prod. Econ.* **12** 283–292.

Snyder, R. D., A. B. Koehler, J. K. Ord. 2002. Forecasting for inventory control with exponential smoothing. *Int. J. Forecasting* **18** 5–18.

Strijbosch, L. W. G., R. M. J. Heuts. 1992. Modelling (s, Q) inventory systems: Parametric versus non-parametric approximations for the lead time demand distribution. *Eur. J. Oper. Res.* **63** 86–101.

Syntetos, A. A., J. E. Boylan. 2006. On the stock control performance of intermittent demand estimators. *Int. J. Prod. Econ.* **103** 36–47.

Watson, R. B. 1987. The effects of demand-forecast fluctuations on customer service and inventory cost when demand is lumpy. *J. Oper. Res. Soc.* **38**(1) 75–82.

Yeh, Q.-J. 1997. A practical implementation of gamma distribution to the reordering decision of an inventory control problem. *Production and Inventory Management Journal* **38**(1) 51–57.