Exponential smoothing and non-negative data

Muhammad Akram
Department of Econometrics and Business Statistics,
Monash University, VIC 3800, Australia.
Email: Muhammad.Akram@buseco.monash.edu

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University, VIC 3800, Australia.
Email: Rob.Hyndman@buseco.monash.edu

J. Keith Ord
McDonough School of Business,
Georgetown University, Washington, DC20057.
Email: ordk@georgetown.edu

RPF Working Paper No. 2008-003
http://www.gwu.edu/~forcpgm/2008-003.pdf

July 22, 2008

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www.gwu.edu/~forcpgm

Exponential smoothing and non-negative data

Muhammad Akram
Department of Econometrics and Business Statistics,
Monash University, VIC 3800, Australia.
Email: Muhammad.Akram@buseco.monash.edu

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University, VIC 3800, Australia.
Email: Rob.Hyndman@buseco.monash.edu

J. Keith Ord
McDonough School of Business,
Georgetown University, Washington, DC20057.
Email: ordk@georgetown.edu

23 May 2008

Exponential smoothing and non-negative data

Abstract: The most common forecasting methods in business are based on exponential smoothing and the most common time series in business are inherently non-negative. Therefore it is of interest to consider the properties of the potential stochastic models underlying exponential smoothing when applied to non-negative data. We explore exponential smoothing state space models for non-negative data under various assumptions about the innovations, or error, process.

We first demonstrate that prediction distributions from some commonly used state space models may have an infinite variance beyond a certain forecasting horizon. For multiplicative error models which do not have this flaw, we show that sample paths will converge almost surely to zero even when the error distribution is non-Gaussian. We propose a new model with similar properties to exponential smoothing, but which does not have these problems, and we develop some distributional properties for our new model.

We then explore the implications of our results for inference, and compare the short-term forecasting performance of the various models using data on the weekly sales of over three hundred items of costume jewelry.

The main findings of the research are that the Gaussian approximation is adequate for estimation and one-step-ahead forecasting. However, as the forecasting horizon increases, the approximate prediction intervals become increasingly problematic. When the model is to be used for simulation purposes, a suitably specified scheme must be employed.

Keywords: forecasting; time series; exponential smoothing; positive-valued processes; seasonality; state space models.

1 Introduction

Positive time series are very common in business, industry, economics and other fields, and exponential smoothing methods are frequently used for forecasting such series. These methods have been developed empirically over the years, a notable example being the Holt-Winters scheme (Winters, 1960). A feature of this method is that it combines a linear trend with a multiplicative seasonal component so that the seasonal effects are proportional to the current level of the series. Such methods have proved extremely successful in short-term forecasting, but they typically lack an underlying statistical foundation. We summarize the progress that has been made in building models for such methods in Section 1.1. Although other classes of models might be considered for non-negative time series, we focus upon models that can be used to underpin these commonly-used methods and allow combinations of additive and multiplicative elements.

Because the Gaussian distribution extends over the whole real line, it clearly cannot provide an exact specification for the error process when the series is constrained to be non-negative. Nevertheless, forecasting practice using the methods just mentioned has almost always accepted that the Gaussian assumption is plausible and the results for short-term forecasting appear to be satisfactory when the process is bounded well away from the origin. However, cases may arise where the prediction intervals include negative values, and as the forecasting horizon is extended, even the point forecasts may become negative.

When the model is purely multiplicative, a logarithmic transformation seems a reasonable option. However, when the model has some additive components, this option is not available. Some authors (e.g., Hyndman et al., 2002) have suggested using a truncated Gaussian distribution for the errors so that the sample space is constrained to take only positive values. Other options include the use of distributions such as the gamma or the lognormal that are defined on the positive half-line. The underlying assumptions of using the non-Gaussian error model for the positive random variable are different from using the log-transformed model. For example, a log-transformation to a linear model implies proportional seasonal effects as well as proportional errors.

The purpose of this paper is to determine how far truncation will resolve the underlying difficulties, at least approximately, and when other distributional assumptions and alternative models will be required. We examine this question using innovations state space models, which are described later in this section. Then, in Section 2, we examine some of the specification problems associated with models defined on the positive half line. In Section 3 we

consider purely multiplicative models and examine how far such a specification resolves the difficulties we have identified. Section 4 provides some specific distributional results when the innovations are from a lognormal distribution. In Section 5, we examine the extent to which the Gaussian distribution can serve as a reasonable approximation, notwithstanding the theoretical objections noted earlier. We need to consider parameter estimation, point forecasting, interval forecasting and finally simulation. We present some empirical results in Section 6, first for a single series on U.S. freight car shipments and then on a set of weekly sales figures for items of costume jewelry. The conclusions appear in Section 7.

Various works, such as West et al. (1985); Harvey and Fernandes (1989) and Grunwald et al. (1993), have used non-Gaussian state space models to describe non-stationary time-series. However, Grunwald et al. (1997) have shown under very mild conditions that, for non-negative series, sample paths of many of these models converge to some constant almost surely, making them unsuitable for modeling in many situations. Finally we note that the well-known GARCH model applies to non-negative series in the sense that it is used to describe volatility, and is not a typical model for non-negative series. An ARIMA model with constraints to ensure non-negativity corresponds to the class of purely additive models, typified by the models listed under Class A in the next section, so that a subset of possible ARIMA models is considered in our analysis. Other ARIMA models share the same properties as those in Class A, with respect to non-negative series.

1.1 Modeling framework

Following Ord et al. (1997), we specify the general innovations state space model as:

$$y_t = w(\boldsymbol{x}_{t-1}) + r(\boldsymbol{x}_{t-1})\varepsilon_t \tag{1a}$$

$$\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_{t-1}) + \boldsymbol{g}(\boldsymbol{x}_{t-1})\varepsilon_t, \tag{1b}$$

where $r(\cdot)$ and $w(\cdot)$ are scalar functions, $\boldsymbol{f}(\cdot)$ and $\boldsymbol{g}(\cdot)$ are vector functions, and ε_t is a white noise process with variance σ^2. Note that we do not specify that the process is Gaussian because such an assumption will conflict with the underlying structure of the data generating process when the series contains only non-negative values.

In the most general case we consider, the state vector may be written as $\boldsymbol{x}_t = (\ell_t, b_t, s_t, s_{t-1}, \ldots, s_{t-m+1})'$ where ℓ_t denotes the local level, b_t is the local trend and the s_{t-j} terms represent local seasonal effects when there are m seasons. We further restrict the

general system (1) to models where the functions represent either additive or multiplicative components. For example, the model with multiplicative error, a (damped) multiplicative trend and a multiplicative seasonal pattern may be written as

$$y_t = \ell_{t-1}\, b_{t-1}^{\phi}\, s_{t-m}(1+\varepsilon_t) \tag{2a}$$

$$\ell_t = \ell_{t-1}\, b_{t-1}^{\phi}(1+\alpha\varepsilon_t) \tag{2b}$$

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

$$s_t = s_{t-m}(1+\gamma\varepsilon_t), \tag{2d}$$

where $0 < \phi < 1$ denotes the dampening factor. We consider these models within the framework proposed in Hyndman et al. (2002) and extended by Taylor (2003). The framework involves 30 different models (15 with additive errors and 15 with multiplicative errors). We call these "ETS models" (following Hyndman et al., 2008) where ETS stands for both ExponenTial Smoothing and Error, Trend, Seasonal. Each ETS model is denoted by a triplet denoting the error, trend and seasonal components. For example, the model (2) may be represented by the triplet ETS(M,M_d,M). Table 1, adapted from Hyndman et al. (2002), shows the 15 ETS models with multiplicative errors.

Trend Component		**Seasonal Component** N (none)	A (additive)	M (multiplicative)
N	(none)	(M,N,N)	(M,N,A)	(M,N,M)
A	(additive)	(M,A,N)	(M,A,A)	(M,A,M)
A_d	(additive damped)	(M,A_d,N)	(M,A_d,A)	(M,A_d,M)
M	(multiplicative)	(M,M,N)	(M,M,A)	(M,M,M)
M_d	(multiplicative damped)	(M,M_d,N)	(M,M_d,A)	(M,M_d,M)

Table 1: *The fifteen ETS state space models with multiplicative errors from the taxonomy of Hyndman et al. (2002) as extended by Taylor (2003).*

In this paper, we divide these ETS models into four classes:

Class M: Purely multiplicative models: (M,N,N), (M,N,M), (M,M,N), (M,M,M), (M,M_d,N) and (M,M_d,M);

Class A: Purely additive models: (A,N,N), (A,N,A), (A,A,N), (A,A,A), (A,A_d,N) and (A,A_d,A);

Class X: Models with additive errors and at least one multiplicative component, and models with multiplicative errors and multiplicative trend but additive seasonality: (A,M,$*$),

(A,M_d,∗), (A,∗,M), (M,M,A), (M,M_d,A), where ∗ denotes any admissible component (11 models);

Class Y: Models with multiplicative errors and additive trend, and the model with multiplicative errors and additive seasonality but no trend: (M,A,∗), (M,A_d,∗) or (M,N,A), where ∗ denotes any admissible component (7 models).

It is evident that only the purely multiplicative models of Class M can guarantee a sample space restricted to the positive half-line with suitable restrictions on the innovations. Class A contains the purely additive models, widely used in practice for short-term forecasting, but they clearly do not conform to the requirements of non-negative processes unless additional conditions are imposed. The remaining models in Classes X and Y all possess both multiplicative and additive components. Holt's linear method (A,N,N) and Holt-Winters method with additive seasonality are members of Class A, and the Holt-Winters method with multiplicative seasonality is a member of Class Y. All have been widely used to model non-negative series for over 40 years. If the observational sample space is not restricted to be strictly positive, the Class X models can have an infinite forecast variances beyond certain forecast horizons, as we show in the next section. This problem does not arise, however, for the Class Y models.

The forecast variance is defined as the variance of y_{t+h} conditional on observations to time t and the initial state:

$$v_{t+h|t} = \mathrm{V}(y_{t+h} \mid y_1, y_2, \ldots, y_t, \boldsymbol{x}_0).$$

We note that Hyndman et al. (2005) provide forecast variance expressions for fifteen of the thirty models; exact expressions are not available for the multi-step-ahead forecast variances for the other models.

2 Problems with the models

We now examine some of the difficulties associated with trying to use the models when the process is strictly positive.

2.1 The infinite variance problem

Any model with the error distribution taking negative values with non-zero probability has the first passage time property that the process will eventually lead to negative values; in practice, the probability is very small if there is a strong upward trend. Thus, we may show

that (Hyndman et al., 2008, Chapter 15) most of the models in class X have undefined means and infinite variances for $h \geq 3$ steps ahead (or $h \geq m+2$ for the three (A,*,M) models).

To see why, consider the ETS(A,M,N) model:

$$\begin{aligned} y_t &= \ell_{t-1} b_{t-1} + \varepsilon_t \\ \ell_t &= \ell_{t-1} b_{t-1} + \alpha \varepsilon_t \\ b_t &= b_{t-1} + \beta \varepsilon_t / \ell_{t-1}. \end{aligned}$$

As soon as the value of ℓ_{t-1} gets close to zero, the sample path becomes very unstable. To see that this problem is general in nature, consider the trend equation at time $t = 2$:

$$b_2 = b_1 + \beta \varepsilon_2 / \ell_1 = b_0 + \beta \left(\frac{\varepsilon_2}{\ell_1} + \frac{\varepsilon_1}{\ell_0} \right) = b_0 + \beta \left(\frac{\varepsilon_2}{\ell_0 b_0 + \alpha \varepsilon_1} + \frac{\varepsilon_1}{\ell_0} \right).$$

If ε_t has a Gaussian distribution, the first term in the brackets is a ratio of two Gaussian variables. When $\ell_0 b_0 = 0$ this term has a Cauchy distribution. In general, for all other values of $\ell_0 b_0$, the distribution is not Cauchy but it still has an infinite variance and undefined expectation (see Stuart and Ord, 1994, pp.400,421). Indeed, these problems arise whenever the level of the series has positive density over an open interval that includes zero. These problems with the trend equation will propagate into the observation equation at time $t = 3$. Similar problems arise with other distributions in Class X.

For ETS models (A,M,N), (A,M,A), (A,M_d,N), (A,M_d,A), (A,M,M), (A,M_d,M), (M,M,A) and (M,M_d,A):

- $V(y_{n+h} \mid \boldsymbol{x}_n) = \infty$ for $h \geq 3$;
- $E(y_{n+h} \mid \boldsymbol{x}_n)$ is undefined for $h \geq 3$.

For ETS models (A,N,M), (A,A,M) and (A,A_d,M):

- $V(y_{n+h} \mid \boldsymbol{x}_n) = \infty$ for $h \geq m+2$;
- $E(y_{n+h} \mid \boldsymbol{x}_n)$ is undefined for $h \geq m+2$.

Essentially, for any model with a Gaussian error process, the first passage time properties will eventually lead to negative values for the series unless there is a strong upward trend. In order to maintain the strictly positive nature of the model, the error process cannot be specified as Gaussian. A Gaussian approximation may work as the basis for computing point forecasts and short-term prediction intervals and, indeed, this method has been widely used over the years.

However, such choices cannot lead to exact distributional results.

To find a possible solution, consider the same simple model ETS(A,M,N). In order for the process to remain strictly positive, we require:

$$\ell_{t-1}b_{t-1} + \varepsilon_t > 0.$$

This condition requires that the distribution of

$$\varepsilon_t^* = 1 + \frac{\varepsilon_t}{\ell_{t-1}b_{t-1}}$$

should be defined on the positive line; that is, $\varepsilon_t^* \in (0, \infty)$. From a practical perspective, a long series may be needed before the positivity condition is violated; the first passage time depends strongly on the parameters.

2.2 The convergence to zero problem

Models with only multiplicative components may appear to be the natural choice for positive data. However, Figure 1 shows three realizations of the ETS(M,N,N) model using the Gaussian distribution (truncated to be positive), all showing a tendency to decay towards zero. The reason for this behavior is discussed in Section 3.1. Again, it is a relatively long-run behavior, and so does not have an immediate impact on short-term forecasting. But for simulations and long-term forecasting, this behavior needs to be understood.

2.3 Non-constant innovations variance

If the error ε_t is to have mean zero and the sample space is restricted to the positive real line, then the variance cannot be constant. This is easily seen for the ETS(M,N,N) model by considering the possible values of ε_t when ℓ_t is close to zero. Further, if the process approaches zero, the mean of a truncated distribution becomes more strongly positive, which may cause an uptick in the series.

Based upon these findings, it would appear that we should consider models with non-negative error structures; we proceed to examine such models in the next section.

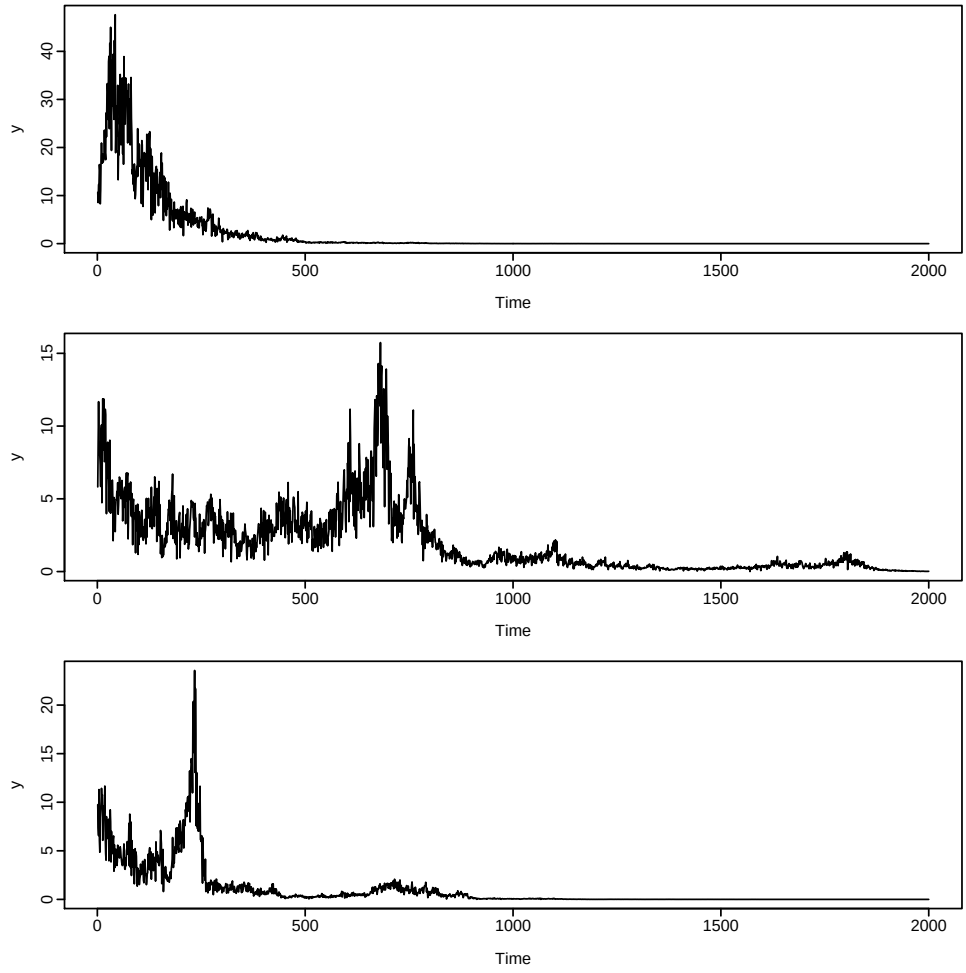

Figure 1: *ETS(M,N,N) simulation:* $\ell_0 = 10, \alpha = 0.3$ *and* $\sigma = 0.3$.

3 Multiplicative error models

In the previous section, we concluded that only models with a multiplicative error structure should be considered for strictly positive data. In this section we show that even in these circumstances, the models may fail to perform satisfactorily.

By way of illustration, we consider the multiplicative simple exponential smoothing model or ETS(M,N,N), as given below:

$$y_t = \ell_{t-1}(1 + \varepsilon_t) \tag{3a}$$

$$\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t), \tag{3b}$$

where ε_t denotes a white noise series with variance σ^2, such that $\varepsilon_t \geq -1$ and $0 < \alpha < 1$ (to ensure the data remain positive). Usually we require ε_t to have mean zero, although later

we will consider more general specifications. Hyndman et al. (2002) consider the model with $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$.

It will be convenient to write the model as

$$y_t = \ell_{t-1}\delta_t \tag{4a}$$

$$\ell_t = \ell_{t-1}(1 + \alpha\delta_t - \alpha). \tag{4b}$$

where $\delta_t = 1 + \varepsilon_t$ are iid with mean 1 and variance σ^2, and defined on the positive half-line. A truncated Gaussian distribution (see Stuart and Ord, 1994) could be used to ensure $\delta_t \geq 0$. When σ^2 is very small, the truncation is almost never needed. Other distributions of interest for δ_t are the lognormal and gamma distributions.

3.1 Kakutani's theorem

We can write the local level state equation of model (3) as

$$\ell_t = \ell_0(1+\alpha\varepsilon_1)(1+\alpha\varepsilon_2)\cdots(1+\alpha\varepsilon_t) = \ell_0 \prod_{j=1}^{t}(1+\alpha\varepsilon_j) = \ell_0 U_t, \tag{5}$$

where $U_t = U_{t-1}(1+\alpha\varepsilon_t)$ and $U_0 = 1$. Therefore U_t is a non-negative product martingale, since $\mathrm{E}(U_{t+1}|U_t) = U_t$.

Kakutani's theorem for product martingales (see Williams, 1991, p.144) may be stated as follows.

Theorem. Let $X_1, X_2, \cdots, X_n$ be positive independent random variables each with mean 1 and let $a_i = E\sqrt{X_i}$. Then for $U_n = \prod_{j=1}^{n} X_j$,

$$U_\infty > 0 \text{ almost surely if } \lim_{n\to\infty} \prod_{i=1}^{n} a_i > 0$$

$$U_\infty = 0 \text{ almost surely if } \lim_{n\to\infty} \prod_{i=1}^{n} a_i = 0.$$

Note that $a_i \geq 0$ and Jensen's inequality (see Shiryaev, 1996, p.192) gives $a_i \leq 1$. Further, provided the distributions of the X_i are not degenerate, $a_i < 1$. Thus, we may apply Kakutani's theorem to equation (5). That is, sample paths for ETS(M,N,N) models with the stated

properties tend to converge stochastically to zero. This is true regardless of the distribution of $1 + \alpha\varepsilon_t$, provided it has mean one and is non-degenerate. Kakutani's theorem is readily extended to other multiplicative error models under similar conditions.

3.2 An alternative approach

Our results so far indicate that the use of non-Gaussian distributions alone does not resolve the problem when we consider long-term forecasting. In order to make progress, we must be willing to relax one or more of the underlying assumptions that were made earlier. The result given by Kakutani's Theorem provides the essential insight. If we are to overcome the tendency to converge to zero, we must allow $\mathrm{E}\sqrt{X_i}$ to take on values equal to or greater than one.

Now let δ_t have mean close to but not necessary equal to one. For example, consider a modified ETS(M,N,N) model, which we write as METS(M,N,N;LN) to indicate both the modified form and the dependence on the lognormal distribution:

$$y_t = \ell_{t-1}\delta_t \tag{6a}$$

$$\ell_t = \ell_{t-1}\delta_t^{\alpha}, \tag{6b}$$

where δ_t is a positive random variable. This form of multiplicative model is chosen primarily for its convenience as it enables us to obtain exact sampling results when we assume that δ_t follows a lognormal distribution. This model also ensures a positive-valued process for all $0 < \alpha < 2$. The model may or may not be an improvement over existing choices, a question we explore in Section 6.3, but its qualitative behavior is similar and it is more easily explored analytically.

Using a log-transformation, (6) can be written as

$$y_t^* = \ell_{t-1}^* + \delta_t^* \tag{7a}$$

$$\ell_t^* = \ell_{t-1}^* + \alpha\delta_t^*, \tag{7b}$$

where $y_t^* = \log(y_t)$, $\ell_t^* = \log(\ell_t)$ and $\delta_t^* = \log(\delta_t)$. Thus the log-transformed model in (7) is identical to the simple exponential smoothing model ETS(A,N,N).

4 Distributional results

We now proceed to develop some distributional results for each of the models (3) and (6). If we denote the mean and variance of $\delta_t = 1 + \varepsilon_t$ by M and V respectively, and $\mathrm{E}(\delta_t^k) = M_k$, then the means and variances of the h-step-ahead prediction distributions may be written as:

Model (3)

$$\mathrm{E}(y_{n+h|n}) = E_{1A} = \ell_n M(1-\alpha+\alpha M)^{h-1} \tag{8a}$$

$$\mathrm{E}(y^2_{n+h|n}) = E_{2A} = \ell_n^2(M^2+V)\left[(1-\alpha+\alpha M)^2+\alpha^2 V\right]^{h-1} \tag{8b}$$

$$\mathrm{V}(y_{n+h|n}) = E_{2A} - E_{1A}^2. \tag{8c}$$

Model (6)

$$\mathrm{E}(y_{n+h|n}) = E_{1M} = \ell_n M M_\alpha^{h-1} \tag{9a}$$

$$\mathrm{E}(y^2_{n+h|n}) = E_{2M} = \ell_n^2(M^2+V)M_{2\alpha}^{h-1} \tag{9b}$$

$$\mathrm{V}(y_{n+h|n}) = E_{2M} - E_{1M}^2. \tag{9c}$$

Here we consider the lognormal distribution; similar results are observed if we use the gamma distribution in place of the lognormal distribution (for detail, see chapter 15 of Hyndman et al., 2008).

4.1 The lognormal distribution

If δ_t^* in (7) is Gaussian with mean μ and variance ω, or $\delta_t^* \sim \mathrm{N}(\mu,\omega)$, we may denote the lognormal assumption by $\delta_t \sim \log\mathrm{N}(\mu,\omega)$. Standard results for the lognormal distribution (see Stuart and Ord, 1994, pp.241–243) yield:

$$\mathrm{E}(\delta_t^k) = \exp(k\mu + k^2\omega/2), \quad \text{for any } k \tag{10a}$$

$$\mathrm{E}(\delta_t) = \exp(\mu+\omega/2) = E_1 \tag{10b}$$

$$\mathrm{V}(\delta_t) = E_1^2[\exp(\omega)-1] \tag{10c}$$

$$\text{and} \qquad \mathrm{E}(\delta_t^{\alpha/2}) = \exp(\alpha\mu/2 + \alpha^2\omega/8). \tag{10d}$$

From Equation (10d) we can see that the expectation of $\delta_t^{\alpha/2}$ will exceed 1 provided $\mu + \alpha\omega/4 > 0$.

If we now consider forecasting h periods ahead, we may set the forecast origin to $t = 0$ without loss of generality to simplify the notation. Then the prediction distribution for $y_h = \ell_0 z_h$ in model (7) is lognormal with $z_h \sim \log N(\mu_h, \omega_h)$, where

$$\mu_h = \mu(1 + (h-1)\alpha) \tag{11a}$$

$$\omega_h = \omega(1 + (h-1)\alpha^2) \tag{11b}$$

$$\mathrm{E}(y_h) = \ell_0 \exp[\mu_h + \omega_h/2] = E_h \tag{11c}$$

$$\text{and} \qquad \mathrm{V}(y_h) = E_h^2[\exp(\omega_h) - 1]. \tag{11d}$$

The distributional result is exact, so that we can explore the behavior of the prediction distribution for long lead-times with the help of Kakutani's Theorem. The possible outcomes for different values of the parameters are summarized in Table 2. The prediction distributions become increasingly skewed as h increases; when $\mathrm{E}(\delta_t^{\alpha/2}) < 1$ and $\mathrm{E}(\delta_t^{\alpha}) \leq 1$, $\Pr(y_h > 0) \downarrow 0$.

It is reasonable to point out that we have added an additional parameter in taking $\mu \neq 0$. However, setting $\mu = 0$ implies a stable median but a declining mean, whereas other choices produce other patterns of behavior. If an additional parameter is to be avoided, it seems equally reasonable to argue for a stable mean and to set $\mu = -\alpha\omega/2$. Similar issues concerning the trend arise in purely additive models, but they do not affect the shape of the predictive distribution in the way that multiplicative elements do.

Individual runs for some parameter combinations are shown in Figure 2. In accordance with Table 2, we observe the drift towards zero when $\mathrm{E}(\delta_t^{\alpha/2}) < 1$ and $\mathrm{E}(\delta_t^{\alpha}) \leq 1$. The reverse is true when $\mu > 0$. Further, the plots show that when the parameter values are close to the boundary conditions, we may need a long series in order to observe the limiting properties. However, we should recall from Figure 1 and the related discussion that different sample realizations may vary considerably.

The sampling distribution for model (3) is not exact, but may be approximated by a lognormal distribution with mean and variance given by (8) using the expectations given in (10).

Range	$\mathrm{E}(\delta_t^{\alpha})$	$\mathrm{E}(\delta_t^{\alpha/2})$	$\mathrm{E}(y_h)$	$\mathrm{V}(y_h)$
$\mu+\alpha\omega<0$	<1	<1	Decreasing	Decreasing
$\mu+\alpha\omega=0$	<1	<1	Decreasing	Finite
$-\alpha\omega<\mu<-\alpha\omega/2$	<1	<1	Decreasing	Increasing
$\mu+\alpha\omega/2=0$	$=1$	<1	Finite	Increasing
$-\alpha\omega/2<\mu<-\alpha\omega/4$	>1	<1	Increasing	Increasing
$\mu+\alpha\omega/4=0$	>1	$=1$	Increasing	Increasing
$\mu+\alpha\omega/4>0$	>1	>1	Increasing	Increasing

Table 2: *Long-term behavior of the prediction distribution for the METS(M,N,N;LN) model, with* $0<\alpha<1$. *The entry 'Finite' means that the term approaches a finite limit.*

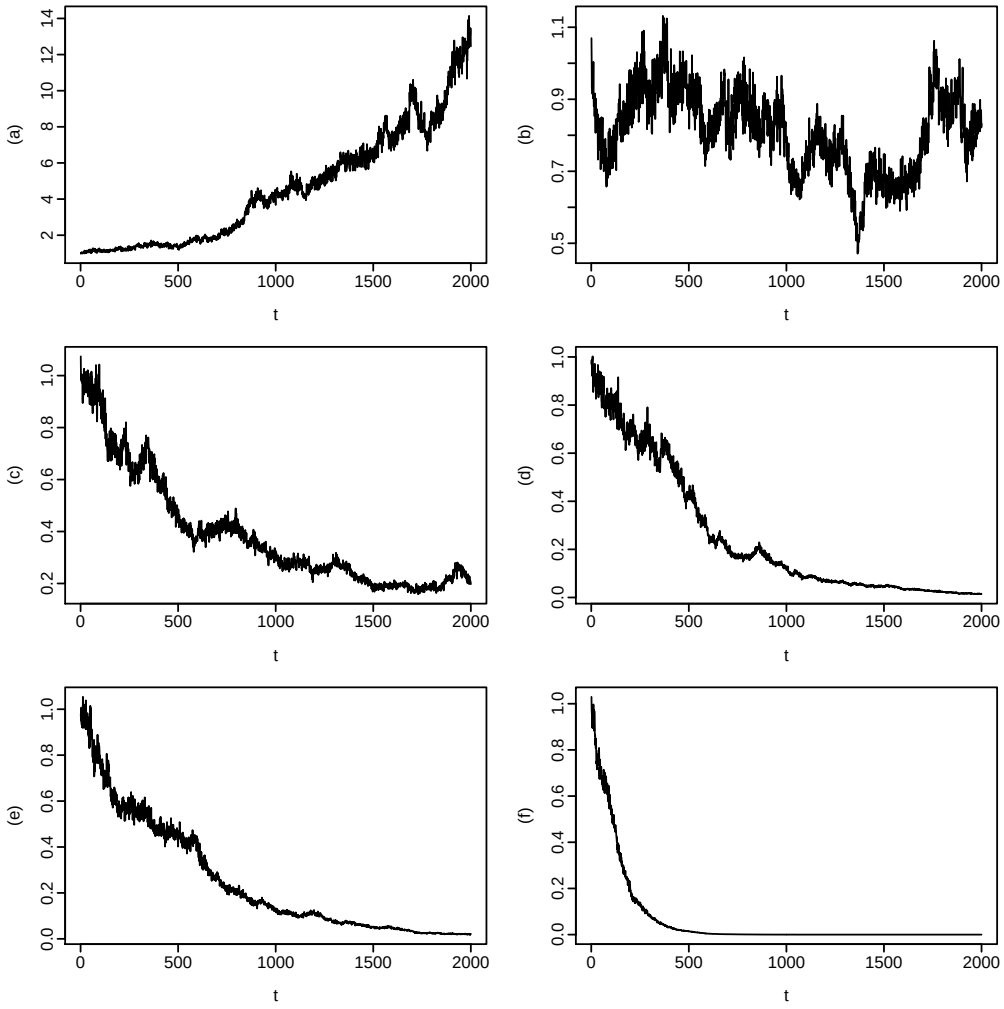

Figure 2: *Simulated data from the model METS(M,N,N;LN) with lognormal errors* $\delta_t \sim \log N(\mu,\omega)$: *(a)* $\mu=\alpha\omega/4$; *(b)* $\mu=0$; *(c)* $\mu=-\alpha\omega/4$; *(d)* $\mu=-3\alpha\omega/8$; *(e)* $\mu=-\alpha\omega/2$; *and (f)* $\mu=-3\alpha\omega/4$; *where* $\ell_0=1$, $\omega^{0.5}=\sigma=0.05$ *and* $\alpha=0.3$.

5 Implications for statistical inference

We now consider the implications of these results for inference. There are three elements to consider: parameter estimation based upon the likelihood function, prediction distributions for a small to moderate number of steps ahead, and the simulation of (potentially) long series.

5.1 The approximate likelihood

Once the error distribution is specified, we may examine the form of the distribution to see how close the approximation is to the true version. It is well-known that the lognormal density function approaches that of the Gaussian distribution as $\omega \to 0$; see Stuart and Ord (1994, p.242) for a graphical representation of this limiting relationship. However, our question is somewhat different in that we are concerned with differences in the maximum likelihood estimates, not the density functions. In order to examine this question, we may compare the estimates obtained by:

(a) applying the Gaussian ML estimators to lognormal data;
(b) evaluating the (correct) estimates using the lognormal likelihood function and then transforming to the mean and variance of the original error process.

In analytical terms, it is straightforward to show that the two approaches produce similar results as $\omega \to 0$; the question is: how good is the first form as an approximation to the second? The value of the lognormal parameter μ does not affect the relative bias or variability of the approximate estimates, so we may focus exclusively upon the effect that the value of $\sigma = \omega^{0.5}$ has upon the approximation. We carried out a small simulation study using $N = 100$ replicates for samples of size $n = 25$ with σ set equal to 0.05, 0.10 and 0.20. Values greater than 0.20 are most unlikely in practice in the present context. The results are summarized in the following table, which examines the ratios of the two estimates for each of the mean and standard deviation of the error. The average bias is measured in percentage terms; the bias for the mean of the error is negligible (less than 0.1% in all cases) and so is omitted from the table. The standard deviations of the percentage biases were also computed across the 100 replicates. Again, those for the mean are very small (less than 0.1%) and are omitted. The figures for the variance of the error are reported in the table and it can be seen that they are of a reasonable magnitude, even for $\sigma = 0.2$. The variances of the estimates themselves are almost equal, indicating that the loss in efficiency is very slight in this region of the parameter space.

		$\alpha = 0.5$		$\alpha = 0.8$	
	h	γ_1	γ_2	γ_1	γ_2
$\sigma = 0.05$	1	0.15	0.04	0.15	0.04
	5	0.21	0.08	0.28	0.14
	10	0.27	0.13	0.39	0.28
$\sigma = 0.10$	1	0.30	0.16	0.30	0.16
	5	0.43	0.33	0.58	0.60
	10	0.55	0.55	0.81	1.19

Table 3: *Standardized skewness and kurtosis coefficients for predictive distributions for the METS(M,N,N) model with lognormal errors.*

σ	0.05	0.10	0.20
Percent bias in variance	0.05	0.32	1.54
SD of percent bias in variance	1.98	3.95	7.96

Clearly, much more extensive simulation studies could be run, but the benefits would be marginal. We can be reasonably confident that when the errors follow the lognormal distribution, the Gaussian likelihood function is a reasonable approximation for the region of the parameter space involved. In turn, since the one-step-ahead error distributions are close to the Gaussian form, the approximate one-step-ahead prediction distributions will also be reasonably close to the underlying forms in most cases.

5.2 Prediction distributions and simulations

We now consider the lognormal model given in (7) and examine the prediction distribution. It follows from (11) that the h-step-ahead prediction distribution is also lognormal, of the form

$$\text{logN}\left(\log(\ell_0) + \mu[(h-1)\alpha + 1]\ ,\ \omega[(h-1)\alpha^2 + 1]\right).$$

As h increases, the divergence between the Gaussian and lognormal models becomes more and more pronounced as the prediction distribution becomes more skewed. In Table 3 we present numerical results for typical values of σ and α. Again, we have focussed upon the modified METS(M,N,N;LN) scheme, but qualitatively similar results will apply more broadly.

We use the standard measures of skewness γ_1 and kurtosis γ_2 based upon the third and fourth moments; $\gamma_1 = \gamma_2 = 0$ for a Gaussian distribution. As expected, the distributions become more skewed and heavy-tailed as the forecasting horizon increases and/or the value of α increases.

For purely multiplicative (Class M) models with lognormal errors, the analytical expressions for point forecasts and prediction intervals for model ETS(A,∗,∗) may be used for the log-transformed ETS(M,∗,∗) model. Otherwise, for Class M models, the best approach is to use simulations based upon a careful specification of the underlying distribution.

In order to apply the analytical approach, we must be sure that the underlying model will produce strictly positive values in any realization of the series. The following example illustrates how we may check whether this requirement is met.

5.3 ETS(M,M,M) model

The model equations for the ETS(M,M,M) model are:

$$\begin{aligned} y_t &= \ell_{t-1} b_{t-1} s_{t-m}(1+\varepsilon_t) \\ \ell_t &= \ell_{t-1} b_{t-1}(1+\alpha\varepsilon_t) \\ b_t &= b_{t-1}(1+\beta\varepsilon_t) \\ s_t &= s_{t-m}(1+\gamma\varepsilon_t). \end{aligned}$$

We will assume that $t = km$ for convenience, to avoid the notational complexities of partial seasonal cycles. Then repeated substitutions result in the reduced form (taking t mod $m = p$):

$$y_t = \ell_0 b_0^t s_{-m+p}(1+\varepsilon_t)\prod_{j=1}^{t-1}\left[(1+\alpha\varepsilon_j)(1+\beta\varepsilon_j)^{t-j}\right]\prod_{i=1}^{k-1}(1+\gamma\varepsilon_i).$$

Inspection of the reduced form shows that the process will remain strictly positive provided all the starting values for the state variables are positive and $\varepsilon_t > \max(-1, -1/\alpha, -1/\beta, -1/\gamma)$ for all t. The most natural way to ensure that this condition is satisfied is to require that $\max(\alpha, \beta, \gamma) < 1$ and that $\varepsilon_t > -1$. Similar conditions apply for the ETS(M,M$_\mathrm{d}$,M) model.

In general, when the model is in Class M, conditions such as those just given will suffice to maintain a positive path for the process. However, when at least one component is additive (as for the Class A models), an unrestricted sample path may eventually hit negative values. When the series has an overall upward trend, the risk is greatly reduced, but cannot be eliminated as a theoretical possibility.

Since the nonlinear models are applied to series that are non-negative, models with an additive component cannot be formally correct. Nevertheless, they have proved extremely useful

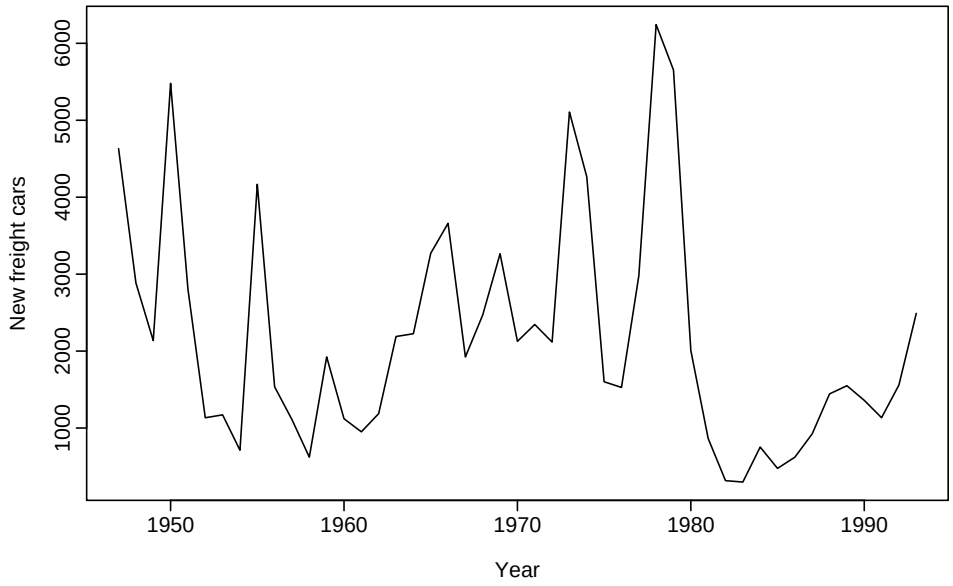

Figure 3: *U.S. freight car shipments, 1947–1993.*

and the implementation problems are minor when considering parameter estimation or predictive statements for relatively short horizons. We only run into difficulties for long horizons or when we are simulating a long series. We may avoid problems either by dropping any realization that goes negative, or by using the modified series $y_t^* = \max(\Delta, y_t)$ for some small $\Delta > 0$. Neither solution is perfect, and should only be applied in circumstances where violations are infrequent. If negative values occur frequently, this is a sign that the proposed model is inappropriate for the specified set of parameters and start values.

6 Empirical comparisons

We will now illustrate some of the points discussed earlier by examining an annual time series on the number of new freight cars shipped in the U.S.A. over the period 1947–1993. (This series is available as Number N0193 in the M3 Competition data.) The data are plotted in Figure 3. A visual inspection of the series suggests a changing local level and the AIC comparison of different local models suggests the ETS(M,N,N) model as the best choice.

6.1 Point forecasts and estimation

We now compare the performance of the Gaussian-based ETS(M,N,N) and ETS(A,N,N) models to those of the lognormal based ETS(M,N,N) models, using fitting samples of 28, 34 and 40

	ETS(A,N,N)	*ETS(M,N,N)*	*L1*	*L2*
$n = 28$				
α	0.32	0.01	0.43	0.40
MAE	1953	1668	2034	2015
MAPE	74	59	79	72
mean	*	*	0.975	1.165
$n = 34$				
α	0.21	0.00	0.38	0.29
MAE	1779	2899	1271	868
MAPE	401	632	286	195
mean	*	*	0.959	1.178
$n = 40$				
α	0.42	0.22	1.01	0.73
MAE	329	243	294	331
MAPE	24	19	23	25
mean	*	*	1.205	1.202

Table 4: *Summary statistics for the U.S. freight cars series: L1=lognormal model (3); L2=lognormal model (6).*

observations and a (non-overlapping) hold-out sample of the next 6 observations in each case. The models were fitted using conditional maximum likelihood.

The results are given in Table 4 and show the Forecast Mean Absolute Error (MAE) for the one-step-ahead errors for the hold-out sample in each case. Only very limited conclusions may be drawn from a single example, but a few points are worth noting. The means for the lognormal models however around 1, reflecting the uncertainty about whether or not the series is declining; otherwise their one-step-ahead performances appear to be similar. However, for longer horizons, the different values of the means imply quite different trajectories. Both models differ somewhat from the ETS(M,N,N) model, but show some similarities with the ETS(A,N,N) results.

These results raise more questions than they resolve, but support the general contention that estimation properties and short-term point forecasts are not seriously affected by the long-run behavior discussed earlier.

6.2 Prediction intervals

One of the principal reasons for the introduction of the lognormal models is the concern about prediction intervals. To illustrate how the positivity constraint affects these intervals, we provide some numerical examples in Table 5. As expected, the prediction intervals based

Distribution	Means			Lower PI			Upper PI		
h:	1	5	10	1	5	10	1	5	10
$\alpha = 0.3$									
Lognormal (3)	100	100	100	54	48	42	186	208	236
Lognormal (6)	100	96.1	91.4	52	44	37	175	182	189
ETS(A,N,N)	100	100	100	38	28	17	162	172	183
ETS(M,N,N)	100	100	100	38	27	14	162	172	186
$\alpha = 0.8$									
Lognormal (3)	100	100	100	54	29	15	186	351	657
Lognormal (6)	100	97.0	93.4	52	26	14	175	256	326
ETS(A,N,N)	100	100	100	38	−17	−61	162	217	261
ETS(M,N,N)	100	100	100	38	−25	−88	162	225	288

Table 5: *Prediction intervals based upon the lognormal distributions using models (3) and (6) with* $\ell_0 = 100$ *and* $V(\delta) = 0.1$.

upon the Gaussian distribution for ETS(A,N,N) and ETS(M,N,N) grow progressively more misleading as α becomes larger or the forecast horizon is extended. The results for models (3) and (6) are fairly similar, although the slightly longer upper tail of the lognormal becomes evident for model (3) at $h = 10$. Note that point forecasts for model (3) are constant since we set $\mathrm{E}(\delta_t) = 1$; this result would not hold otherwise.

6.3 Forecasting jewelry sales

In order to explore further the relative merits of formulations (3) and (6), we fitted these models to 314 series that describe weekly sales of costume jewelry items over the period week 5, 1998 to week 24, 2000. The data were provided by a leading company in that field. Products that were either launched or discontinued during that period were removed from the study. Most products had very high sales over the Christmas period so we partitioned the data as follows:

Estimation sample:	weeks 5–45, 1998 and weeks 2–20, 1999	$(n = 60)$;
Test sample:	weeks 21–45, 1999	$(n^* = 25)$.

The gap in the estimation sample did not cause any problems since the differences in levels before and after the Christmas period were minor; the random fluctuations were generally much larger than any level changes.

Three ETS(M,N,N) models were fitted to each series by maximum likelihood:

- Model 1: (3) assuming a Gaussian error distribution with mean 0;
- Model 2: (3) assuming a lognormal error distribution with median 1;

- Model 3: (6) assuming a lognormal error distribution with median 1.

We calculated the one-step-ahead forecasting errors for each series over the test samples and created summaries using the Mean Squared Error (MSE), the Mean Absolute Percentage Error (MAPE) and the Mean Absolute Scaled Error (MASE) introduced by Hyndman and Koehler (2006). The MASE is defined for a collection of N time series for which there are M potential models for forecasting. The number of observations for time series $y_t^{(j)}$, $j = 1\ldots,N$, is denoted by n_j. The MASE of model i, $i = 1,\ldots,M$, for time series $y_t^{(j)}$ is defined by

$$\text{MASE}(H;i,j) = \frac{1}{H}\sum_{h=1}^{H}\frac{|y_{n_j+h}^{(j)} - \hat{y}_{i,n_j}^{(j)}(h)|}{\text{MAE}_j} \tag{12}$$

where $\text{MAE}_j = (1/(n_j-1))\sum_{i=2}^{n_j}|y_i^{(j)} - y_{i-1}^{(j)}|$, and $\hat{y}_{i,n_j}^{(j)}(h)$ is the h-period-ahead ($h = 1,\ldots,H$) forecast when model i is used for the jth time series.

Although the results sometimes differ for individual series, the overall picture is consistent across the three measures and only the MASE results are reported here. Plots of pairwise comparisons of MASE values for the different models are given in Figure 4. Further study is clearly necessary, but the limited results suggest that model 1 is inferior to the other two. Of the two lognormal models (6) appears to be marginally preferable.

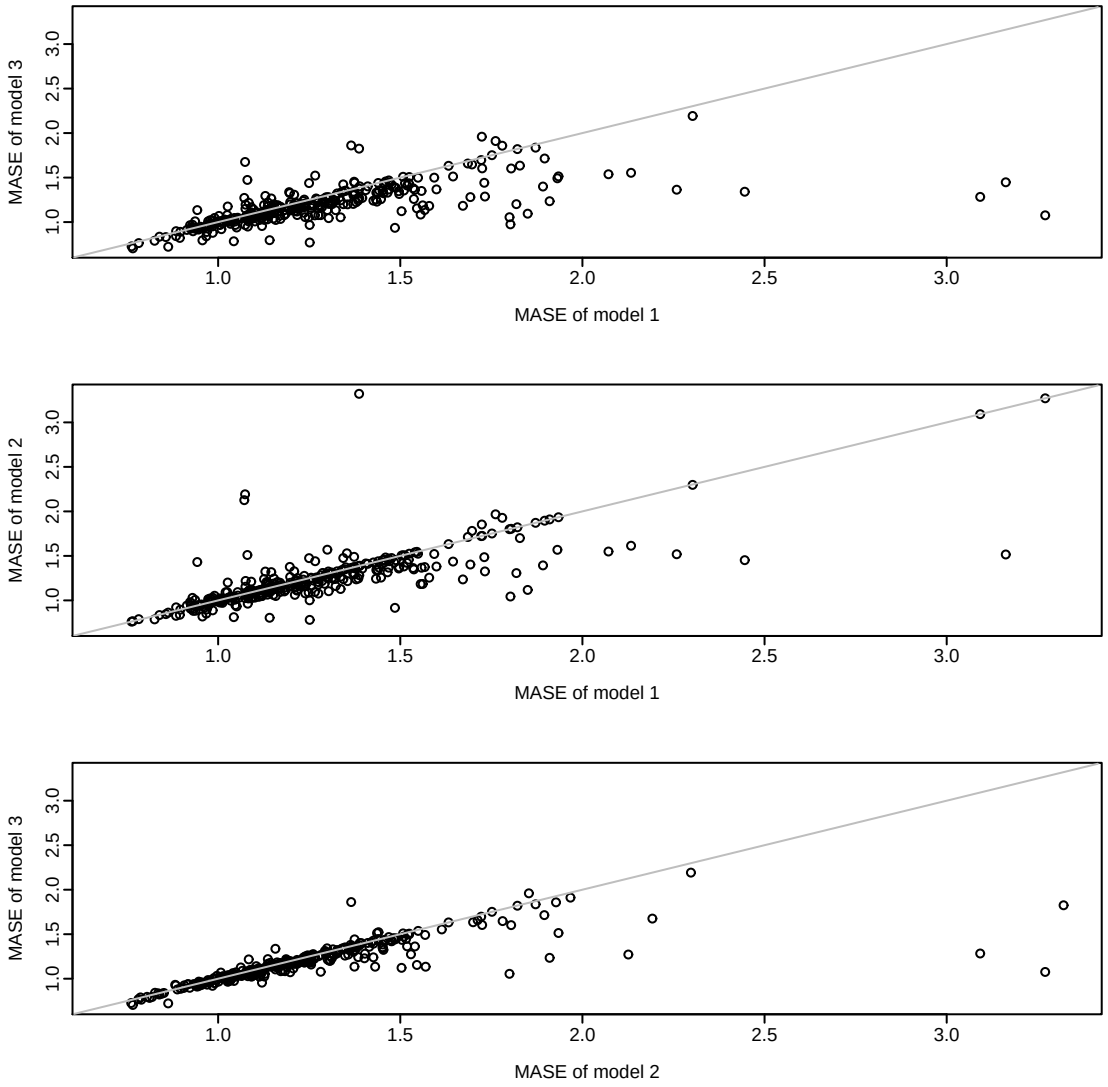

Figure 4: *MASE comparison of the three ETS(M,N,N) models. On the diagonal line the two models have the same MASE.*

7 Conclusions

We have undertaken an exploration of exponential smoothing models defined on the positive half-line. One of the attractions of the innovations approach is that it enables an exact specification of nonlinear models that, in turn, can lead to explicit results for the prediction distribution. Nevertheless, we have uncovered certain properties that make the use of such models more intricate than conventional practice might suggest. We now summarize our findings to date, while we recognize that this is an area where further research is needed.

a. Parameter estimation using the Gaussian likelihood appears to be a viable option for the ranges of the parameters that we are typically likely to encounter.

b. The point forecasts generated from such fitted models appear to be satisfactory, at least for short-term forecasting.

c. When we turn to prediction intervals, the Gaussian approximation becomes progressively less reasonable as h increases.

d. For prediction intervals and simulations, there is no substitute for an appropriate non-Gaussian model. At this stage, we are inclined to recommend the lognormal on the grounds of operational simplicity.

e. Since only the purely multiplicative models have a sample space restricted to the positive half-line, model simulations with other schemes may need to provide a floor below which the series cannot go. Clearly, this is an area where the investigator must proceed with caution.

References

Grunwald, G. K., K. Hamza and R. J. Hyndman (1997) Some properties and generalizations of non-negative Bayesian time series models, *Journal of the Royal Statistical Society, Series B*, **59**, 615–626.

Grunwald, G. K., A. E. Raftery and P. Guttorp (1993) Time series of continuous proportions, *Journal of the Royal Statistical Society, Series B*, **55**, 103–116.

Harvey, A. C. and C. Fernandes (1989) Time series models for count or qualitative observations, *Journal of Business & Economic Statistics*, **7**(4), 407–422.

Hyndman, R. J. and A. B. Koehler (2006) Another look at measures of forecast accuracy, *International Journal of Forecasting*, **22**, 679–688.

Hyndman, R. J., A. B. Koehler, J. K. Ord and R. D. Snyder (2005) Prediction intervals for exponential smoothing using two new classes of state space models, *Journal of Forecasting*, **24**, 17–37.

Hyndman, R. J., A. B. Koehler, J. K. Ord and R. D. Snyder (2008) *Forecasting with exponential smoothing: the state space approach*, Springer-Verlag, Berlin.

Hyndman, R. J., A. B. Koehler, R. D. Snyder and S. Grose (2002) A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, **18**(3), 439–454.

Ord, J. K., A. B. Koehler and R. D. Snyder (1997) Estimation and prediction for a class of dynamic nonlinear statistical models, *Journal of the American Statistical Association*, **92**, 1621–1629.

Shiryaev, A. N. (1996) *Probability*, Springer-Verlag, New York.

Stuart, A. and J. K. Ord (1994) *Kendall's advanced theory of statistics. vol. 1: Distribution theory*, Hodder Arnold, London, 6th ed.

Taylor, J. W. (2003) Exponential smoothing with a damped multiplicative trend, *International Journal of Forecasting*, **19**, 715–725.

West, M., P. J. Harrison and H. S. Migon (1985) Dynamic generalized linear models and Bayesian forecasting (with discussion), *Journal of the American Statistical Association*, **80**, 73–96.

Williams, D. (1991) *Probability with martingales*, Cambridge University Press, Cambridge.

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