Forecasting of Compound Erlang Demand

(Accepted for publication in the *Journal of the Operational Research Society*)

A.A. Syntetos, M.Z. Babai, S. Luo

Intermittent demand items dominate service and repair inventories in many industries and they are known to be the source of dramatic inefficiencies in the defense sector. However, research in forecasting such items has been limited. Previous work in this area has been developed upon the assumption of a Bernoulli or a Poisson demand arrival process. Nevertheless, intermittent demand patterns may often deviate from the memory-less assumption. In this work we extend analytically previous important results to model intermittent demand based on a compound Erlang process, and we provide a comprehensive categorisation scheme to be used for forecasting purposes. In a numerical investigation we assess the benefit of departing from the memory-less assumption and we provide insights into how the degree of determinism inherent in the process affects forecast accuracy. Operationalised suggestions are offered to managers and software manufacturers dealing with intermittent demand items.

Keywords: Demand Forecasting; Demand Categorisation; Erlang Process; Service Logistics

1. INTRODUCTION

1.1. Problem Motivation

Intermittent demand for products appears sporadically, with some time periods showing no demand at all. When demand occurs, the demand size may be constant or variable, perhaps highly so (Johnston et al. 2003) leading to what is often termed as lumpy demand.

Intermittent demand items dominate service and repair parts inventories in many industries (Boylan and Syntetos 2010). A survey by Deloitte (2006) benchmarked the service businesses of many of the world's largest manufacturing companies with combined revenues reaching more than $1.5 trillion; service operations accounted for an average of 25% of revenues. In addition, defense inventories that are predominantly composed from spare parts have been repeatedly identified as a high risk area with a direct impact to a nation's economy. In the US, for example, the Department of Defense (DOD) manages more than 4 million secondary **(Note 1)** items and reported that as of September 2010 the value of its inventory was $95.6 billion (GAO 2012). Similarly, in 2011 the total inventory value for the Ministry of Defense in the UK was (gross value, before adjustments for depreciation) £40.3 billion, £35 billion of which was related to

non-explosive inventories (predominantly spare parts) (Morse 2012). Improvement programs focusing on the management and forecasting of defense inventories are currently in progress in many countries around the globe, including Australia (DMO 2010), Greece (Ziotopoulou 2009), the UK (Morse 2012) and the US (GAO 2011) [(Note 2)].

Intermittent demand is known to be often characterised by increasing failure rate (IFR) distributions (Smith and Dekker 1997). However, all research on forecasting intermittent demand has been based on either the Bernoulli or the memory-less Poisson arrival process. Currently, there is a need to re-appraise the management of intermittent demand SKUs and advance the theoretical state of knowledge with regards to forecasting their requirements (Gardner 2011, Altay and Litteral 2011). This constitutes the focus of our work. For the purpose of our analysis, Erlang distributed inter-demand times will be assumed in order to provide a comprehensive framework for forecasting purposes.

1.2 Research Relevance

Intermittent demand patterns pose considerable difficulties in terms of forecasting due to their compound nature; demand occurrences are interspersed by intervals of no demand at all. Many distributions have been suggested to model the sizes of demand when demand occurs (Axsäter 2006), such as the geometric (Watson 1987, Johnston et al. 2003, Chew and Johnson 2006), logarithmic (Syntetos and Boylan 2006), lognormal (Syntetos et al. 2009a), etc. However, and as discussed in the next section, modeling for forecasting purposes is insensitive to such a distributional assumption. The coefficient of variation (CV) of the demand sizes is a major determinant of forecast performance but the distribution itself is not.

As commonly considered in the inventory control literature, stationary demand arrivals may be modeled as: Bernoulli – implying a discrete treatment of time and resulting in geometrically distributed inter-demand intervals; Poisson, being associated with a continuous treatment of time and resulting in exponentially distributed (Erlang-1) intervals; Censored Poisson, that contains Poisson and Condensed Poisson as a special case, and results in Erlang distributed intervals. (This is further discussed below.)

A comprehensive body of knowledge has been provided for systems accumulating demand information in discrete time intervals as far as forecasting is concerned (Syntetos et al. 2005).

Some analysis has also been conducted emphasising systems where information is recorded (and utilised) at the individual transaction level, under the assumption of Poisson arrivals (Shale et al. 2006). Johnston and Boylan (1996) showed that at the level of individual orders, a Poisson process is more natural than a Bernoulli one (see also Zipkin 2000, Lariviere and van Mieghem 2004). Particularly, if the item being ordered has many different customers who purchase independently or at least not in a scheduled manner, the Poisson process model would have intuitive appeal.

However, a different distribution is required whenever the inter-demand time is 'less variable' than suggested by the memory-less Poisson (or Bernoulli) process (Smith and Dekker 1997). There are many settings where this situation may naturally occur. The most obvious one occurs in spare parts management, where parts used for corrective maintenance may wear.

In addition, demand patterns in Business-to-Business environments (B2B) are all too often determined by the degree of heterogeneity of the client base (Bartezzaghi et al. 1999). Heterogeneous requests occur when the potential market consists of customers with considerably different sizes, for example, few large customers coexist with a number of small customers. The higher the heterogeneity of customers, the higher the demand lumpiness, since periods with high requests from a large customer alternate with periods with low or no requests at all from small customers. Alternatively, following a request from a large customer, it is unlikely that another demand will be received in the near future. The potential correlation between customers' requests further induces lumpiness. Correlation may be due, amongst other reasons, to imitation and fashion which induce similar behaviors in customers. Additionally, business policies such as payment terms, quantity discounts on sales, and flat transaction fees may also influence customers in a way that makes demand lumpy.

Finally, the situation discussed here may emerge in a multi-echelon system, where lot-sizing is applied at lower levels. The case of Erlang distributed inter-demand times is obtained if demand originates from the lot sizing by a single customer experiencing Poisson demand.

Consideration of an Erlang (λ, r) distribution is equivalent to assuming a 'censored' Poisson arrival process in which only every r^{th} event is recorded. This is a realistic representation of consumer purchasing behavior (Herniter 1970). The mean inter-demand time is $\frac{r}{\lambda}$ (or,

equivalently, $\frac{\lambda}{r}$ is the number of demands per time unit) and r is the shape parameter that needs to be integer (see also Larsen 2008). For $r = 1$ we obtain the exponential distribution as a special case. For $r = 2$ we obtain the Erlang-2 distribution which Chatfield and Goodhardt (1973) termed as 'condensed Poisson' because its variance is less than its mean. As r increases the distribution becomes less variable, and the coefficient of variation tends to zero as r tends to infinity. That is, the 'degree of determinism' inherent in the process is associated with the value of r – the higher that value is, the larger the departure from the memory-less assumption resulting eventually to constant inter-demand times.

At this point it should be noted that the degree / nature of intermittence (which, for a given r, is associated with the value of the demand arrival rate λ - the lower that value is, the less often demand occurs) may imply that no standard theoretical distributions fit inter-demand times, in which case data-driven approaches may need to be considered (see, e.g., Scala et al. 2013, 2014).

1.3 Contributions and Organization of the Paper

We extend previous important findings in the area of intermittent demand forecasting to the case of Increasing Failure Rate (IFR) induced demands. This may be justified both from the reliability engineering and a general consumer behavior perspective. We consider the forecasting methods most commonly used in service logistics and analyze their statistical properties under the assumption of Erlang distributed inter-demand times. In doing so, care is exercised to operationalise our theoretical findings to allow for their direct implementation in real world settings. Our analysis leads naturally to the development of a demand classification scheme for the automatic selection of the appropriate estimators.

At this point we should mention that distributions other than the Erlang could, theoretically, have been considered. Both the Weibull and the lognormal distribution are natural candidates; the same is true for the gamma distribution (Pearson type III) which is of the same form as Erlang but the shape parameter is not constrained to being an integer. However, the Erlang has long been shown to be mathematically easier to handle than the other alternatives (Chatfield and Goodhardt 1973) and an *a-priori* case can be made for this distribution due to its explicit and intuitively appealing linkage with the demand generation process (censored Poisson).

The remainder of our paper is structured as follows: in Section 2 we provide the research background for the case of a Bernoulli demand arrival model. Existing findings for the Poisson arrival process are presented (and corrected where necessary) as a special case of the Censored

Poisson one, the analysis of which takes place in Section 3. In Section 4, we conduct a direct theoretical comparison between possible intermittent demand estimators for the purpose of establishing regions of superior performances. A numerical study allows insights to be gained into the sensitivity of the results to the key control parameters (λ, r and CV of the demand sizes). We conclude the paper in Section 5 where we also discuss the empirical implications of our findings.

2. RESEARCH BACKGROUND: THE BERNOULLI CASE

Croston (1972) advocated separating intermittent demand into two components, the inter-demand time and the size of demand, when demand occurs, and analyzing each component separately. He assumed a stationary mean model for representing the underlying demand pattern and a Bernoulli demand process, resulting in geometrically distributed inter-demand intervals. If we let:

z_t = the demand size (accumulation of transaction orders over a discrete time period t), when demand occurs, following any distribution with a mean μ and variance σ^2

p_t = the inter-demand interval that follows the geometric distribution with: $\mathrm{E}(p_t) = p$

$1/p$ = the underlying Bernoulli probability of demand occurrence

then the expected demand per unit time period is: $\mathrm{E}(Y_t) = \mu / p$.

Exponentially Weighted Moving Averages (EWMA) are most often used in practice to forecast intermittent demand requirements (Johnston and Boylan 1996, Syntetos et al. 2014) although the method was designed for fast demands (Brown 1959, Graves 1999). (In practice this method is also referred to as Simple (or Single) Exponential Smoothing, SES.) The EWMA forecast produced at the end of period t (estimate of demand in period t+1) is as follows:

$$Y_t' = Y_{t-1}' + \alpha(Y_t - Y_{t-1}') = \alpha \sum_{i=0}^{\infty} (1-\alpha)^i \, Y_{t-i} \qquad (1)$$

where α is a smoothing constant, with $0 \le \alpha \le 1$.

Under the above demand model, EWMA is unbiased if we consider the estimates made at the end of every forecast review period (*all points in time*). It is biased though if we isolate the estimates made after a demand occurrence (*issue points only*). This is a major problem since stock replenishments are often triggered by an issue point, i.e. a demand occurrence. (For a further discussion on the distinction between *all* and *issue points in time only*, the readers are referred to Johnston and Boylan 1996) **(Note 3)**.

To alleviate the problem, Croston proposed an alternative method that builds demand estimates from constituent elements. According to his method, separate exponential smoothing estimates of the average size of the demand and the average interval between demand incidences are made after demand occurs. If no demand occurs, the estimates remain the same. If we let:

z'_t = the exponentially smoothed size of demand, updated only if demand occurs in period t so that $\mathrm{E}(z'_t)=\mathrm{E}(z_t)=\mu$: $z'_t = z'_{t-1}+\alpha(z_t - z'_{t-1})$, and

p'_t = the exponentially smoothed inter-demand interval, updated only if demand occurs at the end of period t so that $\mathrm{E}(p'_t)=\mathrm{E}(p_t)=p$: $p'_t = p'_{t-1}+\alpha(p_t - p'_{t-1})$

then following Croston's estimation procedure, the forecast, Y'_t for the next time period is given by:

$$Y'_t = \frac{z'_t}{p'_t} \tag{2}$$

Syntetos and Boylan (2001) demonstrated the biased nature of Croston's estimator and showed that there is scope for improving the accuracy of the relevant estimates. The bias stems from the fact that: $\mathrm{E}(1/p'_t)\neq 1/\mathrm{E}(p'_t)$. The emergence of bias is independent of the demand arrival process (e.g., Bernoulli or Poisson) and estimation procedure being used (e.g., EWMA or Simple Moving Average, SMA) although its magnitude is not. Subsequently, Syntetos and Boylan (2005) proposed the following approximately unbiased estimator:

$$Y'_t = \left(1-\frac{\alpha}{2}\right)\frac{z'_t}{p'_t} \tag{3}$$

The above theory-informed heuristic provides a reasonable approximation of the actual demand per period especially for the cases of very low α values and large p inter-demand intervals. This estimator is known in the literature as the SBA method (after Syntetos-Boylan Approximation) and is associated with a small negative bias which is $-(\alpha/2)(\mu/p^2)$. Some other Croston variants have also been discussed recently in the literature (e.g. Teunter et al. 2011) but SBA is the only adaptation to Croston's method with empirical evidence in its support (e.g., Fildes et al. 2008, Syntetos et al. 2009b). The method has been assessed on 18,750 SKUs from the Royal Air Force and found to outperform Croston's estimator (Eaves and Kingsman 2004).

Johnston and Boylan (1996) considered a very important question: what is the degree of intermittence that renders Croston's method more appropriate than EWMA. The researchers worked on the premise that it is preferable to compare directly estimation procedures for the purpose of establishing regions of superior performance and then categorize demand based on the results **(Note 4)**. They compared EWMA and Croston based on their simulated Mean Squared Error (MSE) and under a wide range of realistic assumptions as far as the demand generation process was concerned. They came to the conclusion that Croston should be preferred to EWMA for inter-demand intervals greater than 1.25 forecast revision periods. The contribution of their work lies in the identification of the inter-demand interval as a classification parameter rather than the exact value (cut-off point) assigned to that.

Syntetos et al. (2005) extended the work discussed above to consider theoretically derived MSE expressions. They compared the theoretical performance of EWMA, Croston and SBA through their theoretical MSE over a fixed lead time of duration *L*. Such comparisons resulted in regions where one method performs better than others and enabled the selection of specific methods under specific demand characteristics. All the three methods were assumed to be employed with a common smoothing constant α. The MSE over a fixed lead time of duration *L*, assuming error auto-correlation, is calculated as follows (Strijbosch et al. 2000, Syntetos et al. 2005):

$$MSE_{L.T.} = L\left\{L\,Var(Estimate) + L\,Bias^2 + Var(Actual\ Demand)\right\} \qquad (4)$$

It follows from equation (4) that the MSE over lead time *L* of Method A is greater than the MSE over lead time *L* of Method B if and only if

$$\mathrm{Var}(Estimate)_{\mathrm{A}} + \mathrm{Bias}_{\mathrm{A}}^2 > \mathrm{Var}(Estimate)_{\mathrm{B}} + \mathrm{Bias}_{\mathrm{B}}^2, \qquad (5)$$

where the subscripts refer to the forecasting methods employed.

Considering inequality (5) it is obvious that the comparison between any two estimation procedures is only in terms of the bias and the variance of their one step-ahead estimates. The comparisons resulted in operationalised classification schemes along the lines presented in Figure 1. In addition to the average inter-demand interval (*p*), the squared coefficient of variation of the demand sizes (CV^2) was also identified as an important parameter to distinguish between alternative methods' performance. The empirical utility of the schemes discussed above have been confirmed in many studies (Ghobbar and Friend 2002, Regattieri et al. 2005, Boylan et al. 2008). For a comprehensive account of recent developments in intermittent demand forecasting under the Bernoulli case please refer to Syntetos et al. (2009b).

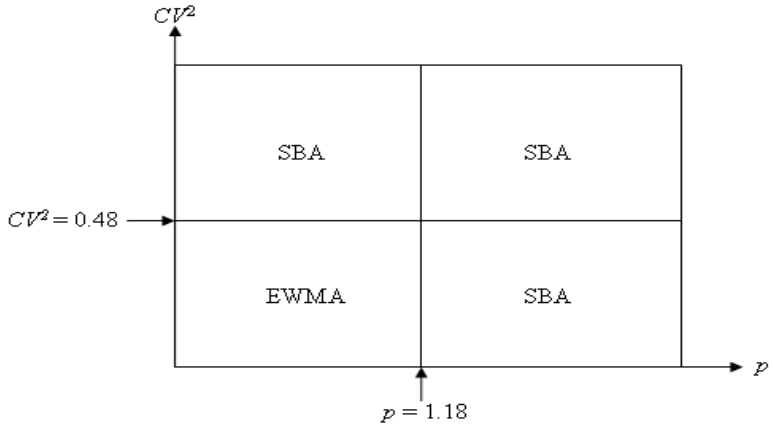


Figure 1: SKU classification based on the comparative performance of SBA and EWMA (all points in time) for $\alpha = 0.15$

3. ANALYSIS OF THE CENSORED POISSON CASE

We assume that demand data is available at the individual transaction level and the inter-arrival (inter-order) intervals are Erlang distributed. Such a distribution arises from a 'censored' Poisson process, i.e. when an ordinary Poisson process of rate λ is censored so that only every r^{th} arrival remains, under the additional assumption that counting begins at an instance independent of the process being counted. An Erlang distribution reduces to Exponential when $r = 1$.

For presentation purposes, let us retain the previous notation but adjust it for the consideration of information being available at the transaction level. Let z_t denote the individual transaction size when it occurs and p_t the inter-order interval. Transaction sizes z_t are assumed to be stationary, *iid* with mean μ and variance σ^2. Inter-order intervals p_t are also assumed to be stationary, *iid* following an Erlang distribution with scale parameter λ and shape parameter *r*. Transaction sizes and inter-order intervals are assumed to be mutually independent.

Let z_t' be the exponentially smoothed transaction size, updated only when a transaction occurs and p_t' the exponentially smoothed inter-order interval, also updated only following a transaction occurrence. Common smoothing constant values α are used for both the transaction size estimates and interval estimates.

The demand and the number of individual transaction orders per unit time period are denoted by y_t and *N* respectively. It follows that the demand in any given time period will be the sum of the individual transaction sizes in that period, i.e.

$$y_t = \sum_{t=1}^{N} z_t \,. \tag{6}$$

The mean and variance of N can be shown to be (Trabka and Marchand 1970):

$$E(N) = \frac{\lambda}{r}, \tag{7}$$

$$\mathrm{Var}(N) = \frac{\lambda}{r^2} + \frac{r^2 - 1}{6r^2} + \frac{\theta_r}{r^2}. \tag{8}$$

where $\theta_r = 2\sum_{k=1}^{r-1} \frac{\varepsilon_k}{(1-\varepsilon_k)^2} \mathrm{e}^{-\lambda(1-\varepsilon_k)}$ (r is an integer larger than 1, $\varepsilon_k = \mathrm{e}^{i(\frac{2\pi}{r})k}$ $(k = 0, 1, 2, \cdots, r-1)$ are the r^{th} roots of unity and θ_1 is 0 for the case of the ordinary Poisson process).

Using equations (7) and (8), we can obtain the mean and variance of the demand per unit time period (e.g., Ross 1996):

$$E(y_t) = E(N)E(z_t) = \frac{\lambda\mu}{r}, \tag{9}$$

$$\begin{aligned} \mathrm{Var}(y_t) &= E(N)\mathrm{Var}(z_t) + (E(z_t))^2 \mathrm{Var}(N) \\ &= \frac{\lambda}{r}\sigma^2 + \left(\frac{\lambda}{r^2} + \frac{r^2 - 1}{6r^2} + \frac{\theta_r}{r^2} \right) \mu^2 \end{aligned}. \tag{10}$$

***Remark* 1:** For $r = 2$, it can be verified that $\theta_2 = -\frac{1}{2}\mathrm{e}^{-2\lambda}$. So the mean and variance of N are $E(N) = \frac{\lambda}{2}$ and $\mathrm{Var}(N) = \frac{\lambda}{4} + \frac{1}{4}\mathrm{e}^{-\lambda}\sinh\lambda$, respectively. The expressions for the mean and variance of y_t then become $\frac{\lambda\mu}{2}$ and $\frac{\lambda}{2}\sigma^2 + \left(\frac{\lambda}{4} + \frac{1}{4}\mathrm{e}^{-\lambda}\sinh\lambda \right)\mu^2$, respectively. For the case of Poisson process ($r = 1$), the expressions for the mean and variance of y_t are λ and $\lambda(\sigma^2 + \mu^2)$, respectively **(Note 5)**.

We provide below an analysis of the statistical properties of the three forecasting methods commonly used to estimate demand in an intermittent context, namely: EWMA, Croston and SBA.

3.1 Exponentially Weighted Moving Average (EWMA)

As discussed in the previous section, the performance of the EWMA method depends on which forecasts are considered for evaluation purposes. That is, the performance of the method is different when we consider all points in time (regardless of a transaction occurrence) (*EWMA-ALL*, hereafter) and only the estimates immediately after an order (issue) has occurred (*EWMA-ISSUE*, hereafter). The former would correspond to the situation of a system employing a re-order

interval or product group review (period inventory control), whilst the latter would be associated with a re-order level stock replenishment system (continuous inventory control). The *EWMA-ALL* estimates are obviously unbiased. Using the relationship between variance of input and output of a simple exponential smoothing forecasting system as given by Brown (1963), we obtain the variance of the *EWMA-ALL* estimates:

$$\begin{aligned}\mathrm{Var}(y_t^{'}(\mathrm{EWMA}-\mathrm{ALL})) &= \frac{\alpha}{2-\alpha}\mathrm{Var}(y_t) \\ &= \frac{\alpha}{2-\alpha}\left[\frac{\lambda}{r}\sigma^2+\left(\frac{\lambda}{r^2}+\frac{r^2-1}{6r^2}+\frac{\theta_r}{r^2}\right)\mu^2\right]\end{aligned}. \tag{11}$$

On the other hand, the *EWMA-ISSUE* estimates are biased, please refer also to Croston (1972):

$$\begin{aligned}\mathrm{E}(y_t^{'}(\mathrm{EWMA}-\mathrm{ISSUE})) &= \alpha\mu+(1-\alpha)\mathrm{E}(y_t) \\ &= \frac{\lambda\mu}{r}+\alpha(1-\frac{\lambda}{r})\mu\end{aligned}. \tag{12}$$

The variance of such estimates may be calculated as follows:

$$\begin{aligned}\mathrm{Var}(y_t^{'}(\mathrm{EWMA}-\mathrm{ISSUE})) &= \alpha^2\sigma^2+\frac{\alpha(1-\alpha)^2}{2-\alpha}\mathrm{Var}(y_t) \\ &= \alpha^2\sigma^2+\frac{\alpha(1-\alpha)^2}{2-\alpha}\left[\frac{\lambda}{r}\sigma^2+\left(\frac{\lambda}{r^2}+\frac{r^2-1}{6r^2}+\frac{\theta_r}{r^2}\right)\mu^2\right]\end{aligned}. \tag{13}$$

3.2 Croston's Method

The estimation equation for Croston's method (CRO method, hereafter) is as follows:

$$y_t'(\mathrm{CRO})=\frac{z_t'}{p_t'}. \tag{14}$$

Using Taylor's theorem, the approximate expressions for the mean and variance of z_t'/p_t' can be shown to be (to the second order term) (Syntetos and Boylan 2010):

$$E\left(\frac{z_t'}{p_t'}\right)\approx\frac{E(z_t')}{E(p_t')}+\frac{1}{2}\cdot\frac{2E(z_t')}{(E(p_t'))^3}\mathrm{Var}(p_t'), \tag{15}$$

$$\begin{aligned}\mathrm{Var}\left(\frac{z_t'}{p_t'}\right)\approx{}&\frac{\mathrm{Var}(z_t')}{(E(p_t'))^2}+\frac{(E(z_t'))^2\mathrm{Var}(p_t')}{(E(p_t'))^4}+\frac{\mathrm{Var}(z_t')\mathrm{Var}(p_t')}{(E(p_t'))^4} \\ &-\frac{2(E(z_t'))^2E(p_t'-E(p_t'))^3}{(E(p_t'))^5}+\frac{(E(z_t'))^2E(p_t'-E(p_t'))^4}{(E(p_t'))^6}-\frac{(E(z_t'))^2(\mathrm{Var}(p_t'))^2}{(E(p_t'))^6}\end{aligned}. \tag{16}$$

The mean and variance of z_t' and p_t' are derived in *Appendix A* (which is presented separately in an electronic companion to this paper):

$$E(p_t')=\frac{r}{\lambda},\ \ \mathrm{Var}(p_t')=\frac{\alpha}{2-\alpha}\cdot\frac{r}{\lambda^2},\ \ E(z_t')=\mu,\ \ \mathrm{Var}(z_t')=\frac{\alpha}{2-\alpha}\sigma^2, \tag{17}$$

Therefore:

$$E(y_t'(\mathrm{CRO}))\approx\frac{\lambda\mu}{r}+\frac{\alpha}{2-\alpha}\frac{\lambda\mu}{r^2} \tag{18}$$

Noting also that

$$E(p_t)=\frac{r}{\lambda},E(p_t^2)=\frac{r(r+1)}{\lambda^2},E(p_t^3)=\frac{r(r+1)(r+2)}{\lambda^3},E(p_t^4)=\frac{r(r+1)(r+2)(r+3)}{\lambda^4},$$

we obtain

$$E(p_t^{'}-E(p_t^{'}))^3=\frac{\alpha^3}{1-(1-\alpha)^3}E(p_t-E(p_t))^3=\frac{\alpha^3}{1-(1-\alpha)^3}\frac{2r}{\lambda^3} \tag{19}$$

and

$$\begin{aligned}E(p_t^{'}-E(p_t^{'}))^4&=\frac{\alpha^4}{1-(1-\alpha)^4}E(p_t-E(p_t))^4+\frac{\alpha^4}{\left[1-(1-\alpha)^2\right]^2}(\mathrm{Var}(p_t))^2\\&=\frac{\alpha^4}{1-(1-\alpha)^4}\frac{3r(r+2)}{\lambda^4}+\left(\frac{\alpha}{2-\alpha}\right)^2\frac{r^2}{\lambda^4}\end{aligned} \tag{20}$$

under the assumption of no auto-correlation for the transaction size and inter-order interval series and homogeneous moments about the mean for both series.

Consequently,

$$\begin{aligned}&\mathrm{Var}(y_t'(\mathrm{CRO}))=\mathrm{Var}\left(\frac{z_t'}{p_t'}\right)\\&\approx\frac{\frac{\alpha}{2-\alpha}\cdot\sigma^2}{\left(\frac{r}{\lambda}\right)^2}+\frac{\mu^2\cdot\frac{\alpha}{2-\alpha}\cdot\frac{r}{\lambda^2}}{\left(\frac{r}{\lambda}\right)^4}+\frac{\frac{\alpha}{2-\alpha}\cdot\sigma^2\cdot\frac{\alpha}{2-\alpha}\cdot\frac{r}{\lambda^2}}{\left(\frac{r}{\lambda}\right)^4}-\frac{2\mu^2\cdot\frac{\alpha^3}{1-(1-\alpha)^3}\cdot\frac{2r}{\lambda^3}}{\left(\frac{r}{\lambda}\right)^5}\\&+\frac{\mu^2\left[\frac{\alpha^4}{1-(1-\alpha)^4}\cdot\frac{3r(r+2)}{\lambda^4}+(\frac{\alpha}{2-\alpha})^2\cdot\frac{r^2}{\lambda^4}\right]}{\left(\frac{r}{\lambda}\right)^6}-\frac{\mu^2\left(\frac{\alpha}{2-\alpha}\cdot\frac{r}{\lambda^2}\right)^2}{\left(\frac{r}{\lambda}\right)^6}\\&=\left[\frac{\alpha}{2-\alpha}+\frac{1}{r}(\frac{\alpha}{2-\alpha})^2\right]\frac{\lambda^2\sigma^2}{r^2}+\left[\frac{\alpha}{2-\alpha}-\frac{1}{r}\frac{4\alpha^3}{1-(1-\alpha)^3}+\frac{r+2}{r^2}\frac{3\alpha^4}{1-(1-\alpha)^4}\right]\frac{\lambda^2\mu^2}{r^3}\end{aligned}. \tag{21}$$

3.3 Syntetos-Boylan Approximation (SBA)

For the Bernoulli case, Syntetos and Boylan (2005) proposed the correction factor $1-\alpha/2$ to Croston's estimates in order to account for the bias of the estimates. In the case of an Erlang process, the correction factor becomes $1-\dfrac{\alpha}{r(2-\alpha)+\alpha}$ (please refer to *Appendix B*, which is presented separately in the electronic companion to this paper). It should be noted that Shale et al. (2006) suggested another correction factor, namely $1-\dfrac{\alpha}{r(2-\alpha)}$ to remove the bias of *CRO*. We argue in *Appendix B* that the result presented by Shale et al. is not correct. The SBA forecast is then expressed as:

$$y_t'(\text{SBA}) = \left(1-\frac{\alpha}{r(2-\alpha)+\alpha}\right)\frac{z_t'}{p_t'}. \tag{22}$$

The SBA estimate is approximately unbiased:

$$\text{E}(y_t^{'}(\text{SBA})) = \left(1-\frac{\alpha}{r(2-\alpha)+\alpha}\right)\text{E}(y_t^{'}(\text{CRO})) \approx \frac{r(2-\alpha)}{r(2-\alpha)+\alpha}\left[\frac{\lambda\mu}{r}+\frac{\alpha}{2-\alpha}\frac{\lambda\mu}{r^2}\right] \approx \frac{\lambda\mu}{r} \tag{23}$$

In addition, using the result for the variance of *CRO* estimate, we obtain:

$$\begin{aligned}\text{Var}(y_t^{'}(\text{SBA})) &= \left(1-\frac{\alpha}{r(2-\alpha)+\alpha}\right)^2 \text{Var}(y_t^{'}(\text{CRO})) \\ &\approx \left(1-\frac{\alpha}{r(2-\alpha)+\alpha}\right)^2 \left\{\left[\frac{\alpha}{2-\alpha}+\frac{1}{r}\left(\frac{\alpha}{2-\alpha}\right)^2\right]\frac{\lambda^2\sigma^2}{r^2}+\left[\frac{\alpha}{2-\alpha}-\frac{1}{r}\frac{4\alpha^3}{1-(1-\alpha)^3}+\frac{r+2}{r^2}\frac{3\alpha^4}{1-(1-\alpha)^4}\right]\frac{\lambda^2\mu^2}{r^3}\right\}\end{aligned} \tag{24}$$

***Remark* 2**: The SBA estimates (22) are approximately unbiased in nature because approximate equation (18) is obtained by keeping only the first three terms in the Taylor series expansion. If more terms in the Taylor expansion are considered, then the approximate expression for $E(y_t'(\text{CRO}))$ will be different, which in turn results in a different approximate expression for $E(y_t'(\text{SBA}))$. For example, if the first four terms are considered, then

$$\begin{aligned} E\left(\frac{z_t'}{p_t'}\right) &\approx \frac{E(z_t')}{E(p_t')}+\frac{1}{2}\cdot\frac{2E(z_t')}{(E(p_t'))^3}\text{Var}(p_t')-\frac{1}{3!}\cdot\frac{6E(z_t')}{(E(p_t'))^4}E(p_t'-E(p_t'))^3 \\ &= \frac{\lambda\mu}{r}+\frac{\alpha}{2-\alpha}\cdot\frac{\lambda\mu}{r^2}-\frac{\mu}{(r/\lambda)^4}\cdot\frac{\alpha^3}{1-(1-\alpha)^3}\cdot\frac{2r}{\lambda^3} \\ &= \frac{\lambda\mu}{r}+\frac{\alpha}{2-\alpha}\cdot\frac{\lambda\mu}{r^2}-\frac{2\alpha^3}{1-(1-\alpha)^3}\cdot\frac{\lambda\mu}{r^3}\end{aligned}, \tag{18′}$$

and then the mean of SBA estimate will be

$$
\begin{aligned}
\mathrm{E}(y_t^{'}(\mathrm{SBA})) &= E\left(\left(1-\frac{\alpha}{r(2-\alpha)+\alpha}\right)\frac{z_t^{'}}{p_t^{'}}\right) \\
&\approx \left(1-\frac{\alpha}{r(2-\alpha)+\alpha}\right)\left(\frac{\lambda\mu}{r}+\frac{\alpha}{2-\alpha}\frac{\lambda\mu}{r^2}-\frac{2\alpha^3}{1-(1-\alpha)^3}\frac{\lambda\mu}{r^3}\right) \\
&\approx \frac{\lambda\mu}{r}-\frac{2(2-\alpha)\alpha^3}{(r(2-\alpha)+\alpha)(1-(1-\alpha)^3)}\frac{\lambda\mu}{r^2}
\end{aligned}
\qquad (23')
$$

Nevertheless, the SBA estimates can significantly improve upon Croston's method for low values of α. In Figure 2 we show the behavior of the bias of Croston's method and SBA with respect to α for λ = 1, μ = 1, 5 and r = 2, 5. The first four terms of the Taylor series are considered to calculate the bias (equations (18') and (23')). Please note that scale of the y-axis differs between the illustrative examples in order to facilitate the better presentation of the results.

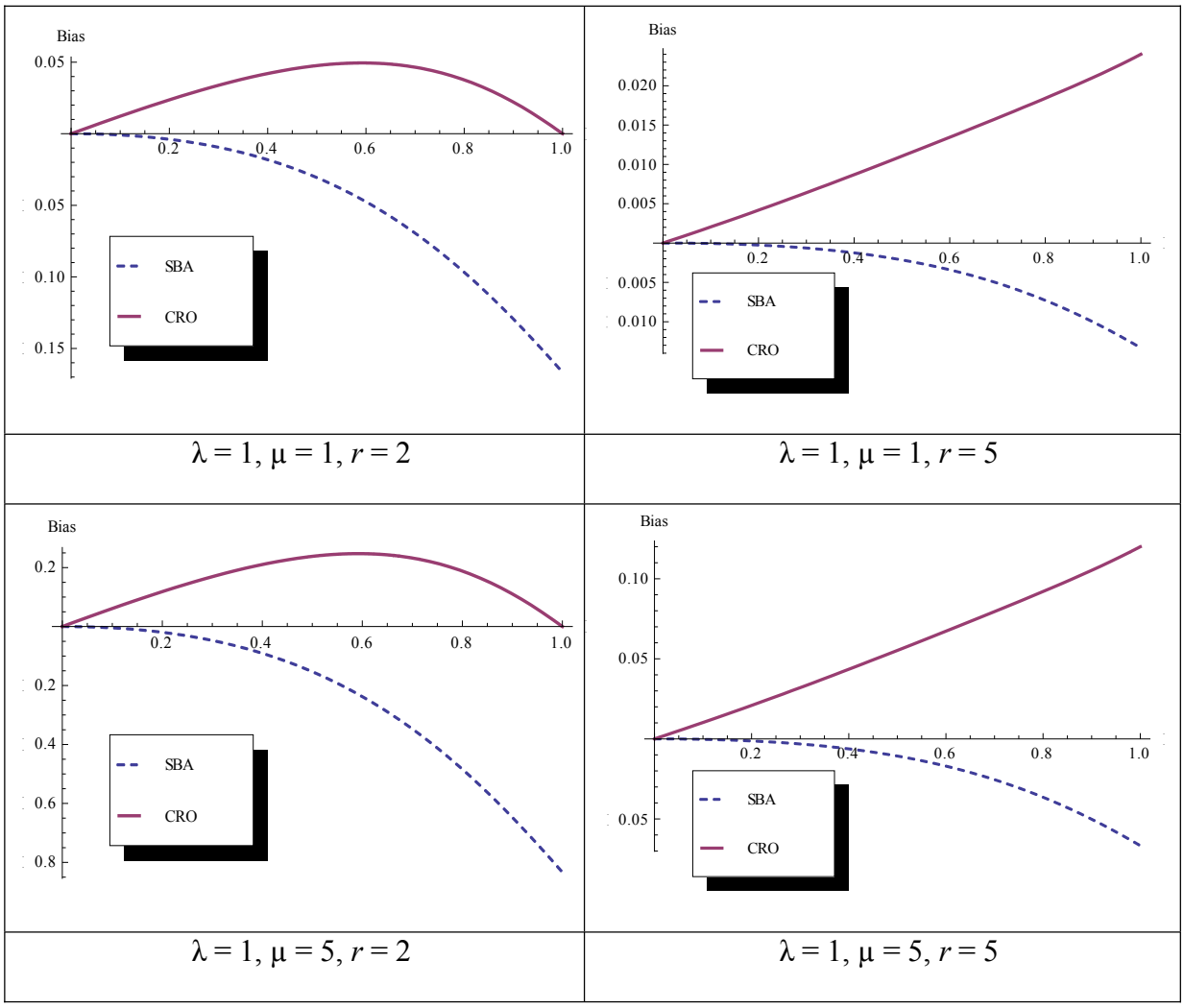


Figure 2: The approximately unbiased nature of the SBA estimator for low values of α

Figure 2 shows clearly that the bias of the SBA estimator is very small as compared to that of Croston for $0 < \alpha \leq 0.2$. This is the typical range of values used in practice. The bias decreases

as the mean demand size decreases and/or the number of Erlang stages increases, which is expected as in both cases the demand per time period decreases. In the next section, a theoretical MSE comparison is undertaken between the three estimators and, through a numerical investigation, we show the areas where each estimator outperforms the others. For the purposes of our analysis a case is being made for the use of low smoothing constant values (≤ 0.2). Given the results presented above, expressions (18) and (23) are being used for $E(y_t'(\text{CRO}))$ and $E(y_t'(\text{SBA}))$, respectively.

4. NUMERICAL INVESTIGATION AND INSIGHTS

In this section we undertake a comparison of the MSE performance of EWMA, Croston and SBA. We first list the sum of the variance and squared bias for each of those estimators.

$$\begin{aligned}&\text{Var}(Estimate)_{\text{EWMA-ALL}} + \text{Bias}^2_{\text{EWMA-ALL}} \\ &= \frac{\alpha}{2-\alpha}\left[\frac{\lambda}{r}\sigma^2 + \left(\frac{\lambda}{r^2} + \frac{r^2-1}{6r^2} + \frac{\theta_r}{r^2}\right)\mu^2\right]\end{aligned} \tag{25}$$

$$\begin{aligned}&\text{Var}(Estimate)_{EWMA-ISSUE} + Bias^2_{EWMA-ISSUE} \\ &= \alpha^2\sigma^2 + \frac{\alpha(1-\alpha)^2}{2-\alpha}\left[\frac{\lambda}{r}\sigma^2 + \left(\frac{\lambda}{r^2} + \frac{r^2-1}{6r^2} + \frac{\theta_r}{r^2}\right)\mu^2\right] + (1-\frac{\lambda}{r})^2\alpha^2\mu^2\end{aligned} \tag{26}$$

$$\begin{aligned}&\text{Var}(Estimate)_{\text{CRO}} + \text{Bias}^2_{\text{CRO}} \\ &\approx \left[\frac{\alpha}{2-\alpha} + \frac{1}{r}(\frac{\alpha}{2-\alpha})^2\right]\frac{\lambda^2\sigma^2}{r^2} + \left[\frac{\alpha}{2-\alpha} - \frac{1}{r}\frac{4\alpha^3}{1-(1-\alpha)^3} + \frac{r+2}{r^2}\frac{3\alpha^4}{1-(1-\alpha)^4}\right]\frac{\lambda^2\mu^2}{r^3} + \left(\frac{\alpha}{2-\alpha}\frac{\lambda\mu}{r^2}\right)^2\end{aligned} \tag{27}$$

$$\begin{aligned}&\text{Var}(Estimate)_{SBA} + Bias^2_{SBA} \\ &\approx \left(1 - \frac{\alpha}{r(2-\alpha)+\alpha}\right)^2 \text{Var}(y_t'(\text{CRO})) \\ &\approx \left(1 - \frac{\alpha}{r(2-\alpha)+\alpha}\right)^2 \left\{\left[\frac{\alpha}{2-\alpha} + \frac{1}{r}(\frac{\alpha}{2-\alpha})^2\right]\frac{\lambda^2\sigma^2}{r^2} + \left[\frac{\alpha}{2-\alpha} - \frac{1}{r}\frac{4\alpha^3}{1-(1-\alpha)^3} + \frac{r+2}{r^2}\frac{3\alpha^4}{1-(1-\alpha)^4}\right]\frac{\lambda^2\mu^2}{r^3}\right\}\end{aligned} \tag{28}$$

Considering (27) and (28) it can be seen by inspection that the inequality $MSE_{\text{CRO}} > MSE_{\text{SBA}}$ always holds, so the SBA method always outperforms the CRO method for all α, r and λ. In that respect there is no need to further consider the performance of CRO against EWMA.

It should be noted that this is an interesting result. Unlike the Bernoulli case where Croston's method performs better than SBA for low inter-demand interval values and low CV^2 values, in this case SBA always performs better than Croston. This is because under the censored Poisson

process (and when considering the first three terms in a Taylor series) SBA is approximately unbiased whereas in the Bernoulli case this method is associated with a small negative bias. Furthermore, SBA has a strictly lower sampling error of the mean as compared to Croston.

Before we conduct the comparisons, we should mention that the choice of the smoothing constant controls the weight assigned to the error of the estimates and, as might be expected, has an effect on the performance of forecasting procedures. In general, α will be somewhere between 0.05 and 0.2. Software packages often use this range of values as the default one. Examples include *RightStock* (DBO Services), *Slim4* (SlimStock) etc (Dawson 2013). Burgin and Wild (1967) found that $\alpha = 0.2$ is suitable for most weekly data but they implicitly recommended lower α values for slow movers. Croston (1972) recommended the use of α values in the range 0.05~0.2, when demand is intermittent. More recently, Syntetos and Boylan (2005) also commented on the empirical relevance of the range 0.05~0.2 which is the one used for this study as well, in increments of 0.05. This is viewed as a realistic assumption from a practitioner perspective.

4.1 Periodic Review Systems: All Points in Time

First we compare the performance of SBA and EWMA for the case of all points in time.

$MSE_{\text{EWMA-ALL}} > MSE_{\text{SBA}}$ if and only if:

$$[r(2-\alpha)+\alpha-(2-\alpha)\lambda]\sigma^2 > \left\{\frac{(2-\alpha)^3}{r(2-\alpha)+\alpha}\left[\frac{1}{2-\alpha}-\frac{1}{r}\frac{4\alpha^2}{1-(1-\alpha)^3}+\frac{r+2}{r^2}\frac{3\alpha^3}{1-(1-\alpha)^4}\right]\lambda-\frac{r(2-\alpha)+\alpha}{r}\left(1+\frac{1}{\lambda}\left(\frac{r^2-1}{6}+\theta_r\right)\right)\right\}\mu^2 \quad (29)$$

We discuss in detail (for demonstration purposes) the case of $r = 2$ and $\alpha = 0.15$ followed by summary results across other control parameter combinations.

In the case of $r = 2$, equation (29) becomes:

$$[4-\alpha-(2-\alpha)\lambda]\sigma^2 > \left[\frac{(2-\alpha)^3}{4-\alpha}\left(\frac{1}{2-\alpha}-\frac{2\alpha^2}{1-(1-\alpha)^3}+\frac{3\alpha^3}{1-(1-\alpha)^4}\right)\lambda-\frac{4-\alpha}{2}\left(1+\frac{e^{-\lambda}\sinh\lambda}{\lambda}\right)\right]\mu^2. \quad (29')$$

For $\alpha = 0.15$, when $0 \le \lambda \le 2.08$, the left-hand-side (LHS) of inequality (29′) is positive but the right-hand-side (RHS) can be shown numerically to be negative, which indicates that inequality (29′) always holds. So in this case, SBA outperforms the *EWMA-ALL* method.

When $\lambda > 2.08$, the LHS of inequality (29′) is strictly negative, resulting in:

$$\frac{\sigma^2}{\mu^2} < \frac{\frac{(2-\alpha)^3}{4-\alpha}\left(\frac{1}{2-\alpha}-\frac{2\alpha^2}{1-(1-\alpha)^3}+\frac{3\alpha^3}{1-(1-\alpha)^4}\right)\lambda-\frac{4-\alpha}{2}\left(1+\frac{e^{-\lambda}\sinh\lambda}{\lambda}\right)}{4-\alpha-(2-\alpha)\lambda}. \quad (30)$$

For $\lambda > 3.06$, the RHS of inequality (30) can be shown numerically to be strictly negative, so inequality (30) does not hold, which indicates that the *EWMA-ALL* method outperforms the SBA method. For $2.08 \le \lambda \le 3.06$, the RHS of inequality (30) can be shown numerically to be positive, then there exists a CV^2 cut off value, below which the SBA method performs better than the *EWMA-ALL* method and above which the opposite is the case. The CV^2 cut off value depends on the value of λ. If we denote the RHS of inequality (30) as $f_1(\lambda)$ and let λ_1 and λ_2 be the extreme cut-off values of CV^2 (2.08 and 3.06 respectively), then for a given $\lambda_1 \le \lambda \le \lambda_2$, the corresponding CV^2 cut off value will be $f_1(\lambda)$. Some examples are listed in Table 1 below.

Table 1: CV² cut-off values for λ between λ_1 and λ_2

λ values	$\lambda_1 = 2.08$	2.1	2.2	2.4	2.6	2.8	2.9	3	$\lambda_2 = 3.06$
CV² cut-off values	∞	23.98	3.39	0.96	0.41	0.16	0.09	0.03	0

It should be noted that, as shown in Figure 3, the CV^2 cut-off values decrease as the value of λ increases. The cut-off value becomes equal to zero when $\lambda = \lambda_2 = 3.06$ and it is asymptotically increasing when λ decreases and tends toward $\lambda_1 = 2.08$.

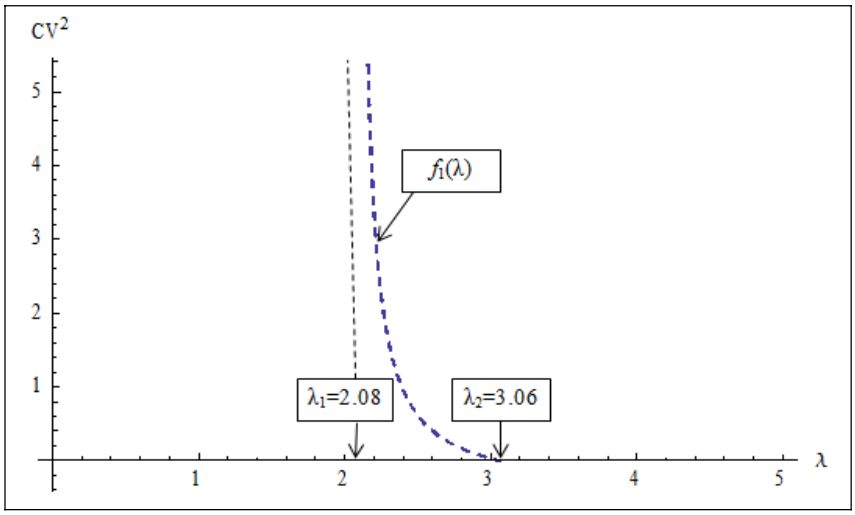


Figure 3: Cut-off values, SBA-EWMA (all point in time), $r = 2$, $\alpha = 0.15$

In Table 2 we present the values of λ_1 and λ_2 (to the second decimal place), for r ranging from 1 to 7 and α from 0.05 to 0.2 (step 0.05).

The results in Table 2 show that the cut-off values increase with r (and moderately with α) which indicate an improvement in the comparative performance of the SBA estimator as these values increase. The increased superiority of the SBA estimator with r can be explained in terms of the increasing degree of intermittence associated with higher values of the shape parameter of the Erlang distribution. Note also that for the extreme case of a process that tends towards a

deterministic state, which corresponds to the case of very high values of *r*, the EWMA estimator suffers from a large variance as opposed to the SBA estimator for which the variance tends towards zero. Furthermore, the cut-off points are not particularly sensitive to variations of the α values which can be explained by the small variance of the SBA estimator for $0 < \alpha \leq 0.2$.

Table 2: Cut-off values, SBA-EWMA (all point in time), for different values of *r* and α

r	Cut-off values	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.2$
$r = 1$	λ_1	1.03	1.05	1.08	1.11
	λ_2	1.20	1.44	1.70	1.96
$r = 2$	λ_1	2.03	2.05	2.08	2.11
	λ_2	2.61	2.83	3.06	3.30
$r = 3$	λ_1	3.03	3.05	3.08	3.11
	λ_2	4.20	4.42	4.64	4.87
$r = 5$	λ_1	5.03	5.05	5.08	5.11
	λ_2	7.83	8.06	8.28	8.51
$r = 7$	λ_1	7.03	7.05	7.08	7.11
	λ_2	11.98	12.21	12.45	12.68

***Remark* 3:** In order to compare the above results with existing schemes for the Bernoulli process (Syntetos et al. 2005), we could transfer the values of λ to those of *p*, the mean of the geometrically-distributed inter-demand intervals for a Bernoulli process. This could be readily done by using the relationship between λ and p: $p = 1 \Big/ \left(1 - (1 + \frac{\lambda}{2})e^{-\lambda}\right)$, which is obtained by letting the probability that demand occurs in a single period under an Erlang order arrival process (Cox 1962), $1 - (1 + \frac{\lambda}{2})e^{-\lambda}$, equal to the corresponding probability under a Bernoulli process, $1/p$. The larger the value of λ is, the smaller the corresponding value of *p* will be. For particular values of λ: 2.08 and 3.06, for instance, the corresponding p values are approximately 1.34 and 1.13, respectively. Accordingly, the above result can be stated as well on *p*: for $\alpha = 0.15$, when $p > 1.34$, the SBA outperforms the *EWMA-ALL* method; when $p < 1.13$, the *EWMA-ALL* method outperforms the SBA; when $1.13 \leq p \leq 1.34$, there exists a CV^2 cut off value, below which the SBA performs better than the EWMA-ALL method and above which the opposite is the case.

4.2 Continuous Review Systems: Issue Points Only

Similar analysis to that presented above shows that in the case of $r = 2$ and α = 0.15, the SBA outperforms the *EWMA-ISSUE* method for any $\lambda < 2.06$ and when $2.06 \leq \lambda \leq 2.34$ for $CV^2 \leq f_2(\lambda)$ where:

$$f_2(\lambda)=\frac{\frac{2(2-\alpha)^3}{4-\alpha}\left(\frac{1}{2-\alpha}-\frac{2\alpha^2}{1-(1-\alpha)^3}+\frac{3\alpha^3}{1-(1-\alpha)^4}\right)-\alpha(2-\alpha)(4-\alpha)(2-\lambda)^2-(4-\alpha)(1-\alpha)^2(\lambda+e^{-\lambda}\sinh\lambda)}{4\alpha(2-\alpha)(4-\alpha)+2(1-\alpha)^2(4-\alpha)\lambda-2(2-\alpha)\lambda^2} \quad (31)$$

In Figure 4, we show the CV^2 cut-off values with respect to λ for $r=2$ and $\alpha=0.15$, i.e. the function $f_2(\lambda)$ in the case of *issue points only*. The results show the same behavior as in the case of *all points in time*. However, λ_1 and λ_2 are much closer in this case.

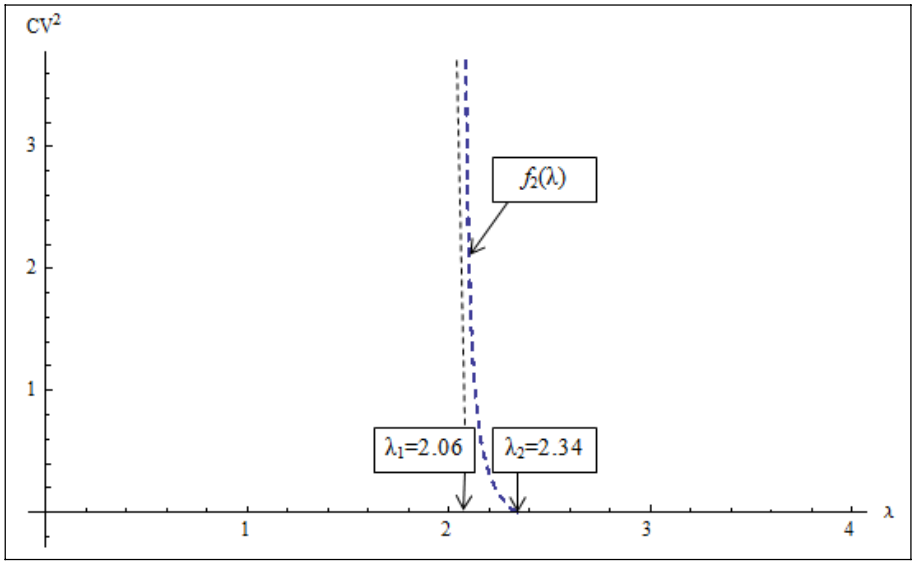


Figure 4: Cut-off values, SBA-EWMA (issue points only), $r=2$, $\alpha=0.15$

The values (to the second decimal place) of λ_1 and λ_2 for r ranging from 1 to 7 and α from 0.05 to 0.2 (step 0.05), when issue points only are considered, are given in Table 3.

Table 3: Cut-off values, SBA-EWMA (issue points only), for different values of r and α

r	Cut-off values	$\alpha=0.05$	$\alpha=0.1$	$\alpha=0.15$	$\alpha=0.2$
$r=1$	λ_1	1.02	1.04	1.06	1.08
	λ_2	1.11	1.20	1.28	1.32
$r=2$	λ_1	2.02	2.04	2.06	2.08
	λ_2	2.40	2.39	2.34	2.27
$r=3$	λ_1	3.02	3.04	3.06	3.08
	λ_2	4.84			
$r=5$	λ_1	5.02	5.04	5.06	5.08
	λ_2				
$r=7$	λ_1	7.02	7.04	7.06	7.08
	λ_2				

For low values of r, the results show a very similar behavior to that discussed for all points in time, with the only exception that λ_1 and λ_2 do not differ by much. However, when $r\geq 3$, the behavior of the cut-off values is different as the CV^2 cut-off value never becomes zero but rather

there is an asymptotic value below which the SBA estimator outperforms the EWMA for all values of λ. Note that in this case, as in the case of all points in time, the cut-off points are not particularly sensitive to variations of the α values.

The case of $r = 5$ and $\alpha = 0.15$ is pictorially presented in Figure 5. The results clearly show that when r is high, demand becomes more intermittent and the superiority of the SBA estimator increases as compared to the EWMA estimator when issue points only are considered. This can be explained in terms of the bias of the EWMA estimator in the case of issue points only, as shown by (12), which increases with r. Note also that the variance of the EWMA estimator persists even for the extreme case of a deterministic process, i.e. infinite number of stages in the Erlang process.

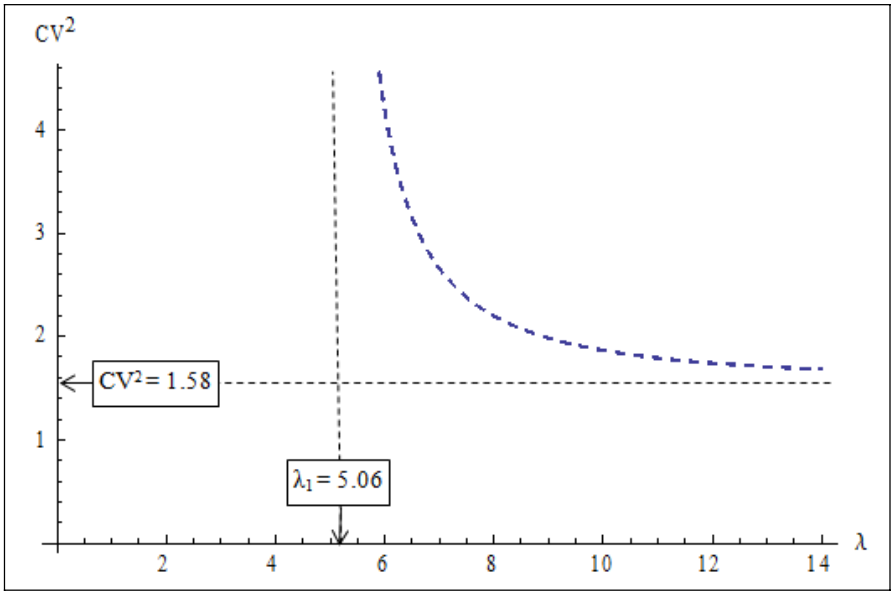


Figure 5: Cut-off values, SBA-EWMA (issue points only), $r = 5$, $\alpha = 0.15$

The asymptotic CV^2 cut-off values (i.e. the CV^2 cut-off values below which the SBA estimator outperforms EWMA for all values of λ) for $r = 3,5,7$ and the α values considered in this research are given in Table 4.

Table 4: Asymptotic CV^2 cut-off values, SBA-EWMA (issue points only)

	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.2$
$r = 3$	0.00	0.14	0.36	0.57
$r = 5$	0.42	1.01	1.58	2.12
$r = 7$	1.06	2.21	3.31	4.35

The results in the case of issue points only show that under the censored Poisson process when r increases, i.e. the average order interval increases, the regions of the superior comparative

performance of SBA and EWMA can be approximated by a simple 4-quadrants solution that is easy to operationalise. This result resembles previously suggested solutions for the Bernoulli case. A censored Poisson process, considering issue points only and for high degrees of intermittence, tends toward the discrete case of the Bernoulli process.

4.3 Operationalisation of the rules

The analysis conducted in sections 4.1 and 4.2 for the case of all points in time and issue points only, respectively, leads naturally to the development of demand classification solutions to be used for forecasting purposes.

For inventory systems operating under a periodic review (all points in time) and for a particular degree of determinism inherent in the system, the rate of transaction occurrence is a useful classification criterion to distinguish between intermittent and non-intermittent demand. This is natural and in line with previous findings under the assumption of a Bernoulli arrival process. The analysis indicates the degree of intermittence that signifies a difference in the performance of estimators built for sporadic demand (SBA) and methods developed for fast demand items (EWMA). The variability associated with the order sizes also has an explanatory power with respect to performance differences but much less so, and in a way that is difficult to capture in simple operational rules.

It seems plausible at this stage to experiment with the accuracy loss resulting from categorizing demand based on the transaction rate only, i.e. not considering the CV^2 as a classification parameter. Operational considerations (e.g., ease of application, and communication of a solution to practitioners) and intuitive appeal suggest that a classification solution consisting only of the transaction rate may be preferable, as long as the forecast accuracy losses resulting from such an approximate solution are not 'considerable'. An additional justification for such an approach is the fact that the cut-off points presented here are anyway the result of an approximate rather than exact analysis.

We present below the accuracy losses resulting from categorizing demand based on the λ_1 values only. We analyse the effect of allocating the range $\lambda_1 < \lambda \leq \lambda_2$ to EWMA, when clearly performance in this area is a function of the CV^2 and either estimator could perform better

depending on the control parameter combinations. To do so, in Table 5 we report for $\lambda = (\lambda_1 + \lambda_2)/2$, the MSE percentage error ($(MSE_{EWMA-ALL} - MSE_{SBA})/MSE_{SBA}$) resulting from the sole consideration of the EWMA for CV^2 values between 0 and $f_1(\lambda)$.

Table 5: Accuracy loss resulting from the introduction of a simple rule for periodic review systems

	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.2$
$r = 1$	2.61%	4.39%	5.48%	6.06%
$r = 2$	2.99%	3.46%	3.81%	4.06%
$r = 3$	3.11%	3.28%	3.41%	3.51%
$r = 5$	3.02%	3.05%	3.07%	3.08%
$r = 7$	2.89%	2.89%	2.89%	2.89%

The results indicate that a classification solution based on the transaction rate only is a very good trade-off between operationalisation and accuracy, as the average error in the mid value of λ in $[\lambda_1; \lambda_2]$ remains relatively low. The rate λ_1 signifies the degree of transaction occurrence, below which ($\leq$) SBA should be used, and above which (>) the EWMA should be the preferred estimator. It should also be noted that, based on the values of λ for which the LHS of (29) is positive, the expression of the cut-off value λ_1 can be easily computed for any values of r and α as follows:

$$\lambda_1 = r + \frac{\alpha}{(2-\alpha)} \tag{32}$$

The exclusion of the CV^2 from the classification solution is not in accordance with previous research, under the assumption of a Bernoulli process (Syntetos et al. 2005). However, it does explain (partly at least) the inconclusive results reached by Boylan et al. (2008) on the empirical validity and utility of the CV^2 as a classification parameter when analyzing the performance of an inventory control software package on about 16,000 intermittent demand SKUs.

The suggested λ_1 values may not be exact (by definition due to the approximate nature of our analysis) but they would constitute a very good starting point for experimentation in a real world system. In the Bernoulli case, Boylan et al. (2008) showed the insensitivity of empirical performance to the exact cut-off value of the inter-demand interval in the range 1.2 - 1.9 periods.

In the case of issue points only, an operationalised classification solution would take the form of schemes previously developed under the Bernoulli case. We analyse the accuracy loss resulting from the introduction of a simple 4-quadrant classification solution where the top right quadrant is allocated to the EWMA. To do so, in Table 6 we report for $r \le 2$ and $\lambda = (\lambda_1 + \lambda_2)/2$, the average MSE percentage error ($(MSE_{EWMA-ISSUE} - MSE_{SBA})/MSE_{SBA}$) resulting from the sole consideration of the EWMA for CV^2 values between 0 and $f_1(\lambda)$. For $r \ge 3$, we report for $\lambda = (\lambda_1 + \lambda_2)/2$ the average MSE percentage error resulting from the sole consideration of the EWMA for CV^2 values between the asymptotic value of CV^2 and $f_1(\lambda)$. However, the value of λ_2 is determined here by fixing a 'reasonable' maximum value of λ such that the average demand arrival is $\lambda / r = 5$. The results in Table 6 indicate that the simple 4-quadrant classification is a good trade-off between operationalisation and accuracy, as the average error is relatively low.

Table 6: Accuracy loss resulting from the introduction of a simple rule for continuous review systems

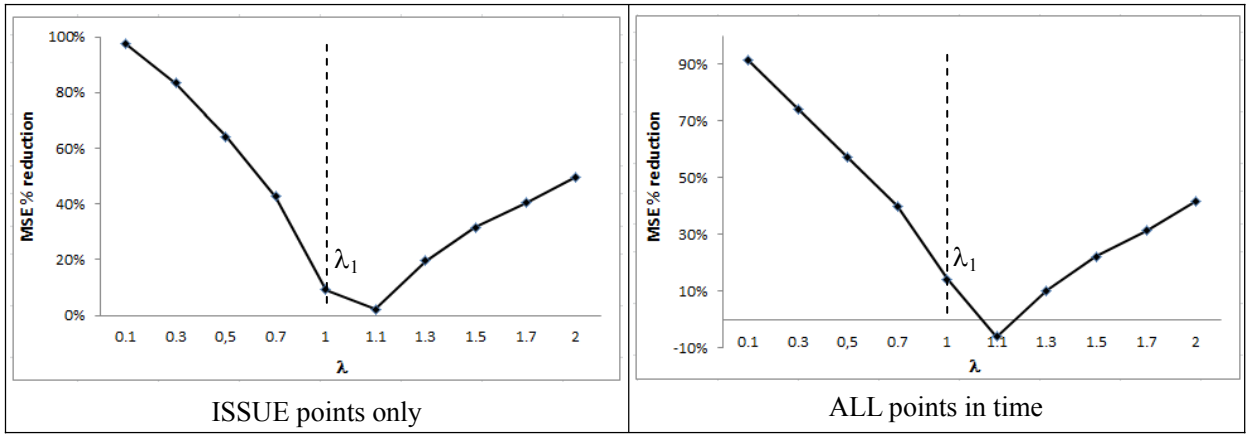
	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.2$
$r = 1$	1.35%	2.20%	2.63%	2.68%
$r = 2$	1.76%	1.23%	0.81%	0.46%
$r = 3$	1.20%	0.15%	0.42%	0.79%
$r = 5$	0.34%	0.28%	0.14%	0.04%
$r = 7$	1.70%	2.17%	2.11%	1.76%

Note also that the cut-off value λ_1 and the asymptotic cut-off value CV^2 can be easily calculated for any values of *r* and *α* as follows:

$$\lambda_1 = \frac{(1-\alpha)^2\left[r(2-\alpha)+\alpha\right] + \sqrt{\left[(1-\alpha)^2\left[r(2-\alpha)+\alpha\right]\right]^2 + 4(2-\alpha)\left[r\alpha(2-\alpha)\left[r(2-\alpha)+\alpha\right]\right]}}{2(2-\alpha)} \tag{33}$$

$$CV^2 = \frac{\frac{\alpha}{4} - \frac{1}{r^3}\left(1 - \frac{\alpha}{r(2-\alpha)+\alpha}\right)^2\left(\frac{1}{2-\alpha} - \frac{1}{r}\frac{4\alpha^2}{1-(1-\alpha)^3} + \frac{r+2}{r^2}\frac{3\alpha^3}{1-(1-\alpha)^4}\right)}{\frac{1}{r^2}\left(1 - \frac{\alpha}{r(2-\alpha)+\alpha}\right)^2\left(\frac{1}{2-\alpha} + \frac{1}{r}\frac{\alpha}{(2-\alpha)^2}\right)} \tag{34}$$

We close this Section with an indication of how large the MSE percentage reductions resulting from the choice of the right method may be in cases other than the borderline ones discussed above. We present two graphs where we show the MSE percentage reductions resulting from using the 'correct' estimator (either SBA or EWMA) as opposed to the alternative. We fix μ, σ and r and we show, for all points in time, the (MSE_{EWMA} - MSE_{SBA}) / MSE_{EWMA} for 5 values of λ on the left hand side of λ_1; similarly we show the (MSE_{SBA} - MSE_{EWMA}) / MSE_{SBA} for another 5 values of λ on the right hand side of λ_1. We do the same for the case of issue points only distinguishing between the two demand categories through the choice of appropriate λ values. More extensive results are included in *Appendix C* presented as part of our electronic companion.



ISSUE points only | ALL points in time

Figure 6. MSE percentage reductions resulting from the choice of the right estimator (SBA, EWMA): $r = 1$, $\mu = 1$, $\sigma = 2$, $\alpha = 0.15$

The percentage differences are considerable, especially for λ values much higher or lower than the cut-off value λ_1. Some (moderately) negative MSE percentage reductions appear only, as expected, for λ values closely positioned to the λ_1 value, i.e. in the disputed areas, the case of which has been discussed in detail in this section.

5. IMPLICATIONS AND CONCLUSION

Henry L. Hinton, Jr, Assistant Comptroller General, National Security and International Affairs Division (GAO 1999) stated: "*Our work continues to show weaknesses in DOD's inventory management practices that are detrimental to the economy*, p. 1". Thirteen years later only minor improvements were reported (GAO 2012) and public announcements on the poor management of

defense inventories and the resulting detrimental impacts on the economy constitute a recurring issue in the news (e.g., BBC 2013). Similarly, the expansion of the after-sales industry and the increasing importance of commercial service operations have not been reflected in the development of ERP and supply chain software packages the functionality of which has often been judged to be inadequate. New scientific research is currently required to inform the inventory improvement programs in the after sales and MRO (Maintenance, Repair and Overhaul) environment and military sector. Service parts are characterized by intermittent demand and any inventory research in this area would also have direct implications for the slower moving (*C*) items in any industrial setting.

Previous work in the area of intermittent demand forecasting was developed upon the assumption of memory-less demand processes (Bernoulli and Poisson). Herniter (1970) noted that an individual's inter-demand pattern is random only if the probability of purchasing in any time interval of given length is constant. Clearly this assumption may not be sustained in a spare parts context or in a more general consumer behavior sense. In the former case, from a reliability engineering perspective, non memory-less distributions may naturally drive demand. In the latter case, a plethora of real world factors such as the very composition of markets, correlation amongst requests etc also signify a departure from the memory-less assumption.

In this work we assumed that demand is driven by a censored Poisson process (that contains Poisson as a special case) resulting in Erlang distributed inter-demand times. The statistical properties of three plausible demand estimators (EWMA, Croston and SBA) were analyzed and a comparison of their theoretical performances resulted in the proposition of some classification solutions that may be used for the purpose of managing the forecasting task.

Demand categorization has received very limited attention in the academic literature despite its importance for many inventory management computerized applications. Managers typically rely upon ad hoc classification rules both with regards to the criteria being used and the cut-off values assigned to them. However, the analysis conducted here suggests conclusively that the degree of intermittence and, in certain cases, the variability associated with the order sizes, are useful classification parameters; both criteria do not only have intuitive appeal but they are also associated with strong theoretical support. Simple classification schemes were presented, easy enough to apply directly in real world inventory control systems, and theoretically-derived cut-off values were proposed that would enable inventory managers to put our propositions in practice.

The results demonstrate the lack of sensitivity of the arrival rate λ cut-off points proposed to the smoothing constant value being used. For re-order level (continuous review) systems the CV^2 cut-off points do show some sensitivity though to the α value. The degree of determinism inherent in the system appears also to affect considerably the comparative performance of forecasting methods. As the number of stages in the Erlang process increases the comparative performance of SBA versus EWMA improves. This is due to the variance inherent in the EWMA estimator which is present even under a deterministic process. Moreover, in the case of issue points only, this is amplified by the positive bias associated with the EWMA estimator. In contrast, the SBA estimator is approximately unbiased and its variance tends towards zero for high values of *r*. In real world applications, simulation with empirical data could lead to the determination of the appropriate *r* value.

This paper advances knowledge in the area of intermittent demand forecasting by relaxing one very important assumption made in previous theoretical studies that related to a memory-less demand process. However, much more needs to be done with regards to the remaining assumptions in the model formulation which admittedly are as important as the one discussed above for demand estimation purposes. The independence assumption underlying all the work presented here (but also most other major contributions in this area) is a key one to be challenged to enable further important developments.

Regardless, the findings of this paper are of a great value to complement the growing body of the inventory control literature that has been developed upon the assumption of a censored Poisson demand process (see, for example, Larsen et al. 2008, Larsen and Thorstenson 2008).

The theoretical case for the Erlang inter-order interval distribution, for an individual consumer, is sound. However, it should be noted that Erlang inter-order intervals at the individual consumer level do not necessarily lead to Erlang purchases at the total consumer/customer level. That is, theoretically at least, the order intervals experienced by a stockist may not be Erlang distributed. Nevertheless, empirical arguments do support this assumption which appears to be intuitively appealing. The empirical validation of the results presented here is a natural next step of research.

Notes

Note 1 DOD defines secondary inventory items to include reparable components, subsystems, and

assemblies other than major end items (e.g., ships & aircrafts), consumable repair parts, bulk items, subsistence, and expendable end items.

Note 2 A practical inventory management application in the defense sector is described by Scala et al. (2013). Relevant studies demonstrate the tremendous scope for improving the control of defense inventories.

Note 3 In continuous review inventory control it is only necessary to consider making replenishment decisions just after a demand has occurred. This is true for Poisson processes, where the time between demands is exponentially distributed, and hence the demand process is associated with the memory-less property from which the above fact follows. However, for renewal demand processes, including the Erlang arrival processes studied in this paper, this is no longer true, in general. Rather, for these processes the passage of time itself may carry information about the demand process. Thus, it may be optimal that a certain time span should trigger a replenishment order, even if a demand has not occurred. Therefore, an order may not only be triggered by a change in the inventory position (defined in the usual way). Heuristically, and for practical purposes, replenishment orders may, of course, be allowed only at the time instances just after a demand has occurred (or at predetermined time intervals, as in a periodic review system). This issue has important implications for the kind of information that is useful for inventory control purposes but is not further pursued in this study.

Note 4 SKU classification for forecasting purposes typically works in the opposite way. That is, ad hoc classification rules are being used to separate SKUs into categories, followed by the specification of a forecasting method for each of the categories. But if the purpose of classification is the selection of forecasting methods, then it makes more sense to compare directly possible estimators and then categorize demand based on regions of superior forecast performance.

Note 5 The calculation of θ_r for high values of r results in complex numbers. As the imaginary part of these numbers is negligible, only the real part is considered in the numerical calculations conducted in this paper.

Acknowledgements

We would like to thank the three anonymous referees for their comments that greatly helped to improve the content of our paper and its presentation.

References

Altay, N., L.A. Litteral. 2011. Service parts management: demand forecasting and inventory control (Handbook), Springer-Verlag, New York.

Axsäter, S. 2006. Inventory control, Springer, New York.

Bartezzaghi, E., R. Verganti, G. Zotteri. 1999. A simulation framework for forecasting uncertain lumpy demand, International Journal of Production Economics, 59 (1-3), 499-510.

BBC (British Broadcasting Corporation). 2013. Communication available at: http://www.bbc.co.uk/news/uk-politics-21608747 (February 28).

Boylan, J.E., A.A. Syntetos. 2010. Spare parts management: a review of forecasting research and extensions, IMA Journal of Management Mathematics, 21 (3), 227-237.

Boylan, J.E., A.A. Syntetos, G.C. Karakostas. 2008. Classification for forecasting and stock control: A case study, Journal of the Operational Research Society, 59 (4), 473-481.

Brown, R.G. 1959. Statistical forecasting for inventory control, McGraw-Hill, New York.

Brown, R.G. 1963. Smoothing, forecasting and prediction of discrete time series, Prentice- Hall, Englewood Cliffs, New Jersey.

Burgin, T.A., A.R. Wild. 1967. Stock control experience and usable theory, Operational Research Quarterly, 18 (1), 35-52.

Chatfield, C., G.J. Goodhardt. 1973. A consumer purchasing model with Erlang inter-purchase times, Journal of the American Statistical Association, 40 (344), 828-835.

Chew, E.P., L.A. Johnson. 2006. Service levels in distribution systems with random customer order size, Naval Research Logistics, 42 (1), 39-56.

Cox, D.R. 1962. Renewal theory, Chapman and Hall, London.

Croston, J.D. 1972. Forecasting and stock control for intermittent demands, Operational Research Quarterly 23 (3), 289-303.

Dawson, P. 2013. Managing Director, DBO Services (http://www.dboservices.com/). Private communications to the corresponding author.

Deloitte. 2006. The service revolution in global manufacturing industries. New York: Deloitte Research, 2006. Available at: http://www.apec.org.au/docs/2011-11_training/deloitte2006.pdf

DMO (Defence Materiel Organisation). 2010. Australian Government, Department of Defence, Fact Sheet, Available at: http://www.defence.gov.au/dmo/asr/ss/Fact_sheet_Feb10.pdf

Eaves, A.H.C., B.G. Kingsman. 2004. Forecasting for the ordering and stockholding of spare parts, Journal of the Operational Research Society, 55 (4), 431-437.

Fildes, R., K. Nikolopoulos, S. Crone, A.A. Syntetos. 2008. Forecasting and operational research: A review, Journal of the Operational Research Society, 59 (9), 1150-1172.

GAO (United States General Accounting Office). 1999. Defense inventory: continuing challenges in managing inventories and avoiding adverse operational effects, GAO/T-NSIAD-99-83, Washington, D.C., February 25. Available at: http://www.gao.gov/products/GAO/T-NSIAD-99-83

GAO (United States General Accounting Office). 2011. DOD's inventory management improvement plan, GAO-11-240R, Washington, D.C., Jan.07. Available at: http://www.gao.gov/products/GAO-11-240R

GAO (United States General Accounting Office). 2012. Defense inventory: actions underway to implement improvement plan, but steps needed to enhance effort, GAO-12-493, Washington, D.C., May 03. Available at: http://www.gao.gov/products/GAO-12-493

Gardner, E.S., Jr. 2011. Forecasting for operations, Keynote, International Symposium on Forecasting (ISF), Prague, Czech Republic.

Ghobbar, A.A. and C.H. Friend. 2002. Sources of intermittent demand for aircraft spare parts within airline operations, Journal of Air Transport Management, 8 (4), 221-231.

Graves, S.C. 1999. A single-item inventory model for a nonstationary demand process, Manufacturing and Service Operations Management, 1 (1), 50-61.

Herniter, J.D. 1970. Probabilistic market models of purchase timing and brand selection, Working Paper, Marketing Science Institute, Cambridge, Mass.

Johnston, F.R., J.E. Boylan. 1996. Forecasting for items with intermittent demand, Journal of the Operational Research Society, 47 (1), 113–121.

Johnston, F.R., J.E. Boylan, E.A. Shale. 2003. An examination of the size of orders from customers, their characterization and the implications for inventory control of slow moving items, Journal of the Operational Research Society, 54 (8), 833-837.

Larivière, M.A., J.A. van Mieghem. 2004. Strategically seeking service: How competition can generate Poisson arrivals, Manufacturing and Service Operations Management, 6 (1), 23-40.

Larsen, C. 2008. A note on Poisson and Erlang processes, Working paper, Cluster of Operational Research and Logistics (CORAL), Department of Economics and Business, Aarhus University, Denmark.

Larsen, C., H.G. Seiding, C. Teller, A. Thorstenson. 2008. An inventory control project in a major Danish company using compound renewal demand models, IMA Journal of Management Mathematics, 19 (2), 145-162.

Larsen, C., A. Thorstenson. 2008. A comparison between the order and the volume fill rate for a base-stock inventory control system under a compound renewal demand process, Journal of the Operational Research Society, 59 (6), 798-804.

Morse, A. 2012. Ministry of defence: managing the defence inventory, National Audit Office (NAO), UK, 2012. Available at: http://www.nao.org.uk/report/managing-the-defence-inventory/

Regattieri, A., M. Gamberi, R. Gamberini, R. Manzini. 2005. Managing lumpy demand for aicraft spaare parts, Journal of Air Transport Management, 11 (6), 426-431.

Ross, S.M. 1996. Stochastic Processes (2nd ed.), John Wiley & Sons, Inc, New York.

Scala, N.M., J. Rajgopal, K.L.S. Needy. 2013. A base stock inventory management system for intermittent spare parts, Military Operations Research, 18 (3), 63-77.

Scala, N.M., J. Rajgopal, K.L.S. Needy. 2014. Managing nuclear spare parts inventories: A data driven methodology, IEEE Transactions on Engineering Management, 61 (1), 28-37.

Shale, E.A., J.E. Boylan, F.R. Johnston. 2006. Forecasting for intermittent demand: the estimation of an unbiased average, Journal of the Operational Research Society, 57 (5), 588-592.

Smith M.A.J., R. Dekker. 1997. On the (S − 1, S) stock model for renewal demand processes, Probability in the Engineering and Informational Sciences, 11, 375-386.

Strijbosch, L.W.G., R.M.J. Heuts, E.H.M. van der Schoot. 2000. A combined forecast-inventory control procedure for spare parts, Journal of the Operational Research Society, 51 (10), 1184-1192.

Syntetos, A.A., M.Z. Babai, Y. Dallery, R. Teunter. 2009a. Periodic control of intermittent demand items: Theory and empirical analysis, Journal of the Operational Research Society, 60 (5), 611-618.

Syntetos, A.A., M.Z. Babai, E.S. Gardner Jr. 2014. Forecasting intermittent inventory demands: Parametric methods VS. Bootstrapping, Journal of Business Research, in press.

Syntetos, A.A., J.E. Boylan. 2001. On the bias of intermittent demand estimates, International Journal of Production Economics, 71 (1-3), 457-466.

Syntetos A.A., J.E. Boylan. 2005. The Accuracy of intermittent demand estimates, International Journal of Forecasting, 21 (2), 303-314.

Syntetos, A.A., J.E Boylan. 2006. On the stock-control performance of intermittent demand estimators, International Journal of Production Economics, 103 (1), 36-47.

Syntetos, A.A., J.E. Boylan. 2010. On the Variance of Intermittent Demand Estimates, International Journal of Production Economics, 128 (2), 546-555.

Syntetos, A.A., J.E. Boylan, J.D. Croston. 2005. On the categorization of demand patterns, Journal of the Operational Research Society, 56 (5), 495-503.

Syntetos, A.A., J.E. Boylan, S.M. Disney. 2009b. Forecasting for inventory planning: A 50-year review, Journal of the Operational Research Society, 60 (S1), 149-160.

Teunter, R., A.A. Syntetos, M.Z. Babai. 2011. Intermittent demand: Linking forecasting to inventory obsolescence, European Journal of Operational Research, 214 (3), 606-615.

Trabka, E.A., E.W. Marchand. 1970. Mean and variance of the number of renewals of a censored Poisson process, Biological Cybernetics, 7 (6), 221-224.

Watson, R.B. 1987. The effects of demand-forecast fluctuations on customer service and inventory costs when demand is lumpy, Journal of the Operational Research Society, 38 (1), 75-82.

Ziotopoulou, D. 2009. Review and analysis of information systems NEMES and PYTHIA for intermittent demand time series forecasting, BSc dissertation, National Technical University of Athens, Greece.

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