International Journal of Production Research

ISSN: 0020-7543 (Print) 1366-588X (Online) Journal homepage: www.tandfonline.com/journals/tprs20

Mean-based error measures for intermittent demand forecasting

Steven Prestwich, Roberto Rossi, S. Armagan Tarim & Brahim Hnich

To cite this article: Steven Prestwich, Roberto Rossi, S. Armagan Tarim & Brahim Hnich (2014) Mean-based error measures for intermittent demand forecasting, International Journal of Production Research, 52:22, 6782-6791, DOI: 10.1080/00207543.2014.917771

To link to this article: https://doi.org/10.1080/00207543.2014.917771

International Journal of Production Research, 2014
Vol. 52, No. 22, 6782–6791, http://dx.doi.org/10.1080/00207543.2014.917771

Taylor & Francis
Taylor & Francis Group

Mean-based error measures for intermittent demand forecasting

Steven Prestwich[a]*, Roberto Rossi[b], S. Armagan Tarim[c] and Brahim Hnich[d]

[a] *Department of Computer Science, University College Cork, Cork, Ireland;* [b] *University of Edinburgh Business School, Edinburgh, UK;* [c] *Department of Management, Hacettepe University, Ankara, Turkey;* [d] *Computer Engineering Department, Izmir University of Economics, Izmir, Turkey*

(*Received 23 October 2013; accepted 15 April 2014*)

To compare different forecasting methods on demand series, we require an error measure. Many error measures have been proposed, but when demand is intermittent some become inapplicable because of infinities, some give counter-intuitive results, and there is no agreement on which is best. We argue that almost all known measures rank forecasters incorrectly on intermittent demand series. We propose several new error measures with almost no infinities, and with correct forecaster ranking on several intermittent demand patterns. We call these 'mean-based' error measures because they evaluate forecasts against the (possibly time-dependent) mean of the underlying stochastic process instead of point demands.

Keywords: forecasting; intermittent demand; error measure

1. Introduction

Inventory management is of great economic importance to industry, but forecasting demand for some inventories is difficult because the demand is *intermittent*: in many time periods it is zero. Intermittent demand (or *count data*) occurs in several industries, for example, in aerospace and military inventories from which spare parts such as wings or jet engines are infrequently required. A survey of forecasting methods for spare parts is given in Boylan and Syntetos (2010). Various methods have been proposed for forecasting, some simple and others statistically sophisticated, but relatively little work has been done on forecasting for intermittent demand. Most work in this area is influenced by that of Croston (1972), who first separated the forecasting of demand size and inter-demand interval.

To choose a good forecasting method, we can test the alternatives empirically on demand series to see which gives the smallest error. For this, we require an *error measure* (or *accuracy measure*). Forecasting methods have been extensively compared on real and simulated data in the well-known M-, M2- and M3-competitions (Makridakis et al. 1982; Makridakis et al. 1993; Makridakis and Hibon 2000). There is no general agreement on which of the many existing error measures is best, so the competitions used several measures. However, the competitions did not deal specifically with intermittent demands, so the experience gained from these competitions cannot be used as a guide.

For intermittent demand series, some error measures are inapplicable because division by zero leads to infinities, but there are still several possibilities. This is an important issue because if researchers are free to choose from a large set of measures, then their results are likely to be incomparable. Moreover, there is a temptation to choose measures that give desired results (Collopy 1992) making experiments less objective. An editor of the International Journal of Forecasting wrote that the choice of error measure *is not a matter of personal preference* and urged researchers to follow contemporary recommendations (Fildes 1992b).

In this paper, we examine the suitability of known error measures for intermittent demand, and propose new improved measures. Section 2 provides background, demonstrates anomalous behaviour in existing error measures and proposes new measures. Section 3 evaluates the new measures on simulated data. Section 4 concludes the paper.

2. Old and new error measures

In this section, we provide some necessary background and describe our contribution. Section 2.1 describes the relevant forecasting methods, Section 2.2 surveys known error measures, Section 2.3 argues that these measures can rank forecasters incorrectly and Section 2.4 proposes new measures.

*Corresponding author. Email: s.prestwich@cs.ucc.ie

2.1 *Forecasting methods*

First, we describe the forecasting methods that will be used in the paper. *Single exponential smoothing* (SES) computes a smoothed series $\tilde{y}_t$ via the formula

$$\tilde{y}_t = \alpha y_t + (1-\alpha)\tilde{y}_{t-1}$$

where $\alpha \in (0, 1)$ is a *smoothing parameter*. The smaller the value of α, the less weight is attached to the most recent observations. An up-to-date survey of exponential smoothing algorithms is given in Gardner (2006). They perform remarkably well, often beating more complex approaches (Fildes et al. 2008). However, SES is known to perform poorly on intermittent demand, at least under some error measures.

The standard method for handling intermittency is *Croston's method* (Croston 1972), which applies SES to the non-zero demand sizes y and inter-demand intervals τ independently, using smoothing factors α and β, respectively. Given smoothed demand $\tilde{y}_t$ and smoothed interval $\tilde{\tau}_t$ at time t, the forecast is $f_t = \tilde{y}_t/\tilde{\tau}_t$. Both $\tilde{y}_t$ and $\tilde{\tau}_t$, and hence f_t, are updated at each time t for which $y_t \neq 0$. Alternative versions were proposed by Levén and Segerstedt (2004), Syntetos (2001), Syntetos and Boylan (2005) and we shall use the variant of Syntetos and Boylan (2005) which is known to have low bias and variance on stochastic demand. This variant is commonly referred to as SBA, but we shall refer to it and other variants as CR.

We also mention two trivial forecasting methods. Firstly, the *random walk* method (RW), also known as the *naive method*: take the previous period's demand as a forecast. RW is often used as a baseline for evaluating other methods. Secondly, the forecaster that always forecasts 0, which following Teunter and Duncan (2009) we call ZF. It was proposed by Croston, mentioned by Venkitachalam et al. (2003) and studied by Chatfield and Hayyab (2007) and Teunter and Duncan (2009).

2.2 *Existing error measures*

Next, we survey existing error measures, largely based on de Gooijer and Hyndman (2005) and Hyndman and Koehler (2006). Other measures exist, for example, Periods in Stock, Number of Shortages (Wallström and Segerstedt 2010) and Average Ranking (Makridakis et al. 1982), but we restrict our attention to those in the two surveys.

No one error measure is generally accepted as best on intermittent demand, and opinion is highly divided — see for example the debate in Ahlburg et al. (1992). Wallström and Segerstedt (2010) use principal components analysis to show that no single measure is sufficient, Ghobbar and Friend (2003) recommend using different measures for different types of demand, and in forecasting competitions several measures are used.

2.2.1 *Scale-dependent measures*

The most common are:

- Mean [Signed] Error (ME): mean(e_t)
- Mean Square Error (MSE): mean(e_t^2)
- Root-Mean Square Error (RMSE): $\sqrt{\text{MSE}}$
- Mean Absolute Error (MAE): mean($|e_t|$)
- Median Absolute Error (MdAE): median($|e_t|$)

where e_t is the error $y_t - \hat{y}_t$. These are useful for comparing methods on one series, but not for comparing over several series that are on different scales (Hyndman and Koehler 2006). Doubt has been cast on the suitability of MAE for intermittent demand (Wallström and Segerstedt 2010).

2.2.2 *Percentage errors*

These are also popular:

- Mean Absolute Percentage Error (MAPE): mean($|p_t|$)
- Median Absolute Percentage Error (MdAPE): median($|p_t|$)
- Root-Mean Square Percentage Error (RMSPE): $\sqrt{\text{mean}(p_t^2)}$
- Root-Median Square Percentage Error (RMdSPE): $\sqrt{\text{median}(p_t^2)}$
- Symmetric Mean Absolute Percentage Error (sMAPE): mean($200|e_t|/(y_t + \hat{y}_t)$)
- Symmetric Median Absolute Percentage Error (sMdAPE): median($200|e_t|/(y_t + \hat{y}_t)$)

where $p_t = 100e_t/y_t$. The last two measures are motivated by the fact that MAPE and MdAPE penalise positive errors more than negative ones.

However, percentage errors are undefined if any $y_t = 0$ (and if $\hat{y}_t = 0$ in the last two cases) and have very skewed distributions when $y_t \approx 0$. It is also pointed out in Hyndman and Koehler (2006) that they assume a meaningful zero, which is not the case for some data such as temperatures. Despite these drawbacks MAPE is recommended by most textbooks and was the main error measure used in the M-competition, while MdAPE is recommended by Fildes (1992a), and sMAPE and sMdAPE were used in the M3-competition.

It is pointed out by Kolassa and Schütz (2007) that many commercial software packages report a MAPE even when a series contains zeros, although the MAPE is technically undefined in this case. This is done by simply excluding periods with zero demands, which does not reflect the true errors of a forecast. This MAPE variant does not seem to have its own name so we shall denote it by iMAPE.

2.2.3 *Relative error-based measures*

We may also rescale by using errors from other measures:

- Mean Relative Absolute Error (MRAE): mean($|r_t|$)
- Median Relative Absolute Error (MdRAE): median($|r_t|$)
- Geometric Mean Relative Absolute Error (GMRAE): gmean($|r_t|$)

where $r_t = e_t/e_t^*$ and e_t^* is the error from a baseline method which is often RW.

GMRAE is also known as Relative Geometric Root-Mean Square and has desirable statistical properties (Fildes 1992a). It is used by Syntetos and Boylan (2005), and Armstrong and Collopy (1992) recommends the use of relative error-based measures. It has been proposed for intermittent demand in particular (Syntetos and Boylan 2005). However, these measures have the drawback of infinite variance because e_t^* can be arbitrarily small (Chatfield and Hayyab 2007; Hyndman and Koehler 2006). In the particular case of intermittent demand with RW as baseline, e_t^* is often zero so these measures are undefined. Extreme values can be trimmed (Armstrong and Collopy 1992) but this introduces some arbitrariness (Hyndman and Koehler 2006).

2.2.4 *Relative measures*

Instead of computing an absolute quantity to measure the accuracy of a method, we may compare it with another method. This can be done for many types of error measure, for example

- Relative Mean Absolute Error (RelMAE): MAE/MAE$_b$
- Relative Mean Squared Error (RelMSE): MSE/MSE$_b$
- Relative Root Mean Squared Error (RelRMSE): RMSE/RMSE$_b$ etc.

where MAE$_b$, MSE$_b$ and RMSE$_b$ are the MAE, MSE and RMSE of a baseline measure. The most popular baseline is RW, in which case RelRMSE is Thiel's U2 statistic (Thiel 1966) and log RelMSE is Thompson's LMR measure (Thompson 1990). RelMAE was recommended for intermittent demand by Syntetos and Boylan (2005) and called CumMAE by Armstrong and Collopy (1992). It is rare for these measures to become infinite, because the denominator is only zero if the baseline forecaster gives perfect results.

Another relative measure is Percent Better (PB) where a method is compared to another, usually RW, by how often its absolute error is smaller. This is recommended by Kolassa and Schütz (2007). A related measure is Percent Best (PBt) which compares several methods and computes the percentage of times each is most accurate. A drawback with PB and PBt is that they give no indication of the size of errors, so one large error is considered to be less serious than two tiny errors.

2.2.5 *Scaled errors*

We mention two of these. Firstly, the MAD/Mean Ratio (Kolassa and Schütz 2007), also called the Weighted MAPE, which we shall abbreviate to MMR. MAD (Mean Absolute Deviation) is another name for MAE so MMR = MAE/ME. Secondly, the MASE (Hyndman and Koehler 2006) (Mean Absolute Scaled Error) defined by MASE = mean($|q_t|$) where q_t is a scaled error defined by

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n} |y_i - y_{i-1}|}$$

and $t = 1 \ldots n$ is the set of sample periods used for forecasting. MASE effectively evaluates a forecasting method against RW. The only case in which it is undefined is when all in-sample demands are identical. Other scaled error measures

defined analogously to MASE include the Root-Mean Squared Scaled Error (RMSSE) and the Median Absolute Scaled Error (MdASE). MASE has been argued to be superior to several other methods used in forecasting competitions. An advantage of MASE over MMR is that it is more reliable on demand with seasonality, trends or other forms of non-stationarity. However, Kolassa and Schütz (2007) note that the MASE of two series with identical forecasts and identical demands during the forecast horizon will differ if the two series differed in their historical demands. This is counter-intuitive so MASE is not always easy to interpret.

2.3 *Ranking forecasters*

It is natural to ask: which forecaster is best for intermittent demand? There is no universally agreed ranking but CR is often applied in practice to intermittent demand (Fildes et al. 2008), and versions of CR are used in leading statistical forecasting software packages such as SAP and Forecast Pro (Teunter, Syntetos, and Babai 2011), so we might expect it to be ranked first. However, there is some debate on this issue.

Chatfield and Hayyab (2007) investigate whether SES, CR, ZF or a moving average forecaster is best for intermittent demand. They find that ZF does surprisingly well, beating SES and CR on lumpy demand under a modified MAPE but losing under MSE and U2. However, Teunter and Duncan (2009) note that ZF is of no practical use for inventory control. Syntetos and Boylan (2005) found that SES beat CR on intermittent demand, using more than one error measure, though CR was better if issue point only were considered. According to Gardner (2006) it is hard to conclude from the various studies that CR is best, because the results depend on the data and error measures used. Bacchetti and Saccani (2012) also note that there is no conclusive evidence pointing to a best method. But Teunter and Duncan (2009) conclude that the apparently poor performance of CR in some studies is caused by the use of inappropriate error measures, while Ghobbar and Friend (2003), Teunter and Duncan (2009), Willemain et al. (1994) found that CR beats SES on intermittent series.

Our position is that it is both reasonable and consistent with current wisdom to rank CR above SES, and SES above ZF, on intermittent demand series. CR has maintained its popularity over several decades, and if practitioners prefer a method based on experience, then any error measure that disagrees with this preference is of little use to them. We shall therefore take it as axiomatic that any error measure that fails this test, which we denote by CR $\succ$ SES $\succ$ ZF, should not be applied to intermittent demand. Of course some researchers will disagree with this position, which is entirely reasonable, but we hope that our work will be of use to those who agree with our ranking axiom.

2.4 *Mean-based error measures*

In Section 3, we shall test existing error measures against this axiom. But first we propose several new measures: in fact one for each existing measure, obtained by evaluating forecasts against the *mean of the underlying stochastic process* of the demand, which we denote y_t^m, instead of against the point demand y_t. So instead of the usual error $e_t = y_t - \hat{y}_t$ we use $e_t^m = y_t^m - \hat{y}_t$, instead of $p_t = 100e_t/y_t$ we use $p_t^m = 100e_t^m/y_t^m$, and similarly for baseline measures. The error e_t^m measures how well a forecaster ignores noise and estimates the underlying demand rate. We shall call these *mean-based* measures, and denote them by adding the prefix 'm' to the measure they are based on (mMAE, mMSE, etc).

We expect mean-based measures to give more appropriate results on intermittent demand than their standard versions. CR and SES arguably aim to provide an estimate of average demand rate, rather than accurate per-period forecasts. For example, if demands of mean size d occur with probability p, then CR and SES will both forecast demands of approximately pd, which is the average demand rate. Therefore, it seems reasonable to evaluate them against the actual demand rate rather than against point demands. CR should beat SES under this approach, as its forecasts remain close to the demand rate during periods of no demand, while the SES forecasts have greater variance. Finally, ZF should perform poorly under this approach as its forecasts are always below the demand rate. These predictions are tested below.

For artificial data, it is easy to find y_t^m. For stationary demand, we can use $y_t^m = \text{mean}(y_t)$, where the mean is either derived analytically (assuming that we use a distribution with a well-defined mean) or simply computed over the entire series. For non-stationary demand the mean of the stochastic process is a function of time, but we can still use our knowledge of the data to obtain the dynamic underlying demand rate. For example, the obsolescence experiments of Teunter, Syntetos, and Babai (2011) use artificial data whose non-zero demand probability drops either linearly or abruptly to 0, whereas demand sizes follow a fixed distribution: in either case we can multiply the current probability by the fixed distribution mean to obtain y_t^m.

On real-world series, the stochastic process is of course unknown, though one can assume a particular form (for example, a Poisson process) and then estimate its parameters. A simple approach is to estimate the current mean demand via standard techniques used to estimate seasonal components. It is common to estimate a seasonal component by taking a moving average over a window, stretching forward and backward in time. We can use the same moving window technique to obtain

a smoothed version of the demand series, and use this as y_t^m. If the demand series is too short to use a moving window, we can estimate the changing mean by regression. Or if we assume demand to be stationary, we can take the series mean as y_t^m. As with seasonal component estimation, there are several reasonable approaches.

Mean-based error measures have wider applicability than their original counterparts. Percentage errors such as MAPE are undefined whenever $y_t = 0$, whereas mMAPE is only undefined when $y_t^m = 0$ which is much rarer: for example, on stationary demand y_t^m is the series mean, which is only zero when $y_t = 0$ for all t. Relative error-based measures such as GMRAE with RW as baseline are undefined on intermittent demand, because both the demand and RW's forecast are often zero so the RW error e_t^* (which is the denominator) is also zero. However, a measure such as mGMRAE is only undefined when the denominator $e_t^{*m} = y_t^{*m} - \hat{y}_t^*$ is zero: the RW forecast $\hat{y}_t^*$ will often be zero, but again y_t^{*m} is rarely zero.

We now have a large number of new error measures, none of which is likely to give infinite answers on intermittent demand. This allows us to measure forecasting deviations using absolute or squared values, scaled or unscaled or based on percentages, and on one or multiple demand series. In the next section, we shall evaluate them with respect to our forecaster ranking axiom.

3. Experiments

The error measures we compare are selected from the various classes described in Section 2.2. To represent the scale-dependent measures, we use MAE, MdAE and MSE and their mean-based equivalents mMAE, mMdAE and mMSE; from the percentage errors, we use iMAPE and mMAPE. Note that the mMAPE of ZF is always $100e_t^m/y_t^m = 100(y_t^m - 0)/y_t^m = 100$ and the iMAPE of ZF is always $100e_t/y_t = 100(y_t - 0)/y_t = 100$ (MAPE itself is undefined on intermittent demand). To represent the relative error-based measures, we use only mGMRAE with RW as baseline (GMRAE with RW as baseline is undefined for intermittent demand). To represent the relative measures, we use PB and mPB with RW as baseline.

We start with data based on that used in experiments of Teunter, Syntetos, and Babai (2011). Demands occur with some probability in each period; hence, inter-demand intervals are distributed geometrically, and we use a logarithmic distribution for demand sizes. Geometrically, distributed intervals are a discrete version of Poisson intervals, and the combination of Poisson intervals and logarithmic demand sizes yields a negative binomial distribution, for which there is theoretical and empirical evidence: see, for example, the recent discussion in Syntetos et al. (2011). Teunter et al. generate demand data that is nonzero with probability p_0, where p_0 is either 0.2 or 0.5, and whose size is logarithmically distributed. The logarithmic distribution is characterised by a parameter $\ell \in (0, 1)$ and is discrete with $\Pr[X = k] = -\ell^k/k \log(1 - \ell)$ for $k \geq 1$. They use two values: $\ell = 0.001$ to simulate low demand and $\ell = 0.9$ to simulate lumpy demand.

Tables 1–4 show best results for SES, CR and ZF using α and β values chosen from $\{0.1, 0.2, 0.3\}$. We initialise the forecasters by choosing arbitrary initial values $\hat{y}_0 = \hat{\tau}_0 = 1$ then running them for 10^4 periods using demand probability p_0. Results are then computed over 10^5 time periods. To estimate the stochastic process mean, we simply compute the mean of all 10^5 demands (including zeros). The results show that MAE, MdAE and iMAPE are unreliable error measures for some types of intermittent demand because they incorrectly rank the three forecasters. PB is more reliable but the differences are sometimes very small, and in one case PB ranked CR and SES equally. Among existing measures, only MSE behaves correctly. In contrast, *all* mean-based measures behave correctly, though mPB still scores CR and SES quite similarly. We also tried geometrically distributed demand sizes as in Teunter, Syntetos, and Babai (2011), and regular intermittent demand as in Croston (1972), with similar results.

However, in further experiments MSE was also unreliable. Willemain et al. (1994) point out that demand in industrial data is often autocorrelated: demand may occur in streaks, with longer sequences of zero or nonzero values than one would expect. This is a positive autocorrelation on demand intervals, but they also observed negative autocorrelation: frequent alternation between zero and nonzero demand. Following Willemain et al., we model autocorrelation by a first-order 2-state Markov process. Let all demands be 0 or 1, and denote the transition probability from 0 to 1 by p_{01}, and from 1 to 0 by p_{10}. On negatively autocorrelated demand MSE ranks forecasters correctly, but results for positively autocorrelated demand with $p_{01} = p_{10} = 0.3$ are shown in Table 5. Here MAE, MdAE, MSE and iMAPE are all unreliable while mMAE, mMdAE, mMSE, mMAPE, mGMRAE and mPB give correct rankings, as does PB. We found similar results for other values of p_{01} and p_{10}.

Collectively, these results imply that almost all tested error measures are unreliable. They also imply that other untested error measures are unreliable, because they are monotonic functions of MAE or MSE and hence rank forecasters in the same way. These include RMSE, relative measures such as RelMAE and RelMSE and as their special cases U2 and LMR, and scaled errors such as MMR and MASE.

We conclude that all known error measures (except ME which measures bias, not deviation) are unreliable on some types of intermittent demand, even when they do not incur infinities, so there is currently no reliable way of measuring deviation. These results reinforce and complement those of Teunter and Duncan (2009), who show that MAE and RMSE rank ZF above

Table 1. Results for artificial demand with $p_0 = 0.2$ and $\ell = 0.001$.

Error	SES		CR			ZF					
measure	α	Error	α	β	Error	error	Forecaster ranking				
MAE	0.3	0.32134	0.1	0.3	0.31846	0.20141	ZF	$\succ$	CR	$\succ$	SES
MdAE	0.3	0.23740	0.3	0.3	0.20867	0.00000	ZF	$\succ$	CR	$\succ$	SES
MSE	0.1	0.16931	0.1	0.1	0.16271	0.20151	CR	$\succ$	SES	$\succ$	ZF
iMAPE	0.3	79.77339	0.1	0.1	80.07155	100.00000	SES	$\succ$	CR	$\succ$	ZF
PB	0.1	32.52000	0.1	0.1	32.52000	16.26000	CR	$=$	SES	$\succ$	ZF
mMAE	0.1	0.07434	0.1	0.1	0.03225	0.20122	CR	$\succ$	SES	$\succ$	ZF
mMdAE	0.3	0.13379	0.3	0.3	0.04651	0.20160	CR	$\succ$	SES	$\succ$	ZF
mMSE	0.1	0.00856	0.1	0.1	0.00167	0.04054	CR	$\succ$	SES	$\succ$	ZF
mMAPE	0.1	36.91216	0.1	0.1	16.01403	100.00000	CR	$\succ$	SES	$\succ$	ZF
mPB	0.3	97.83000	0.1	0.3	98.75000	20.15000	CR	$\succ$	SES	$\succ$	ZF
mGMRAE	0.1	0.30005	0.1	0.1	0.13512	0.84936	CR	$\succ$	SES	$\succ$	ZF

Table 2. Results for artificial demand with $p_0 = 0.5$ and $\ell = 0.001$.

Error	SES		CR			ZF					
measure	α	Error	α	β	Error	error	Forecaster ranking				
MAE	0.3	0.49945	0.1	0.3	0.49962	0.49963	SES	$\succ$	CR	$\succ$	ZF
MdAE	0.3	0.50463	0.3	0.3	0.50064	0.00000	ZF	$\succ$	CR	$\succ$	SES
MSE	0.1	0.26335	0.1	0.1	0.25643	0.50003	CR	$\succ$	SES	$\succ$	ZF
iMAPE	0.3	49.97213	0.1	0.1	24.98300	100.00000	CR	$\succ$	SES	$\succ$	ZF
PB	0.3	50.63000	0.1	0.1	50.65000	25.31000	CR	$\succ$	SES	$\succ$	ZF
mMAE	0.1	0.09310	0.1	0.1	0.06324	0.49993	CR	$\succ$	SES	$\succ$	ZF
mMdAE	0.3	0.15114	0.3	0.3	0.09643	0.49870	CR	$\succ$	SES	$\succ$	ZF
mMSE	0.1	0.01339	0.1	0.1	0.00614	0.24985	CR	$\succ$	SES	$\succ$	ZF
mMAPE	0.1	18.63340	0.1	0.1	12.65683	100.00000	CR	$\succ$	SES	$\succ$	ZF
mPB	0.3	99.95000	0.1	0.1	100.00000	49.84000	CR	$\succ$	SES	$\succ$	ZF
mGMRAE	0.1	0.17928	0.1	0.1	0.12201	0.99721	CR	$\succ$	SES	$\succ$	ZF

Table 3. Results for artificial demand with $p_0 = 0.2$ and $\ell = 0.9$.

Error	SES		CR			ZF					
measure	α	Error	α	β	Error	error	Forecaster ranking				
MAE	0.1	1.25741	0.1	0.3	1.23205	0.77191	ZF	$\succ$	CR	$\succ$	SES
MdAE	0.3	0.53944	0.3	0.3	0.67425	0.00000	ZF	$\succ$	SES	$\succ$	CR
MSE	0.1	7.10279	0.1	0.1	6.83258	7.35755	CR	$\succ$	SES	$\succ$	ZF
iMAPE	0.1	68.79125	0.1	0.1	59.83256	100.00000	CR	$\succ$	SES	$\succ$	ZF
PB	0.3	31.80000	0.1	0.3	33.09000	17.31000	CR	$\succ$	SES	$\succ$	ZF
mMAE	0.1	0.44085	0.1	0.1	0.21087	0.77266	CR	$\succ$	SES	$\succ$	ZF
mMdAE	0.3	0.56279	0.3	0.3	0.28997	0.76560	CR	$\succ$	SES	$\succ$	ZF
mMSE	0.1	0.36904	0.1	0.1	0.07450	0.59660	CR	$\succ$	SES	$\succ$	ZF
mMAPE	0.1	57.11138	0.1	0.1	27.31830	100.00000	CR	$\succ$	SES	$\succ$	ZF
mPB	0.2	87.12000	0.3	0.3	89.02000	12.42000	CR	$\succ$	SES	$\succ$	ZF
mGMRAE	0.1	0.58574	0.1	0.1	0.29300	1.09061	CR	$\succ$	SES	$\succ$	ZF

SES (and above a moving average), and SES above CR, on a large data-set of intermittent demand from an air force. Our results apply to more error measures and are more easily reproducible, being based on simple artificial data. But their result

Table 4. Results for artificial demand with $p_0 = 0.5$ and $\ell = 0.9$.

Error measure	SES α	SES Error	CR α	CR β	CR Error	ZF error	Forecaster ranking				
MAE	0.1	2.38007	0.1	0.3	2.28662	1.93788	ZF	$\succ$	CR	$\succ$	SES
MdAE	0.3	1.45094	0.3	0.3	1.38990	0.00000	ZF	$\succ$	CR	$\succ$	SES
MSE	0.1	16.01983	0.1	0.1	15.59856	18.97148	CR	$\succ$	SES	$\succ$	ZF
iMAPE	0.1	72.02938	0.1	0.3	65.44701	100.00000	CR	$\succ$	SES	$\succ$	ZF
PB	0.1	50.78000	0.1	0.1	50.89000	31.69000	CR	$\succ$	SES	$\succ$	ZF
mMAE	0.1	0.68463	0.1	0.1	0.48617	1.93752	CR	$\succ$	SES	$\succ$	ZF
mMdAE	0.3	0.95963	0.3	0.3	0.72759	1.90290	CR	$\succ$	SES	$\succ$	ZF
mMSE	0.1	0.79809	0.1	0.1	0.38147	3.75220	CR	$\succ$	SES	$\succ$	ZF
mMAPE	0.1	35.32882	0.1	0.1	25.08728	100.00000	CR	$\succ$	SES	$\succ$	ZF
mPB	0.3	78.51000	0.3	0.3	79.11000	15.88000	CR	$\succ$	SES	$\succ$	ZF
mGMRAE	0.1	0.94599	0.1	0.1	0.72287	2.83995	CR	$\succ$	SES	$\succ$	ZF

Table 5. Results for autocorrelated demand with $p_{01} = p_{10} = 0.3$.

Error measure	SES α	SES Error	CR α	CR β	CR Error	ZF error	Forecaster ranking				
MAE	0.3	0.41673	0.1	0.3	0.49992	0.49880	SES	$\succ$	ZF	$\succ$	CR
MdAE	0.3	0.37776	0.3	0.3	0.49221	0.00000	ZF	$\succ$	SES	$\succ$	CR
MSE	0.2	0.24507	0.1	0.1	0.26352	0.49880	SES	$\succ$	CR	$\succ$	ZF
iMAPE	0.3	41.77340	0.1	0.1	49.86910	100.000	SES	$\succ$	CR	$\succ$	ZF
PB	0.3	43.55000	0.1	0.3	44.29000	21.82000	CR	$\succ$	SES	$\succ$	ZF
mMAE	0.1	0.13732	0.1	0.1	0.09385	0.49902	CR	$\succ$	SES	$\succ$	ZF
mMdAE	0.3	0.23571	0.3	0.3	0.14783	0.49940	CR	$\succ$	SES	$\succ$	ZF
mMSE	0.1	0.02808	0.1	0.1	0.01344	0.24865	CR	$\succ$	SES	$\succ$	ZF
mMAPE	0.1	27.52970	0.1	0.1	18.81399	100.000	CR	$\succ$	SES	$\succ$	ZF
mPB	0.3	96.82000	0.3	0.3	98.65000	30.02000	CR	$\succ$	SES	$\succ$	ZF
mGMRAE	0.1	0.28674	0.1	0.1	0.20195	0.99883	CR	$\succ$	SES	$\succ$	ZF

shows that the inappropriateness of at least some current error measures extends to real data. In contrast, our new measures gave correct results in all cases.

Finally, we test our method on real data obtained from an Irish inventory control company. The data is from a spare parts inventory and contains 24 monthly demands for each of 8727 products, with high intermittency: the mean probability of a nonzero demand is 0.238. We would prefer to use long series, but it is common to have access only to relatively short series and companies must often make forecasts based on short demand histories (Syntetos and Boylan 2005); for methods to deal with short histories see Dolgui and Pashkevich (2008a, 2008b).

The forecasters we compare are CR, SES and ZF with the same smoothing factors as Syntetos and Boylan (2005): 0.05, 0.1, 0.15 and 0.2. They also used several thousand series each with 24 demands and we treat our series in a similar way: initialise SES and CR in the first year, then evaluate them in the second year. The estimates of the SES demand size, the CR demand size and the CR inter-demand interval are initialised to their averages over the first year. If no demand occurs in the first year, the inter-demand interval estimate is initialised to 12 and the demand size estimate to 1.

In our initial experiments, we found that the mean-based measures did not behave as well as expected, often ranking SES $\succ$ CR (though ZF was ranked correctly). This turned out to be caused by large differences between the mean demands over the first and second years. On short series, SES appears to cope with non-stationary demand better than CR, which takes longer to adjust its forecasts. We therefore applied a data-cleaning phase in which we discard any series whose first year mean demand is more than three times that of its second year, or vice-versa, leaving 2202 series.

We choose the relative measures RelMAE, U2 and PB, and compare their performance against their mean-based variants mRelMAE, mU2 and mPB. For the latter, we test two different methods for estimating the underlying stochastic process:

Table 6. Results on real spare parts inventory data.

Forecaster	RelMAE	U2	PB	mRelMAE fixed	mU2 fixed	mPB fixed	mRelMAE trend	mU2 trend	mPB trend
SES 0.05	1.011	0.728	31.49	0.119	0.100	98.89	0.291	0.206	87.14
SES 0.10	1.010	0.735	31.54	0.166	0.139	98.66	0.291	0.214	87.59
SES 0.15	1.010	0.744	31.58	0.228	0.191	97.04	0.313	0.240	87.27
SES 0.20	1.010	0.753	31.58	0.288	0.243	94.74	0.345	0.275	87.21
CR 0.05	1.001	0.721	31.55	0.129	0.098	98.77	0.307	0.217	86.58
CR 0.10	0.990	0.719	31.61	0.120	0.093	98.83	0.288	0.204	87.42
CR 0.15	0.980	0.718	31.70	0.118	0.094	98.54	0.274	0.196	88.01
CR 0.20	0.971	0.718	31.80	0.122	0.099	98.13	0.264	0.191	88.34
ZF	0.623	0.788	16.07	0.722	0.535	14.16	0.695	0.510	14.61

- *fixed*: under the assumption of stationarity, we use the mean of all 24 demands as the estimated means y_t^m of the underlying stochastic process for a series;
- *trend*: under the assumption of non-stationarity, we estimate the trend τ_t by linear least-squares over the 24 periods then take $y_t^m = \max(\tau_t, 0)$.

The results are shown in Table 6. Taking the best result for each forecaster, we find that U2, PB and all the mean-based measures except mPB/fixed rank the forecasters correctly: mPB/fixed ranks them as SES $\succ$ CR $\succ$ ZF, while RelMAE ranks them as ZF $\succ$ CR $\succ$ SES. We take the imperfect performance of mPB/fixed to indicate the incorrectness of the stationarity assumption, showing the importance of finding a good method for estimating the stochastic process. These results confirm our simulation results: that existing error measures may incorrectly rank forecasters; that MSE- and PB-based measures behave better than those based on MAE; and that mean-based measures are well behaved if we use an appropriate estimate of the demand means.

Based on our experiments, and on inapplicabilities pointed out by Hyndman and Koehler (2006), we make two recommendations regarding measures of deviation for intermittent demand. Firstly, we do not recommend *any* existing error measures. Secondly, we recommend several new error measures: mMSE, mRMSE, mMAE, mMdAE, mMAPE, mRelMSE (including special cases mU2 and mLMR plus other relative measures such as mRelMAE), mRelRMSE, mMMR, mMASE and mGMRAE. Which of these is best depends on user preference, and considerations such as whether errors are to be compared on one or across several series. However, care must be taken when estimating real demand means, and underlying trends should be taken into account.

4. Conclusion

We have shown that almost all known error measures rank forecasting methods incorrectly on some intermittent demand series. Even combining several measures does not cure the problem, because if they all rank forecasters incorrectly, then so will their combination (though this is not an argument against using multiple measures). Given this result, and the well-known fact that several error measures are inapplicable to intermittent demand because of infinities, there is currently no reliable way of measuring forecast deviation errors on such demands.

To address this problem, we described a simple way of modifying error measures so that they are more widely applicable and behave more correctly. This yields many new *mean-based* error measures that can be used to compare forecasters on intermittent demand. They are unlikely to be plagued by infinities, and in tests they consistently ranked forecasters correctly.

It might be argued that defining new error measures only adds to the confusion, but we believe that the improved behaviour and wider applicability of our measures make them worth considering when faced with intermittent demand. To simplify matters, we should perhaps recommend a small number of new measures. Based on popularity, the recommendations of experts and the desire to use measures based on both absolute and squared errors, we choose mMAPE, mGMRAE and mU2.

In future work, we shall evaluate other error measures that can be modified by our technique, and experiment with non-stationary demand. The statistical properties of the new error measures should also be investigated. Finally, the new error measures can also be applied to non-intermittent demand, and we shall evaluate their usefulness using series from the forecasting competitions.

Funding

This work was partially funded by Enterprise Ireland Innovation Voucher IV-2009-2092. S. Armagan Tarim is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under [grant number 110M500]. R. Rossi is supported by the University of Edinburgh CHSS Challenge Investment Fund and by the European Community's Seventh Framework Programme (FP7) under [grant number 244994] (project VEG-i-TRADE).

References

Ahlburg, D. A., C. Chatfield, S. J. Taylor, P. A. Thompson, R. L. Winkler, A. H. Murphy, and R. Fildes. 1992. "A Commentary on Error Measures." *International Journal of Forecasting* 8: 99–111.

Armstrong, J. S., and F. Collopy. 1992. "Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons." *International Journal of Forecasting* 8: 69–80.

Bacchetti, A., and N. Saccani. 2012. "Spare Parts Classification and Demand Forecasting for Stock Control: Investigating the Gap Between Research and Practice." *Omega* 40 (6): 722–737.

Boylan, J. E., and A. A. Syntetos. 2010. "Spare Parts Management: A Review of Forecasting Research and Extensions." *IMA Journal of Management Mathematics* 21 (3): 227–237.

Chatfield, D. C., and J. C. Hayyab. 2007. "All-zero Forecasts for Lumpy Demand: A Factorial Study." *International Journal of Production Research* 45 (4): 935–950.

Collopy, F. 1992. "Generalization and Communication Issues in the Use of Error Measures: A Reply." *International Journal of Forecasting* 8: 107–109.

Croston, J. D. 1972. "Forecasting and Stock Control for Intermittent Demands." *Operational Research Quarterly* 23: 289–304.

Dolgui, A., and M. Pashkevich. 2008a. "On the Performance of Binomial and Beta-binomial Models of Demand Forecasting for Multiple Slow-Moving Inventory Items." *Computers and Operations Research* 35 (3): 893–905.

Dolgui, A., and M. Pashkevich. 2008b. "Extended Beta-binomial Model for Demand Forecasting of Multiple Slow-moving Inventory Items." *International Journal of Systems Science* 39 (7): 713–726.

Fildes, R. 1992a. "The Evaluation of Extrapolative Forecasting Methods." *International Journal of Forecasting* 8 (1): 81–98.

Fildes, R. 1992b. "On Error Measures: A Response to the Commentators - The Best Error Measure?" *International Journal of Forecasting* 8: 109–111.

Fildes, R., K. Nikolopoulos, S. F. Crone, and A. A. Syntetos. 2008. "Forecasting and Operational Research: A Review." *Journal of the Operational Research Society* 59: 1150–1172.

Gardner Jr, E. S. 2006. "Exponential Smoothing: The State of the Art - Part II." *International Journal of Forecasting* 22 (4): 637–666.

Ghobbar, A. A., and C. H. Friend. 2003. "Evaluation of Forecasting Methods for Intermittent Parts Demand in the Field of Aviation: A Predictive Model." *Computers & Operations Research* 30: 2097–2114.

de Gooijer, J. D., and R. J. Hyndman. 2005. "25 Years of IIF Time Series Forecasting: A Selective Review." Tinbergen Institute Discussion Paper No 05–068/4. Tinbergen Institute.

Hyndman, R. J., and A. B. Koehler. 2006. "Another Look at Measures of Forecast Accuracy." *International Journal of Forecasting* 22 (4): 679–688.

Kolassa, S., and W. Schütz. 2007. "Advantages of the MAD/Mean Ratio Over the MAPE." *Foresight: The International Journal of Applied Forecasting* 6: 40–43.

Levén, E., and A. Segerstedt. 2004. "Inventory Control With a Modified Croston Procedure and Erlang Distribution." *International Journal of Production Economics* 90 (3): 361–367.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler. 1982. "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition." *Journal of Forecasting* 1: 111–153.

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord, and L. F. Simmons. 1993. "The M-2 Competition: A Real-Time Judgmentally Based Forecasting Study." *International Journal of Forecasting* 9: 5–23.

Makridakis, S., and M. Hibon. 2000. "The M3-Competition: Results, Conclusions and Implications." *International Journal of Forecasting* 16 (4): 451–476.

Syntetos, A. A. 2001. "Forecasting for Intermittent Demand." Unpublished PhD thesis., Buckinghamshire Chilterns University College, Brunel University, UK.

Syntetos, A. A., and J. E. Boylan. 2005. "The Accuracy of Intermittent Demand Estimates." *International Journal of Forecasting* 21: 303–314.

Syntetos, A., Z. Babai, D. Lengu, and N. Altay. 2011. "Distributional Assumptions for Parametric Forecasting of Intermittent Demand." In *Service Parts Management: Demand Forecasting and Inventory Control*, edited by N. Altay and A. Litteral, 31–52. New York, NY: Springer Verlag.

Teunter, R. H., and L. Duncan. 2009. "Forecasting Intermittent Demand: A Comparative Study." *Journal of the Operational Research Society* 60: 321–329.

Teunter, R. H., A. A. Syntetos, and M. Z. Babai. 2011. "Intermittent Demand: Linking Forecasting to Inventory Obsolescence." *European Journal of Operations Research* 214 (3): 606–615.

Thiel, H. 1966. *Applied Economic Forecasting*. Chicago: Rand McNally.

Thompson, P. A. 1990. "An MSE Statistic for Comparing Forecast Accuracy Across Series." *International Journal of Forecasting* 6: 219–227.

Venkitachalam, G. H. K., D. B. Pratt, C. F. DeYoung, S. A. Morris, M. L. Goldstein. 2003. "Forecasting and Inventory Planning for Parts With Intermittent Demand – A Case Study." Presented at the Industrial Engineering Research Conference, Portland, OR, USA.

Wallström, P., and A. Segerstedt. 2010. "Evaluation of Forecasting Error Measurements and Techniques for Intermittent Demand." *International Journal of Production Economics* 128 (2): 625–636.

Willemain, T. R., C. N. Smart, J. H. Shockor, and P. A. DeSautels. 1994. "Forecasting Intermittent Demand in Manufacturing: A Comparative Evaluation of Croston's Method." *International Journal of Forecasting* 10 (4): 529–538.