Available online at www.sciencedirect.com
SCIENCE DIRECT®
International Journal of Forecasting 21 (2005) 303–314
international journal of forecasting
www.elsevier.com/locate/ijforecast

The accuracy of intermittent demand estimates

Aris A. Syntetos[a,*], John E. Boylan[b,1]

[a]The School of Management, University of Salford, Maxwell Building, The Crescent, Salford M5 4WT, UK
[b]Buckinghamshire Business School, Buckinghamshire Chilterns University College, Chalfont Campus, Gorelands Lane, Chalfont St. Giles, Bucks HP8 4AD, UK

Abstract

Intermittent demand appears sporadically, with some time periods showing no demand at all. In this paper, four forecasting methods, Simple Moving Average (SMA, 13 periods), Single Exponential Smoothing (SES), Croston's method, and a new method (based on Croston's approach) recently developed by the authors, are compared on 3000 real intermittent demand data series from the automotive industry. The mean signed and relative geometric root-mean-square errors are shown to meet the theoretical and practical requirements of intermittent demand, as do the Percentage Better and Percentage Best summary statistics based on these measures. These measures are subsequently applied in a simulation experiment. The out-of-sample comparison results indicate superior performance of the new method. In addition, the results show that the mean signed error is not strongly scale dependent and the relative geometric root-mean-square error is a well-behaved accuracy measure for intermittent demand.

Keywords: Demand forecasting; Intermittent demand; Accuracy measures; Croston's method; Exponential smoothing; Forecasting competition

1. Research background

Intermittent demand appears sporadically, with some time periods showing no demand at all. Moreover, when demand occurs, it may not be for a single unit or a constant size. Such demand is difficult to predict, and errors in prediction may be costly in terms of obsolescent stock or unmet demand.

Single Exponential Smoothing (SES) and Simple Moving Averages (SMA) are often used in practice to deal with intermittent demand. Both methods have been shown to perform satisfactorily on real intermittent demand data. However, the "standard" forecasting method for intermittent demand items is considered to be Croston's method (Croston, 1972, as corrected by Rao, 1973). Croston built demand estimates from constituent elements, namely, the demand size when demand occurs (z_t) and the interdemand interval (p_t). Both demand sizes and

* Corresponding author. Tel.: +44 161 295 5804; fax: +44 161 295 3821.
E-mail addresses: a.syntetos@salford.ac.uk (A.A. Syntetos), john.boylan@bcuc.ac.uk (J.E. Boylan).
[1] Tel.: +44 1494 60 51 30; fax: +44 1494 87 42 30.

0169-2070/$ - see front matter © 2004 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2004.10.001

intervals are assumed to be stationary. Demand is assumed to occur as a Bernoulli process. Subsequently, the interdemand intervals are geometrically distributed (with mean p). The demand sizes are assumed to follow the normal distribution (with mean μ and variance σ^2). These assumptions have been challenged in respect of their realism (Willemain, Smart, Shockor, & DeSautels, 1994) and in respect of their theoretical consistency with Croston's method (Snyder, 2002; Shenstone & Hyndman, 2003). Regarding the latter issue, it is important to note that there may be some merit in working backwards from Croston's method to a model, as it would also be possible to experiment with different model assumptions and then design new methods, but these issues are not explored in our paper.

According to Croston's method, separate exponential smoothing estimates of the average size of the demand (z_t', $E(z_t)=E(z_t')=\mu$) and the average interval between demand incidences (p_t', $E(p_t)=E(p_t')=p$) are made after demand occurs (using the same smoothing constant value). If no demand occurs, the estimates remain exactly the same. The forecast, Y_t', for the next time period is given by $Y_t'=z_t'/p_t'$, and, according to Croston, the expected estimate of demand per period in that case would be $E(Y_t')=E(z_t'/p_t')=E(z_t')/E(p_t')=\mu/p$ (that is, the method is unbiased). If demand occurs in every time period, Croston's estimator is identical to SES.

The method is, intuitively at least, superior to SES and SMA. Croston's method is currently used by leading statistical forecasting software packages and has motivated a substantial amount of research work over the years. Syntetos and Boylan (2001) showed that Croston's estimator is biased. The bias arises because, if it is assumed that estimators of demand size and demand interval are independent, then

$$E\left(\frac{z_t'}{p_t'}\right) = E(z_t')E\left(\frac{1}{p_t'}\right) \tag{1}$$

but

$$E\left(\frac{1}{p_t'}\right) \neq \frac{1}{E(p_t')} \tag{2}$$

and therefore, Croston's method is biased. It is clear that this result does not depend on Croston's assumptions of stationarity and geometrically distributed demand intervals. The magnitude of the error depends on the smoothing constant value being used. We show in Appendix A that the bias associated with Croston's method, in practice, can be approximated, for all smoothing constant values, by $\frac{\alpha}{2-\alpha}\mu\frac{(p-1)}{p^2}$ (where α is the smoothing constant value used for updating the interdemand intervals.)

The bias can be conveniently expressed as a percentage of the average demand, and it is easily shown to be: $100\frac{\alpha}{2-\alpha}\left(1-\frac{1}{p}\right)$.

A new intermittent demand forecasting method is proposed in this paper, incorporating this bias approximation. The method has been developed based on Croston's idea of building demand estimates from constituent elements. It is approximately unbiased and has been shown (Syntetos, 2001) to outperform Croston's method on theoretically generated data. The new estimator of mean demand is as follows:

$$Y_t' = \left(1-\frac{\alpha}{2}\right)\frac{z_t'}{p_t'} \tag{3}$$

where α is the smoothing constant value used for updating the interdemand intervals. In this paper, the same smoothing constant is used for updating demand sizes as for demand intervals, although, following the suggestion of Schultz (1987), a different smoothing constant may be used.

The derivation of the new estimator is based on Croston's assumptions of stationary, identically, independently distributed series of demand sizes and demand intervals, geometrically distributed interdemand intervals, and independence of demand sizes and intervals. There are no restrictions regarding the demand size distribution. There is clearly evidence in support of these assumptions (e.g., Janssen, 1998; Eaves, 2002). Nevertheless, as mentioned previously, these assumptions have also been questioned, and therefore, it is important to test the performance of the new estimator on real demand data.

1.1. An empirical accuracy comparison exercise

In recent decades, empirical and experimental studies have taken the form of a "forecasting competition", where expert participants analysed and forecasted many real-life time series coming from entirely different populations (M-Competition: Mak-

ridakis et al., 1982; M2-Competition: Makridakis et al., 1993; M3-Competition: Makridakis & Hibon, 2000; as summarised by Fildes & Ord, 2002). No intermittent data have been considered in any of these competitions.

In this paper, the forecasting accuracy of alternative intermittent demand estimators is compared. The empirical data sample consists of the monthly demand histories (over a 2-year period) of 3000 Stock Keeping Units (SKUs). This data set is not part of the set of 17,000 series described earlier by Syntetos and Boylan (2001). These series contained many "slower intermittent" items (demand series with large p values) but none that are "faster intermittent" (p values close to 1), which are required for a wider study to determine what actually constitutes intermittence. Consequently, a separate sample of 3000 "faster intermittent" data from the automotive industry have been provided by a forecasting and stock control software package manufacturer.

All SKUs are treated as "single" units as opposed to being sold in packs of a certain size. The average interdemand interval ranges from 1.04 to 2 months, and the average demand per unit time period, from 0.5 to 120 units. The average demand size, when demand occurs, is between 1 and 194 units, and the variance of the demand sizes ranges from 0 to 49,612 (the squared coefficient of variation ranges from 0 to 14). The sample contains slow movers (very low demand per unit time period due to infrequent demand arrivals, low average demand sizes, or both), lumpy demand items [demand is zero in some time periods with (highly) variable demand sizes, when demand occurs], as well as intermittent demand series with a constant (or approximately constant) size of demand, when demand occurs. The distribution of the real demand data files, with respect to their average interdemand interval and the squared coefficient of variation of the demand sizes, is indicated in Fig. 1. The demand data sample is available upon request from the first author.

Series with average interdemand interval less than two are often considered to be nonintermittent. Johnston and Boylan (1996) reconceptualised the term "intermittence" by considering the mean interdemand intervals, for which Croston's method outperformed SES. The authors recommended a rule that if the mean interdemand interval is greater than 1.25 forecast revision periods, then Croston's method should be used rather than the SES. This form of rule is based on the premise that it is preferable to identify conditions for superior forecasting performance and then to categorise demand based on these results, rather than the other way around. The essence of the reconceptualisation lies in this approach and the identification of the mean interdemand interval as the categorisation parameter. Syntetos (2001) took this approach forward by identifying theoretically sound demand categorisation schemes based on both the mean interdemand interval and the coefficient of variation of demand sizes. According to the results of

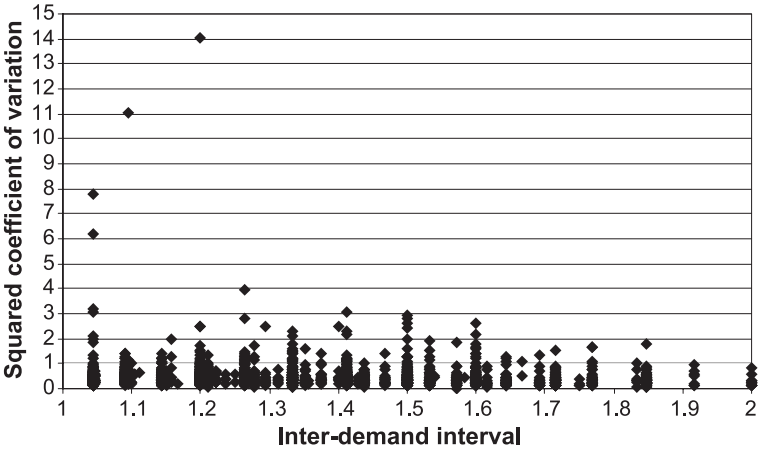

Fig. 1. The demand series characteristics.

Johnston and Boylan and of Syntetos, many series with mean interdemand intervals less than 2 are best characterised as intermittent.

It will be shown in this paper that "faster intermittent" data can benefit from the application of methods originally designed for slower data. Further research is currently ongoing on the accuracy of forecasting methods for more highly intermittent data. Nevertheless, the conclusions on the selection of accuracy measures apply for all intermittent data, regardless of the degree of intermittence.

1.2. Structure of the paper

The rest of our paper is structured as follows. In Section 2, the selection of appropriate accuracy measures in an intermittent demand context is reviewed, from both practitioner and academic perspectives. In Section 3, the notation to be used in this paper is presented. In Section 4, the accuracy measures to be used in our empirical exercise are presented, along with a justification for their selection. In Section 5, some issues related to the structure of the empirical experiment are discussed. In Section 6, the empirical results are presented and analysed. The conclusions are presented in Section 7 of the paper.

2. Accuracy measures for intermittent demand

It is the very nature of intermittent demand data and, in particular, the existence of some zero-demand time periods, that creates significant difficulties in selecting an appropriate accuracy measure. Nevertheless, those special properties of intermittent demand series seem to have been underestimated or completely ignored in the past by both practitioners and academicians.

Commercial software packages often report the mean absolute percentage error (MAPE), with the better packages distinguishing between "within-sample" and "out-of-sample" evaluations. Some packages also report the relative absolute error of a method compared with the random walk or naive 1 forecast, which is simply the demand in the previous time period. Neither of these measures is appropriate for intermittent demand because zero observations may yield 'division by zero' problems. Ad hoc procedures, such as excluding zero-demand time periods or adding a small amount to zero demands, do little to improve confidence in these measures.

From an academic perspective, Willemain et al. (1994) compared exponential smoothing and Croston's method using:

1. Monte Carlo Simulation. Theoretical intermittent demand data were generated for different scenarios that violated Croston's assumptions. The comparison with exponential smoothing was made with respect to MAPE, Median APE (MdAPE), root-mean-square error (RMSE), and mean absolute deviation (MAD) for issue points in time only.
2. Industrial intermittent data, focusing on the MAPE for one step ahead forecasts (for all points in time). The error statistic used was a ratio estimate, the numerator being the sum of absolute errors over every time period and the denominator being the sum of actual demands. This formulation guaranteed a nonzero denominator.

Another academic paper that considers accuracy measures for comparing alternative intermittent demand estimation procedures is by Johnston and Boylan (1996). In this paper, the relative arithmetic mean square error (RAMSE) was selected to compare the Size-Interval method and SES on theoretically generated demand data. It is important to note that in neither paper is any justification given by the authors for the choice of the accuracy measures that were used in their research.

Willemain, Smart, and Schwarz (2004) argue that "... we need to assess the quality not of a point forecast of the mean but of a forecast of the entire distribution", but they concede that it is impossible to compare this on an item-specific basis. Instead, the authors recommend pooling percentile estimators across items and measuring the conformance of the observations (expressed using the corresponding percentiles) to a uniform distribution.

In a stock control context, order quantities can be calculated using the average demand per period (i.e., our forecast), but the reorder points' (or order levels') calculation requires estimates of the demand distribution itself. For a parametric approach, one needs to distinguish between testing (i) the estimation method and (ii) the goodness of fit of the distribution.

Consequently, the use of pooled percentile estimators could be used for comparing a forecasting method, combined with a distribution (e.g., Poisson), against a bootstrapping approach. However, to test the accuracy of the estimator itself, direct measures of forecasting accuracy are required.

3. Notation

The notation to be used in the remainder of the paper is as follows:

L	The lead time, $L \geq 1$
$Y'_{t,L}$	The estimate (made in period t) of demand in period $t+L$, obtained by any of the forecasting methods considered
Y_{t+L}	The actual demand in period $t+L$
e_{t+L}	The forecast error in period $t+L$
n	The number of demand time periods considered for the purpose of comparison; $n=m-r-L$, where m is the total number of demand time periods contained in any of the series, and r is the number of periods that was used for initialisation purposes, i.e., not considered for generating results

4. Selection of accuracy measures

Before considering accuracy measures for the purpose of comparing alternative methods in an intermittent demand context, it would be wise to determine which ones can be computed, taking into account that there are some zero-demand time periods. All relative-to-the-series accuracy measures (e.g., MAPE and MdAPE) must be excluded because actual demand may be zero. The symmetric MAPE could be calculated, but it is known to suffer from asymmetry problems (Goodwin & Lawton, 1999). Finally, all relative-to-a-base accuracy measures must also be eliminated. In this last case, the forecast error in a particular time period is related to some sort of benchmark, usually the error produced by the naive 1 method for the same period, which would often be zero.

4.1. Absolute accuracy measures

Absolute error measures are calculated as a function of the forecast errors alone; there are no ratio calculations with respect to the series itself or to other forecasting methods. Examples include the mean square error (MSE) and the mean absolute error (MAE). Although such measures can be computed for intermittent demand, when averaged across many time series, they do not take into account the scale differences between them.

One possible exception to the exclusion of absolute measures may be the mean error (ME), also known as the mean signed error. The mean error is defined as:

$$ME = \frac{\sum_{t=1}^{n} \left(Y_{t+L} - Y'_{t,L}\right)}{n} = \frac{\sum_{t=1}^{n} e_{t+L}}{n} \quad (4)$$

From a practical perspective, the ME is very easy to calculate and has a very straightforward interpretation: The difference between the absolute average ME given by any two methods indicates how much more (or less) biased is one method in comparison with another.

Because the ME measure takes into account the sign of the error, it is conjectured that it is less scale dependent than are other absolute accuracy measures, such as the MAE and the MSE. The extent of its scale dependence will be assessed by comparing its performance to that of the scaled ME. To generate accuracy results using this latter measure, the originally calculated MEs, per method, per series, will be divided by the average demand per unit time period (for the series under concern) so that scale dependencies are eliminated.

In addition, to account for a potentially skewed distribution of the MEs, a nonparametric procedure (percentage of times better—PB) is also introduced in the following subsection, which considers only the relationship between the MEs, to generate results without taking into account their actual size, again eliminating scale effects. It should be noted that the new intermittent demand estimator discussed in this paper is a bias-reduced version of Croston's method, and in that respect, it is expected to lead to improvements in any bias-related measure.

4.2. Accuracy measures relative to another method

The "percentage of times better" summarises the performance of one method relative to another method

across a set of time series, where the forecast error of the methods for each individual series is assessed using one of the "absolute accuracy measures". The percentage of times better (PB) is easily interpreted (Makridakis, 1993) and is regarded as an intuitive nonparametric procedure (Makridakis & Hibon, 1995).

The PB tells us how many times a method performs better than another, but not by how much. For this to be done, a descriptive accuracy measure needs to be selected.

In the following paragraph, the relative geometric root-mean-square error (RGRMSE) is presented as a means of discussing the advantages and disadvantages associated with all relative-to-another-method accuracy measures.

The RGRMSE for methods A and B in a particular time series is defined as:

$$RGRMSE = \frac{\left(\prod_{t=1}^{n}\left(Y_{t+L} - Y'_{\mathrm{A},t,L}\right)^2\right)^{\frac{1}{2n}}}{\left(\prod_{t=1}^{n}\left(Y_{t+L} - Y'_{\mathrm{B},t,L}\right)^2\right)^{\frac{1}{2n}}} \quad (5)$$

where the subscripts A and B refer to the forecasting methods.

Fildes (1992) discussed the theoretical properties of the RGRMSE. According to his analysis, assume that the squared errors produced by a particular method at different points in time (for a particular series) are of the form:

$$\left(Y_{t+L} - Y'_{M,t,L}\right)^2 = e^2_{M,t+L} = \varepsilon^2_{M,t+L} u_{t+L} \quad (6)$$

where u_{t+L} are assumed to be positive and can be thought of as errors due to the particular time period affecting all methods equally, while the $\varepsilon_{M,t+L}{}^2$ are the method's (M) specific errors.

According to Fildes, such a model corresponds to the case where the data and, subsequently, the errors are contaminated by occasional outliers. Fildes showed that the use of a geometrically (rather than arithmetically) averaged RMSE (GRMSE) expressed in a relative way (RGRMSE of one method compared with another) is independent of the u_{t+L}.

Considering its desirable statistical properties (Fildes, 1992), the RGRMSE will be used to report summary error results in this research. The following definition applies across a range of series (s=1, . . .,κ):

$$RGRMSE = \frac{\left(\prod_{s=1}^{\kappa}\left(GRMSE_{\mathrm{A},s}\right)^2\right)^{\frac{1}{2\kappa}}}{\left(\prod_{s=1}^{\kappa}\left(GRMSE_{\mathrm{B},s}\right)^2\right)^{\frac{1}{2\kappa}}} = \left(\prod_{s=1}^{\kappa}\frac{GRMSE_{\mathrm{A},s}}{GRMSE_{\mathrm{B},s}}\right)^{\frac{1}{\kappa}}$$

where the $\mathrm{GRMSE}_{i,s}$, per series, for method i is calculated as:

$$GRMSE_{i,s} = \left(\prod_{t=1}^{n}\left(Y_{t+L} - Y_{i,t,L'}\right)^2\right)^{\frac{1}{2n}} \quad (8)$$

PB is defined as the percentage of times (observations) that one method performs better than one other. The PB is an easy measure to calculate, with a simple interpretation. For intermittent demand data, PB is particularly meaningful because all series and all data periods within each series are considered to generate results. In the context of this research, this accuracy measure is applied to generate results across all demand data series. For this to be done, one or more descriptive accuracy measures are needed that will provide results about the alternative methods' performance in each one of the series. The ME and GRMSE are to be used for that purpose.

The PB measure reports the proportion of times that one method performs better than one other method. When more than two estimators are involved in the accuracy comparison exercise, it is useful to report the proportion of times that one method performs better than all other methods, i.e., the proportion of times that each method performs best. In this case, the measure is referred to as Percentage Best (PBt) rather than Percentage Better.

5. Experimental structure

The objective of this research is to compare the empirical forecasting accuracy of SES, Croston's method, and the new estimator discussed in this paper. The estimation procedure employed by the software

manufacturer that provided the empirical data series used in this research, when dealing with intermittence, is the Simple Moving Average method (SMA). The theoretical properties of SMA for intermittent data have not been explored in the academic literature. It has been reported that for systems employing monthly forecast revisions, the length of the SMA is most commonly somewhere between 3 and 12 points, whereas for weekly revisions, it is between 8 and 24 points (Johnston, Boylan, Shale, & Meadows, 1999). In the software package, developed by the manufacturer that provided the demand data sample, the number of periods is set to 13 (independent of the length of the review period, i.e., weeks or months). This value has been reported by the software manufacturer to provide the best results.

The 13-period moving average (SMA(13)) is the estimator that has been used in practice to deal with the intermittent nature of the series available for this research. Consequently, it has been decided to consider the performance of this estimator also, for comparison purposes. The length of the Simple Moving Average is set to 13, and the method can be viewed as a benchmark for the purpose of analysing the simulation results. The use of the SMA(13) method for simulation purposes necessitates the exclusion of the first 13 periods of data for generating results. That is, to initialise the particular method's application, the first 13-period demand data need to be used. The other estimators can be initialised by withholding fewer demand data periods. However, in an intermittent demand context, the number of demand data periods will invariably be greater than the number of demand occurrences. The consideration of size-interval estimators in our simulation experiment necessitates the existence of some demand occurrences (and subsequently of some interdemand intervals) in the initialisation subsample, if reasonable initial values are to be obtained. Therefore, a large initialisation subsample is required, irrespective of which size-interval estimators are considered.

Given (i) the relevance of SMA(13) as a benchmark for our empirical data sample and (ii) consistency requirements, it has been decided to initialise all methods' application by using the same number of periods, i.e., 13. Therefore, the out-of-sample comparison results will refer to the latest 11 monthly demand data. The first SMA(13) and SES estimate is taken to be the average demand over the first 13 periods. In a similar way, the exponentially smoothed estimates of demand size and interdemand interval are initialised as their averages over the first 13 periods. If no demand occurs in the first 13 periods, the initial SES and moving average estimates are set to zero and the interdemand interval estimate to 13. As far as the demand size is concerned, it is more reasonable to assign an initial estimate of 1 rather than 0. Optimisation of the smoothing constant values used for the smoothing methods is not considered because of the very few (if any) demand occurring periods in the data subset withheld for initialisation purposes.

Because the available data sets were so short (24 periods), the initialisation effect is carried forward by all estimators on all out-of-sample point estimates. Therefore, the empirical findings on the forecasting performance of the methods considered may be strongly affected by their initialisations, given that there are only 11 out-of-sample data points. Clearly, longer demand data series would have been welcomed. However, longer histories of data are not necessarily available in real-world applications. Decisions often need to be made considering samples similar to the one used for this research.

The smoothing constant value was introduced as a control parameter. In an intermittent demand context, low smoothing constant values are recommended in the literature. Smoothing constant values in the range 0.05–0.2 are viewed as realistic (Croston, 1972; Willemain et al., 1994; Johnston & Boylan, 1996). From an empirical perspective, this range covers the usual values of practical usage. In consequence, this is the range of values that we focus upon during this research. In particular, four values will be simulated: 0.05, 0.10, 0.15, and 0.20.

The lead time has also been introduced as a control parameter. The lead times considered are one, three, and five periods. Finally, two cases relating to the timing of forecasts should be considered: analysis of all time periods for generating results (all points in time) or focusing on the time periods immediately after a demand occurrence (issue points only). The former scenario corresponds to a reorder interval or product group review inventory control system, while the latter to a reorder level/continuous review system. Both scenarios have been simulated in the experiment. In summary six pair-wise comparative assessments were

made across 24 control parameter combinations over 3000 series for each of the accuracy measures used.

6. Empirical analysis

In this section of the paper, mean error (ME), scaled mean error, relative geometric root-mean-square error (RGRMSE), Percentage Better (PB), and Percentage Best (PBt) summary results are presented. The results are based on forecasting simulations using real data.

The scaled ME statistic, which scales by the average demand per unit time, was introduced to check the scale dependence of the ME. Statistical significance for the ME and scaled ME results has been checked using the two-sample *t*-test (assuming unequal variances). The *z*-test statistic (difference between population proportions) was reported for the PB and PBt measures. Regarding the RGRMSE, the two-sample *t*-test (assuming unequal variances) was used to assess whether the difference between the arithmetic means of the logarithm of the geometric root-mean-square errors produced by two methods differs significantly from zero. Testing whether that difference deviated significantly from zero is equivalent to testing whether the RGRMSE across series is significantly different from one. Due to the length of our paper, detailed statistical test results have not been presented. Nevertheless, detailed tabulations are available upon request from the first author.

6.1. Mean errors and relative geometric root-mean-square errors

The analysis starts by considering a scale-dependent error measure, the ME, and then the scaled ME (to eliminate any scale dependencies) and the RGRMSE (to eliminate the effect of outliers).

The ME results have shown that the variance of the average (across all series) ME increases rapidly with the lead time. This, in turn, implies nonindependence of the error values. As one of the referees pointed out, if the error values were independent, then the error variance should increase linearly with lead time, which is not the case.

Overall, there are no great discrepancies between the ME and scaled ME results in terms of which method performs better or worse. When the former measure was used, no significant differences between forecasting methods were indicated for α=0.05. However, as the smoothing constant increases, it becomes easier to demonstrate statistically significant accuracy differences. The scaled ME results confirm the results obtained on the original MEs, but most of the accuracy differences are now statistically significant for all α values. The consistency of the conclusions drawn from the ME and scaled ME results indicates that the former measure is not strongly scale dependent on the data examined in this paper. This is regarded as an important finding because empirical evidence is now provided, in an intermittent demand context, to show that the ME, despite its absolute nature as an error measure, does not suffer from the serious scale-dependence problem that afflicts other absolute measures such as the MSE and RMSE. Considering the scaled ME results, the estimators discussed in this paper can be ordered as follows in terms of their bias (the first method being the one with least bias; see Table 1).

When the RGRMSE is considered, the methods are ordered as follows in terms of their forecasting accuracy (the majority of the accuracy comparison differences that involve the SMA method are statistically significant, independently of the α value being used for the smoothing estimators. For the rest of the comparisons, statistical significance becomes easier to demonstrate as the smoothing constant value increases; see Table 2).

The above ordering is accurate for all simulated scenarios, with the exception of α=0.15 and 0.2 (L.T.=3, 5). This issue is further discussed in the following subsection.

Table 1
Ordering of methods by scaled mean errors

Scaled mean errors	
All points in time	Issue points only
1. Syntetos and Boylan	Syntetos and Boylan
2. SES	Croston
3. SMA(13)	SMA(13)
4. Croston	SES

The overall measure of bias is calculated for each estimator as an equally weighted average of the absolute scaled MEs across all the control parameter combinations.

Table 2
Ordering of methods by RGRMSE

Relative geometric root-mean-square error	
All points in time	Issue points only
1. Syntetos and Boylan	Syntetos and Boylan
2. SES	Croston
3. Croston	SES
4. SMA(13)	SMA(13)

The overall measure of accuracy is calculated for each estimator as an equally weighted average of the GRMSEs across all the control parameter combinations.

The results obtained on the ME, scaled ME, and RGRMSE measure indicate that, in both reorder interval and a reorder level contexts, the new method discussed in this paper is the most accurate estimator. When all points in time are considered, SES is shown to perform more accurately than Croston's method does. This finding does not agree with the empirical results presented in Willemain et al. (1994), where Croston's estimator was found to outperform SES in a reorder interval context. However, it is consistent with the results presented for the more generalised error assessment framework (see also Section 2) in Willemain et al. (2004). For issue points only, Croston's method performs better than SES does, and this is in accordance with theoretical expectations (at least as far as the bias measure is concerned; see Syntetos, 2001).

6.2. Percentage Better and Percentage Best

In this subsection, the empirical results have been synthesised to generate an ordering of the performance of all estimators. For each particular combination of the control parameter values, the estimators were ranked in terms of the number of pair-wise comparisons that outperform any of the other estimators considered (the first method being the best). With the exception of very few occasional cases, the accuracy differences observed were statistically significant, irrespective of the control parameter combination and specific pair-wise comparison. The results are shown in Table 3.

The above ordering is not valid for α=0.1, L.T.=5 (all points in time), where Croston's method performs better than SES, and α=0.1, L.T.=5 (issue points only), where SMA(13) and SES perform identically.

There is a remarkable consistency between the results generated based on the RGRMSE (across series)

Table 3
Ordering of methods by percentage better (RGRMSE)

Percentage better (RGRMSE)	
All points in time	Issue points only
α=0.05, 0.1, (0.15, 0.2, 1 step-ahead forecasts)	
1. Syntetos and Boylan	Syntetos and Boylan
2. SES	Croston
3. Croston	SES
4. SMA(13)	SMA(13)
α=0.15, 0.2 (L.T.=3, 5)	
1. Syntetos and Boylan	Syntetos and Boylan
2. SMA(13)	SMA(13)
3. Croston	Croston
4. SES	SES

applied as a descriptive measure and the Percentage Better results generated on the GRMSE per series. In the previous subsection, summary results were presented regarding the comparative forecasting accuracy performance of the alternative estimators, based on the RGRMSE, and few simulated scenarios were identified where the accuracy differences were not very well marked. The Percentage Better results confirm the validity of those conclusions and demonstrate considerable accuracy differences in the scenarios discussed above.

The same is not true when a comparison of results is undertaken based on the ME applied as a descriptive measure across series and the Percentage Better results generated based on the ME given by alternative estimators in a single series. In fact, the two sets of results are considerably different (Table 4).

Table 4
Ordering of methods by percentage better (ME)

Percentage better (mean signed error)			
All points in time		Issue points only	
α=0.05	α=0.1	α=0.05	α=0.1
1. SMA(13)	SES	SMA(13)	S-B
2. SES	SMA(13)	S-B	SES
3. S-B	S-B	SES	SMA(13)
4. Croston	Croston	Croston	Croston
α=0.15, 0.2			
1. SES		Syntetos and Boylan	
2. Syntetos and Boylan		SES	
3. Croston		Croston	
4. SMA(13)		SMA(13)	

S-B stands for Syntetos and Boylan.

The above ordering is not valid for α=0.1, L.T.=1 (issue points only), where the ordering is as follows: (1) SES, (2) SMA(13), (3) Syntetos and Boylan, and (4) Croston.

When Percentage Best results were generated, there were only 5 out of the 144 cases that were tested (six pair-wise comparisons, 24 control parameter combinations) where the observed differences were not statistically significant. The Percentage Best results on the GRMSE indicate that the new method performs better than all the other three estimators. It is outperformed only by SMA(13) for α=0.05. The SMA(13) performs better than Croston's method and SES do, and finally, SES performs better than Croston's method does. Those findings are valid for the majority (if not all) of the simulated scenarios.

Very similar conclusions were obtained when the Percentage Best results on bias are analysed (in that case only six nonstatistically significant differences were observed). In fact, the only difference is that the new method is outperformed by the SMA(13) for both α=0.05 and 0.1 (all points in time). Considering the Percentage Best results, on the RGRMSE, the estimators discussed in this paper can be ordered based on the following scheme. For each particular combination of the control parameter values, the estimators were ranked in terms of the number of pair-wise comparisons that outperform any of the other estimators considered. The first method is the best (Table 5).

The above ordering is also valid for the Percentage Best results on ME, but with the following differences: (i) the SMA(13) performs better than the new method for α=0.1, all points in time; (ii) the SES estimator performs better than the new method for α=0.15, 0.2, L.T.=1, all points time; and (iii) the SES estimator performs better than SMA(13) for α=0.2, issue points only.

Table 5
Ordering of methods by percentage best (RGRMSE)

Percentage best (RGRMSE)			
All points in time		Issue points only	
α=0.05	α=0.1, 0.15, 0.2	α=0.05	α=0.1, 0.15, 0.2
1. SMA(13)	S-B	SMA(13)	S-B
2. S-B	SMA(13)	S-B	SMA(13)
3. SES	SES	SES	SES
4. Croston	Croston	Croston	Croston

The Percentage Best results are more meaningful than the Percentage Better results from a practical perspective because, ultimately, only the best method will be used. What matters for a practitioner is the accuracy in determining the first place (best estimator) rather than the accuracy associated with the determination of the lower places.

It has been argued in the literature that the Percentage Better is an easily interpreted and a very intuitive nonparametric approach to generate comparative forecasting accuracy results. The Percentage Better accuracy measure has been recommended for use not only on a pair-wise comparison level but also for generating overall (across all estimators) accuracy results (Makridakis & Hibon, 1995, p. 7): "The percentage better measure requires (as a minimum) the use of two methods. . ..If more than two methods are to be compared the evaluation can be done for each pair of them."

Nevertheless, the empirical evidence presented in this section demonstrates that the use of the Percentage Better measure for more than two estimators leads to rather complex ordering schemes. In addition, the PBt results appear to be insensitive to the descriptive accuracy measure chosen for generating results in each of the series included in the empirical sample. Therefore, the PBt is recommended for use in large-scale empirical exercises.

It should be noted that a few of the series examined in our empirical study are particularly erratic (see Fig. 1). Such series may appear to be insignificant, but they can cause scale-dependence problems in comparing the performance of forecasting methods. However, by choosing four scale-independent measures (scaled ME, RGRMSE, PB, and PBt) and one measure, ME, that has been found not be strongly scale dependent, any such problems have been avoided.

The results generated in this subsection indicate that accuracy differences based on a descriptive accuracy measure across series are not necessarily reflected on the number of series for which one estimator performs better than one or all other estimators.

The new method proposed earlier in the paper has been shown in this subsection to be the most accurate estimator. In an industrial context, the proportion of

the stock range that is devoted to intermittent demand items is often considerable. In that respect, any improvements in a company's system regarding those items may be translated to substantial cost savings. The simplicity of the new estimator also favours its application in a real system.

7. Conclusions

In this paper, SMA(13), SES, Croston's method, and a new method developed by the authors based on Croston's approach have been compared with respect to their forecasting accuracy. Results have been generated by considering the following: (i) mean errors, scaled mean errors, and relative geometric root-mean-square errors and (ii) Percentage Better and Percentage Best results, based on the ME and RGRMSE per series. The selection of these particular error measures has been justified from theoretical and practical perspectives.

The new method can be claimed to be the most accurate estimator for the "faster intermittent" demand data investigated. SES performs better than Croston's method does in a reorder interval context, but when issue points only are considered, the comparison results are inconclusive. The SMA(13) compares favourably with the smoothing methods for low α values. This estimator is also found to be robust to the presence of outliers, but is the least accurate according to the relative geometric root-mean-square error.

The empirical analysis shows, as one would expect, that different accuracy measures can lead to different conclusions in an intermittent demand context. In this paper, it has been argued that the Percentage Best (PBt) measure should be preferred to the Percentage Better (PB) measure because of its relevance to the choice of the best forecasting method and the fact that it gives more consistent results. Finally, the empirical results demonstrate that the RGRMSE is a very well-behaved accuracy measure for intermittent demand.

Appendix A. Derivation of the new estimator

Let the exponentially smoothed demand size estimate be $x_1=z_t'$, and suppose $E(x_1)=\mu$. Let the exponentially smoothed demand interval estimate be $x_2=p_t'$ and suppose $E(x_2)=p$.

Applying Taylor's theorem to a function $g(x_1,x_2)=x_1/x_2$ and assuming independence between estimates of demand size and demand intervals,

$$E[g(x_1,x_2)] = \frac{\mu}{p} + \frac{1}{2}\frac{\partial^2 g(\mu,p)}{\partial x_2^2}\mathrm{Var}(x_2) + \ldots$$

$$= \frac{\mu}{p} + \frac{\mu}{p^3}\mathrm{Var}(x_2) + \ldots$$

Because x_2 is an exponentially smoothed estimate (smoothing constant α) of interdemand intervals that are geometrically distributed with variance $p(p-1)$:

$$E\left(\frac{z_t'}{p_t'}\right) \approx \frac{\mu}{p} + \frac{\alpha}{2-\alpha}\mu\frac{(p-1)}{p^2}$$

Hence,

$$E\left(\left(1-\frac{\alpha}{2}\right)\frac{z_t'}{p_t'}\right) \approx \frac{\mu}{p}.$$

This demonstrates the approximate unbiasedness of the new estimator, which improves as the value of p increases.

References

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, *23*, 289–304.

Eaves, A. H. C. (2002). *Forecasting for the ordering and stock holding of consumable spare parts*, PhD Thesis, Lancaster University, UK.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, *8*, 81–98.

Fildes, R., & Ord, K. (2002). Forecasting competitions—Their role in improving forecasting practice and research. In M. P. Clements, & D. F. Hendry (Eds.), *A companion to economic forecasting* (pp. 322–353). Oxford: Blackwell.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*, 405–408.

Janssen, F. B. S. L. P. (1998). *Inventory management systems; control and information issues*, PhD Thesis, Tilburg University, The Netherlands.

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

Johnston, F. R., Boylan, J. E., Shale, E., & Meadows, M. (1999). A robust forecasting system, based on the combination of two

simple moving averages. *Journal of the Operational Research Society*, *50*, 1199–1204.

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, *9*, 527–529.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., et al. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., et al. (1993). The M2-competiton: A real time judgementally based forecasting study. *International Journal of Forecasting*, *9*, 5–22.

Makridakis, S., & Hibon, M. (1995). *Evaluating accuracy (or error) measures*, Working paper 95/18/TM, INSEAD, France.

Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.

Rao, A. V. (1973). A comment on "Forecasting and stock control for intermittent demands". *Operational Research Quarterly*, *24*, 639–640.

Schultz, C. R. (1987). Forecasting and inventory control for sporadic demand under periodic review. *Journal of the Operational Research Society*, *38*, 453–458.

Shenstone, L., & Hyndman, R. J. (2003). *Stochastic models underlying Croston's method for intermittent demand forecasting*, Working Paper 1/2003, Department of Econometrics and Business Statistics, Monash University, Australia.

Snyder, R. (2002). Forecasting sales of slow and fast moving inventories. *European Journal of Operational Research*, *140*, 684–699.

Syntetos, A. A. (2001). *Forecasting of intermittent demand*, PhD Thesis, Buckinghamshire Chilterns University College-Brunel University, UK.

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

Willemain, T. R., Smart, C. N., & Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting*, *20*, 375–387.

Willemain, T. R., Smart, C. N., Shockor, J. H., & DeSautels, P. A. (1994). Forecasting intermittent demand in manufacturing: A comparative evaluation of Croston's method. *International Journal of Forecasting*, *10*, 529–538.

Biographies: Aris A. SYNTETOS holds a BA degree from the University of Athens, an MSc degree from Stirling University, and in 2001, he completed a PhD at Brunel University-Buckinghamshire Business School. In the period 2001–2003, he did his compulsory military service in Greece, and currently, he is a lecturer at the University of Salford. His research interests include demand forecasting and management processes.

John E. BOYLAN is Reader in Management Science at Buckinghamshire Business School, Buckinghamshire Chilterns University College. He completed a PhD at Warwick University in 1997 and has published papers on short-term forecasting in a variety of journals. His research interests relate to demand forecasting in an inventory management context.