1991-11

Exponential smoothing: The effect of initial values and loss functions on post-sample forecasting accuracy

Makridakis, Spyros

Elsevier Ltd

Exponential smoothing: The effect of initial values and loss functions on post-sample forecasting accuracy

Spyros Makridakis and Michèle Hibon
INSEAD, 77305 Fontainebleau Cedex, France

Abstract: This paper describes an empirical investigation aimed at measuring the effect of different initial values and loss functions (both symmetric and asymmetric) on the post-sample forecasting accuracy. The 1001 series of the M-competition are used and three exponential smoothing methods are employed. The results are compared over various types of data and forecasting horizons and validated with additional data. The paper concludes that contrary to expectations, post-sample forecasting accuracies are not affected by the type of initial values used or the loss function employed in the great majority of cases.

Keywords: Forecasting, Time series, Exponential smoothing, Accuracy, *M*-competition.

1. Introduction

Exponential smoothing methods are widely used in many industrial applications including production planning, production scheduling and inventory control [Brown (1959); Brown (1963); Brown (1967); Gardner (1985); Holt et al. (1960); Johnson and Montgomery (1974); Makridakis and Wheelwright (1989); Winters (1960)]. Although extremely simple and easy to model, such methods have been found by many studies to be as accurate as more complex and statistically sophisticated alternatives [Groff (1973); Chatfield (1978); Koehler and Murphree (1988); Makridakis and Hibon (1979); Makridakis et al. (1982); Martin and Witt (1989)]. Furthermore exponential smoothing methods are robust, easy to program, require a minimum of historical data while the cost of running them on the computer is the smallest of all available alternatives.

The purpose of this paper is to empirically investigate the effect of various initial values and loss functions on the post-sample forecasting accuracy of three of the most widely used (Single, Holt's and Dampened) exponential smoothing methods. The fourth widely used method (Winters') was not utilized as empirical findings have shown it to produce forecasts very similar to those of Holt's [see Makridakis et al. (1982)]. Section 2 of the paper reviews the literature and provides the reasoning for undertaking this study. Section 3 describes the methodology used and formulates various hypotheses to be studied. Section 4 presents and analyses the results. There is a concluding section which discusses the implications of the findings, validates such findings with another set of data provided by Fildes (1989) and presents possible avenues for further research.

2. Literature review and reasons for undertaking this study

Since the introduction of exponential smoothing methods the question of how to initialize the first smoothed value(s) has always been posed [Cogger (1973); McClain (1981); Taylor (1981); Wade (1967)]. Several alternatives have been proposed in the literature, but there is little advice on which of these alternatives to use [see Chatfield and Yar (1988)]. The most common among them are the following (see Appendix for more details):

(1) *Least squares estimates*: The historical data available is used to estimate ordinary least squares estimates of the initial value(s) [Brown (1959)]. In practice this is the most widely used approach for computing them.

(2) *Backcasting*: The data is inverted and forecasting starts using the most recent data and going backwards forecasting the less recent ones. The forecast, or smoothed values, at period 1 are then used as initial values to start the usual forecasting [Ledolter and Abraham (1984)].

(3) *Training set*: The data is divided into two parts. The first part (usually the smaller of the two) is used to estimate the initial values for the exponential smoothing equation(s) used with the second part where the final forecasts [see Makridakis et al. (1983)] are being based.

(4) *Convenient initial values*: Some convenient values can be used to initialize the smoothing equation(s). For instance, the first data value can be used to initialize the level, while the difference between the first and the second actual value (or the average of the second minus the first and the fourth minus the third) can be used to initialize the trend. [Makridakis and Wheelwright (1978).]

(5), (6) and (7) *Zero values*: The initial values can be all set to zero, or alternatively one can be set to zero and the other(s) can be initialized using one of the alternatives described above. This set of value(s) can be used as benchmarks to judge the improved accuracy of approaches 1 to 4 above. Although it seems an unreasonable alternative it provides an advantage in terms of large initial errors which force the estimated values to approach the actual ones much faster than alternative initialization procedures.

Because of the widespread applications of exponential smoothing methods even small reductions in their forecasting errors can bring big improvements in terms of lower costs and/or better customer services [Gardner (1990a)]. At present few guide-lines and no empirical evidence exist to help users decide upon the best initialization procedure [see Chatfield and Yar (1988); Gardner (1985)]. The present study aims to provide such empirical evidence and propose guidelines, if any exist, for selecting appropriate initialization approaches.

Forecasting and, in general, statistical models can be optimized using a number of loss functions such as linear, quadratic, or higher order. The rationale behind such choice is that the negative consequence of forecasting errors are not necessarily proportional. Thus higher order loss functions which penalize bigger errors, in a quadratic or cubic fashion, can be used. On the other hand, when forecasting errors are considered to be proportional then a linear loss can be employed. As in the case of initial values there is not much help or empirical evidence to guide the choice of the best loss function to optimize a model's parameters [Cogger (1979); Granger (1969); Montgomery and Johnson (1976)], although, in practice the great majority of computer programs employ a quadratic loss that minimizes the sum of square errors when a model is fitted to historical data. The aim of this paper is to study the influence of the five most widely used loss functions on the post-sample forecasting accuracy of the three exponential smoothing methods utlilized in the present study.

Finally, the effects of non-symmetric loss-functions are investigated as in practice the cost of negative errors (i.e., underestimating demand) is usually considered more critical than that of positive ones (i.e., overestimating demand). Although alternative forms of modeling non-symmetric loss functions might be possible in the present paper our purpose is to simply determine the influence of non-symmetric losses on the post-sample forecasting errors and suggest guide-lines, if any exist, in using non-symmetric loss functions to balance the cost of negative versus positive forecasting errors.

3. Experimental design and methodology

The three (Single, Holt's and Dampened) most commonly used exponential smoothing methods were selected for the study (see Appendix for a description of the models involved). Seven types of initial values (see last section and Appendix) were used for Holt's and Dampened smoothing and five for Single. In addition five optimization criteria (loss functions) were employed. They range from a linear to a cubic power one (see Appendix). The optimization of the model parameter(s) was done using *a grid search algorithm* which found the optimal smoothing constants through finer and finer searches around a global optimum initially identified through the grid search. In total 35 possibilities were tested for each of the three smoothing methods. A non-symmetric loss function was also applied by weighting positive errors less than negative ones. Such weighting was done at five levels (0.35, 0.50, 0.65, 0.80, 0.95) while computing the model fitted errors. Consequently the post sample forecasting accuracy of each horizon and method was recorded and compared to that of symmetric optimization.

The methodology employed consisted of using the 1001 series of the M-competition [see Makridakis et al. (1982)] for each of the applicable possiblities. The procedure used was exactly the same as utilized in the M-competition. This means that when a data series was seasonal its values were first deseasonalized using the classical decomposition method [the post-sample forecasting accuracy when seasonal series were deseasonalized using other decomposition approaches were not different than those of the classical decomposition, see Makridakis et al (1982)], a forecasting model was subsequently estimated and forecasts from this model obtained. Finally, these forecasts were reseasonalized using the seasonal indices found by the classical decomposition method if the data was indeed seasonal. If the data series was not seasonal the model was estimated directly on the original data and forecasts were directly found. Following the above-mentioned procedure, optimal model parameters were estimated and subsequently used to forecast for periods $1, 2, \ldots, m$ (where $m = 6$ for yearly data, $m = 8$ for quarterly data and $m = 18$ for monthly data). These forecasts were then compared to the actual values (known but obviously not used in developing the forecasting model) so as to compute the post-sample forecasting errors for each of the m forecasting horizons. Three post-sample accuracy measures were computed from such errors: the Mean absolute deviations (MAD), the Mean absolute percentage errors (MAPE) and the mean square errors (MSE). These accuracy measures were calculated separately for yearly, quarterly and monthly data and were also summarized for all data and forecasting horizons. Similarly, the same accuracy measures were also computed when a non-symmetric loss function was used to optimize the parameter(s) in the model fitting phase.

The approach used in this study is not different to that of real life applications where m forecasts are made at period t (present) even though their accuracy can only be found in the future when the actual data becomes available.

4. Presentation and analysis of the results

Table 1 shows the MAPE of the best and worst initialization alternatives together with that of least square estimates (the most widely used approach) for various forecasting horizons. The optimization alternative used was that of minimizing a symmetric quadratic (MSE) loss function. As it can be seen

Table 1
Comparison of initialization.

Optimization by MSE			Forecasting horizons						
			1	3	6	8	12	18	Average (all horizons)
All data	Single	Least squares	8.7	13.3	19.7	18.0	16.9	26.1	17.0
		Best: Convenient	8.5	13.1	19.4	17.9	16.9	26.1	16.9
		Worst: Zero	8.8	13.2	19.6	18.4	16.9	25.8	17.0
All data	Holt	Least squares	8.7	12.9	21.3	22.7	21.3	33.6	19.8
		Best: Both zero	8.6	12.5	19.4	20.6	19.4	34.8	18.7
		Worst: Convenient	8.8	13.4	22.1	25.4	26.0	42.5	22.8
All data	Dampen	Least squares	8.5	12.4	18.5	18.1	17.2	27.4	17.0
		Best: Convenient	8.4	12.5	18.7	18.2	17.2	27.5	17.0
		Worst: Both zero	8.7	12.9	19.2	18.8	17.2	27.1	17.3

Table 2
Comparison of optimization.

Initialization by prevalent values			Forecasting horizons						
			1	3	6	8	12	18	Average (all horizons)
All data	Single	MSE	8.7	13.3	19.7	18.0	16.9	26.1	17.0
		Best: Median	8.	13.0	19.4	17.9	16.8	25.9	16.9
		Worst: e_t^3	8.7	13.4	19.8	18.0	16.9	26.1	17.0
All data	Holt	MSE	8.7	12.9	21.3	22.7	21.3	33.6	19.6
		Best: MSE	8.7	12.9	21.3	22.7	21.3	33.6	19.6
		Worst: Median	9.5	15.5	24.6	28.3	33.0	50.0	27.6
All data	Dampen	MSE	8.5	12.4	18.5	18.1	17.2	27.4	17.0
		Best: MAPE	8.7	12.2	18.6	17.8	17.5	24.8	16.8
		Worst: Median	8.5	12.4	19.1	19.3	17.9	26.7	17.4

in Table 1 the differences in average forecasting accuracy between the best and worst alternatives are extremely small. The same type of results can be observed in Table 2 which shows the MAPE of the best and worst symmetric optimization alternative together with that of the MSE (the most widely used approach) for various forecasting horizons when the initialization approach employed was that of ordinary least squares. None of the differences are large except for those of Holt's for longer than six forecasting horizons.

In conclusion, there is no evidence from the empirical results, except in the case of Holt's smoothing for periods longer than six horizons, to suggest that differences in the initialization procedures and/or loss functions affect the post-sample forecasting accuracies. The same conclusions can be drawn when the data are separated into yearly, quarterly and monthly. All of the observed differences when various initial approaches and loss functions are used are extremely small except in the case of Holt's smoothing for periods longer than six horizons. Table 3, for instance, shows the results of an analysis of variance for yearly data ($m = 6$ for such data) when Dampened smoothing was used. None of the differences between initialization procedures (columns), optimization criteria (rows) or their interaction is statistically significant (the smallest P-value is equal to 0.34). When the same analysis of variance is conducted for Holt's smoothing the only statistically significant differences come from the long horizon effect.

Table 4 summarizes the MAPE for the best and worst alternatives for the various initialization values and loss functions. 'B' on the top right corner of each box means 'Best' among the horizontal alternatives (i.e., optimization criteria) while 'W' signifies 'Worst'. Similarly, 'B' and 'W' on the left, lower corner of each box mean 'Best' and 'Worst' alternative among the vertical ones (i.e., initialization values). Table 4(a) presents the results of Single smoothing, 4(b) presents those of Holt's while 4(c) presents those of Dampen.

The differences in Table 4(a) are extremely

Table 3
Two-way analysis of variance. [a]

Source	Sum of squares	d.f.	Mean square	Computed F-value	p-value
Columns	2266.0	5.0	453.0	0.87	0.498
Rows	2940.0	5.0	587.0	1.13	0.340
Row × columns	1199.0	25.0	47.0	0.09	1.000
Error	3062580.0	5904.0	519.0		
Totals	3068990.0	5939.0			

[a] 165 series of yearly data with Dampen-trend method. Average Error on 6 forecasting horizons. Horizontal values: Errors for each Optimization criteria. Vertical values: Errors for each Starting value. The differences are not statistically significant.

small (the largest is only 0.3%). Moreover, none of the alternatives perform consistently 'best' or 'worst' in terms of the initialization and optimization alternatives experimented with.

The differences in Table 4(b) are considerably bigger than those in Table 4(a). In addition the median is consistently the worst optimization approach while MSE is consistently the best. Among the different initialization alternatives the best results are found when both initial values are set to zero, except in one case where the best result is when only one of the two is set to zero.

Differences in Table 4(c) are small and less consistent than those in Table 4(b). The loss function which does best most of the time is that of MAPE while the corresponding 'best' for initialization is that of least square estimates and convenient values.

There are no consistent results that hold across Tables 4(a), 4(b), and 4(c). Thus, we cannot say that some specific initialization procedure or loss function holds true across all three exponential smoothing methods studied. In Single smoothing, Table 4(a), the best results are found when the loss

Table 4(a)
Single smoothing, average MAPE for all forecasting horizons and time series. [a]

Initial values	Symmetric loss functions				
	I MAD	II MAPE	III Median	IV MSE 2nd power	V Cubic power
1 Least squares estimates	16.9^{B}	16.9^{B}	${}_{W}16.9^{B}$	${}_{W}17.0^{W}$	17.0^{W}
2 Backcasting	16.9	${}_{B}16.8$	${}_{B}16.7^{B}$	${}_{B}16.9$	17.0^{W}
3 Training set	16.9	${}_{B}16.8^{B}$	16.8^{B}	${}_{W}17.0$	${}_{W}17.1$
4 Convenient values	${}_{B}16.8^{B}$	${}_{B}16.8^{B}$	16.8^{B}	${}_{B}16.9^{W}$	${}_{B}16.9^{W}$
5 $s = 0$, or $s = 0$ and $t = 0$	${}_{W}17.0$	${}_{W}17.1^{W}$	${}_{W}16.9^{B}$	${}_{W}17.0$	17.0

[a] 'B' at the upper, right hand side of each box signifies best while 'W' signifies worst accuracy. 'B' at the lower, left hand side of each box signifies best, while 'W' signifies worst accuracy.

Table 4(b)
Holt's smoothing, average MAPE for all forecasting horizons and time series. [a]

Initial values	Symmetric loss functions				
	I MAD	II MAPE	III Median	IV MSE 2nd power	V Cubic power
1 Least squares estimates	19.8^{B}	19.9	27.6^{W}	19.8^{B}	20.3
2 Backcasting	20.7	21.5	27.8^{W}	20.3^{B}	21.6
3 Training set	21.4	22.8	30.5^{W}	20.8^{B}	22.5
4 Convenient values	${}_{W}23.5$	${}_{W}23.6$	${}_{W}29.6^{W}$	${}_{W}22.8^{B}$	${}_{W}24.8$
5 $s = 0$, or $s = 0$ and $t = 0$	${}_{B}19.0$	${}_{B}18.9$	27.0^{W}	${}_{B}18.7^{B}$	${}_{B}19.7$
6 $s = 0$ t = least squares	19.1	19.1	27.1^{W}	18.9^{B}	${}_{B}19.7$
7 s = least squares $T = 0$	19.2^{B}	19.3	${}_{B}26.8^{W}$	21.8	25.0

[a] 'B' at the upper, right hand side of each box signifies best while 'W' signifies worst accuracy. 'B' at the lower, left hand side of each box signifies best, while 'W' signifies worst accuracy.

Table 4(c)
Dampen smoothing, average MAPE for all forecasting horizons and time. [a] Series.

Initial values	Symmetric loss functions				
	I MAD	II MAPE	III Median	IV MSE 2nd power	V Cubic power
1 Leat squares estimates	16.9	16.8^{B}	17.4^{W}	${}_{B}17.0$	${}_{B}17.2$
2 Backcasting	17.0	${}_{B}16.7^{B}$	17.4^{W}	17.1	17.3
3 Training set	17.0^{B}	17.0^{B}	17.4^{W}	17.3	17.4^{W}
4 Convenient values	${}_{B}16.8$	${}_{B}16.7^{B}$	17.4^{W}	${}_{B}17.0$	${}_{B}17.2$
5 $s = 0$, or $s = 0$ and $t = 0$	${}_{W}17.3$	${}_{W}17.2^{B}$	${}_{W}17.7^{W}$	${}_{W}17.3$	17.4
6 $s = 0$ t = least squares	17.1^{B}	${}_{W}17.2$	17.6^{W}	17.1^{B}	${}_{B}17.2$
7 s = least squares $T=0$	16.9^{B}	17.1	${}_{B}17.3$	17.2	${}_{W}17.5^{W}$

[a] 'B' at the upper, right hand side of each box signifies best while 'W' signifies worst accuracy. 'B' at the lower, left hand side of each box signifies best, while 'W' signifies worst accuracy.

function is the median and the worst the cubic power. There is no best initialization procedure although the worst is when the first value is set to zero. However, it must be emphasized that all differences are extremely small. In Holt's smoothing, Table 4(b), the worst loss function is the median [the opposite of Table 4(a)] and the best is the MSE in all but one case when the MAD provides the best results. In terms of initial values the best post-sample accuracies are found when the first values are both set to zero, except in one case when only one of the two is set to zero. The worst results are when convenient values are used to initialize.

For Dampened smoothing, Table 4(c), the results are closer to those of Single. Thus the worst loss function is the cubic power (although not in all cases), the worst initialization is with zero values (not in all cases) while the best is found with convenient values (again not in all cases).

Table 5a
Single smoothing, average MAD for all forecasting horizons and time series. [a] (Values have been divided by 1000)

Initial values	Loss functions				
	I MAD	II MAPE	III Median	IV MSE 2nd power	V Cubic power
1 Least squares estimates	15.4	15.5^{W}	15.3^{B}	${}_{W}15.4$	${}_{W}15.4$
2 Backcasting	15.4^{W}	15.4^{W}	${}_{B}15.1$	15.3	${}_{B}15.0^{B}$
3 Training set	${}_{B}15.3^{W}$	${}_{B}15.2$	15.3^{W}	${}_{B}15.2$	${}_{B}15.0^{B}$
4 Convenient values	15.5^{W}	15.4	15.5^{W}	15.3	${}_{B}15.0^{B}$
5 $s = 0$, or $s = 0$ and $t = 0$	${}_{W}15.6^{W}$	${}_{W}15.6^{W}$	${}_{W}15.6^{W}$	${}_{W}15.4$	15.1^{B}

[a] 'B' at the upper, right hand side of each box signifies best while 'W' signifies worst accuracy. 'B' at the lower, left hand side of each box signifies best, while 'W' signifies worst accurcy.

Table 5(b)
Holt's smoothing, average MAD for all forecasting horizons and time series [a] (Values have been divided by 1000)

Initial values	Loss functions				
	I MAD	II MAPE	III Median	IV MSE 2nd power	V Cubic power
1 Least squares estimates	12.4^{B}	13.2	${}_{B}17.3^{W}$	12.4^{B}	15.1
2 Backcasting	11.9	${}_{B}12.7$	21.9^{W}	11.8^{B}	${}_{W}15.8$
3 Training set	${}_{W}15.5$	${}_{W}18.7$	${}_{W}28.1^{W}$	${}_{W}14.3^{B}$	14.9
4 Convenient values	13.6	15.6	23.5^{W}	11.7^{B}	14.4
5 $s =$ or $s = 0$ and $t = 0$	${}_{B}11.3^{B}$	12.9	20.4^{W}	${}_{B}11.5$	${}_{B}13.9$
6 $s = 0$ $t =$ least squares	12.2^{B}	13.6	22.8^{W}	12.2^{B}	14.7
7 $s =$ least squares $T = 0$	13.0^{B}	13.9	18.6^{W}	13.6	${}_{B}13.9$

[a] 'B' at the upper, right hand side of each box signifies best, while 'W' signifies worst accuracy. 'B' at the lower, left hand side of each box signifies best, while 'W' signifies worst accuracy.

The great majority of differences are extremely small while in Single smoothing such differences are even smaller – practically zero.

Table 5 shows results similar to those of Table 4 except the post-sample accuracy is that of MAD instead of MAPE. Concerning Single smoothing the differences in post-sample MADs are extremely small as was the case in Table 4(a). However, there are no other consistent patterns between Tables 4(a) and 5(a). For instance, in Table 5(a) there is a consistent improvement in post-sample MAD when the model parameters are optimized through a cubic loss function while in Table 4(a) this is not the case. Similarly, the initial values

Table 5(c)
Dampenend smoothing, average MAD for all forecasting horizons and time series. [a] (Values have been divided by 1000)

Initial values	Loss functions				
	I MAD	II MAPE	III Median	IV MSE 2nd power	V Cubic power
1 Least squares estimates	13.4	14.9^{W}	14.2	${}_{B}12.2^{B}$	12.9
2 Backcasting	13.3	14.3	14.4^{W}	12.8^{B}	${}_{W}13.2$
3 Training set	13.5	${}_{B}13.2$	15.8^{W}	13.1^{B}	13.1^{B}
4 Convenient values	${}_{B}13.2$	14.3	${}_{W}18.7^{W}$	12.9^{B}	${}_{W}13.2$
5 $s = 0$, or $s = 0$ and $t = 0$	13.6	${}_{W}15.6^{W}$	15.1	12.4^{B}	12.9
6 $s = 0$ $t =$ least squares	13.9	14.8	${}_{B}15.0^{W}$	12.6^{B}	${}_{B}12.4$
7 $s =$ least squares $T = 0$	${}_{W}15.0$	15.3^{W}	${}_{B}15.0$	${}_{W}14.1$	13.0^{B}

[a] 'B' at the upper, right hand side of each box signifies best, while 'W' signifies worst accuracy. 'B' at the lower, left hand side of each box signifies best, while 'W' signifies worst accuracy.

that provide the most accurate results are not the same in Tables 4(a) and 5(a).

In Holt's smoothing the differences in MADs are bigger than those of single smoothing, however the cubic power loss function does not improve the results. Moreover, the median continues to be the worst optimization alternative while the MSE is the best. There is also consistency in initialization procedures where the results of Table 4(b) and 5(b) are similar. Thus, setting both initial values at zero provides the best results most of the time while the worst results are found when the initialization is done through a training set.

Finally, there is little consistency between Tables 4(c) and 5(c) – Dampened smoothing. In Table 5(c) the best optimization criterion is MSE in all but one case, while in Table 4(c) the best was MAPE (in all but two cases). Finally, there is no initialization procedure which is consistently best in Table 5(c) while the best in 4(c) was that of convenient values (in all but two cases).

Thus, it can be concluded that few consistent results can be reported between Tables 4(a) and 5(a) and 4(c) and 5(c). That is whatever, if anything, influences post-sample MAPEs does not consistently influence MADs. This is not, however, the case with Tables 4(b) and 5(b) – referring to Holt's smoothing – where the results are fairly consistent.

Although, the authors are well aware of the problems of using MSE over many series of unequal values they also computed post-sample accuracies using such measure in order to provide a complete range of results and anticipate possible criticism that a widely-used measure such as the MSE was not used. As it could be expected, the values found were large and extremely unstable. The averages were often reduced by a factor of 10,000 by excluding as few as six series. Given the large number of series and forecasting horizons involved (almost 14,000 in total) such large fluctuations make the use of MSE inappropriate as a comparative measure [see also Chatfield (1988)]. Furthermore, no consistent or insightful results could be deduced by examining the various tables of post-sample MSE values even when large errors were excluded. This is why tables using MSEs are not reported in this paper.

The non-symmetric optimization was done using ordinary least square estimates for initial values and a quadratic (MSE) loss function. Five levels of non-symmetric losses were used by adding to the sum of model fitting square errors 35%, 50%, 65%, 80% or 95% of the square error at period t, when such error was positive while adding the entire square error when it was negative (see Appendix for more details). As usual the parameter(s) that minimized the sum of square errors of model fitted were chosen and were used to make m forecasts and subsequently compute the post-sample accuracies.

The differences in past-sample accuracies when a non-symmetric loss function was used were extremely small for *all three* exponential smoothing methods. In Single smoothing the great majority of such differences were in the second decimal. In Holt's smoothing there were some small improvements in post-sample accuracies for longer than twelve forecasting horizons when the non-symmetric loss function, at the 35% level, was used. However, such differences were not statistically significant while the best overall results were *still* obtained when a symmetric loss function was employed. In Dampened smoothing the best overall results were found with a non-symmetric loss function at the 50% level. Furthermore, for twelve or longer forecasting horizons the improvements were considerably larger than those of Single or Holt's smoothing *and* consistent but small in absolute values.

4.1. Discussion

The purpose of this paper is *not* to enter into the debate of which accuracy measure is the most appropriate or what is the value of empirical competitions. Such issues have been debated elsewhere [Chatfield (1988); Fildes and Makridakis (1990); Zellner (1986); Armstrong and Lusk (1983)]. Instead, it aims at investigating the issues of the various initial values proposed in the literature and a range of loss functions used at present. At the same time, the authors are well aware that relative measures such as MAPE are more appropriate when averaging over many series and this is why MAPEs were used to express the results of this study in all tables except 5 which uses MADs. At the same time MSE were also computed for reasons of completeness.

If the median is excluded as a loss function to base the optimization of model parameters, few consistent differences can be found in past-sample forecasting accuracies whether such accuracies are measured in terms of MAPE, MAD or MSE. Moreover, no consistent patterns could be found when MAPES, MADS or MSES criteria were used to optimize the model's parameters and MAPES, MADS or MSES measures were employed to compute post-sample accuracies. Thus, there was no correspondence between the type of loss function used during the model fitting and the accuracy measure employed to compute the post-sample errors.

These results are surprising. In the forecasting literature the initialization procedure and the optimiziation criteria have been considered to influence post-sample forecasting accuracies. It has been also advocated that there must be a correspondence between the loss function used in the model fitting and the corresponding post-sample accuracy employed to measure forecasting errors [Zellner (1986)].

From a practical point of view the prevalent approaches of using MSE as a loss function and ordinary least square estimates to initialize the starting values seems adequate as the differences between such approaches and the best of the alternatives are small and statistically non-significant, in the great majority of cases. Furthermore, as these approaches (MSE as the loss function and least square estimates for initialization) are easy to program and require little computer time to apply there is no motivation to change them. On the other hand, it makes no sense to consider more elaborate alternatives such as backcasting for initial values or medians for optimizing the model's parameter(s) since such alternatives are more difficult to program and require more computer time when used to obtain forecasts.

In addition to the various results reported in the last section several other ideas were tested during our study. For instance we found that *sample size did not exhibit any consistent influence on the magnitude of post-sample forecasting errors or the choice of the best initialization or optimization alternatives*. This finding is consistent with that reported in Makridakis and Hibon (1979) and, is no doubt, due to the fact that the pattern of the series changes even abruptly in some cases. In addition, if frequency distributions of the differences in post-sample errors between the various approaches were made it was found that the great majority of them were less than 1% – this was in particular true with Single and Dampened smoothing. Furthermore, no obvious patterns of such differences could be deduced and no important factors could be found that could explain the larger than 1% errors.

Another idea tested was whether a specific set of initial values or loss functions was best for yearly, quarterly or monthly data. But again no consistent conclusion that holds among the three methods could be reached. Similarly, no forecasting horizons could be better predicted than others by the appropriate choice of specific initial values or loss functions. Finally, loss functions not used either in theory or practice (e.g., the power of 1.5, 2.5 and 4) were also tried and again the results showed few consistent differences except possibly when optimizing using a 4th power loss function which produced the worst results in most of the cases studied.

The practical implications of our study suggest that there are few benefits, if any, in attempting to find optimal ways to initialize the values of exponential smoothing methods (at least the three we studied). Moreover, the choice of a best loss function is of no consequence as long as the median is excluded. As Gardner (1990b) explained 'the reason that starting values and loss functions don't make any difference is that the optimal smoothing parameter(s) found *compensate* for various starting values and different loss functions'. However, we must emphasize that our results apply to the average of forecasts that have been found mechanically (i.e., using an automatic approach) without studying each series separately to determine the best initial values or optimal loss function. In our view additional research will be required to determine if our findings also apply when single series are studied and optimized individually [e.g., see Chatfield (1978); Chatfield and Yar (1988)].

Our findings suggest that the prevalent approach of initializing by ordinary least squares and optimizing by a quadratic loss (MSE) function provide satisfactory results which, on average when the methods are run mechanically, cannot be improved in any consistent way that holds constant across methods, data types, forecasting horizons or sample sizes. These conclusions are both good and bad news. The good news is that exponential smoothing methods (and in particular Simple and Dampened) are easy, accurate and robust forecasting techniques that can be readily used across a

Table 6(a)
Single smoothing, average MAPE for all forecasting horizons and time series.

Initial values	Fildes data				
	I MAD	II MAPE [a]	III Median	IV MSE 2nd power	V Cubic power
1 Least squares estimates	18.1	18.1	18.1	18.2	18.2
2 Backcasting	18.1	18.1	18.2	18.1	18.2
3 Training set	18.1	17.9	18.1	18.1	18.1
4 Convenient values	18.1	17.9	18.2	18.1	18.2
5 $s = 0$, or $s = 0$ and $t = 0$	18.1	18.1	18.1	18.1	18.1

[a] Practically all the MAPES are the same in this table. This is why no best and worst approach has been indicated.

wide range of actual forecasting applications. The bad news is that theoretical expectations do not seem to hold empirically for reasons that are not always clear apart from saying that the pattern of series is changing. Thus, research efforts must concentrate on better understanding such reasons and in developing alternative methods and approaches that can more accurately predict real life time series whose pattern, we know, change over time. Somehow it must be possible to beat Single smoothing for longer forecasting horizons and Dampened for shorter and medium ones. Moreover, research efforts must be directed in better understanding the effects of one-period-ahead versus two, three,...,m-period optimization and their consequence on post-sample forecasting accuracy [Makridakis (1990)]. Finally, more research needs to be done to better understand the lack of consistency between various loss functions used in model optimization and the resulting post-sample accuracies. For instance, one would have expected a correspondence between the type of loss function used in model fitting and the best results found when post-sample accuracies were mea-

Table 6(b)
Holt's smoothing, average MAPE for all forecasting horizons and time series. [a]

Initial values	Fildes data				
	I MAD	II MAPE	III Median	IV MSE 2nd power	V Cubic power
1 Least squares estimates	9.1^{B}	12.4^{W}	${}_{B}11.5$	9.7	10.4
2 Backcasting	8.8^{B}	12.5	17.1^{W}	9.2	10.3
3 Training set	8.9^{B}	10.2	18.0^{W}	10.1	10.7
4 Convenient values	9.6^{B}	${}_{W}12.7$	${}_{W}20.0^{W}$	10.3	10.5
5 $s = 0$, or $s = 0$ and $t = 0$	${}_{W}18.1$	11.6^{B}	16.9	${}_{W}18.9$	${}_{W}19^{W}$
6 $s = 0$ t = least squares	${}_{B}7.3^{B}$	9.3	15.4^{W}	${}_{B}8.2$	${}_{B}8.8$
7 s = least squares $t = 0$	10.4	${}_{B}9.0$	17.7^{W}	8.9^{B}	9.6

[a] 'B' at the upper, right hand side of each box signifies best while 'W' signifies worst accuracy. 'B' at the lower, left hand side of each box signifies best, while 'W' signifies worst accuracy.

Table 6(c)
Dampened smoothing, average MAPE for all forecasting horizons and time. [a] Series.

Initial values	Fildes data				
	I MAD	II MAPE	III Median	IV MSE 2nd power	V Cubic power
1 Least squares estimates	12.6	13.4	13.6^{W}	$_{B}10.4$	$_{B}9.4^{B}$
2 Backcasting	$_{B}12.5^{B}$	$_{B}13.2$	13.8^{W}	12.6	13.5
3 Training set	14.3^{W}	13.9	$_{W}14.3^{W}$	13.3	10.8^{B}
4 Convenient values	13.4^{W}	$_{B}13.2$	$_{B}13.1$	12.3^{B}	13.4^{W}
5 $s = 0$, or $s = 0$ and $t = 0$	$_{W}18.0$	$_{W}17.4$	13.9^{B}	$_{W}18.1^{W}$	18.0
6 $s = 0$ $t =$ least squares	$_{W}18.0$	$_{W}17.4$	13.9^{B}	$_{W}18.1$	$_{W}18.2$
7 $s =$ least squares $t = 0$	15.0^{W}	14.3	$_{B}13.1$	12.6	11.5^{B}

[a] 'B' at the upper, right hand side of each box signifies best while 'W' signifies worst accuracy. 'B' at the lower, left hand side of each box signifies best, while 'W' signifies worst accuracy.

sured in the same fashion; however, none were found in our study. Finally, additional work is needed to determine whether or not our findings also apply to single series when an expert forecaster attempts to minimize post-sample errors.

4.2. Validation

An interesting question in all types of empirical work is whether or not the results found can be generalized and can also hold with other types of data. In order to validate the generality of the findings it was therefore decided, after the present results were found, to test the various possibilities we experimented in this study with the data of Fildes (1989) Such data are *not* at all similar to those of the M-competiton. They consist of 261 monthly series all coming from a Single source (AT&T). Moreover, all series exhibit a strong negative trend and include little or no seasonality.

Table 6 presents the best and worst alternatives (except for Single Smoothing where the 'Best' and 'Worst' alternatives are practically the same) for the Fildes data. The similarities between the results shown in Table 6 and the corresponding ones in Table 5 which uses the M-competition data is considerable as far as Single and Holt's smoothing are concerned. That is the magnitude of the difference between the various experimental cases is very similar while the best and the worst alternatives are practically the same. With Dampened smoothing the best initialization procedure, for the Fildes data, most of the time, was that of the least squares (this was not so with the M-competition data) while there was no loss function which provided in a consistent way the best or the worst results as it was also the case with the M-competition data.

5. Conclusion

This study has shown few differences in post sample forecasting accuracies when different initialization values and optimization (loss) functions have been used. In addition non-symmetric loss functions did not change in any significant fashion the post-sample results. Apart from the conclusion that the median produced inferior results, no other pervasive finding holds across the experimental possibilities tested. Finally, concerning the differences observed the biggest ones were for longer than six forecasting horizons and were mostly concentrated to Holt's exponential smoothing. All differences between the various experimental cases in Single smoothing were extremely small while the magnitude of those very few of

Dampened which were larger were a small fraction of those of Holt's.

The practical implications of this study suggest dropping existing concerns about initial values and loss functions at least when the various methods are run on a push button basis and instead concentrating on more important issues affecting post-sample forecasting accuracy such as optimizing for more than one-step-ahead forecasting horizons and using actual post-sample measures to base the model selection process.

To allow replication and/or extensions of the present study both the M-competition and the Fildes data can be obtained at no cost by writing to Spyros Makridakis at INSEAD.

Appendix A: Exponential methods used

Single

$$e_t = X_t - \hat{X}_{t-1}(1),$$

where X_t is the actual data at period t and $\hat{X}_{t-1}(1)$ is the one-step-ahead forecast at period $t-1$ for period t.

$$S_t = S_{t-1} + \alpha e_t,$$

where α is the smoothing constant whose value is $0 \leqslant \alpha \leqslant 1$ and $\hat{X}_t(m) = S_t$, where the maximum m is six for yearly data, eight for quarterly and eighteen for monthly.

Holt's smoothing

$$e_t = X_t - \hat{X}_{t-1}(1),$$
$$S_t = S_{t-1} + T_{t-1} + \alpha e_t,$$
$$T_t = T_{t-1} + \beta e_t,$$

where β is a smoothing constant whose value is $0 \leqslant \beta \leqslant 1$ and $\hat{X}_t(m) = S_t + mT_t$.

Dampened trend

$$e_t = X_t - \hat{X}_{t-1}(1),$$
$$S_t = S_{t-1} + \phi T_{t-1} + \alpha e_t,$$
$$T_t = \phi T_{t-1} + \beta e_t,$$
$$\hat{X}_t(m) = S_t + \sum_{i=1}^{m} \phi^i T_t.$$

Appendix B: Initial values used

Least squares estimates

For single exponential smoothing the initial value S_1 was found as

$$S_1 = \frac{1}{n} \sum_{t=1}^{n} X_t$$

where n is the number of historical data available.

For Holt's and dampened exponential smoothing S_1 and T_1 are found as

$$T_1 = \frac{n \sum_{t=1}^{n} tX_t - \sum_{t=1}^{n} t \sum_{t=1}^{n} X_t}{n \sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t \right)^2}$$

and

$$S_1 = \frac{1}{n} \sum_{t=1}^{n} X_t - T_1 \frac{1}{n} \sum_{t=1}^{n} t.$$

This initialization approach is referred to as the 'prevalent' one as it is the most widely used in forecasting applications [Brown (1959); Johnson and Montgomery (1974)].

2. Backcastings

The data is inverted and the most recent data value becomes period 1 while the least recent (i.e., period) becomes the last one (i.e., period n). Consequently the values of S_1, or S_1 and T_1 are found as above and the appropriate equation(s) is(are) used to forecast. The last values of S_n, or S_n and T_n are used for initial estimates in the regular forecastings except that the sign of the value of T_n is reversed. Thus, in single smoothing

$$S_1 = S_n,$$

while in Holt's and Dampened,

$$S_1 = S_n, \qquad T_1 = T_n.$$

3. Training set

The data is separated into two sets (the first set makes up one third of the historical data while the second makes up the remaining two thirds). The initial values for S_1 or S_1 and T_1 for the training set are found as in 1 above.

If f is the last period of the training (first) set then the values of the S_1 or S_1 and T_1 for the remaining data are found as:

$$S_1 = S_f,$$

or

$$S_1 = S_f, \qquad T_1 = T_f.$$

4. Convenient values

The value of S_1 or S_1 and T_1 are simply set as follows:

$$S_1 = X_1,$$

or

$$S_1 = X_1, \qquad T_1 = (X_2 - X_1 + X_4 - X_3)/2.$$

5. Zero values

The initial values are set as follows:

$$S_1 = 0,$$

or

$$S_1 = 0, \qquad T_1 = 0.$$

6. Zero Value (for Holt's and Dampened smoothing only)

$S_1 = 0$

$T_1 =$ Least square estimate (see 1 above).

7. Zero value (for Holt's and Dampened smoothing only)

$S_1 =$ Least square estimate (see 1 above)

$T_1 = 0.$

Appendix C: Symmetric loss functions

The one-period-ahead forecasting errors e_t were computed as:

$$e_t = X_t - \hat{X}_{t-1}(1).$$

Consequently the smoothing parameters α, α and β, or α, β and ϕ were chosen in such a way as to minimize the corresponding model fitting loss function,

(i) *Mean absolute deviation* (MAD):

$$\frac{1}{n}\sum_{t=1}^{n} |e_t|.$$

(ii) *Mean absolute percentage error* (MAPE):

$$\frac{1}{n}\sum_{t=1}^{n} \frac{|e_t|}{X_t}.$$

(iii) *Median absolute percentage error (Median)*: The middle value (median) when all absolute percentage errors were arranged from the smallest to the largest.

(iv) *Mean square error* (MSE):

$$\frac{1}{n}\sum_{t=1}^{n} e_t^2.$$

(v) *3th power*:

$$\frac{1}{n}\sum_{t=1}^{n} |e_t|^3.$$

Appendix D: Non-symmetric loss functions

The prevalent initialization procedure (least square estimates see 1 above) and the prevalent optimization function (MSE, (I) above) were used with the following non-symmetric loss functions:

$$\frac{1}{n}\sum \psi e_t^2,$$

where

$$\psi = 1 \quad \text{when } e_t > 0,$$
$$\;\;= c \quad \text{when } e_t < 0,$$

where c took the values of 0.35, 0.50, 0.65, 0.80 and 0.95.

References

Armstrong, J.C. and E.J. Lusk, 1983, "Commentary on the Makridakis time series competition (M-competition)", *Journal of Forecasting*, 2, 259–311.

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, NJ).

Brown, R.G., 1967, *Decision Rules for Inventory Management* (Holt, Rinehart and Winston, New York) Part 2.

Chatfield, C., 1978, "The Holt–Winters forecasting procedure", *Applied Statistics*, 27, 264–279.

Chatfield, C., 1988, "Apples, oranges and mean square error", *International Journal of Forecasting*, 4, 515–518.

Chatfield, C. and M. Yar, 1988, "Holt–Winters forecasting: Some practical issues", *The Statistician*, 37, 129–140.

Cogger, K.O., 1973, "Extensions of the fundamental theorem of exponential smoothing", *Management Science*, 19, 547–554.

Cogger, K.O., 1979, "Time series analysis and forecasting with an absolute error criterion", *TIMS Studies in Management Science*, 12, 189–201.

Fildes, R., 1989, "Evaluation of aggregate and individual forecast method selection selection rules", *Management Science*, 35, 1056–1965.

Fildes, R. and S. Makridakis, 1988, "Forecasting and loss functions", *International Journal of Forecasting*, 545–550.

Gardner, E.S., 1985, "Exponential smoothing: The state of the art", *Journal of Forecasting*, 4, 1–28.

Gardner, E.S., 1990a, "Evaluating forecasting performance in inventory control systems", *Management Science*, 36, 490–499.

Gardner, E.S., 1990b, Private correspondence.

Gardner, E.S. and E. McKenzie, 1985, "Forecasting trends in time series", *Management Science*, 31, 1237–1246.

Granger, C.W.J., 1969, "Prediction with a generalized cost of error function", *Operational Research Quarterly*, 20, 150–161.

Groff, G.K., 1973, "Empirical comparison of models for short-range forecasting", *Management Science*, 20, 22–31.

Harrison, P.J., 1967, "Exponential smoothing and short-term sales forecasting", *Management Science*, 13, 821–842.

Holt, C.C. et al., 1960, *Planning Production, Inventories, and Work Force* (Prentice-Hall, Englewood Cliffs, NJ) ch. 14.

Johnson, L.A. and D.C. Montgomery, 1974, *Operations Research In Production Planning, scheduling, and inventory Control* (Wiley, New York).

Koehler, A.B. and E.S. Murphree, 1988, "A comparison of results from state space forecasting with forecasts from the Makridakis competition", *International Journal of Forecasting*, 4, 45–58.

Ledolter, J. and B. Abraham, 1984, "Some comments on the initialization of exponential smoothing", *Journal of Forecasting*, 3, 79–84.

Makridakis, S., 1990, "A new approach to times series forecasting", *Management Science*, 36, 505–512.

Makridakis, S. and S.C. Wheelwright, 1989, *Forecasting Methods for Management*, 5th ed. (Wiley, New York).

Makridakis, S. and S.C. Wheelwrith, 1978, *Interactive Forecasting: Univariate and Multivariate Methods*, 2nd ed. (Holden-Day, San Francisco).

Makridakis, S. and M. Hibon, 1979, "Accuracy of forecasting: An empirical investigation (with discussion)", *Journal of the Royal Statistical Society (A)*, 142, 97–145.

Makridakis, S., A. Anderson, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, R. Parzen and R. Winkler, 1983, "The accuracy of extrapolation (time series) methods: results of a forecasting competition", *Journal of Forecasting*, 1, 111–153.

Makridakis, S., S.C. Wheelwright and V. McGee, 1983, *Forecasting: Methods and Applications*, 2nd ed. (Wiley, New York).

Martin, C.A. and S.F. Witt, 1989, "Forecasting tourist demand: A comparison of the accuracy of several quantitative methods", *International Journal of Forecasting*, 5, 7–20.

McClain, J.O., 1981, "Restarting a forecasting system when demand suddenly changes", *Journal of Operations Management*, 2, 53–61.

Montgomery, C.D. and L.A. Johnson, 1976, *Forecasting and Time Series Analysis* (McGraw-Hill, New York).

Taylor, S.G., 1981, "Initialization of exponential smoothing forecasts", *AIIE Transactions*, 13, 199–205.

Wade, R.C., 1967, "A technique for initializing exponential smoothing forecasts", *Management Science*, 13, 601–602.

Winters, P.R., 1960, "Forecasting sales by exponentially weighted moving averages", *Management Science*, 6, 324–342.

Zellner, A., 1986, "A tale of forecasting 1001 series: The Bayesian knight strikes again", *International Journal of Forecasting*, 2, 491–494.