

A New Bayesian Formulation for Holt's Exponential Smoothing

ROBERT R. ANDRAWIS^{1*} AND AMIR F. ATIYA²

¹ *Data Mining Center of Excellence, MCIT, Cairo, Egypt*

² *Department of Computer Engineering, Cairo University, Giza, Egypt*

ABSTRACT

In this paper we propose a Bayesian forecasting approach for Holt's additive exponential smoothing method. Starting from the state space formulation, a formula for the forecast is derived and reduced to a two-dimensional integration that can be computed numerically in a straightforward way. In contrast to much of the work for exponential smoothing, this method produces the forecast density and, in addition, it considers the initial level and initial trend as part of the parameters to be evaluated. Another contribution of this paper is that we have derived a way to reduce the computation of the maximum likelihood parameter estimation procedure to that of evaluating a two-dimensional grid, rather than applying a five-variable optimization procedure. Simulation experiments confirm that both proposed methods give favorable performance compared to other approaches. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS forecasting; data mining; computer engineering

INTRODUCTION

The last few decades have witnessed significant advances in the topic of exponential smoothing. It has established itself as one of the leading forecasting strategies. Some of the exponential smoothing methods have occupied close to the top spots in the M3 forecasting competition rankings (see M3 Competition, 2005; Makridakis and Hibon, 2000). Also, from the theoretical point of view our understanding of exponential smoothing methods has also improved significantly; see the thorough and extensive review of Gardner (2006). Unfortunately, the parameter estimation aspect of exponential smoothing has not been adequately tackled in the literature. The main approach has been to apply some general-purpose optimization procedure (for minimizing the mean square error). The number of parameters can be as much as five or six parameters in non-seasonal methods (including the initial level and trend) and the optimization procedure cannot guarantee reaching the global minimum. For example Farnum (1992) showed that the response surface is not necessarily convex. Makridakis and Hibon (1991) argued that the initial level and trend are less influential than the other

* Correspondence to: Robert R. Andrawis, Data Mining Center of Excellence, MCIT, Cairo 11221, Egypt. E-mail: robertrezk@yahoo.ca

parameters and found that several simple heuristic rules to set their values are comparable. We have found that these initial values do indeed matter, especially if evaluated using optimal methods. A theoretical breakthrough in modeling exponential smoothing methods has been the introduction of single source of error state space models by Ord *et al.* (1997) and Hyndman *et al.* (2002). This has essentially put them on a solid statistical foundation that has allowed other aspects to be derived from first principles, for example the forecast variance (Hyndman *et al.*, 2001). Another aspect that can be based on the state space formulation has been to estimate the parameters using the maximum likelihood framework. This has also been developed by Hyndman *et al.* (2002), but there have been some precursors in the work of Broze and Mélard (1990). Maximum likelihood also leads to a similar tedious multi-variable optimization formulation. While the computational issue is a consideration, the more pressing issue is estimation accuracy. Naturally, an inaccurate estimate of the parameters begets inaccurate forecasts.

The other competing methodology for parameter estimation is the Bayesian paradigm (West and Harrison, 1989). In this approach the parameters are considered to obey some form of a priori distribution. Then, the posterior of the parameters given the observables is evaluated. This posterior is then used to compute the density of the point to be forecast (see the excellent review by Geweke and Whiteman, 2006). Essentially this posterior is used to weight the forecast according to the plausibility of the parameter set producing this forecast. The Bayesian concept has also been generalized to the level of models (rather than only parameters) in the so-called Bayesian model averaging concept (Hoeting *et al.*, 1999; Chatfield, 1995). One advantage of the Bayesian approach is that it gives the full distribution of the forecast (and hence also confidence bands). Moreover, this distribution reflects the uncertainty due to parameter estimation. This could possibly combat the fact that the estimated prediction intervals for many methods tend to be too narrow (Chatfield, 2001). Obtaining the distribution of the forecast is considered a very favorable feature, as it would lead to more sound decision making, and could give estimates of risk. The exacting aspect of the Bayesian methodology is the computational issue. The encountered multi-variable integral over the parameter values is usually very hard to evaluate analytically (even for some simple and established time series models an analytical formula is not available). The common approach has been to apply some form of Monte Carlo procedure, especially the MCMC method. In spite of the vastness of the Bayesian topic there have been very few applications to the exponential smoothing models. Forbes *et al.* (2000) apply a Monte Carlo procedure to estimate the point forecast and the forecast variance. The method is based on the state space formulation of Ord *et al.* (1997) and Hyndman *et al.* (2002).

As a step towards possibly improving the parameter estimation aspect, in this paper we consider the Holt additive model and develop a Bayesian forecasting method. Like Forbes *et al.* (2000) we also use the state space formulation (Ord *et al.*, 1997; Hyndman *et al.*, 2002) as a starting point. However, we apply a different methodology in manipulating the probabilities that makes it possible to evaluate many of the integrals analytically, and therefore no Monte Carlo or MCMC procedures are needed. We reduce the final solution to that of evaluating a two-dimensional integral, from a starting point of a five-dimensional integral (to be evaluated for a grid of values of the variable to be forecast). This two-dimensional integral can be evaluated numerically in a straightforward manner by simply constructing a 2D grid.

Another contribution of this work concerns a simplification of the maximum likelihood procedure, also for Holt's additive model. We convert the five-parameter optimization procedure to a two-parameter optimization problem, thus leading to significantly faster implementation. In addition, the forecasting accuracy improves because for the new formulation one is guaranteed to obtain the exact

maximum. On the other hand, for the traditional approach (of using a general-purpose optimizer for the five-parameter problem) the optimization process is vulnerable to local maxima or very flat surfaces that preclude getting to the exact maximum. These problems can also impact the forecasting performance, as we shall see.

The paper is organized as follows. The next section is a brief overview of Holt's model, laying out some terminology and definitions. The third section briefly reviews the Bayesian concept, and proceeds by detailing the derivations of the new Bayesian method. The fourth section presents the proposed maximum likelihood work. The fifth section gives the simulations results, followed by the conclusion in the final section.

HOLT'S METHOD

Let y_t be the time series to be forecast. Holt's model is based on estimating smoothed versions of the level and the trend of the time series. The level plus the trend is then extrapolated forward to obtain the forecast. The equations governing the update of the trend and the level are given by (Gardner, 2006):

$$l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1}) \quad (1)$$

$$b_t = \gamma(y_t - l_{t-1}) + (1 - \gamma)b_{t-1} \quad (2)$$

where l_t is the estimated level and b_t is the estimated trend of the time series. The forecast is given by

$$\hat{y}_{t+m} = l_t + mb_t \quad (3)$$

In an important contribution, Hyndman *et al.* (2001, 2002) and Ord *et al.* (1997) have derived a single source of error state space formulation for Holt's model, given by

$$y_t = l_{t-1} + b_{t-1} + \epsilon_t \quad (4)$$

$$l_t = l_{t-1} + b_{t-1} + \alpha \epsilon_t \quad (5)$$

$$b_t = b_{t-1} + \gamma \epsilon_t \quad (6)$$

where ϵ_t is the error term. Assume that it is normally distributed with zero-mean and variance σ^2 . It is often convenient to work on the basis of this state space formulation, as it allows deriving the distributions of the quantities of interest. In this work we use this formulation as a basis for all subsequent analyses. Note that from (4) we get

$$\epsilon_t = y_t - l_{t-1} - b_{t-1} \quad (7)$$

Substituting in (5) and (6), we get the two original level and trend equations of (1) and (2), respectively, thus confirming the validity of these state equations.

THE PROPOSED METHOD

The general idea

The proposed model can be described generally as follows. Let y_t be the time series, and assume that the estimation dataset is the time series portion from 1 to T . For any time series x_t let us denote $x_{i:j}$ as the time series portion from time i to time j . Let us group all model parameters in one vector a . Consider that we would like to forecast m steps ahead. Therefore we need the probability density $p(y_{T+m} | y_{1:T})$ and this will yield the point forecast and the confidence interval of the forecast. This density can be evaluated as

$$p(y_{T+m} | y_{1:T}) = \int p(y_{T+m} | y_{1:T}, a) p(a | y_{1:T}) da \tag{8}$$

The term $p(y_{T+m} | y_{1:T}, a)$ represents the probability density of the future value of the time series, given the parameters are fixed at value a . The term $p(a | y_{1:T})$ represents the probability that parameter set a is a valid set given the data. It serves to weight the forecasts obtained by the fixed-parameter models (as given by the probability density $p(y_{T+m} | y_{1:T}, a)$), according to the plausibility of their respective parameters.

Computation of the posterior probability of the parameters

To apply formula (8) we need to evaluate two terms: $p(y_{T+m} | y_{1:T}, a)$ and $p(a | y_{1:T})$. In this section we derive a formula for the latter term. In the next section we consider the other term.

Let us define the vector of level and trend variables: $v_t = (l_t \ b_t)'$, where $'$ denotes the transpose operation. In matrix form, the formulas (1) and (2) become

$$v_t = Av_{t-1} + wy_t \tag{9}$$

where

$$A = \begin{pmatrix} 1-\alpha & 1-\alpha \\ -\gamma & 1-\gamma \end{pmatrix} \tag{10}$$

and

$$w = \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} \tag{11}$$

Iterating Equation (9) gives

$$v_t = A^t v_0 + \sum_{i=0}^{t-1} A^i w y_{t-i} \tag{12}$$

There are five parameters available: the initial values for the level and the trend, the α and the γ parameters, and the variance σ^2 of the error term, so $a' = (l_0, b_0, \alpha, \gamma, \sigma^2) = (v_0', \alpha, \gamma, \sigma^2)$. Using Bayes' formula, we get

$$p(a|y_{1:T}) = \frac{p(y_{1:T}|a)p(a)}{p(y_{1:T})} \tag{13}$$

$$\propto p(y_{1:T}|a)p(a) \tag{14}$$

where the proportionality sign means proportionality up to a constant multiplier that is independent of the variables of relevance. The density $p(a)$ is the prior density of the parameters in the considered Bayesian formulation.

Let us focus now on v_0 . For simplicity of notation, we keep only the dependence on v_0 in the expressions. It is understood, though, that every density is also conditional on the other parameters α, γ, σ :

$$p(y_{1:T}|v_0) = p(y_T|y_{1:T-1}, v_0)p(y_{T-1}|y_{1:T-2}, v_0) \dots p(y_1|v_0) \tag{15}$$

$$= p(y_T|y_{1:T-1}, v_{T-1}, v_0)p(y_{T-1}|y_{1:T-2}, v_{T-2}, v_0) \dots p(y_1|v_0) \tag{16}$$

$$= p(y_T|v_{T-1})p(y_{T-1}|v_{T-2}) \dots p(y_1|v_0) \tag{17}$$

The term v_{T-i} could be inserted in the conditioning portion of each of the terms in (16) due to the fact that $y_{1:T-i}$ and v_0 uniquely determine v_{T-i} due to formula (12). Equation (17) follows from the Markov property that is satisfied in the state space formulation of Holt's method.

The density should be $p(y_t|v_{t-1})$ can be obtained in a straightforward way from (4) as a normal density. Let $\mathcal{N}(x; \mu_t, \sigma^2)$ denote a normal density for a variable x having mean μ and variance σ^2 . Then

$$p(y_t|v_{t-1}) = \mathcal{N}(y_t; \mu_t, \sigma^2) \tag{18}$$

The variance σ^2 is the same as the variance of the error term ϵ_t and the mean μ_t is given by

$$\mu_t = l_{t-1} + b_{t-1} \tag{19}$$

$$= e' \left[A^{t-1}v_0 + \sum_{i=0}^{t-2} A^i w y_{t-i-1} \right] \tag{20}$$

where e is the vector of all ones. Thus, the final expression becomes

$$p(a|y_{1:T}) = \frac{\prod_{t=1}^T \mathcal{N}(y_t; \mu_t, \sigma^2)p(a)}{p(y_{1:T})} \tag{21}$$

The forecast step

Recall that the probability density that yields the forecast is given by

$$p(y_{T+m}|y_{1:T}) = \int p(y_{T+m}|y_{1:T}, a)p(a|y_{1:T}) da \tag{22}$$

We have evaluated the term $p(a | y_{1:T})$ in the last subsection, and here we will derive the formula for $p(y_{T+m} | y_{1:T}, a)$. Consider the case $m > 1$. We can write

$$p(y_{T+m} | y_{1:T}, a) = \int p(y_{T+m} | \epsilon_{T+1:T+m-1}, y_{1:T}, a) p(\epsilon_{T+1:T+m-1}) d\epsilon_{T+1:T+m-1} \tag{23}$$

Let us evaluate v_{T+m-1} in terms of $\epsilon_{T+1:T+m-1}$ (we use (5) and 6):

$$v_{T+i} = Qv_{T+i-1} + w\epsilon_{T+i} \tag{24}$$

where

$$Q = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{25}$$

and

$$w = \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} \tag{26}$$

Performing the recursion $m - 1$ times starting $T + 1$ to $T + m - 1$, we get

$$v_{T+m-1} = Q^{m-1}v_T + \sum_{j=0}^{m-2} Q^j w \epsilon_{T+m-j-1} \tag{27}$$

Since from (4)

$$y_{T+m} = l_{T+m-1} + b_{T+m-1} + \epsilon_{T+m} \tag{28}$$

$$= e'v_{T+m-1} + \epsilon_{T+m} \tag{29}$$

we get

$$p(y_{T+m} | y_{1:T}, \epsilon_{T+1:T+m-1}, a) = p(y_{T+m} | v_{T+m-1}, y_{1:T}, \epsilon_{T+1:T+m-1}, a) \tag{30}$$

$$= \mathcal{N}(y_{T+m}; e'v_{T+m-1}, \sigma^2) \tag{31}$$

where the first equation follows from the fact that $\epsilon_{T+1:T+m-1}$ and $y_{1:T}$ uniquely determine v_{T+m-1} using (27) and (12). Let us now evaluate the integral in (23). We substitute the numeric values of (25) and (26) into (27) to get

$$e'v_{T+m-1} = l_T + mb_T + u'\epsilon \tag{32}$$

where

$$u = \begin{pmatrix} \alpha + (m-1)\gamma \\ \alpha + (m-2)\gamma \\ \vdots \\ \alpha + \gamma \end{pmatrix} \tag{33}$$

and

$$\epsilon = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} \\ \vdots \\ \epsilon_{T+m-1} \end{pmatrix} \quad (34)$$

Using the fact that $p(\epsilon_{T+i}) = \mathcal{N}(\epsilon_{T+i}; 0, \sigma^2)$ and substituting the formula (31) into (23), we obtain an integral of a product of normal density functions. After some manipulation, it can be evaluated as

$$p(y_{T+m}|y_{1:T}, a) = \mathcal{N}(y_{T+m}; l_T + mb_T, \sigma^2(1 + \|u\|^2)), \quad m > 1 \quad (35)$$

The case $m = 1$ can be obtained by simply observing (29)

$$p(y_{T+1}|y_{1:T}, a) = \mathcal{N}(y_{T+1}; l_T + b_T, \sigma^2), \quad m = 1 \quad (36)$$

We then obtain the requested formula by substituting (21) and 35 into (22). We get

$$p(y_{T+m}|y_{1:T}) = \frac{\int \mathcal{N}(y_{T+m}; l_T + mb_T, \sigma^2(1 + \|u\|^2)) \prod_{t=1}^T \mathcal{N}(y_t; \mu_t, \sigma^2) p(a) da}{p(y_{1:T})} \quad (37)$$

The forecast would then be the mean of the previous equation (w.r.t. the variable y_{T+m}):

$$\hat{y}_{T+m} = \frac{\int (l_T + mb_T) \prod_{t=1}^T \mathcal{N}(y_t; \mu_t, \sigma^2) p(a) da}{p(y_{1:T})} \quad (38)$$

Moreover, we can obtain the variance of the forecast:

$$\text{var}(y_{T+m}|y_{1:T}) = \frac{\int [\sigma^2(1 + \|u\|^2) + (l_T + mb_T)^2] \prod_{t=1}^T \mathcal{N}(y_t; \mu_t, \sigma^2) p(a) da}{p(y_{1:T})} - \hat{y}_{T+m}^2 \quad (39)$$

It might be conceivable to evaluate equations (38) and (39) using some form of Monte Carlo sampling. However, this is very difficult to achieve because the product of normal densities in the integrands yield terms with very high standard deviation. As an alternative approach we propose to evaluate the five-dimensional integral (37) by reducing it to a two-dimensional integral, and then solving the two-dimensional integral numerically. This is explained in detail in the next subsection.

The integration step

In Bayesian parameter estimation, a well-known rule of thumb is that the prior density for k parameters counts as roughly k data points, so its influence is not as great as the observations themselves. Thus the practice has been to select a form for the prior that is amenable to theoretical simplification, while selecting the moments of this form to suit the considered problem or roughly fit the considered data. We will follow this practice here, and this results in a simplification of the formulas. Consider that the priors are given as follows:

$$p(l_0|\sigma) = \mathcal{N}(l_0; \mu_l, k_l\sigma^2) \tag{40}$$

$$p(b_0|\sigma) = \mathcal{N}(b_0; \mu_b, k_b\sigma^2) \tag{41}$$

$$p(\sigma) = \left[\frac{2}{\sigma_0^{c_1+1} \Gamma\left(\frac{c_1+1}{2}\right)} \right] \sigma^{c_1} e^{-\frac{\sigma^2}{\sigma_0^2}} I(\sigma > 0) \tag{42}$$

where $\mu_l, k_l, \mu_b, k_b, c_1$ and σ_0 are constants that control the moments of the a priori density, and I is the indicator function. Concerning α and γ , they are assumed to have any arbitrary priors. Note that the prior for l_0 has a standard deviation proportional to σ . This is a realistic assumption, because as the added noise increases in magnitude the interval for possibly locating the starting value l_0 gets wider. The same is true for the case of b_0 .

In the subsequent analysis it is easier to work with the variance $z = \sigma^2$ in stead of the standard deviation σ . Using z instead of σ , the prior for z becomes a chi-square density. Computing the total prior, we get

$$p(a) = c_2 \mathcal{N}(l_0; \mu_l, k_l z) \mathcal{N}(b_0; \mu_b, k_b z) z^{\frac{c_1-1}{2}} e^{-\frac{z}{\sigma_0^2}} p(\alpha) p(\gamma) \tag{43}$$

where c_2 is given by

$$c_2 = \frac{1}{\sigma_0^{c_1+1} \Gamma\left(\frac{c_1+1}{2}\right)} \tag{44}$$

Substituting into (37):

$$p(y_{T+m} | y_{1:T}) \propto \int_0^1 \int_0^1 \int_0^\infty \int_0^\infty \mathcal{N}(y_{T+m}; l_T + mb_T, (1 + \|u\|^2)z) \left(\prod_{i=1}^T \mathcal{N}(y_i; \mu_i, z) \right) \mathcal{N}(l_0; \mu_l, k_l z) \mathcal{N}(b_0; \mu_b, k_b z) z^{\frac{c_1-1}{2}} e^{-\frac{z}{\sigma_0^2}} p(\alpha) p(\gamma) dl_0 db_0 dz d\alpha d\gamma \tag{45}$$

Consider the innermost two integrals, w.r.t. l_0 and b_0 . They can be evaluated in closed form, because the integrand is in the form of a normal density in terms of these two variables. Define the following:

$$\Lambda = \begin{pmatrix} \frac{1}{k_l} & 0 \\ 0 & \frac{1}{k_b} \end{pmatrix} \tag{46}$$

$$\mu_{v_0} = \begin{pmatrix} \mu_l \\ \mu_b \end{pmatrix} \tag{47}$$

$$g = y_{T+m} - e'_m \sum_{i=0}^{T-1} A^i w y_{T-i} \tag{48}$$

$$h_t = y_t - e' \sum_{i=0}^{t-2} A^i w y_{t-i-1} \tag{49}$$

$$D = \frac{A'^T e_m e_m' A^T}{1 + \|u\|^2} + \sum_{t=1}^T A'^{t-1} e e' A^{t-1} + \Lambda \tag{50}$$

$$d = \frac{A'^T e_m g}{1 + \|u\|^2} + \sum_{t=1}^T h_t A'^{t-1} e + \Lambda \mu_{v_0} \tag{51}$$

where $e_m = (1, m)'$, $e = (1, 1)'$ and A and w are as defined in (10) and (11). Then the integrand of (45) can be manipulated into the following form:

$$\text{Integrand} = \mathcal{N}(v_0; D^{-1}d, D^{-1}z) \frac{e^{-\frac{1}{2z} \left[\frac{g^2}{1 + \|u\|^2} + \sum_{t=1}^T h_t^2 + \mu'_{v_0} \Lambda \mu_{v_0} - d' D^{-1} d \right]}}{(2\pi z)^{\frac{T+1}{2}} \det^{\frac{1}{2}}(D) \sqrt{k_t k_b (1 + \|u\|^2)}} z^{\frac{c_1-1}{2}} e^{-\frac{z}{\sigma_0^2}} p(\alpha) p(\gamma) \tag{52}$$

Then, integrating the innermost two integrals for the above integrand (w.r.t. l_0 and b_0) leads to

$$p(y_{T+m} | y_{1:T}) \propto \int_0^1 \int_0^1 \int_0^\infty \frac{e^{-\frac{1}{2z} \left[\frac{g^2}{1 + \|u\|^2} + \sum_{t=1}^T h_t^2 + \mu'_{v_0} \Lambda \mu_{v_0} - d' D^{-1} d \right]}}{\det^{\frac{1}{2}}(D) \sqrt{1 + \|u\|^2}} z^{-\frac{(T-c_1+2)}{2}} e^{-\frac{z}{\sigma_0^2}} p(\alpha) p(\gamma) dz d\alpha d\gamma \tag{53}$$

where all constant multipliers in the above equation are not included due to the existence of the proportionality operation. Consider now the integral w.r.t. z . This integral can be evaluated in closed form. From [9], we get

$$p(y_{T+m} | y_{1:T}) \propto \int_0^1 \int_0^1 \frac{\rho^{-\left(\frac{T-c_1}{4}\right)} K_{\left(\frac{T-c_1}{2}\right)} \left(\frac{2\sqrt{\rho}}{\sigma_0} \right) p(\alpha) p(\gamma)}{\det^{\frac{1}{2}}(D) \sqrt{1 + \|u\|^2}} d\alpha d\gamma \tag{54}$$

where

$$\rho = \frac{1}{2} \left[\frac{g^2}{1 + \|u\|^2} + \sum_{t=1}^T h_t^2 + \mu'_{v_0} \Lambda \mu_{v_0} - d' D^{-1} d \right] \tag{55}$$

and $K_\nu(x)$ denotes the modified Bessel function of the second kind. For many scientific software products, such as Matlab, this function is available as a built-in function. We can see that we have now reduced the five-dimensional integral to a two-dimensional one. This renders the problem feasible from the computational point of view. Of course, this integral has to be evaluated for all values

of a grid of y_{T+m} , and then the resulting function has to be normalized so that it integrates to 1. Computationally faster methods can be obtained from this formula if only the mean and the variance of the forecast need to be evaluated. In fact, a simple re-derivation of the steps presented above, starting from (38) and (39), will yield such quantities using only four 2D integrals similar in form to (54). This analysis, however, will not be presented here to avoid sidetracking into too many issues.

Applicability to other models

Exponential smoothing methods have two main features: trend and seasonality. This analysis applies only to methods with trend. If seasonality is added the model becomes much more complex and probably not feasible. We believe though that this is not a problem. In Gardner's review (2006, p. 658) he reviews the application studies that appeared in the literature covering many types of time series. The majority of the studies consider only trend (Holt's method) but not seasonality (Holt-Winters' method), even though the tested time series most often possess seasonality. It seems that researchers and practitioners prefer to deseasonalize, then apply Holt's method, rather than apply Holt-Winters' method outright. Concerning trend, let us classify the trend using the standard abbreviations as 'N' (non-existent), 'A' (additive), 'D' (damped), 'M' (multiplicative). For 'N', an analogous but simpler analysis can be performed, as the equations in this case are much simpler than what we have handled. For 'D', the state equations (4), (5) and (6) are the same, except that in (6) the b_{t-1} in the RHS is multiplied by the damping coefficient ϕ . Everything will proceed as presented, except that we have the parameter vector possessing six components now (ϕ is added) and the final formula will be generally similar to (54), except that it will be a three-dimensional integral (the third integral being over ϕ). Concerning 'M', our method is not applicable, as the nice linear nature of the state space equations does not apply to this situation.

MAXIMUM LIKELIHOOD

The maximum likelihood approach as a tool for parameter estimation for Holt's model has been proposed by Hyndman *et al.* (2002). Their model, however, is based on using some nonlinear optimization method to maximize the likelihood with respect to the five-dimensional parameter vector. We propose here a much simpler solution by reducing the problem to a simple two-dimensional optimization problem that can be solved using a simple 2D search in a fixed range.

The log-likelihood function is defined as

$$L(a) = \log(p(y_{1:T} | a)) \tag{56}$$

From (17), (18) and (20), we get

$$L(a) = \sum_{t=1}^T \log p(y_t | v_{t-1}) \tag{57}$$

$$= -T \log(\sqrt{2\pi}\sigma) - \sum_{t=1}^T \frac{\left(y_t - e' \left[A^{t-1} v_0 + \sum_{i=0}^{t-2} A^i w y_{t-i-1} \right] \right)^2}{2\sigma^2} \tag{58}$$

Maximizing this expression w.r.t. v_0 , we get

$$v_0 = \left(\sum_{t=1}^T A'^{t-1} e e' A'^{t-1} \right)^{-1} \sum_{t=1}^T A'^{t-1} e h_t \quad (59)$$

where

$$h_t = y_t - e' \sum_{i=0}^{t-2} A^i w y_{t-i-1} \quad (60)$$

The optimal σ can be obtained likewise, but it is not needed at this point. Substituting from (59) back into (58), we obtain the final objective function that should be maximized:

$$J = \left(\sum_{t=1}^T A'^{t-1} e h_t \right)' \left(\sum_{t=1}^T A'^{t-1} e e' A'^{t-1} \right)^{-1} \left(\sum_{t=1}^T A'^{t-1} e h_t \right) - \sum_{t=1}^T h_t^2 \quad (61)$$

The previous equation corresponds to the log-likelihood after removing the constant terms that will not influence the maximization process. We have to maximize J w.r.t. the two variables α and γ . The range of the two variables is from 0 to 1, so a simple 2D grid in $[0, 1] \times [0, 1]$ can be performed to obtain the maximum. Such a computation can be instantly done and is much more computationally efficient than the older procedure of using an optimizer w.r.t. the five-dimensional parameter vector. In addition, more insight is gained by having such a compact formula (for example, we can see that the standard deviation of the error term does not affect the maximum likelihood solution).

SIMULATION EXPERIMENTS

To obtain an idea about the comparative advantage of the proposed methods, we have compared these methods to some standard benchmark methods for evaluating the parameters of the Holt's method. Specifically, we consider the following three benchmark methods:

- Set the initial level equal to the first time series point and the initial trend as the difference between the first two time series points, i.e., $l_0 = y_1$, $b_0 = y_2 - y_1$. Set α and γ so that the forecast error on the estimation time series portion is minimized. We call this method the classical method, or in short CLASSIC1.
- On the first 10 points of the time series we perform a linear regression. The intercept gives l_0 , while the slope gives b_0 . Set α and γ so that the forecast error on the estimation data portion is minimized (see Hyndman *et al.*, 2002). We call this the CLASSIC2.
- Maximize the likelihood of (56) w.r.t. to all five parameters α , γ , l_0 , b_0 , σ using the Matlab `fmincon` numerical optimization routine. The `fmincon` routine is based on solving a quadratic programming subproblem at each iteration. An estimate of the Hessian of the Lagrangian is updated at each iteration using the BFGS formula. The quadratic programming part is solved using an efficient active set strategy. In our case we have four constraints: $0 \leq l_0 \leq 1$ and $0 \leq b_0 \leq 1$ (the $\sigma \geq 0$ constraint is enforced by using $|\sigma|$ instead). The other parameters such as tolerance, maximum number of function evaluations and maximum number of iterations, are selected respectively as 10^{-10} , 10,000,

and 1,000,000 (beyond these selected values there is generally no further improvement). The initial values are selected as follows: the parameters l_0 and b_0 are selected as in the previous method by fitting a regression line to the first 10 points. The standard deviation is selected similar to the parameter σ_0 for the Bayesian model (see the paragraph after next for details). Concerning α and γ , each is initialized as zero. We call this approach ML NUM (NUM stands for numerical).

Against these three benchmarks we apply the proposed Bayesian model based method (in short BAYES), as well as the maximum likelihood method as proposed in the previous section. We abbreviate this by the letters ML. In the BAYES method we assume that the form of the a priori distributions is as in (40), (41), and (42) for l_0 , b_0 , and σ , respectively. Concerning α and β , we select their a priori distributions as simply uniform in $[0, 1]$.

As for the parameters of the other a priori distributions, the key ones are estimated from the data. On the first 10 points of the time series we perform a linear regression. The intercept is taken as an estimate of μ_l , the mean of the distribution $p(l_0)$. The slope is taken as an estimate of μ_b , the mean of the distribution $p(b_0)$. The parameters k_l and k_b are set equal to 1. We considered the variance to be exponentially distributed, that is, $c_1 = 1$. The parameter σ_0 is estimated by subtracting a centered moving average of the time series from the time series itself, and obtaining the standard deviation of the resulting residual time series. All estimates of the parameters of the priors are of course only rough estimates. In Bayesian formulation an accurate estimate of the prior is usually not crucial, as most of the influence on the forecast will ultimately come from the observables.

We considered the *NN3 Artificial Neural Networks and Computational Intelligence Forecasting Competition, 2007* see NN3 2007. This is one of the major competitions that was held recently. It contains 111 monthly business type time series. The time series vary in length between 50 and 126. By inspecting the individual time series we found that it possesses quite similar features to the M3 time series, such as seasonality and trend.

We pre-processed the data before applying the competing algorithms. First, a log transformation is applied, by simply taking the log of the time series. Then a seasonality test is administered to determine whether the time series contains a seasonal component. The test is made by taking the autocorrelation with a lag of 12 months, to test the hypothesis 'no seasonality' using Bartlett's formula for the confidence interval (see Box and Jenkins, 1976). If the test indicates the presence of seasonality, then we use the classical additive decomposition approach (Makridakis *et al.*, 1998). In this approach a centered moving average is performed, then a month-by-month average is computed on the smoothed series. This average will then be the seasonal average. We subtract that from the original series to create the deseasonalized series. The competing algorithms are applied on the pre-processed series. We held out the last 10 points of each time series, which represent the time horizon that has to be forecast.

There are two possible ways to apply the BAYES model. In the first one we compute the density of the value to be forecast according to (54). Then we compute the mean of this density and transform that mean back to the original space by adding the seasonal component and applying an exponential transformation (to invert the log step). The other way is to transform the density to the original space using the formula for random variable transformations. We used the former method, as in our earlier experimentation with the model it gave decidedly better results.

We used as error measure the symmetric mean absolute percentage error, defined as

$$\text{SMAPE} = \frac{1}{M} \sum_m \frac{|\hat{y}_m - y_m|}{(|\hat{y}_m| + |y_m|)/2} \quad (62)$$

where y_m is the actual time series value, \hat{y}_m is the forecast, and M is the number of points that are forecast. The summation is over all points to be forecast in a time series and over all time series. Since it is a relative error measure it is possible to combine the errors for the different time series into one number. We have 10 forecasts per time series times 111 time series, which makes over 1000 points, thus giving confidence in the validity of the findings of the comparison study.

To test the statistical significance of the observed average ranking of the compared methods we implemented a test proposed by Koning *et al.* (2005). It is a test based on the rankings of the different methods. It is termed multiple comparisons with the best (MCB), and is based on the work of McDonald and Thompson (1967). It essentially tests whether some methods perform significantly worse than the best method. In this method we compute the rank of each method k on each time series n , say $R_k(n)$, with 1 being the best and 5 being the worst. Let $K \equiv 5$ denote the number of methods compared, and let P be the number of time series (in our case $P = 111$). The average rank of each method k , or \bar{R}_k , is computed by averaging $R_k(n)$ over all time series. The $\alpha\%$ confidence limits (we used $\alpha = 90\%$) will then be

$$\bar{R}_k \pm 0.5q_{\alpha K} \sqrt{\frac{K(K+1)}{12P}} \quad (63)$$

where $q_{\alpha K}$ is the upper α percentile of the range of K independent standard normal variables. Further details of this test can be found in Koning *et al.* (2005). Table I gives a summary of all the SMAPE and the rank results, including the 90% ranking confidence bands. Figure 1 displays the average rankings and their confidence bands. Also, Figure 2 shows as an example one of the time series, together with the forecasts produced by BAYES. Also shown in the figure are the 90% confidence bands for the forecasts as a function of the forecast horizon.

To investigate the robustness of the BAYES and ML methods to outliers, we have examined the NN3 time series and identified that 25 of them have outliers in the estimation period. An outlier is defined by the occurrence of a sudden jump in time series (after log transformation and deseasonalization) by 10% or more. The ML method turned out to beat BAYES (ML giving 0.322 versus 0.379 for BAYES).

We have also measured the speed (in CPU time) of the three major contenders: the BAYES, ML, and ML NUM. Per time series the computation time on average turned out to be respectively 95.1 s, 6.1 s, and 27.0 s; thus the ranking in terms of speed is ML, then ML NUM, then BAYES. The reason BAYES is computationally demanding is that the two-dimensional integral is evaluated for every possible value of y_{T+m} (in order to obtain the full conditional distribution of y_{T+m}). If only the point forecast and the forecast variance are needed, then as mentioned above one can evaluate these using

Table I. The performance of the proposed BAYES and ML methods in comparison with the three other benchmark methods on the NN3 competition time series problems

Method	SMAPE	Mean rank	Rank interval (90%)
BAYES	0.175	2.60	(2.34, 2.86)
ML	0.156	2.45	(2.19, 2.71)
ML NUM	0.164	2.68	(2.42, 2.95)
CLASSIC1	0.289	3.70	(3.44, 3.96)
CLASSIC2	0.196	3.51	(3.25, 3.77)

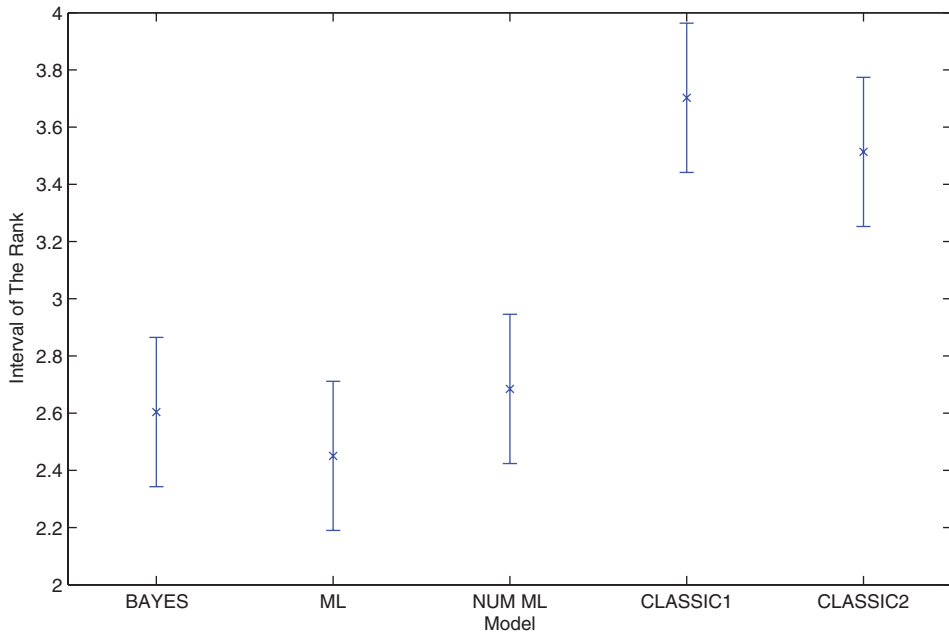


Figure 1. The average ranks with 90% confidence limits for the multiple comparison with the best test

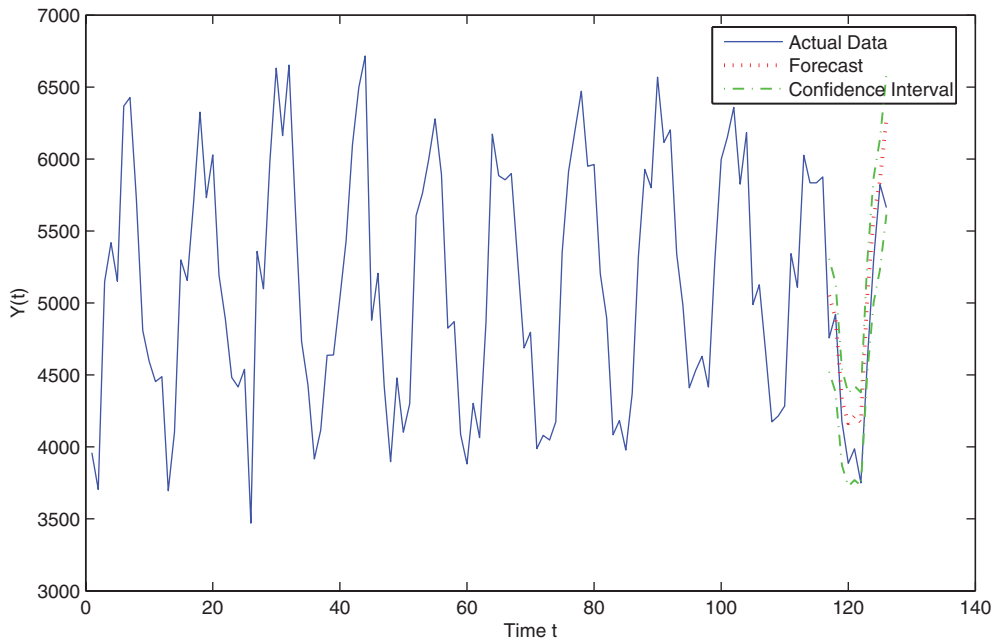


Figure 2. The forecast for the BAYES method for a time series, together with the 90% confidence limits for the forecast. The figure shows all the points of the time series (including the estimation period), and the last 10 points represent the period to be forecast

four 2D integrals, and so it will be computationally much faster than computing the full density of the forecasts.

Comments on the results

One can deduce the following observations:

- One can see from the table that the proposed maximum likelihood method as derived in Section 4 gives the best performance. Both ML and BAYES models significantly outperform the conventional bench-marks CLASSIC1 and CLASSIC2, both in terms of rank and in terms of SMAPE. ML also outperform ML NUM, but the difference in rank is not significant at the 90% level. BAYES is mixed against ML NUM, somewhat better in the rank measure and somewhat worse in the SMAPE measure. If we consider the combination of performance and speed criteria, then the ML is the clear winner: it yields superior performance with very little computational cost.
- The proposed ML is a little better than the traditional ML NUM performance-wise, and much more superior computation-wise, making ML the way to go if a maximum likelihood based approach is to be considered.
- The BAYES method offers the added advantage of obtaining the density of the forecast value. This is the strong point for BAYES that is not present in any of the other approaches. This leads to the added computational cost.
- One can conclude that the influence of the initial level and trend values, l_0 and b_0 respectively, is very significant, and they are important to set in an optimal fashion. The essential difference between the CLASSIC1 and the CLASSIC2 methods on one hand, and the BAYES, ML, and ML NUM on the other, is the way l_0 and b_0 are handled. Most practice so far has been to set l_0 and b_0 in a heuristic fashion (as in CLASSIC1 and CLASSIC2). But this practice evidently should change, especially since the alternative (i.e., setting these parameters in an optimal fashion) can be done by a simple two-dimensional search (as proposed in this article for the ML and BAYES models).

CONCLUSIONS

In this paper we have introduced two novel methods for the parameter estimation problem for Holt's additive method. The first proposed solution is based on the Bayesian approach. The method is reduced to computing a two-dimensional integral, which can be done numerically in a straightforward way. The strong point about this approach is that it yields the forecast density, unlike the majority of the available exponential smoothing methods. The other method utilizes the state space formulation to derive a computational approach for evaluating the maximum likelihood parameter estimates. The proposed approach renders the optimization very simple and practical to implement, which leads to improvements in performance and significant gains in speed.

Overall, we feel the proposed BAYES and ML approaches are a good addition to the repertoire of different exponential smoothing methods, and there could be some good fraction of time series where these methods would hold an edge.

ACKNOWLEDGEMENTS

We would like to acknowledge the help of Nesreen Ahmed and Athanasius Youhanna of Cairo University, who helped in the pre-processing portion of the NN3 time series. We would also like to

acknowledge the useful discussions with Professor Ali Hadi of the American University of Cairo and Cornell University, Dr Hisham El-Shishiny of IBM Egypt, and Dr Neamat El Gayar of Cairo University. This work is part of the *Data Mining for Improving Tourism Revenue in Egypt* research project within the Egyptian Data Mining and Computer Modeling Center of Excellence.

REFERENCES

- Box G, Jenkins G. 1976. *Time Series Analysis, Forecasting and Control*. Holden-Day: San Francisco, CA.
- Broze L, Mélard G. 1990. Exponential smoothing: estimation by maximum likelihood. *Journal of Forecasting* **9**: 445–455.
- Chatfield C. 1995. Model uncertainty, data mining, and statistical inference (with discussion). *Journal of the Royal Statistical Society A* **158**: 419–466.
- Chatfield C. 2001. Prediction intervals for time-series forecasting. In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Armstrong JS (ed.). Springer: Berlin; 475–494.
- Farnum NR. 1992. Exponential smoothing: behavior of the ex-post sum of squares near 0 and 1. *Journal of Forecasting* **11**: 47–56.
- Forbes C, Snyder R, Shami R. 2000. Bayesian exponential smoothing. Econometrics and Business Statistics Working Paper, Monash University.
- Gardner E. 2006. Exponential smoothing: the state of the art—part II. *International Journal of Forecasting* **22**: 637–666.
- Geweke J, Whiteman C. 2006. Bayesian forecasting. In *Handbook of Economic Forecasting*, Elliot G, Granger C, Timmermann A. (eds). North-Holland: Amsterdam; 3–79.
- Gradshteyn I, Ryzhik I. 1980. *Table of Integrals, Series, and Products*. Academic Press: New York.
- Hyndman R, Koehler A, Ord K, Snyder R. 2001. Prediction intervals for exponential smoothing state space models. Working Paper 11/2001, Department of Econometrics and Business Statistics, Monash University, Australia.
- Hyndman R, Koehler A, Ord K, Snyder R, Grose S. 2002. A state space formulation for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* **18**: 439–454.
- Hoeting JA, Madigan D, Raftery AE, Volinsky CT. 1999. Bayesian model averaging: a tutorial. *Statistical Science* **14**(4): 382–417.
- Koning AJ, Franses PH, Hibon M, Stekler HO. 2005. The M3 competition: statistical tests of the results. *International Journal of Forecasting* **21**: 397–409.
- M3 Competition. 2005. <http://www.forecasters.org/data/m3comp/m3comp.htm> [22 August 2008].
- Makridakis S, Hibon M. 1991. Exponential smoothing: the effect of initial values and loss functions on post-sample forecasting accuracy. *International Journal of Forecasting* **7**: 317–330.
- Makridakis S, Hibon M. 2000. The M3-competition: results, conclusions and implications. *International Journal of Forecasting* **16**: 451–476.
- Makridakis S, Wheelwright SC, Hyndman RJ. 1998. *Forecasting: Methods and Applications* (3rd edn). Wiley: Chichester; Ch. 3.
- McDonald BJ, Thompson WA. 1967. Rank sum multiple comparisons in one and two way classifications. *Biometrika* **54**: 487–497.
- NN3. 2007. Artificial neural networks and computational intelligence forecasting competition. <http://www.neural-forecasting-competition.com/> [22 August 2008].
- Ord K, Koehler A, Snyder R. 1997. Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association* **92**: 1621–1629.
- West M, Harrison J. 1989. *Bayesian Forecasting and Dynamic Models*. Springer: Berlin.

Authors' biographies:

Robert R. Andrawis, Received his B.S. degree in 2006 from the Department of Computer Engineering, Cairo University, Egypt. Since 2006, he has been a researcher with the Data Mining and Computer Modeling Center of Excellence, Ministry of Information and Telecommunications (MCIT). He is currently also pursuing the M.S.

degree at Cairo University. His interests are in the theory of forecasting and data mining. He participated and obtained the first rank in the NN5 Forecasting Competition for Artificial Neural Networks and Computational Intelligence. This is a major international competition whose results will appear in a special issue of International Journal of Forecasting.

Amir F. Atiya, Received his B.S. degree in 1982 from Cairo University, and the M.S. and Ph.D. degree in 1986 and 1991 from Caltech, Pasadena, CA, all in electrical engineering. Dr. Atiya is currently a Professor at the Department of Computer Engineering, Cairo University. He recently held several visiting appointments, such as in Caltech and in Chonbuk National University, S. Korea. His research interests are in the areas of neural networks, machine learning, theory of forecasting, computational finance, and Monte Carlo Methods. He obtained several awards, such as the Kuwait Prize in 2005. Currently, he is an associate editor for the IEEE Transactions on Neural Networks.

Authors' addresses:

Robert R. Andrawis, Data Mining Center of Excellence, Ministry of Communication and Information Technology, Cairo 11221, Egypt.

Amir F. Atiya, Department of Computer Engineering, Cairo University, Giza, Egypt.