A state space framework for automatic forecasting using exponential smoothing methods

Rob J. Hyndman
Department of Econometrics and Business Statistics
Monash University, VIC 3800, Australia.
Telephone: +613 9905 2358
Fax: +613 9905 5474
Email: Rob.Hyndman@buseco.monash.edu.au

Anne B. Koehler
Department of Decision Sciences and Management Information Systems,
Miami University, Oxford, OH 45056, USA.
Telephone: (513) 529 4826
Fax: (513) 529 9689
Email: koehleab@muohio.edu

Ralph D. Snyder
Department of Econometrics and Business Statistics
Monash University, VIC 3800, Australia.
Telephone: +613 9905 2366
Fax: +613 9905 5474
Email: Ralph.Snyder@buseco.monash.edu.au

Simone Grose
Department of Econometrics and Business Statistics
Monash University, VIC 3800, Australia.
Telephone: +613 9905 2964
Fax: +613 9905 5474
Email: Simone.Grose@buseco.monash.edu.au

A state space framework for automatic forecasting using exponential smoothing methods

Abstract:

We provide a new approach to automatic forecasting based on an extended range of exponential smoothing methods. Each method in our taxonomy of exponential smoothing methods provides forecasts that are equivalent to forecasts from a state space model. This equivalence allows (1) easy calculation of the likelihood, the AIC and other model selection criteria; (2) computation of prediction intervals for each method; and (3) random simulation from the underlying state space model. We demonstrate the methods by applying them to the data from the M-competition and the M3-competition. The method provides forecast accuracy comparable to the best methods in the competitions; it is particularly good for short forecast horizons with seasonal data.

Keywords:

automatic forecasting, exponential smoothing, prediction intervals, state space models.

1 Introduction

In business, there is a frequent need for fully automatic forecasting that takes into account trend, seasonality and other features of the data without need for human intervention. In supply chain management, for example, forecasts of demand are required on a regular basis for very large numbers of time series, so that inventory levels can be planned to provide an acceptable level of service to customers. Current methodology employs either highly complicated and often poorly understood techniques such as automatic Box-Jenkins procedures (e.g., Libert, 1984), or exponential smoothing methods (Brown, 1959) that do not adequately capture the range of data, and for which there are often no prediction intervals provided.

Although the exponential smoothing methods have been around since the 1950s, there has not been a well-developed modelling framework incorporating stochastic models, likelihood calculation, prediction intervals and procedures for model selection. In this paper, we fill that gap.

Some important steps toward this framework were established by Gardner (1985), and Ord, Koehler & Snyder (1997). Earlier work in establishing prediction intervals for exponential smoothing methods appeared in Chatfield and Yar (1991), Ord, Koehler and Snyder (1997) and Koehler, Snyder and Ord (1999).

The work of Brown (1959) and Gardner (1985) led to the use of exponential smoothing in automatic forecasting (e.g., Stellwagen and Goodrich, 1999). However, we develop a more general class of methods with a uniform approach to calculation of prediction intervals, maximum likelihood estimation and the exact calculation of model selection criteria such as Akaike's Information Criterion.

Makridakis, Wheelwright and Hyndman (1998) advocate the methods in the taxonomy proposed by Pegels (1969) and extended by Gardner (1985). We shall adopt the same taxonomy (with some modifications) as a framework for selecting amongst exponential smoothing methods. Each method has a trend component and a seasonal component as given in the following table.

Trend Component	Seasonal Component N (none)	 A (additive)	 M (multiplicative)
N (none)	NN	NA	NM
A (additive)	AN	AA	AM
M (multiplicative)	MN	MA	MM
D (damped)	DN	DA	DM

Cell NN describes the simple exponential smoothing method, cell AN describes Holt's linear method. The additive Holt-Winters' method is given by cell AA and the multiplicative Holt-Winters' method is given by cell AM. The other cells correspond to less commonly used but analogous methods.

Following the general approach of Ord, Koehler and Snyder (1997), hereafter referred to as OKS, we can derive an equivalent state space formulation with a single source of error for each of the 12 methods in the framework. This enables easy calculation of the likelihood, and provides facilities to compute prediction intervals for each model. A single source of error model is preferable to the more usual multiple source of error state space model because it allows the state space formulation of non-linear as well as linear cases, and allows the state equations to be expressed in a form that coincides with the error-correction form of the usual smoothing equations. A state space formulation for methods NN, AN, AA and AM has previously been derived (Snyder, 1985; OKS) and we now provide an analogous formulation for the other methods in our framework as well.

We show in Section 3 that, for each of the 12 methods in the above table, there are two possible state space models, one corresponding to the additive error assumption and one the other to the multiplicative error assumption. These two models give equivalent point forecasts although different prediction intervals and different likelihoods. One of the interesting results from our framework and methodology is that we can distinguish multiplicative seasonality (or trend) from a multiplicative error term.

Additive errors are commonly used in statistical analysis and software. A fully multiplicative model (i.e., NN, NM, MN or MM with multiplicative errors) has been traditionally modelled by first applying the log transformation to the data and then fitting a fully additive model. Our alternative strategy of employing a multiplicative error term allows for mixed additive and multiplicative models.

We propose an automatic forecasting procedure that tries each of the 24 state space models on a given time series and selects the "best" method using the AIC.

In Section 2 we describe a general approach to the point forecast equations for each of the methods, and in Section 3 we give the state space equations for both the additive error and multiplicative error versions of each method. We discuss estimation and model selection in Section 4 and use the results to formulate the automatic forecasting algorithm outlined in Section 4.2. We experiment with several variations on the algorithm by applying it to the 1001 series from the M-Competition (Makridakis, et al., 1982). The results of these experiments are summarized in Section 5 from which we select the best variation of the algorithm. In Section 6, we describe the results of applying our algorithm to the 3003 series from the M3-competition (Makridakis and Hibon, 2000), and, in Section 7, we present a Monte Carlo case study of the automatic forecasting algorithm.

2 Point forecast equations

Following Makridakis, Wheelwright and Hyndman (1998), we can write each of the 12 exponential smoothing methods as follows.

$$\ell_t = \alpha P_t + (1-\alpha) Q_t \tag{1}$$

$$b_t = \beta R_t + (\phi - \beta) b_{t-1} \tag{2}$$

$$s_t = \gamma T_t + (1-\gamma) s_{t-m} \tag{3}$$

where ℓ_t denotes the series level at time t, b_t denotes the slope at time t, s_t denotes the seasonal component of the series at time t and m denotes the number of seasons in a year; the values of P_t, Q_t, R_t, and T_t vary according to which of the cells the method belongs, and α, β, γ and ϕ are constants.

Table 1 shows the values of P, Q, R, and T and the formulae for computing point forecasts h periods ahead.

Trend component	**Seasonal component** **N** (none)	 **A** (additive)	 **M** (multiplicative)
N (none)	$P_t = Y_t$ $Q_t = \ell_{t-1}$ $\phi = 1$ $F_{t+h} = \ell_t$	$P_t = Y_t - s_{t-m}$ $Q_t = \ell_{t-1}$ $T_t = Y_t - Q_t$ $\phi = 1$ $F_{t+h} = \ell_t + s_{t+h-m}$	$P_t = Y_t / s_{t-m}$ $Q_t = \ell_{t-1}$ $T_t = Y_t / Q_t$ $\phi = 1$ $F_{t+h} = \ell_t s_{t+h-m}$
A (additive)	$P_t = Y_t$ $Q_t = \ell_{t-1} + b_{t-1}$ $R_t = \ell_t - \ell_{t-1}$ $\phi = 1$ $F_{t+h} = \ell_t + hb_t$	$P_t = Y_t - s_{t-m}$ $Q_t = \ell_{t-1} + b_{t-1}$ $R_t = \ell_t - \ell_{t-1}$ $T_t = Y_t - Q_t$ $\phi = 1$ $F_{t+h} = \ell_t + hb_t + s_{t+h-m}$	$P_t = Y_t / s_{t-m}$ $Q_t = \ell_{t-1} + b_{t-1}$ $R_t = \ell_t - \ell_{t-1}$ $T_t = Y_t / Q_t$ $\phi = 1$ $F_{t+h} = (\ell_t + hb_t) s_{t+h-m}$
M (multiplicative)	$P_t = Y_t$ $Q_t = \ell_{t-1} b_{t-1}$ $R_t = \ell_t / \ell_{t-1}$ $\phi = 1$ $F_{t+h} = \ell_t b_t^h$	$P_t = Y_t - s_{t-m}$ $Q_t = \ell_{t-1} b_{t-1}$ $R_t = \ell_t / \ell_{t-1}$ $T_t = Y_t - Q_t$ $\phi = 1$ $F_{t+h} = \ell_t b_t^h + s_{t+h-m}$	$P_t = Y_t / s_{t-m}$ $Q_t = \ell_{t-1} b_{t-1}$ $R_t = \ell_t / \ell_{t-1}$ $T_t = Y_t / Q_t$ $\phi = 1$ $F_{t+h} = \ell_t b_t^h s_{t+h-m}$
D (damped)	$P_t = Y_t$ $Q_t = \ell_{t-1} + b_{t-1}$ $R_t = \ell_t - \ell_{t-1}$ $\beta < \phi < 1$ $F_{t+h} = \ell_t +$ $(1 + \phi + \cdots + \phi^{h-1}) b_t$	$P_t = Y_t - s_{t-m}$ $Q_t = \ell_{t-1} + b_{t-1}$ $R_t = \ell_t - \ell_{t-1}$ $T_t = Y_t - Q_t$ $\beta < \phi < 1$ $F_{t+h} = \ell_t +$ $(1 + \phi + \cdots + \phi^{h-1}) b_t + s_{t+h-m}$	$P_t = Y_t / s_{t-m}$ $Q_t = \ell_{t-1} + b_{t-1}$ $R_t = \ell_t - \ell_{t-1}$ $T_t = Y_t / Q_t$ $\beta < \phi < 1$ $F_{t+h} = [\ell_t +$ $(1 + \phi + \cdots + \phi^{h-1}) b_t] s_{t+h-m}$

Table 1: *Formulae for recursive calculations and point forecasts.*

These equations differ slightly from the usual equations given, for example, in Makridakis, Wheelwright and Hyndman (1998, page 171). First, we consider the damped trend methods. Second, we use Q_t in place of ℓ_t in the equations for T_t. The effect of using the equations in the form given in Table 1 is that, when we update the seasonal component, we use the level ℓ_{t-1} and growth rate b_{t-1} from the previous time period rather than the newly revised level ℓ_t from the current time period. This alternative form of the equations is designed to allow the models to be written in state space form (see Section 3). The equations we use for AM are not the usual multiplicative Holt-Winters equations, but are equivalent to those used by OKS. It should be noted that this change makes no difference for the models with additive seasonality, but it does change the forecasts slightly for models with multiplicative seasonality.

The formulas for damped trend are appropriate when there is trend in the time series, but one believes that continuing to use the final estimate for the growth rate at the end of the historical data would lead to unrealistic forecasts. Thus, the equations for damped trend do what the name indicates: dampen the trend as the length of the forecast horizon increases. In Table 1, one can see that the forecast for h-periods-ahead is $F_{t+h} = \ell_t + (1 + \phi + \cdots + \phi^{h-1})b_t$. The trend is dampened by a factor of ϕ for each additional future time period. Our formulas for damped trend differ from those of Gardner (1985) by a factor of ϕ. Gardner begins the dampening immediately for the forecast one-period-ahead and his forecast function is $F_{t+h} = \ell_t + (\phi + \phi^2 + \cdots + \phi^h)b_t$.

Writing (1)–(3) in their error-correction form we obtain

$$\begin{aligned}
\ell_t &= Q_t + \alpha(P_t - Q_t) && (4)\\
b_t &= \phi b_{t-1} + \beta(R_t - b_{t-1}) && (5)\\
s_t &= s_{t-m} + \gamma(T_t - s_{t-m}). && (6)
\end{aligned}$$

The method with fixed level (constant over time) is obtained by setting $\alpha = 0$, the method with fixed trend (drift) is obtained by setting $\beta = 0$, and the method with fixed seasonal pattern is obtained by setting $\gamma = 0$. Note also that the additive trend methods are obtained by letting $\phi = 1$ in the damped trend methods.

3 State space models

Ord, Koehler and Snyder (1997) discuss special cases of the "single source of error" state space models that underlie some of the exponential smoothing methods. We expand their work to cover all the methods in the classification outlined in Section 1. For each method, we obtain two models—a model with additive errors and a model with multiplicative errors. The pointwise forecasts for the two models are identical, but prediction intervals will differ.

The general OKS framework involves a state vector $\boldsymbol{x}_t$ and state space equations of the form

$$\begin{aligned}
Y_t &= h(\boldsymbol{x}_{t-1}) + k(\boldsymbol{x}_{t-1})\varepsilon_t && (7)\\
\boldsymbol{x}_t &= f(\boldsymbol{x}_{t-1}) + g(\boldsymbol{x}_{t-1})\varepsilon_t && (8)
\end{aligned}$$

where $\{\varepsilon_t\}$ is a Gaussian white noise process with mean zero and variance σ^2. We define $\boldsymbol{x}_t = (\ell_t, b_t, s_t, s_{t-1}, \dots, s_{t-(m-1)})$, $e_t = k(\boldsymbol{x}_{t-1})\varepsilon_t$ and $\mu_t = h(\boldsymbol{x}_{t-1})$. Then $Y_t = \mu_t + e_t$.

Trend component	**Seasonal component**		
	N (none)	**A** (additive)	**M** (multiplicative)
N (none)	$\mu_t = \ell_{t-1}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$\mu_t = \ell_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = \ell_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A (additive)	$\mu_t = \ell_{t-1} + b_{t-1}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t$	$\mu_t = \ell_{t-1} + b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = (\ell_{t-1} + b_{t-1})s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
M (multiplicative)	$\mu_t = \ell_{t-1}b_{t-1}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t/\ell_{t-1}$	$\mu_t = \ell_{t-1}b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = \ell_{t-1}b_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1})$
D (damped)	$\mu_t = \ell_{t-1} + b_{t-1}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \alpha\beta\varepsilon_t$	$\mu_t = \ell_{t-1} + b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \alpha\beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = (\ell_{t-1} + b_{t-1})s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \alpha\beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$

Table 2: *State space equations for each additive error model in the classification. Multiplicative error models are obtained by replacing ε_t by $\mu_t\varepsilon_t$ in the above equations.*

The model with additive errors is written as $Y_t = \mu_t + \varepsilon_t$ where $\mu_t = F_{(t-1)+1}$ denotes the one-step forecast made at time $t-1$. So, in this case, $k(\boldsymbol{x}_{t-1}) = 1$. The model with multiplicative errors is written as $Y_t = \mu_t(1+\varepsilon_t)$. Thus, $k(\boldsymbol{x}_{t-1}) = \mu_t$ for this model and $\varepsilon_t = e_t/\mu_t = (Y_t - \mu_t)/\mu_t$ and hence ε_t is a relative error for the multiplicative model.

All the methods in Table 1 can be written in the form (7) and (8). The underlying equations are given in Table 2. The models are not unique. Clearly, any value of $k(\boldsymbol{x}_{t-1})$ will lead to identical point forecasts for Y_t. For example, Koehler, Snyder and Ord (1999) and Archibald (1994) give several models for AM by altering the value of $k(\boldsymbol{x}_{t-1})$.

The only difference between the additive error and multiplicative error models is in the observation equation (7). The state equation (8) can be put in exactly the same form by substituting $\varepsilon_t = e_t/k(\boldsymbol{x}_{t-1})$ into each state equation. For example, consider cell NN. For the additive error model

$$\varepsilon_t = e_t \qquad \text{and} \qquad \ell_t = \ell_{t-1} + \alpha e_t.$$

For the multiplicative error model

$$\varepsilon_t = e_t/k(\boldsymbol{x}_{t-1}) = e_t/\ell_{t-1} \qquad \text{and} \qquad \ell_t = \ell_{t-1}(1+\alpha\varepsilon_t) = \ell_{t-1} + \alpha e_t.$$

Thus the state equations are identical in form.

Note that not all of the 24 state space models are appropriate for all data. The multiplicative error models are not well defined if there are zeros or negative values in the data. Similarly, we don't consider the additive error models with multiplicative trend or multiplicative seasonality if any observations are zero. Further, if the data are not quarterly or monthly (and do not have some other obvious seasonal period), then we do not consider any of the seasonal methods.

4 Estimation and model selection

Let

$$L^*(\boldsymbol{\theta}, \boldsymbol{X}_0) = n \log\left(\sum_{t=1}^{n} e_t^2/k^2(x_{t-1})\right) + 2\sum_{t=1}^{n} \log|k(x_{t-1})|. \tag{9}$$

Then L^* is equal to twice the negative logarithm of the conditional likelihood function in OKS with constant terms eliminated.

The parameters $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \phi)$ and initial states $\boldsymbol{X}_0 = (\ell_0, b_0, s_0, s_{-1}, \dots, s_{-m+1})$ can be estimated by minimizing L^*. Alternatively, estimates can be obtained by minimizing the one-step MSE, minimizing the one-step MAPE, minimizing the residual variance σ^2 or via some other criterion for measuring forecast error. We shall experiment with each of these estimation approaches in Section 5.

We constrain the estimation by restricting the parameters to lie within the following intervals

$$0.1 \le \alpha \le 0.9, \quad 0.1 \le \beta \le 0.9, \quad 0.1 \le \gamma \le 0.9, \quad \beta \le \phi \le 1.$$

Usually, α, β and γ are restricted to values in $(0, 1)$. However we use a smaller range to avoid instabilities occurring. We also constrain the initial states $\boldsymbol{X}_0$ so that the seasonal indices add to zero for additive seasonality, and add to m for multiplicative seasonality.

Models are selected using Akaike's Information Criterion:

$$\text{AIC} = L^*(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{X}}_0) + 2p$$

where p is the number of parameters in $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{X}}_0$ denote the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{X}_0$. We select the model that minimizes the AIC amongst all of the 24 models that are appropriate for the data. Using the AIC for model selection is preferable to other measurements of forecast error such as the MSE or MAPE as it penalizes against models containing too many parameters.

The AIC also provides a method for selecting between the additive and multiplicative error models. Point forecasts from the two models are identical so that standard forecast accuracy measures such as MSE or MAPE are unable to select between the error types. The AIC is able to select between the error types because it is based on likelihood rather than one-step forecasts.

Obviously, other model selection criteria (such as the BIC) could also be used in a similar manner.

4.1 Initialization

The non-linear optimization requires some initial values. We use $\alpha = \beta = \gamma = 0.5$ and $\phi = 0.9$. The initial values of ℓ_0, b_0 and s_k ($k = -m+1, \dots, 0$) are obtained using the following heuristic scheme.

- For seasonal data, compute a $2 \times m$ moving average through the first few years of data (we use up to four years if the data are available). Denote this by $\{f_t\}$, $t = m/2+1, m/2+2, \dots$.
- For additive seasonality, we detrend the data to obtain $Y_t - f_t$. For multiplicative seasonality, we detrend the data to obtain Y_t/f_t. Then compute initial seasonal indices, $s_{-m+1}, \dots, s_0$, by averaging the detrended data for each season over the first 3 years available (from $t = m/2+1$ to $t = 7m/2$). We normalize these seasonal indices so they add to zero for additive seasonality, and add to m for multiplicative seasonality.
- For seasonal data, compute a linear trend using OLS regression on the first 10 seasonally adjusted values (using the seasonal indices obtained above) against a time variable $t = 1, \dots, 10$.
- For non-seasonal data, compute a linear trend on the first 10 observations against a time variable $t = 1, \dots, 10$.
- Then set ℓ_0 to be the intercept of the trend.

- For additive trend, set b_0 to be the slope of the trend.
- For multiplicative trend, set $b_0 = 1 + b/a$ where a denotes the intercept and b denotes the slope of the fitted trend.

These heuristic values of the initial state $\boldsymbol{X}_0$ are then refined by estimating them as parameters along with the elements of $\boldsymbol{\theta}$.

4.2 Automatic forecasting

We combine the preceding ideas to obtain a robust and widely applicable automatic forecasting algorithm. The steps involved are summarized below.

- For each series, we apply the models that are appropriate, optimizing the parameters of the model in each case.
- We select the best of the models according to the AIC.
- We produce forecasts using the best model (with optimized parameters) for as many steps ahead as required.
- To obtain prediction intervals, we use a bootstrap method by simulating 5000 future sample paths for $\{Y_{n+1}, \ldots, Y_{n+h}\}$ and finding the $\alpha/2$ and $1 - \alpha/2$ percentiles of the simulated data at each forecasting horizon. The sample paths are generated using the normal distribution for errors (parametric bootstrap) or using the resampled errors (ordinary bootstrap).

With the linear state space models, it is possible to compute analytical prediction intervals. However, this is difficult with the non-linear models, and so we prefer to use a simulation approach which is applicable to all the models and is very easy to implement.

5 Application to M-competition data

To test the algorithm, and to experiment with the various estimation approaches possible, we applied the algorithm to the 1001 series of the M-competition data (Makridakis, et al., 1982). We tested the following five estimation methods:

1 MLE: minimizing L^*;

2 MSE: Minimizing MSE;

3 MAPE: Minimizing MAPE;

4 AMSE: Minimizing $(\text{MSE}_1 + \text{MSE}_2 + \text{MSE}_3)/3$ where MSE_k denotes the mean square of the k-step forecast errors, $k = 1, 2, 3$;

5 Sigma: Minimizing the residual variance σ^2.

For each of the 5 methods of estimation, we computed forecasts up to 18 steps ahead (the number of steps as specified in the M-competition). Then we computed the MAPE for each forecast horizon, by averaging the absolute percentage errors across all 1001 series. Note that the number of forecasts at each horizon varies as the number of forecasts for each series was specified in the M-competition and depended on the seasonality of the data (e.g., annual, quarterly, monthly).

Table 3 shows the average MAPE across all forecast horizons. Similar results for the 504 non-seasonal series, 89 quarterly series and 406 monthly series are given in Table 4. We note that the these are out-of-sample forecast accuracy measures. Overall, AMSE estimation seems to perform the best, closely followed by MSE estimation.

We also compared the performance of the methods on how frequently prediction intervals contained the true values of the series. For each combination of methods, we computed the percentage of true values contained in the (nominally) 95% prediction intervals. We did this using both parametric intervals (PPI) based on normally distributed errors and nonparametric intervals (NPPI) based on resampling the fitted errors. The results are reported in Tables 3 and 5. All five estimation methods overestimate the coverage probability of prediction intervals. (This is a well-known forecasting phenomenon—see Makridakis, Wheelwright and Hyndman, 1998, p.470.) Interestingly, the methods resulting in the best MAPE values seem to give the worst coverage probabilities, and vice-versa.

Estimation Method	MAPE	PPI	NPPI
AMSE	17.63	81.9%	73.2%
MSE	17.73	83.4%	74.5%
Sigma	18.49	83.2%	74.2%
MLE	18.55	83.1%	73.5%
MAPE	19.08	85.6%	77.1%

Table 3: *Average MAPE from the five estimation methods using all 1001 series. PPI gives coverage of nominal 95% parametric prediction intervals and NPPI gives coverage of nominal 95% nonparametric prediction intervals.*

Estimation Method	**Non-seasonal**	**Quarterly**	**Monthly**
AMSE	23.06	16.36	13.32
MSE	23.20	17.43	13.51
Sigma	24.12	17.06	14.26
MLE	24.61	16.84	13.99
MAPE	25.64	15.23	14.34

Table 4: *Average MAPE for each seasonal subset of series.*

Estimation Method	**Non-seasonal**	**Quarterly**	**Monthly**
AMSE	79.4%	71.1%	84.7%
MSE	82.0%	72.8%	85.5%
Sigma	81.9%	72.2%	85.2%
MLE	81.9%	72.6%	85.1%
MAPE	84.1%	78.7%	87.5%

Table 5: *Coverage of parametric prediction intervals for each seasonal subset of series.*

Figure 1 shows the MAPE for individual forecast horizons and for different subsets of the series, using the AMSE method of estimation. Note that for quarterly data, only 8 forecasts were required by the M-competition.

For the AMSE method, we now compare our results with those obtained by other methods in the M-competition. Figure 2 shows the MAPE for each forecast horizon for our method and three of the best-performing methods in the M-competition. Clearly, our method is comparable in performance to these methods. Table 6 shows the MAPE across various forecast horizons, and demonstrates that our method performs better than the others shown for shorter forecast horizons, but not so well for longer forecast horizons.

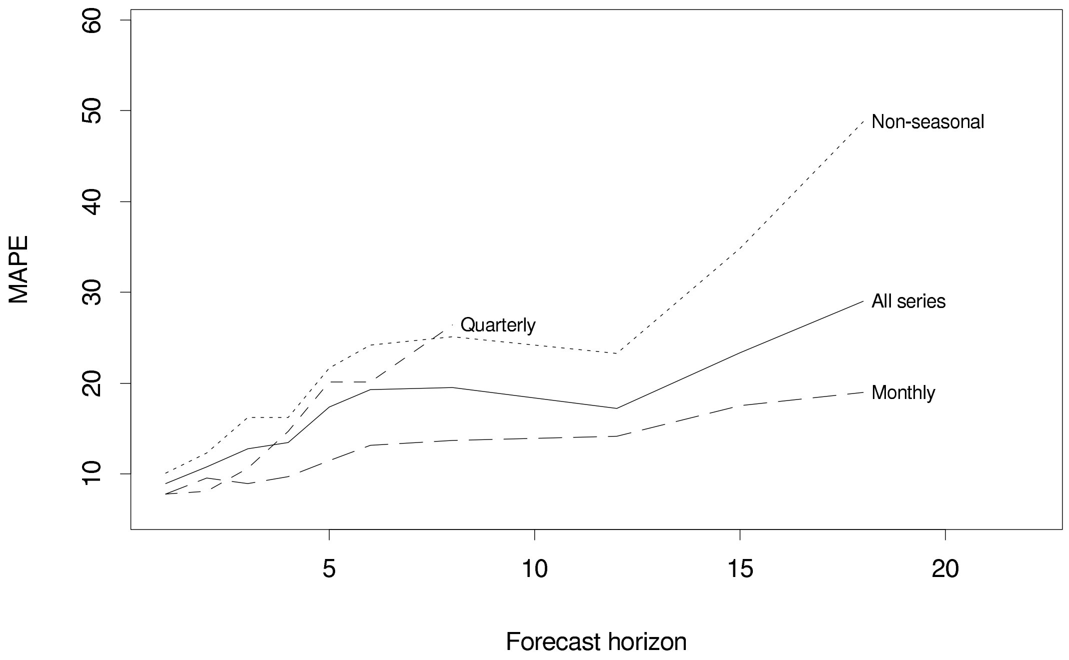

Figure 1: *MAPE across different forecast horizons for all series (1001 series), non-seasonal data (504 series), quarterly data (89 series) and monthly data (406 series).*

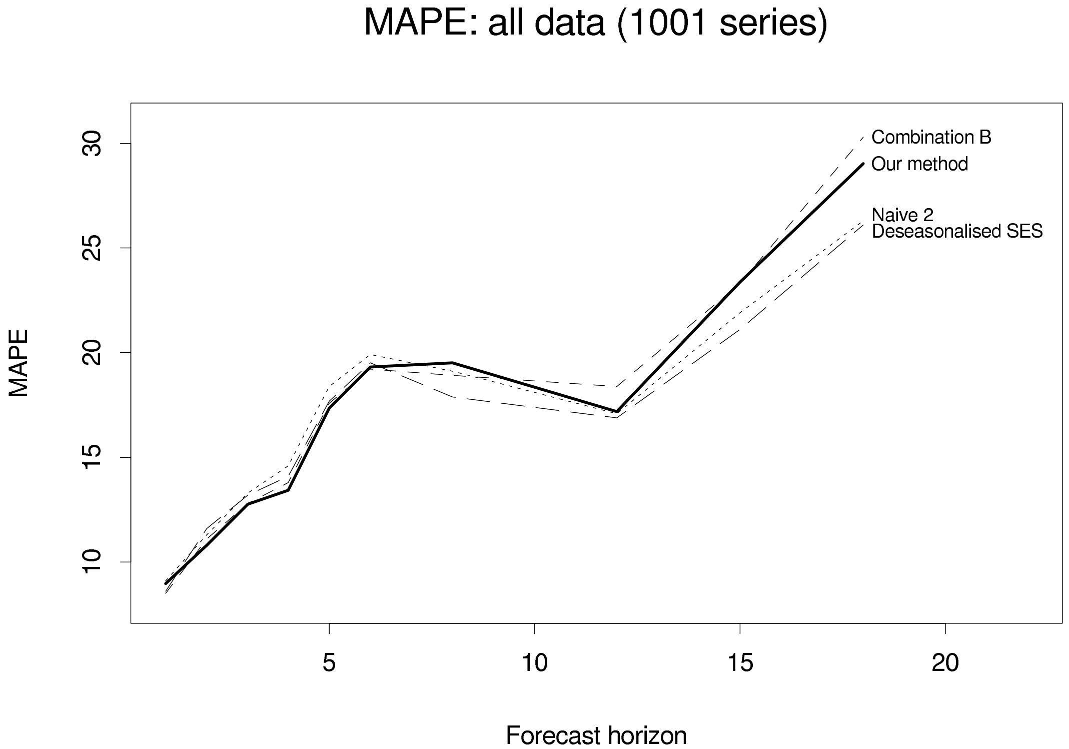

Figure 2: *MAPE across different forecast horizons (1001 series) comparing our method with some of the best methods from the M-competition (Makridakis, et al., 1982).*

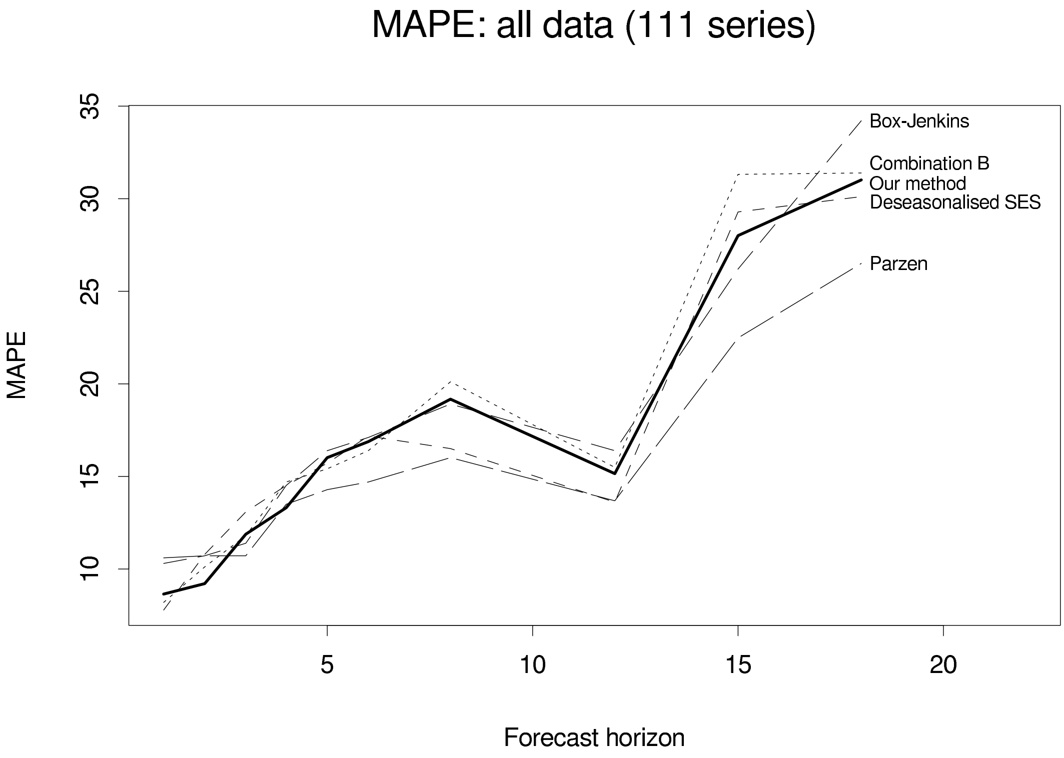

Figure 3: *MAPE across different forecast horizons (111 series) comparing our method with some of the best methods from the M-competition (Makridakis, et al., 1982).*

	Forecasting horizons										Average of forecasting horizons					
Method	1	2	3	4	5	6	8	12	15	18	1–4	1–6	1–8	1–12	1–15	1–18
Naive2	9.1	11.3	13.3	14.6	18.4	19.9	19.1	17.1	21.9	26.3	12.4	14.4	15.2	15.7	16.4	17.4
Deseasonalised SES	8.6	11.6	13.2	14.1	17.7	19.5	17.9	16.9	21.1	26.1	11.9	14.1	14.8	15.3	16.0	16.9
Combination B	8.5	11.1	12.8	13.8	17.6	19.2	18.9	18.4	23.3	30.3	11.6	13.8	14.8	15.6	16.5	17.8
Our method	9.0	10.8	12.8	13.4	17.4	19.3	19.5	17.2	23.4	29.0	11.5	13.8	14.7	15.4	16.4	17.6

Table 6: *Average MAPE across different forecast horizons (1001 series).*

	Forecasting horizons										Average of forecasting horizons					
Method	1	2	3	4	5	6	8	12	15	18	1–4	1–6	1–8	1–12	1–15	1–18
Naive2	8.5	11.4	13.9	15.4	16.6	17.4	17.8	14.5	31.2	30.8	12.3	13.8	14.9	14.9	16.4	17.8
Deseasonalised SES	7.8	10.8	13.1	14.5	15.7	17.2	16.5	13.6	29.3	30.1	11.6	13.2	14.1	14.0	15.3	16.8
Combination B	8.2	10.1	11.8	14.7	15.4	16.4	20.1	15.5	31.3	31.4	11.2	12.8	14.4	14.7	16.2	17.7
Box-Jenkins	10.3	10.7	11.4	14.5	16.4	17.1	18.9	16.4	26.2	34.2	11.7	13.4	14.8	15.1	16.3	18.0
Lewandowski	11.6	12.8	14.5	15.3	16.6	17.6	18.9	17.0	33.0	28.6	13.5	14.7	15.5	15.6	17.2	18.6
Parzen	10.6	10.7	10.7	13.5	14.3	14.7	16.0	13.7	22.5	26.5	11.4	12.4	13.3	13.4	14.3	15.4
Our method	8.7	9.2	11.9	13.3	16.0	16.9	19.2	15.2	28.0	31.0	10.8	12.7	14.3	14.5	15.7	17.3

Table 7: *Average MAPE across different forecast horizons (111 series).*

A smaller set of 111 series was used in the M-competition for comparisons with some more time-consuming methods. Table 7 shows a MAPE comparison between our method and these other methods. Again, this demonstrates that our method performs better than the others shown for shorter forecast horizons, but not so well for longer forecast horizons. Figure 3 shows the MAPE for each forecast horizon for our method and the methods given in Table 7. Note that our method out-performs all other methods when averaged over forecast horizons 1–4.

	Model	Non-seasonal	Quarterly	Biannual	Monthly	Total
Additive errors	NN	70				70
	NA		1		40	41
	NM		10		54	64
	AN	46				46
	AA		3	1	10	14
	AM		5		35	40
	MN	57				57
	MA		3		6	9
	MM		4		13	17
	DN	53				53
	DA		3		16	19
	DM		8		28	36
Multiplicative errors	NN	93		1	1	95
	NA		4		37	41
	NM		7		38	45
	AN	89				89
	AA		10		19	29
	AM		11		39	50
	MN	45			1	46
	MA		4		4	8
	MM		7		9	16
	DN	51			1	52
	DA		2		17	19
	DM		7		38	45
	Total	504	89	2	406	1001

Table 8: *Number of times each model chosen using the AIC.*

Table 8 shows the models selected for each of the 1001 series using AMSE estimation. The commonly used models NN (simple exponential smoothing), and AN (Holt's method), were chosen most frequently, providing some justification for their popularity. Interestingly, the non-trended seasonal models (NA and NM) were selected much more frequently than the popular Holt-Winters' models (AA and AM). Damped trend models were selected a total of 224 times compared to 268 times for additive trend, 153 times for multiplicative trend and 356 times for no trend. Amongst seasonal series, additive seasonality was selected 180 times, multiplicative seasonality 313 times, and no seasonal component 4 times. Of the 1001 series, an additive error model was chosen 466 times and a multiplicative model was chosen 535 times.

For some models, the time taken for estimation of parameters was considerable (of the order of several minutes). This particularly occurred with monthly data (where there are 13 initial states to estimate) and a full trend/seasonal model (giving 4 parameters to estimate). Searching for optimal values in a space of 17 dimensions can be very time-consuming!

Consequently, we propose the following two-stage procedure to speed up the computations:

1 Estimate $\boldsymbol{\theta}$ while holding $\boldsymbol{X}_0$ at the heuristic values obtained in Section 4.1.

2 Then estimate $\boldsymbol{X}_0$ by minimizing AMSE while holding $\hat{\boldsymbol{\theta}}$ fixed.

This procedure speeds the algorithm by reducing the number of dimensions over which to optimize.

The following table gives the average MAPE and computation time for the 1001 series from the M-competition using AMSE estimation.

Initialization Method	MAPE	Time for 1001 series
Heuristic only	18.14	16 min
Two-stage	17.85	22 min
Full optimization	17.63	2 hours, 20 min

The "Heuristic only" method simply uses the initial values obtained in Section 4.1, and the "Full optimization" method optimizes the initial values along with the parameters (as was done in all of the preceding computations). Clearly, a great deal of time can be saved using the two-stage method without sacrificing much forecast accuracy.

6 Application to M3 data

Next, we applied our methodology to the M3-competition data (Makridakis and Hibon, 2000). Based on the results from the M-competition data, we used AMSE estimation and optimal initialization. The results are given in Tables 9–15 along with some of the methods from the M3-competition. (See Makridakis and Hibon, 2000, for details of these methods.) For each forecast horizon, we have also provided a ranking of our method compared to the 24 methods used in the M3-competition. These are based on the symmetric MAPEs averaged across series for each forecast horizon.

As with the M-competition data, our method performs best for short forecast horizons (up to 4–6 steps ahead). It seems to perform especially well on seasonal data, particularly monthly data. On the other hand, it seems to perform rather poorly on annual, non-seasonal time series.

	Forecasting horizons										Average of forecasting horizons					
Method	1	2	3	4	5	6	8	12	15	18	1–4	1–6	1–8	1–12	1–15	1–18
Naive2	10.5	11.3	13.6	15.1	15.1	15.8	14.5	16.0	19.3	20.7	12.62	13.55	13.74	14.22	14.80	15.46
B-J automatic	9.2	10.4	12.2	13.9	14.0	14.6	13.0	14.1	17.8	19.3	11.42	12.39	12.52	12.78	13.33	13.99
ForecastPRO	8.6	9.6	11.4	12.9	13.3	14.2	12.6	13.2	16.4	18.3	10.64	11.67	11.84	12.12	12.58	13.18
THETA	8.4	9.6	11.3	12.5	13.2	13.9	12.0	13.2	16.2	18.2	10.44	11.47	11.61	11.94	12.41	13.00
RBF	9.9	10.5	12.4	13.4	13.2	14.1	12.8	14.1	17.3	17.8	11.56	12.26	12.40	12.76	13.24	13.74
ForcX	8.7	9.8	11.6	13.1	13.2	13.8	12.6	13.9	17.8	18.7	10.82	11.72	11.88	12.21	12.80	13.48
Our method	8.8	9.8	12.0	13.5	13.9	14.7	13.0	14.1	17.6	18.9	11.04	12.13	12.32	12.66	13.14	13.77
Rank	4	3	7	7	12	12	8	6	7	6	4	8	6	6	6	7

Table 9: *Average symmetric MAPE across different forecast horizons: all 3003 series.*

	Forecasting horizons										Average of forecasting horizons					
Method	1	2	3	4	5	6	8	12	15	18	1–4	1–6	1–8	1–12	1–15	1–18
Naive2	8.0	8.1	9.5	9.5	9.9	11.5	12.1	11.0	14.0	15.5	8.77	9.41	10.12	10.54	10.91	11.40
B-J automatic	7.1	7.4	8.0	8.8	9.2	10.3	10.5	10.5	13.3	14.5	7.82	8.46	9.03	9.31	9.79	10.37
ForecastPRO	6.2	6.6	7.5	8.1	8.4	9.7	10.0	9.6	11.5	13.1	7.12	7.76	8.38	8.64	8.98	9.45
THETA	6.5	6.9	7.8	8.0	8.9	10.2	9.9	10.2	12.0	13.6	7.30	8.05	8.64	9.03	9.37	9.84
RBF	8.0	8.0	8.7	8.6	8.7	10.1	10.5	10.6	12.4	13.3	8.30	8.68	9.23	9.59	9.92	10.29
ForcX	6.4	6.8	7.6	8.3	8.6	10.0	10.5	10.0	12.5	13.7	7.26	7.93	8.63	8.93	9.35	9.86
Our method	6.2	6.4	7.7	8.2	8.9	10.2	10.6	10.1	12.0	14.0	7.12	7.93	8.67	9.01	9.35	9.87
Rank	1	1	3	5	6	4	8	3	2	6	1	2	4	3	2	4

Table 10: *Average symmetric MAPE across different forecast horizons: 862 seasonal series.*

	Forecasting horizons										Average of forecasting horizons					
Method	1	2	3	4	5	6	8	12	15	18	1–4	1–6	1–8	1–12	1–15	1–18
Naive2	11.5	12.6	15.3	17.3	17.1	17.5	15.9	19.2	22.8	24.1	14.17	15.22	15.32	15.97	16.73	17.54
B-J automatic	10.0	11.6	13.9	15.9	16.0	16.4	14.4	16.4	20.7	22.4	12.87	13.97	14.04	14.43	15.09	15.85
ForecastPRO	9.6	10.8	13.0	14.9	15.3	15.9	14.1	15.6	19.5	21.7	12.05	13.25	13.34	13.78	14.37	15.09
THETA	9.2	10.6	12.7	14.3	14.9	15.4	13.2	15.1	19.0	21.2	11.71	12.85	12.90	13.32	13.91	14.62
RBF	10.6	11.6	13.9	15.3	15.0	15.6	14.1	16.3	20.4	20.7	12.87	13.69	13.78	14.27	14.88	15.51
ForcX	9.6	11.1	13.2	15.1	15.1	15.4	13.8	16.5	21.2	22.0	12.25	13.24	13.29	13.77	14.51	15.34
Our method	9.9	11.2	13.7	15.6	15.9	16.6	14.4	16.7	21.3	22.2	12.61	13.83	13.91	14.39	15.03	15.77
Rank	8	5	11	8	14	15	11	11	13	7	9	11	10	10	10	10

Table 11: *Average symmetric MAPE across different forecast horizons: 2141 nonseasonal series.*

	Forecasting horizons						Average of forecasting horizons	
Method	1	2	3	4	5	6	1–4	1–6
Naive2	8.5	13.2	17.8	19.9	23.0	24.9	14.85	17.88
B-J automatic	8.6	13.0	17.5	20.0	22.8	24.5	14.78	17.73
ForecastPRO	8.3	12.2	16.8	19.3	22.2	24.1	14.15	17.14
THETA	8.0	12.2	16.7	19.2	21.7	23.6	14.02	16.90
RBF	8.2	12.1	16.4	18.3	20.8	22.7	13.75	16.42
ForcX	8.6	12.4	16.1	18.2	21.0	22.7	13.80	16.48
Our method	9.3	13.6	18.3	20.8	23.4	25.8	15.48	18.53
Rank	19	18	19	19	17	19	19	19

Table 12: *Average symmetric MAPE across different forecast horizons: 645 annual series.*

	Forecasting horizons							Average of forecasting horizons		
Method	1	2	3	4	5	6	8	1–4	1–6	1–8
Naive2	5.4	7.4	8.1	9.2	10.4	12.4	13.7	7.55	8.82	9.95
B-J automatic	5.5	7.4	8.4	9.9	10.9	12.5	14.2	7.79	9.10	10.26
ForecastPRO	4.9	6.8	7.9	9.6	10.5	11.9	13.9	7.28	8.57	9.77
THETA	5.0	6.7	7.4	8.8	9.4	10.9	12.0	7.00	8.04	8.96
RBF	5.7	7.4	8.3	9.3	9.9	11.4	12.6	7.69	8.67	9.57
ForcX	4.8	6.7	7.7	9.2	10.0	11.6	13.6	7.12	8.35	9.54
Our method	5.0	6.6	7.9	9.7	10.9	12.1	14.2	7.32	8.71	9.94
Rank	4	1	7	13	17	10	20	7	9	12

Table 13: *Average symmetric MAPE across different forecast horizons: 756 quarterly series.*

	Forecasting horizons										Average of forecasting horizons					
Method	1	2	3	4	5	6	8	12	15	18	1–4	1–6	1–8	1–12	1–15	1–18
Naive2	15.0	13.5	15.7	17.0	14.9	14.4	15.6	16.0	19.3	20.7	15.30	15.08	15.26	15.55	16.16	16.89
B-J automatic	12.3	11.7	12.8	14.3	12.7	12.3	13.0	14.1	17.8	19.3	12.78	12.70	12.86	13.19	13.95	14.80
ForecastPRO	11.5	10.7	11.7	12.9	11.8	12.0	12.6	13.2	16.4	18.3	11.72	11.78	12.02	12.43	13.07	13.85
THETA	11.2	10.7	11.8	12.4	12.2	12.2	12.7	13.2	16.2	18.2	11.54	11.75	12.09	12.48	13.09	13.83
RBF	13.7	12.3	13.7	14.3	12.3	12.5	13.5	14.1	17.3	17.8	13.49	13.14	13.36	13.64	14.19	14.76
ForcX	11.6	11.2	12.6	14.0	12.4	12.0	12.8	13.9	17.8	18.7	12.32	12.28	12.44	12.81	13.58	14.44
Our method	11.5	10.6	12.3	13.4	12.3	12.3	13.2	14.1	17.6	18.9	11.93	12.05	12.43	12.96	13.64	14.45
Rank	2	1	5	3	4	5	7	6	7	6	3	3	3	4	4	4

Table 14: *Average symmetric MAPE across different forecast horizons: 1428 monthly series.*

	Forecasting horizons							Average of forecasting horizons		
Method	1	2	3	4	5	6	8	1–4	1–6	1–8
Naive2	2.2	3.6	5.4	6.3	7.8	7.6	9.2	4.38	5.49	6.30
B-J automatic	1.8	3.0	4.5	4.9	6.1	6.1	7.5	3.52	4.38	5.06
ForecastPRO	1.9	3.0	4.0	4.4	5.4	5.4	6.7	3.31	4.00	4.60
THETA	1.8	2.7	3.8	4.5	5.6	5.2	6.1	3.20	3.93	4.41
RBF	2.7	3.8	5.2	5.8	6.9	6.3	7.3	4.38	5.12	5.60
ForcX	2.1	3.1	4.1	4.4	5.6	5.4	6.5	3.42	4.10	4.64
Our method	2.0	3.0	4.0	4.4	5.4	5.1	6.3	3.37	3.99	4.51
Rank	14	12	6	2	2	1	4	12	4	4

Table 15: *Average symmetric MAPE across different forecast horizons: 174 other series.*

7 Model selection accuracy

We carried out simulations from the underlying stochastic state space models to see how well one can identify the underlying model using the procedure outlined in Section 4.2. For these simulations, we used non-seasonal models and generated 5000 series for each model. The results are summarized in Table 16.

The parameters used in generating these models are shown in Table 17. These parameters were chosen to generate data that look reasonably realistic.

Clearly, the algorithm has a very high success rate at determining whether the errors should be additive or multiplicative. The main source of error in model selection is mis-selecting the trend component, especially for damped trend models. That is not surprising given the value of ϕ chosen was very close to 1.

	Additive error				Multiplicative error			
	NN	AN	MN	DN	NN	AN	MN	DN
Correct model selections	78.6	77.6	73.1	43.7	87.6	76.5	45.9	23.5
Correct additive/multiplicative selections	88.0	99.8	99.4	99.3	95.7	98.7	99.4	98.0
Correct trend selections	89.2	77.7	73.5	44.1	91.6	76.5	45.9	23.7

Table 16: *Percentage of correct model selections based on 5000 randomly generated series of each type.*

	Additive error				Multiplicative error			
	NN	AN	MN	DN	NN	AN	MN	DN
α	0.50	0.20	0.08	0.20	0.70	0.70	0.70	0.70
β		0.10	0.10	0.10		0.03	0.03	0.03
ϕ				0.98				0.98
σ	0.10	1.00	0.15	1.00	0.11	0.11	0.11	0.11
ℓ_0	1.00	1.00	0.10	1.00	1.00	1.00	0.03	1.00
b_0		0.20	1.05	0.20		0.10	1.04	0.10

Table 17: *Parameters and initial states used in generating random data from each model.*

8 Conclusions

We have introduced a state space framework that subsumes all the exponential smoothing models and which allows the computation of prediction intervals, likelihood and model selection criteria. We have also proposed an automatic forecasting strategy based on the model framework.

Application of the automatic forecasting strategy to the M-competition data and IJF-M3 competition data has demonstrated that our methodology is particularly good at short term forecasts (up to about 6 periods ahead). We note that we have not done any preprocessing of the data, identification of outliers or level shifts, or used any other strategy designed to improve the forecasts. These results are based on a simple application of the algorithm to the data. We expect that our results could be improved further if we used some sophisticated data preprocessing techniques as was done by some of the competitors in the M3 competition (e.g., Reilly, 1999).

For several decades, exponential smoothing has been considered an ad hoc approach to forecasting, with no proper underlying stochastic formulation. That is no longer true. The state space framework we have described brings exponential smoothing into the same class as ARIMA models, being widely applicable and having a sound stochastic model behind the forecasts.

9 References

ARCHIBALD, B.C. (1994) "Winters Model: three versions, diagnostic checks and forecast performances", Working paper WP-94-4, School of Business Administration, Dalhousie University, Halifax, Canada.

BROWN, R.G. (1959) *Statistical forecasting for inventory control*, McGraw-Hill: New York.

CHATFIELD, C. and M. YAR (1991) Prediction intervals for multiplicative Holt-Winters, *Int. J. Forecasting*, **7**, 31–37.

GARDNER, E.S. (1985) Exponential smoothing: the state of the art, *Journal of Forecasting*, **4**, 1–28.

KOEHLER, A.B., R.D. SNYDER and J.K. ORD (1999) "Forecasting models and prediction intervals for the multiplicative Holt-Winters method", Working paper 1/99, Department of Econometrics and Business Statistics, Monash University, Australia.

LIBERT, G. (1984) The M-competition with a fully automatic Box-Jenkins procedure, *J. Forecasting*, **3**, 325–328.

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., and M. HIBON (2000) The M3-competition: results, conclusions and implications *Int. J. Forecasting*, **16**, 451–476.

MAKRIDAKIS, S., S.C. WHEELWRIGHT and R.J. HYNDMAN (1998) *Forecasting: methods and applications*, John Wiley & Sons: New York.

ORD, J.K., A.B. KOEHLER and R.D. SNYDER (1997) Estimation and prediction for a class of dynamic nonlinear statistical models, *J. Amer. Statist. Assoc.*, **92**, 1621–1629.

PEGELS, C.C. (1969) Exponential forecasting: some new variations, *Management Science*, **12**, 311–315.

REILLY, D. (1999) *Autobox 5.0 for Windows: reference guide*, Automatic Forecasting Systems: Hatboro, Pennsylvania.

SNYDER, R.D. (1985) Recursive estimation of dynamic linear statistical models, *J. Roy. Statist. Soc.*, B **47**, 272–276.

STELLWAGEN, E.A., and R.L. GOODRICH (1999) *Forecast Pro 4.0 manual*, Business Forecast Systems: Belmont, Massachusetts.