# Forecasting using robust exponential smoothing with damped trend and seasonal components

Crevits R, Croux C.

KBI_1714

# Forecasting using Robust Exponential Smoothing with Damped Trend and Seasonal Components

Ruben Crevits and Christophe Croux

*Faculty of Economics and Business, KU Leuven*

**Abstract.** We provide a framework for robust exponential smoothing. For a class of exponential smoothing variants, we present a robust alternative. The class includes models with a damped trend and/or seasonal components. We provide robust forecasting equations, robust starting values, robust smoothing parameter estimation and a robust information criterion. The method is implemented in the R package `robets`, allowing for automatic forecasting. We compare the standard non-robust version with the robust alternative in a simulation study. Finally, the methodology is tested on data.

**Keywords.** Automatic Forecasting, Outliers, R package, Time series

# 1 Introduction

In time series analysis exponential smoothing methods are popular because they are easy to use and the forecasting procedure can be made automatic. Simple exponential smoothing, or sometimes called single exponential smoothing is the most basic method. It is a suitable method if the time series has no trend or seasonality, but a slowly varying mean. For a time series $y_1, \ldots, y_t$, the forecasts are

$$\begin{aligned} \hat{y}_{t+h|t} &= \ell_t \\ \ell_t &= \alpha y_t + (1-\alpha)\ell_{t-1} \end{aligned} \tag{1.1}$$

with $\hat{y}_{t+h|t}$ the $h$-step ahead forecast. The time series $\ell_t$ gives the 'level' of the series. The degree of smoothing is determined by the smoothing parameter $\alpha$, which is usually estimated by minimizing the sum of squared prediction errors.

For trending and seasonal time series there is the Holt-Winters method (Hyndman and Athanasopoulos, 2013). It is also referred to as double exponential smoothing or exponential smoothing with additive trend and seasonal component. It has additional parameters $\beta$ and $\gamma$ which determine the smoothing rate of the trend and the seasonal component. Pegels (1969) suggested a model with *multiplicative* trend and seasonal component. Gardner (1985) proposed exponential smoothing with *damped* trend. Taylor (2003) showed that exponential smoothing with a damped multiplicative trend is a very competitive forecasting method. Damping a trend has in particular an advantage for long forecasting horizons $h$. The forecast doesn't go to infinity as with the regular additive or multiplicative trend, but converges to a finite value. An extra parameter $\phi$ determines the rate at which this happens.

A disadvantage of exponential smoothing methods is that they are not outlier robust. An observation may have an unbounded influence on each subsequent forecast. The selection of the smoothing parameters is also affected by outliers, since these are estimated by minimizing a sum of squared forecasting errors. In the past there have been efforts to make exponential smoothing methods robust. Gelper et al. (2010) proposed a methodology for robust exponential smoothing. They also provided a way to estimate the smoothing parameters robustly. Cipra and Hanzak (2011) have an alternative robust exponential smoothing scheme; Croux et al. (2008) supplied a numerically stable algorithm of their earlier proposal. A multivariate version of the robust simple exponential smoothing recursions is given in Croux et al. (2010).

In this paper we extend the existing robust methods to a more general class exponential smoothing variants, including (damped) additive trends and additive or multiplicative seasonal components. The outline of the paper is as follows. In Section 2 we review the class of exponential smoothing methods. In Section 3 we propose the robust method. For each variant we robustify the recursions, smoothing parameter estimation and choice of the starting values. In Section 4 we present the R package `robets`, which is an implementation of the method in Section 3. In Section 5 the robust method is tested in a simulation study. In the last section we evaluate the forecasting performance of the method on the time series of the M3-competition of Makridakis and Hibon (2000).

# 2 Exponential smoothing methods

We use the taxonomy of Hyndman et al. (2005) to describe the class of fifteen exponential smoothing models. Each model can be described by three letters:

E, underlying error model: A (additive) or M (multiplicative),

T, type of trend: N (none), A (additive) or $\mathrm{A_d}$ (damped) and

S, type of seasonal: N (none), A (additive) or M (multiplicative).

For example: MAN is exponential smoothing with additive trend without seasonal component and a multiplicative underlying model. All considered combinations are shown in Table 1. The combinations ANM, AAM and $\mathrm{AA_dM}$ are omitted since the corresponding prediction intervals are not derived in Hyndman et al. (2005). In the next subsections we describe the considered models in more detail.

Table 1: The fifteen considered exponential smoothing methods.

| | Seasonal | | |
|---|---|---|---|
| Trend | N (none) | A (additive) | M (multiplicative) |
| N (none) | ANN/MNN | ANA/MNA | MNM |
| A (additive) | AAN/MAN | AAA/MAA | MAM |
| $\mathrm{A_d}$ (damped) | $\mathrm{AA_dN/MA_dN}$ | $\mathrm{AA_dA/MA_dA}$ | $\mathrm{MA_dM}$ |
| forecasting formula | (2.1) | (2.2) | (2.3) |

## 2.1 Trend (T)

The forecasting equations of simple exponential smoothing are given in equation (1.1). However some time series have trending behavior. For such series a full trend (A) or a damped trend ($\mathrm{A_d}$) might be useful. Suppose we have a time series $y_t$, which is observed at $t = 1, \ldots, T$. The forecasts of $\mathrm{AA_dN}$ and $\mathrm{MA_dN}$ can be computed with a recursive scheme:

$$\begin{aligned} \hat{y}_{t+h|t} &= \ell_t + \sum_{j=1}^{h} \phi^j b_t \\ \ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\ b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1} \end{aligned} \tag{2.1}$$

with $\hat{y}_{t+h|t}$ the forecast at horizon $h$ made at time $t$. By setting $\phi = 0$, we have the forecasting equations of ANN/MNN or simple exponential smoothing without trend. Setting $\phi = 1$ gives the equations of AAN/MAN or exponential smoothing with a full additive trend. The smoothing parameter $\alpha$ determines the rate at which the level $\ell_t$ is allowed to change. If it is close to zero the level stays almost constant and if it is one the level follows the observations perfectly. The parameter $\beta$ determines the rate at which the trend may change. The extra parameter $\phi$ is related with how fast the local trend is damped. Indeed, the longterm forecast converges to $\ell_t + \frac{\phi}{1-\phi} b_t$ if $h \to \infty$. The parameters $\alpha$, $\beta$ and $\phi$ take values between zero and one.

While the forecasts corresponding to the additive or multiplicative model are the same, the underlying data generating process is different, resulting in different confidence intervals (see Subsection 2.3) and parameter estimates (see Subsection 3.3).

## 2.2 Seasonal component (S)

It is also possible to model slowly changing seasonality effects. In the models ANA/MNA, AAA/MAA and $\mathrm{AA_dA/MA_dA}$ we add a seasonal component:

$$
\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + \sum_{j=1}^{h} \phi^j b_t + s_{t-m+h_m^+} \\
\ell_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1} \\
s_t &= \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m},
\end{aligned}
\tag{2.2}
$$

with $h_m^+ = \lfloor (h-1) \bmod m \rfloor + 1$. The number of seasons is $m$. Typically, for monthly data, $m = 12$. The seasonal smoothing parameter is $\gamma$. If the value is high, the seasonal components will quickly follow changes in seasonality.

It turns out that for many time series a multiplicative seasonal component is more suitable. The forecasting equations for methods MNM, MAM and $\mathrm{MA_dM}$ are given below:

$$
\begin{aligned}
\hat{y}_{t+h|t} &= (\ell_t + \sum_{j=1}^{h} \phi^j b_t) s_{t-m+h_m^+} \\
\ell_t &= \alpha \tfrac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + \phi b_{t-1}) \\
b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)\phi b_{t-1} \\
s_t &= \gamma \tfrac{y_t}{\ell_{t-1} + \phi b_{t-1}} + (1-\gamma)s_{t-m}.
\end{aligned}
\tag{2.3}
$$

## 2.3 Underlying models (E)

Hyndman et al. (2002) present the underlying models for each exponential smoothing variant. Making a model assumption is necessary to obtain prediction intervals. We will declare an observation $y_{t+1}$ as an outlier if it does not belong to the prediction interval constructed around $\hat{y}_{t+1|t}$. It is then also possible to set up a likelihood function to estimate smoothing parameters. For simple exponential smoothing as in (1.1), the additive error model is

$$\begin{aligned} y_t &= \ell_{t-1} + \epsilon_t \\ \ell_t &= \ell_{t-1} + \alpha\epsilon_t \end{aligned} \tag{2.4}$$

and the multiplicative error model is

$$\begin{aligned} y_t &= \ell_{t-1}(1 + \epsilon_t) \\ \ell_t &= \ell_{t-1}(1 + \alpha\epsilon_t). \end{aligned} \tag{2.5}$$

It is possible to check that both underlying models have the same optimal point forecasts (Hyndman et al., 2002, Section 3). The single source of error is $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. For the multiplicative model the lower tail is truncated such that $1 + \epsilon_t$ remains positive. Because $\sigma^2$ is usually small in multiplicative models, this truncation is negligible. The model with multiplicative errors is used when the observations are strictly positive and when we expect that the error grows proportionally with the observation value. For a description of the models of other exponential smoothing methods, we refer to Hyndman et al. (2002).

The underlying models are needed for constructing prediction intervals. For the additive error models the prediction interval at forecast horizon $h = 1$ is

$$\left[\hat{y}_{t+1|t} - q\sigma, \hat{y}_{t+1|t} + q\sigma\right]$$

with $q = z_{1-\alpha/2}$ for a $100(1-\alpha)\%$ interval. Here $z_{1-\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal; for a 95% prediction interval $q \approx 2$. For the multiplicative error models the interval is

$$\left[\hat{y}_{t+1|t}(1 - q\sigma), \hat{y}_{t+1|t}(1 + q\sigma)\right].$$

Exponential smoothing models are a subclass of ARIMA models. The ANN, AAN, $\mathrm{AA_dN}$, ANA, AAA, and $\mathrm{AA_dA}$ models can be rewritten as ARIMA models, see Hyndman and Athanasopoulos (2013). It should be noted, however, that the parameterization of these models is different and more cumbersome. The ARIMA counterparts for the seasonal models require additional parameter restrictions. Exponential smoothing models are less complex than their equivalent ARIMA. Furthermore, the multiplicative models can not be represented as ARIMA models. For robust ARMA estimation we refer to Muler et al. (2009).

# 3 Robust exponential smoothing methods

We make a robust version of the exponential smoothing forecaster of Hyndman and Khandakar (2008). We robustify the forecasting equations, the estimation of the starting values, the estimation of the smoothing parameters and the information criterion.

## 3.1 Robust forecasting equations

We follow the procedure of Gelper et al. (2010, p. 288). For all considered exponential smoothing models we robustify the forecasting equations by replacing each observation $y_t$ with a cleaned version $y_t^*$. If the one-step ahead forecast error $y_t - \hat{y}^*_{t|t-1}$ exceeds $k$ times the scale of the errors, we consider the observation to be an outlier. The one-step ahead prediction $\hat{y}^*_{t|t-1}$ is based on the cleaned observations $y_1^*$, $y_2^*$, …, $y_{t-1}^*$. Our choice for $k$ is 3: if the one-step prediction error follows a normal distribution then only 0.03 % of the observations are falsely indicated as outliers. The cleaned observations are given by the formula:

$$y_t^* = \psi\left(\frac{y_t - \hat{y}^*_{t|t-1}}{\hat{\sigma}_t}\right)\hat{\sigma}_t + \hat{y}^*_{t|t-1} \tag{3.1}$$

with $\psi$ the Huber function

$$\psi(x) = \begin{cases} x & \text{if } |x| < k \\ \operatorname{sign}(x)k & \text{otherwise} \end{cases}$$

and with $\hat{\sigma}_t$ an estimate of the scale of the one-step ahead prediction error.

We model the scale as slowly varying and estimate it recursively in a robust way as in

Gelper et al. (2010),

$$\hat{\sigma}_t^2 = \lambda_\sigma \rho \left( \frac{y_t - \hat{y}^*_{t|t-1}}{\hat{\sigma}_{t-1}} \right) \hat{\sigma}^2_{t-1} + (1 - \lambda_\sigma) \hat{\sigma}^2_{t-1} \tag{3.2}$$

with $\rho$ the bounded biweight function

$$\rho_{\text{biweight}}(x) = \begin{cases} c_k(1 - (1 - (x/k)^2)^3) & \text{if } |x| < k \\ c_k & \text{otherwise} \end{cases} \tag{3.3}$$

with $k = 3$, $c_k = 4.12$ and with $\lambda_\sigma = 0.1$. This function is close to the quadratic function if the prediction error is small, but bounded for large errors. If we assume an underlying multiplicative model, the cleaned observations are given by

$$y_t^* = (1 + \psi \left( \frac{y_t - \hat{y}^*_{t|t-1}}{\hat{y}^*_{t|t-1} \hat{\sigma}_t} \right) \hat{\sigma}_t) \; \hat{y}^*_{t|t-1} \tag{3.4}$$

with $\hat{\sigma}_t$ the scale of the relative errors. This scale is updated as follows:

$$\hat{\sigma}_t^2 = \lambda_\sigma \rho \left( \frac{y_t - \hat{y}^*_{t|t-1}}{\hat{y}^*_{t|t-1} \hat{\sigma}_{t-1}} \right) \hat{\sigma}^2_{t-1} + (1 - \lambda_\sigma) \hat{\sigma}^2_{t-1}. \tag{3.5}$$

## 3.2 Robust starting values

The forecasting equations given in Section 3.1 are defined recursively and require starting values $\ell_0$, $b_0$, $s_{-m+1}$, $\ldots$, $s_0$. Although effects of the starting values decay exponentially, it still matters to select them in a robust way. The estimation of the smoothing parameters may be affected by non-robustly chosen starting values, in particular since exponential smoothing methods are often used for short time series.

The starting values are found by using a short startup period of observations $y_1, \ldots, y_S$. We take $S = 5m$ with $m$ the number of seasons. If there is no seasonal component, then $S = 10$. The standard non-robust way to select $\ell_0$ and $b_0$ is by regressing $y$ for $t = 1 \ldots S$, resulting in a least squares estimate of the intercept $\hat{\ell}_0$ and of the coefficient $\hat{b}_0$. We do a robust regression, and use the repeated median. Fried (2004) applied it to discover trends in short time series. The estimates are then

$$\hat{\ell}_0 = \text{med}_i(y_i - \hat{b}_0 i) \qquad \text{and} \qquad \hat{b}_0 = \text{med}_i \ \text{med}_{i \neq j} \frac{y_i - y_j}{i - j}$$

with $i, j = 1 \ldots S$.

The starting values for the seasonal components are typically found by taking the average difference from the regression line for each season. We take instead the median difference:

$$s_{q-m} = \text{med}\,(y_q - \hat{y}_q, y_{q+m} - \hat{y}_{q+m}, \ldots, y_{q+S-m} - \hat{y}_{q+S-m})$$

for $q = 1, \ldots, m$. If the seasonal component is multiplicative, the computation is slightly different:

$$s_{q-m} = \text{med}\left(\frac{y_q}{\hat{y}_q}, \frac{y_{q+m}}{\hat{y}_{q+m}}, \ldots, \frac{y_{q+S-m}}{\hat{y}_{q+S-m}}\right).$$

Finally, to start up the recursive equation (3.2) of the local scale estimate $\hat{\sigma}_t$, a starting value $\hat{\sigma}_0$ is needed. Let $\tilde{t} = (t \mod m) - m$. For a model with additive errors and additive seasonality,

$$\hat{\sigma}_0 = \text{MAD}(y_t - \hat{\ell}_0 - \hat{b}_0 t - s_{\tilde{t}}).$$

with MAD, the Median Absolute Deviation, defined as

$$\underset{t}{\text{MAD}}(e_t) = 1.4826 \ \underset{t}{\text{med}}|e_t|.$$

Similarly, for a model with additive errors and multiplicative seasonality we take $\hat{\sigma}_0 = \text{MAD}_t(y_t - (\hat{\ell}_0 - \hat{b}_0 t)/s_{\tilde{t}})$. For a model with multiplicative errors and additive seasonality, a good choice is

$$\hat{\sigma}_0 = \underset{t}{\text{MAD}}\left(\frac{y_t - \hat{\ell}_0 - \hat{b}_0 t - s_{\tilde{t}}}{\hat{\ell}_0 - \hat{b}_0 t - s_{\tilde{t}}}\right),$$

and for multiplicative seasonality

$$\hat{\sigma}_0 = \underset{t}{\mathrm{MAD}} \left( \frac{y_t - (\hat{\ell}_0 - \hat{b}_0 t)/s_{\tilde{t}}}{(\hat{\ell}_0 - \hat{b}_0 t)/s_{\tilde{t}}} \right).$$

## 3.3 Robust parameter estimation

The parameter vector to be optimized is

$$\boldsymbol{\theta} = (\alpha, \beta, \phi, \gamma).$$

Depending on the model being estimated, the parameters involving a (damped) trend ($\phi$, $\beta$) or a seasonal component ($\gamma$) are not included in $\boldsymbol{\theta}$. We propose a robust way to estimate the parameters, but first we review the non-robust estimators.

### 3.3.1 Maximum likelihood

We follow Ord et al. (1997) and Hyndman et al. (2002). If an additive error model is assumed the maximum likelihood estimate is

$$\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}\right) = \underset{\boldsymbol{\theta},\sigma}{\mathrm{argmax}} -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})\right)^2.$$

with $\hat{y}_{t|t-1}(\boldsymbol{\theta})$ the one-step ahead prediction using the parameter vector $\boldsymbol{\theta}$. The likelihood is maximal for

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})\right)^2,$$

so the parameters can simply be estimated by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}} -\frac{T}{2}\log\left(\frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})\right)^2\right), \tag{3.6}$$

hence minimizing the sum of squared one-step ahead prediction errors.

With a multiplicative error model, the maximum likelihood estimate is

$$\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}\right) = \underset{\boldsymbol{\theta},\sigma}{\operatorname{argmax}} -\frac{T}{2}\log\sigma^2 - \sum_{t=1}^{T}\log\left|\hat{y}_{t|t-1}(\boldsymbol{\theta})\right| - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(\frac{y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})}{\hat{y}_{t|t-1}(\boldsymbol{\theta})}\right)^2.$$

By setting the derivate to $\sigma$ equal to zero, we find

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})}{\hat{y}_{t|t-1}(\boldsymbol{\theta})}\right)^2,$$

yielding

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} -\frac{T}{2}\log\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{y_t - \hat{y}_{t|t-1}(\boldsymbol{\theta})}{\hat{y}_{t|t-1}(\boldsymbol{\theta})}\right)^2\right) - \sum_{t=1}^{T}\log\left|\hat{y}_{t|t-1}(\boldsymbol{\theta})\right|. \tag{3.7}$$

All of these estimators are not outlier robust.

### 3.3.2 Robust estimation

We propose to replace the sum of squares by a $\tau^2$ estimator in the likelihood functions. The $\tau^2$ is a robust estimator of scale introduced by Yohai and Zamar (1988). It is consistent and has a breakdown point of 50%. For a given set of residuals $e_1, \ldots, e_T$, it is computed as

$$\tau^2\left(e_1, \ldots, e_T\right) = \frac{s_T^2}{T}\sum_{t=1}^{T}\rho\left(\frac{e_t}{s_T}\right) \tag{3.8}$$

with $s_T = \mathrm{MAD}_t(e_t)$ and with $\rho$ the biweight function from (3.3). This version of the $\tau^2$ scale is easy to compute. The boundedness of $\rho$ makes the $\tau^2$ estimator robust to outlying observations.

For the additive model, the robust version of (3.6) is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathrm{roblik}_{\mathrm{A}}(\boldsymbol{\theta}) \tag{3.9}$$

with

$$\text{roblik}_{\text{A}}(\boldsymbol{\theta}) = -\frac{T}{2} \ \log \tau^2 \left(y_1 - \hat{y}^*_{1|0}(\boldsymbol{\theta}), \ldots, y_T - \hat{y}^*_{T|T-1}(\boldsymbol{\theta})\right) \tag{3.10}$$

Notice that this estimator is similar to the robust estimator for ARMA models of Maronna et al. (2006, section 8.8.3).

For the multiplicative model, denote the relative errors as $e_t(\boldsymbol{\theta}) = (y_t - \hat{y}^*_{t|t-1}(\boldsymbol{\theta}))/\hat{y}^*_{t|t-1}(\boldsymbol{\theta})$, for $t = 1, \ldots, T$. The robust version of the likelihood is

$$\text{roblik}_{\text{M}}(\boldsymbol{\theta}) = -\frac{T}{2} \ \log \tau^2 \left(e_1(\boldsymbol{\theta}), \ldots, e_T(\boldsymbol{\theta})\right) \ - \ \sum_{t=1}^{T} \log \left|\hat{y}^*_{t|t-1}(\boldsymbol{\theta})\right|. \tag{3.11}$$

For computing the estimator, we do not take the second term in (3.11) along, since for a parameter $\boldsymbol{\theta}$ such that one prediction $\hat{y}^*_{t|t-1}(\boldsymbol{\theta})$ is close to zero, the robust likelihood can become unbounded due to the robustness of the $\rho$-function. Such a degenerate solution should be avoided. Therefore we minimize a robust version of the mean squared relative error

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, \tau^2 \left(e_1(\boldsymbol{\theta}), \ldots, e_T(\boldsymbol{\theta})\right). \tag{3.12}$$

The numerical optimization problems of (3.9) and (3.12) are solved with the Nelder-Mead solver of the function `optim` in the statistical software package R (R Core Team, 2017). This solver requires an initial guess for $\boldsymbol{\theta}$. In Hyndman and Khandakar (2008) the initial values of the smoothing parameters $(\alpha, \beta, \phi, \gamma)$ are chosen in a data independent way. Therefore there is no robustness issue here, and we choose the initial values in exactly the same way.

## 3.4 Robust information criterion

We use a robust information criterion to compare different models, similar as in Hyndman and Khandakar (2008), but with the robustified likelihood. The definition of the robust Akaike information criterion is

$$\text{robAIC} = -2 \ \text{roblik} + 2p, \tag{3.13}$$

with $p$ the number of parameters of the model. The formulas for the robust Bayesian information and the robust corrected Akaike information criterion are

$$\text{robBIC} = -2 \text{ roblik} + \log(T)p \tag{3.14}$$

and

$$\text{robAICc} = -2 \text{ roblik} + 2\frac{pT}{T-p-1}. \tag{3.15}$$

The expressions for the robustified likelihood are given in (3.10) for the additive model, and in (3.11) for the multiplicative model. In the remainder of the text, we follow the suggestion of Hyndman and Khandakar (2008) and use the (robust) AICc to compare several exponential smoothing variants. For a given time series, all models from Table 1 are estimated and the one with the lowest robust AICc will be selected and used for forecasting.

# 4 Implementation

In this section we discuss the implementation of the robust exponential smoothing method in the R language. We converted the function `ets` in the `forecast` package of Hyndman and Khandakar (2008) to a robust version called `robets`. This function has the same functionalities as `ets`, and can be found in the R package `robets` we developed. Given a time series object `y`, predictions can be made as follows:

```
model <- robets(y)
plot(forecast(model, h = 8))
```

The above R commands perform automatic forecasting of the time series, with a forecast horizon of $h = 8$. It is also possible to add an additional argument specifying a single model, supplying the acronym of the model by a three-letter string as in Table 1.

The function `robets` works as `ets`, except that the robust methodology is applied instead. This means that the forecasting equations use cleaned values, as in (3.1), where the tuning constant of the Huber $\psi$-function is set to its default value $k = 3$, but any other value can be provided as an additional argument. Furthermore, the parameters are estimated solving criterion (3.9) if the underlying model is additive, and solving (3.12) if multiplicative. This is our proposal in Section 3.3. The user can also choose among different robust information criterions: robAICc (3.15), robAIC (3.13) or robBIC (3.14), with the former as default.

The `robets` function provides additional output useful for detecting outliers. As such, the following outlyingness measure is computed for every time point:

$$\frac{y_t - \hat{y}^*_{t|t-1}}{\hat{\sigma}_t}.$$

If the absolute value of the outlyingness of an observation is larger than $k$, then it is considered to be an outlier. Outliers may be highlighted on the plot of the time series.

As an example, consider a quarterly time series of length 39, that can be found on `http://stats.stackexchange.com/questions/146098`. Figure 1 (left) shows the result. The title of the plot indicates the selected model (based on the AICc), which is here a multiplicative model with no trend and multiplicative seasonality. The prediction intervals are constructed as in Subsection 2.3. In this time series the last observation is outlying, indicated by a red dot. This last observation is an outlier since, for the given season, its value is much higher than in previous years. In Figure 1 (right) the results for the non-robust method are given. Note that a different model has been selected. The forecasted values are much higher due to the outlier. The robust method gives less importance to the outlier, resulting in forecasts closer to the original level of the time series.

The package is practical, but also fast. The computation time mainly depends on the length of the time series. We compare the execution time[1] of `robets` with `ets` for different lengths $T$ in Table 2, averaged over 100 generated time series. We see from Table 2 that the robust method takes about twice as much time as `ets`.

[1] We use the `microbenchmark` package in R to compute the timings (Mersmann, 2015).

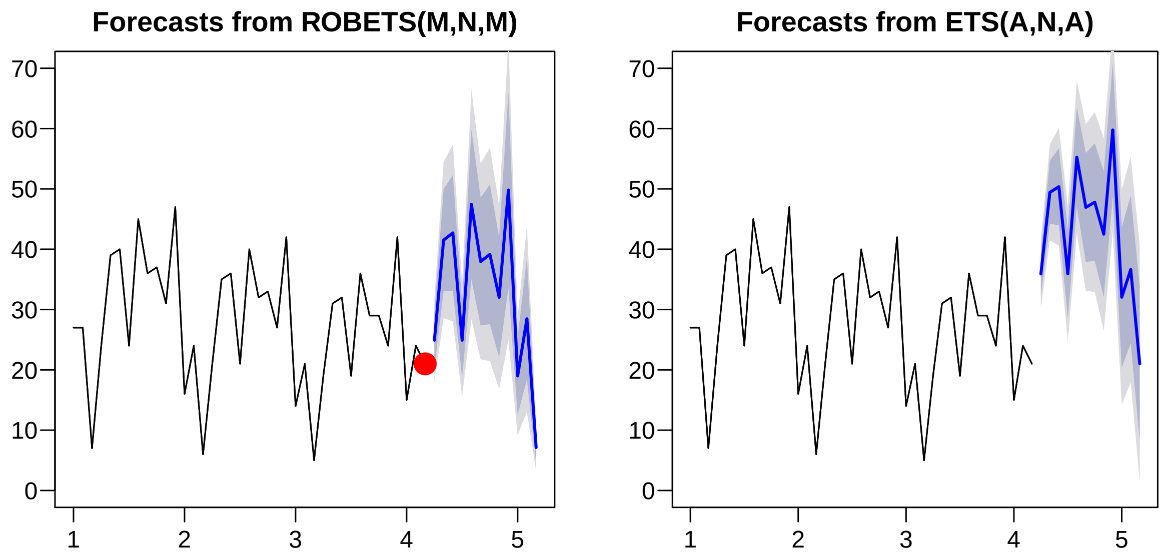


Figure 1: A quarterly time series of length 39, together with 8 forecasted values. The left plot uses the robust method, the right plot the non-robust. Detected outliers are indicated by a red dot (left plot). The dark and light gray zone are the 80 and 95% prediction intervals respectively.

Table 2: Average computation time in milliseconds for time series of length $T$.

| $T$ | 25 | 50 | 75 | 100 | 200 |
|---|---|---|---|---|---|
| `ets` | 5 | 8 | 9 | 9 | 18 |
| `robets` | 9 | 12 | 20 | 20 | 51 |

# 5 Simulation study

In this section we study the effect of outliers on the robust and non-robust method. We generate time series with the underlying models of the fifteen variants of Table 1. The time series length is $T = 40$ and the number of seasons per period is $m = 4$. We choose $\sigma = 0.05$ for all models, additive or multiplicative. The other choices are $\alpha = 0.36$, $\beta = 0.21$, $\phi = 0.9$ and $\gamma = 0.2$ and starting values $l_0 = 1$ and $b_0 = 0.05$. The starting seasonal component is $s_{-3:0} = (s_{-3}, s_{-2}, s_{-1}, s_0) = (-0.01, 0.01, 0.03, -0.03)$ for models with additive seasonality, and $s_{-3:0} = (0.99, 1.01, 1.03, 0.97)$ for models with multiplicative seasonality.

To generate time series with outliers, we adapt the underlying model. To generate time series with outliers, we replace the single source of error $\epsilon_t$ by $\epsilon_t + u_t$ in the observation equation. The distribution of the contamination $u_t$ is $u_t = 0$ with probability 1-$\epsilon$ and $u_t \overset{\text{iid}}{\sim} \mathcal{N}(0, K^2\sigma^2)$ with probability $\epsilon$. Hence $\epsilon$ is the fraction of outliers. Unless mentioned otherwise we set $(\epsilon, K) = (0.05, 20)$. For instance, for simple exponential smoothing with the additive error model, the contaminated version of model (2.4) is

$$\begin{aligned} y_t &= \ell_{t-1} + \epsilon_t + u_t \\ \ell_t &= \ell_{t-1} + \alpha\epsilon_t. \end{aligned} \tag{5.1}$$

For the multiplicative error model, we have

$$\begin{aligned} y_t &= \ell_{t-1} f(\epsilon_t + u_t) \\ \ell_t &= \ell_{t-1}(1 + \alpha\epsilon_t), \end{aligned} \tag{5.2}$$

with $f(x) = 1 + x$ for $x > 0$ and $f(x) = e^x$ for $x \leq 0$. The function $f$ is introduced to avoid outliers with negative values. For other exponential smoothing variants the contaminated models are analogous.

We simulate time series and estimate the model parameters and starting values. We

[2]Starting values are not estimated with the maximum likelihood estimator, but with the non-robust heuristic described in Section 3.2 and used by Hyndman and Khandakar (2008) as initial values of the numerical optimizer for finding the starting values.

Table 3: The simulated RMSE ($\times 100$) for each model for $h = 1$. The model is known, but the parameters are estimated. Standard errors are in parenthesis.

| generating model | no outliers non-robust | | no outliers robust | | outliers non-robust | | outliers robust | |
|---|---|---|---|---|---|---|---|---|
| ANN | 5.00 | (0.15) | 4.98 | (0.15) | 8.50 | (0.88) | 5.22 | (0.16) |
| ANA | 5.39 | (0.17) | 5.42 | (0.18) | 12.44 | (1.39) | 5.58 | (0.18) |
| AAN | 4.94 | (0.16) | 4.98 | (0.16) | 12.60 | (0.86) | 5.65 | (0.22) |
| AAA | 5.46 | (0.18) | 5.74 | (0.20) | 17.90 | (3.31) | 5.98 | (0.20) |
| $\mathrm{AA_dN}$ | 5.38 | (0.17) | 5.48 | (0.18) | 12.56 | (1.77) | 5.41 | (0.17) |
| $\mathrm{AA_dA}$ | 5.19 | (0.16) | 5.27 | (0.16) | 15.93 | (1.57) | 5.94 | (0.21) |
| MNN | 5.29 | (0.19) | 5.27 | (0.19) | 14.43 | (3.36) | 5.37 | (0.20) |
| MNA | 5.02 | (0.15) | 5.31 | (0.16) | 9.44 | (0.49) | 5.78 | (0.18) |
| MAN | 17.15 | (0.84) | 16.99 | (0.86) | 55.82 | (10.30) | 17.50 | (0.75) |
| MAA | 17.27 | (0.67) | 17.73 | (0.70) | 46.12 | (8.20) | 18.89 | (0.97) |
| $\mathrm{MA_dN}$ | 7.78 | (0.28) | 7.71 | (0.28) | 22.59 | (3.65) | 8.67 | (0.32) |
| $\mathrm{MA_dA}$ | 8.09 | (0.30) | 8.28 | (0.32) | 22.11 | (2.72) | 9.18 | (0.36) |
| MNM | 5.03 | (0.17) | 5.10 | (0.18) | 12.54 | (2.43) | 5.78 | (0.20) |
| MAM | 16.05 | (0.69) | 17.13 | (0.77) | 44.06 | (4.32) | 18.79 | (0.79) |
| $\mathrm{MA_dM}$ | 7.56 | (0.30) | 7.55 | (0.30) | 18.74 | (2.69) | 7.82 | (0.28) |

generate $h_{\max} = 8$ additional observations to be predicted. Since it is impossible to predict outliers, we don't allow outliers in the out-of-sample period. We compare the robust method with the non-robust exponential smoothing method[2] of Hyndman and Khandakar (2008).

For each model in Table 1 we compute the out-of-sample squared error at each horizon, and take the root of the average over $N = 500$ simulations:

$$\mathrm{RMSE}_h = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{T+h} - \hat{y}_{T+h|T})^2}. \tag{5.3}$$

In Table 3 we report the RMSE for horizon $h = 1$. For the clean simulations the RMSE is slightly larger with the robust method than with the non-robust method. In the contaminated setting we see a large increase in RMSE with the non-robust method. This is not happening for the robust method.

We repeat the previous simulation study, but now the model is unknown. For every generated time series all fifteen models are estimated and the one with the lowest information

Table 4: The simulated RMSE ($\times 100$) for each model for $h = 1$. The model is unknown and is selected automatically. Standard errors are in parenthesis.

| generating model | no outliers non-robust | | no outliers robust | | outliers non-robust | | outliers robust | |
|---|---|---|---|---|---|---|---|---|
| ANN | 5.18 | (0.16) | 5.37 | (0.16) | 12.55 | (1.60) | 5.44 | (0.18) |
| ANA | 5.51 | (0.17) | 5.83 | (0.20) | 14.12 | (1.78) | 5.57 | (0.19) |
| AAN | 5.10 | (0.17) | 5.40 | (0.19) | 14.77 | (0.86) | 6.08 | (0.26) |
| AAA | 5.56 | (0.18) | 6.00 | (0.21) | 19.53 | (3.84) | 6.17 | (0.21) |
| $\mathrm{AA_dN}$ | 5.72 | (0.17) | 5.70 | (0.19) | 18.36 | (3.16) | 6.10 | (0.26) |
| $\mathrm{AA_dA}$ | 5.29 | (0.17) | 5.43 | (0.16) | 15.52 | (1.17) | 6.28 | (0.25) |
| MNN | 5.46 | (0.19) | 5.73 | (0.21) | 19.39 | (3.91) | 5.64 | (0.18) |
| MNA | 5.18 | (0.16) | 5.35 | (0.17) | 10.67 | (0.95) | 5.80 | (0.20) |
| MAN | 17.42 | (0.86) | 17.70 | (0.84) | 54.06 | (11.30) | 17.65 | (0.75) |
| MAA | 17.46 | (0.68) | 17.93 | (0.78) | 46.58 | (9.54) | 18.44 | (0.73) |
| $\mathrm{MA_dN}$ | 7.95 | (0.29) | 8.07 | (0.29) | 23.61 | (2.71) | 8.84 | (0.35) |
| $\mathrm{MA_dA}$ | 8.29 | (0.31) | 8.68 | (0.37) | 19.61 | (1.77) | 9.36 | (0.34) |
| MNM | 5.03 | (0.17) | 5.35 | (0.18) | 10.20 | (0.90) | 5.65 | (0.19) |
| MAM | 16.08 | (0.72) | 16.81 | (0.75) | 42.30 | (3.98) | 17.54 | (0.73) |
| $\mathrm{MA_dM}$ | 7.65 | (0.31) | 8.00 | (0.30) | 20.45 | (2.84) | 7.67 | (0.27) |

criterion gets selected. The non-robust and robust methods are compared. In Table 4 the RMSE over 500 simulations at horizon $h = 1$ is computed. As expected, the numbers are the slightly larger than in Table 3, but the conclusions are similar. The robust method is slightly worse than the non-robust method for time series without outliers, but clearly better for time series with outliers, for all models.

We replicated the simulation for different lengths of the time series and changing the order of seasonality. Even if the time series length is small and the number of seasons is large ($m = 12$), the RMSE is much lower for the robust method than for the non-robust in a contaminated setting. We also computed the RMSE for other prediction horizons (unreported), yielding comparable results.

# 6 Data

We apply the robust method to the 3003 time series of the M3 competition of (Makridakis and Hibon, 2000). These data can be found in the R package `Mcomp`. The median length of the time series is 69, the smallest is of length 20 and the longest 144. The data are yearly, quarterly or monthly. The first part of the time series is used for estimation and the last $h$ data points are used as an out-of-sample period. For yearly time series, $h = 6$, for quarterly $h = 8$ and for monthly, $h = 18$. Denote $N_h$ the number of time series for which a prediction at horizon $h$ is made (for $h = 1$, $N_1 = 3003$).

We compute the out-of-sample symmetric mean absolute percentage error (sMAPE) as in Makridakis and Hibon (2000):

$$\text{sMAPE}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \left| \frac{y_{T_i+h,i} - \hat{y}_{T_i+h|T_i,i}}{(y_{T_i+h,i} + \hat{y}_{T_i+h|T_i,i})/2} \right| \times 100. \tag{6.1}$$

with $T_i$ the time stamp of the last point of the estimation period for the $i$-th series. This metric is scale independent, which is useful since every time series has a different scale. We also compute the symmetric median absolute percentage error:

$$\text{sMedAPE}_h = \operatorname*{median}_i \left| \frac{y_{T_i+h,i} - \hat{y}_{T_i+h|T_i,i}}{(y_{T_i+h,i} + \hat{y}_{T_i+h|T_i,i})/2} \right| \times 100. \tag{6.2}$$

In Table 5 these forecasting accuracy measures are reported for several forecast horizons. The results of the $\text{sMAPE}_h$ can directly be compared with Table 6 of Makridakis and Hibon (2000), where 24 different forecasting methods are compared. It turns out that the `ets`

Table 5: Forecasting accuracy for the time series of the M3-competion.

| | | Forecasting horizon $h$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 12 | 15 | 18 |
| $\text{sMAPE}_h$ | non-robust (`ets`) | 8.5 | 9.5 | 11.4 | 13.2 | 15.4 | 14.7 | 12.6 | 13.5 | 17.0 | 19.1 |
| | robust (`robets`) | 9.5 | 10.7 | 12.2 | 15.0 | 15.1 | 15.3 | 14.1 | 15.1 | 21.5 | 20.1 |
| $\text{sMedAPE}_h$ | non-robust (`ets`) | 3.0 | 3.8 | 4.7 | 5.9 | 6.3 | 6.7 | 6.2 | 7.0 | 9.0 | 10.1 |
| | robust (`robets`) | 3.2 | 4.0 | 4.6 | 6.0 | 6.4 | 6.7 | 6.3 | 7.0 | 9.3 | 10.1 |

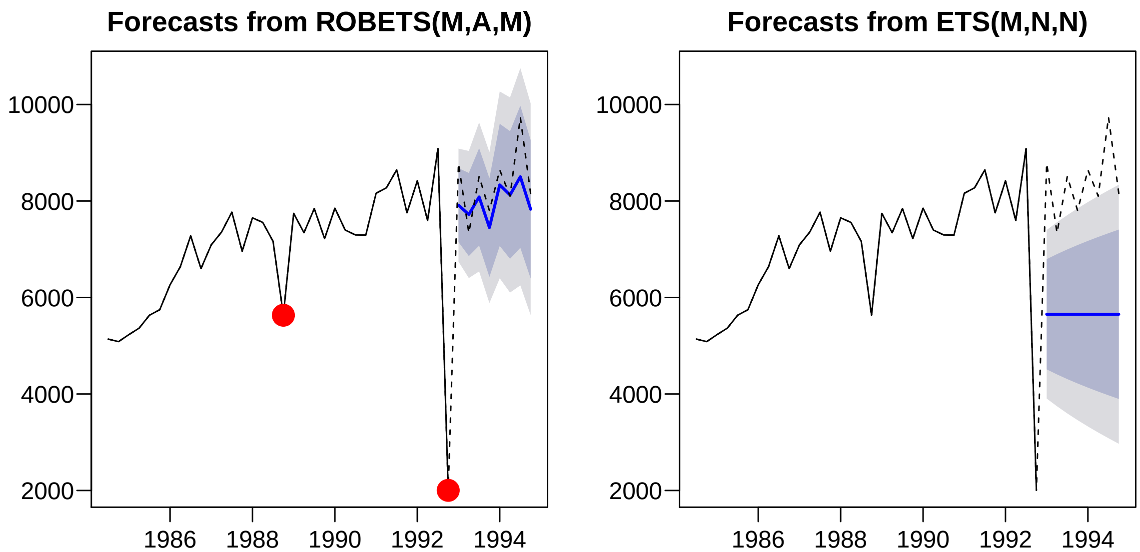


Figure 2: Time series 819 from the M3-competition. The last 8 observations are used as out-of-sample period (dashed line). Forecasts using the robust (left plot) and non-robust (right plot) are given in blue.

method is among the best methods at every horizon $h$, while `robets` is only slightly behind `ets`. In Table 5 we also reported the sMedAPE$_h$. With this measure there is almost no difference between `robets` and `ets`, which means that for the majority of the time series, the robust and non-robust method perform about the same.

We give an example where the robust method has a much better forecasting performance. Take the quarterly time series number 819 from the M3-competition. In Figure 2 we plot the forecasts with the robust and non-robust method. The last 8 points are not used in the estimation of the model, but are forecasted. The robust method gives less weight to the last observation of the estimation period, which appears to be an outlier, resulting in a much better forecasting performance.

# 7 Conclusion

We propose a robust version of the exponential smoothing framework of Hyndman and Khandakar (2008). The method is outlier robust: it has robust forecasting equations, robust smoothing parameter estimation and performs robust model selection. It is implemented in the R package `robets`.

The proposed robust method can also be used for outlier detection. If a time series contains outliers, they can be labeled as possible anomalies and investigated further. The `robets` package is easy to use and fast, making it possible to forecast large numbers of univariate time series in an automatic way. Furthermore, it can be used as a tool to find outliers in large data sets with many time series.

We cannot expect that the robust forecasting procedure consistently outperforms the non-robust counterpart. Indeed, in Section 6 we saw that the non-robust procedure has on average, over a large number of short time series, a better forecast performance. However, for an important number of time series, as the one in Figure 2, the robust method is best. We advise to check for the presence of outliers in each series: if outliers are present, then use the robust method; otherwise the standard method is recommended.

**Acknowledgments.** The authors gratefully acknowledge financial support of the KU Leuven grant GOA/12/14.

# References

Cipra, T., Hanzak, T., 2011. Exponential smoothing for time series with outliers. Kybernetika 47 (2), 165–178.

Croux, C., Gelper, S., Fried, R., 2008. Computational aspects of robust Holt-Winters smoothing based on M-estimation. Applications of Mathematics 53 (3), 163–176.

Croux, C., Gelper, S., Mahieu, K., 2010. Robust exponential smoothing of multivariate time series. Computational Statistics and Data Analysis 54, 2999–3006.

Fried, R., 2004. Robust filtering of time series with trends. Nonparametric Statistics 16, 313–328.

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

Gelper, S., Fried, R., Croux, C., 2010. Robust forecasting with exponential and holt-winters smoothing. Journal of Forecasting 29, 285–300.

Hyndman, R. J., Athanasopoulos, G., 2013. Forecasting: Principles & practice. Https://www.otexts.org/fpp/. Accessed on August 2, 2017.

Hyndman, R. J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software 27 (3).

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

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

Makridakis, S., Hibon, M., 2000. The M3-competition: results, conclusions and implications. International Journal of Forecasting 16, 451–476.

Maronna, R., Martin, R., Yohai, V., 2006. Robust Statistics. Wiley, UK.

Mersmann, O., 2015. microbenchmark: Accurate Timing Functions. R package version 1.4-2.1. https://CRAN.R-project.org/package=microbenchmark.

Muler, N., Peña, D., Yohai, V. J., 2009. Robust estimation for ARMA models. The Annals of Statistics 37 (2), 816–840.

Ord, J. K., Koehler, A. B., Snyder, R. D., 1997. Estimation and prediction for a class of dynamic nonlinear statistical models. Journal of the American Statistical Association 92 (440), 1621–1629.

Pegels, C. C., 1969. Exponential forecasting: some new variations. Management Science 15, 311–320.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

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

Yohai, V. J., Zamar, R. H., 1988. High breakdown-point estimates of regression by means of the minimization of an efficient scale. Journal of the American Statistical Association 83 (402), 406–413.

**FACULTY OF ECONOMICS AND BUSINESS**
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be