Robust forecasting with exponential and holt-winters smoothing

Journal of Forecasting
J. Forecast. **29**, 285–300 (2010)
Published online 24 June 2009 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.1125

Robust Forecasting with Exponential and Holt–Winters Smoothing

SARAH GELPER,[1]* ROLAND FRIED[2] AND
CHRISTOPHE CROUX[3†]

[1] *Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, the Netherlands*
[2] *Department of Statistics, University of Dortmund, Dortmund, Germany*
[3] *Faculty of Business and Economics, Katholieke Universiteit Leuven, Leuven, Belgium*

ABSTRACT
Robust versions of the exponential and Holt–Winters smoothing method for forecasting are presented. They are suitable for forecasting univariate time series in the presence of outliers. The robust exponential and Holt–Winters smoothing methods are presented as recursive updating schemes that apply the standard technique to pre-cleaned data. Both the update equation and the selection of the smoothing parameters are robustified. A simulation study compares the robust and classical forecasts. The presented method is found to have good forecast performance for time series with and without outliers, as well as for fat-tailed time series and under model misspecification. The method is illustrated using real data incorporating trend and seasonal effects. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS forecasting; Holt–Winters smoothing; robust methods; time series

INTRODUCTION

Exponential smoothing is a simple technique used to smooth and forecast a time series without the necessity of fitting a parametric model. It is based on a recursive computing scheme, where the forecasts are updated for each new incoming observation. Exponential smoothing is sometimes considered as a naive prediction method. Yet it is often used in practice where it shows good performance, as illustrated, for example, in Makridakis *et al.* (1998, Ch. 4) and in Kotsialos *et al.* (2005). The Holt–Winters method, also referred to as double exponential smoothing, is an extension of exponential smoothing designed for trended and seasonal time series. Holt–Winters smoothing is a widely used tool for forecasting business data that contain seasonality, changing trends and seasonal correlation.

The exponential and Holt–Winters techniques are sensitive to unusual events or outliers. Outliers affect the forecasting methods in two ways. First, the smoothed values are affected since the update

* Correspondence to: Sarah Gelper, Erasmus School of Economics, Burgemeester Oudlaan 50, 3000 Rotterdam, the Netherlands. E-mail: gelper@ese.eur.nl
† This article was published online on 24 June 2009. An error was subsequently identified (the affiliation of the third author was incorrect). This notice is included in the online and print versions to indicate that both have been corrected on 20 July 2009.

equations involve current and past values of the series including the outliers. The second effect of outliers is on the selection of the parameters used in the recursive updating scheme. These parameters regulate the degree of smoothing and are chosen to minimize the sum of squared forecast errors. This approach is non-robust and may result in less accurate forecasts in the presence of outliers. We propose robust versions of the exponential and Holt–Winters smoothing techniques which make both the smoothing and the parameter selection robust to outliers. The robustification of the smoothing algorithm is based on a pre-cleaning mechanism that identifies and downweights outlying values. The robust recursive scheme applies the standard smoothing techniques to the pre-cleaned data and is thus very easy to implement. Data pre-cleaning, in particular for non-stationary time series, is in practice rarely done, yet it significantly improves forecast performance, as is shown below ('Comparison of the forecast performance'). There we also show that our proposed cleaning method outperforms a more standard approach based on a 2-sigma rule.

Although methodology for robust time series analysis has been developed (e.g. Maronna *et al.*, 2006, Ch. 8, for a review), surprisingly little attention has been given to robust alternatives to exponential and Holt–Winters smoothing and forecasting. A first attempt was made by Cipra (1992), who presents the Holt–Winters method as a weighted regression problem, and then replaces the traditional least squares fit by a robust regression fit. This yields a reasonable method, but it can be improved upon. Another proposal was made by Taylor (2007), who uses quantile regression to obtain robust forecasts. Kirkendall (2006) proposes a method that classifies each observation as being either regular, an outlier, or initiating a level shift. However, no recursive formulae were obtained in these approaches. The recursive formulation of the Holt–Winters method is very important, though, as it is one of the reasons why it is so attractive and easy to use.

The approach we recommend applies a particular robust version of the Kalman filter to the state-space model associated with exponential and Holt–Winters smoothing. Robust Kalman filtering has already been considered in the literature, e.g. Masreliez and Martin (1977), Martin and Thomson (1982), Romera and Cipra (1995), Ruckdeschel (2001), and Cipra and Romera (1997). The latter authors briefly mention the possibility of using their robust Kalman filter for Holt–Winters smoothing. This Kalman filter yields a simple recursive updating scheme and results in more robust forecasts. However, several crucial issues have not yet been addressed, like the robust selection of the smoothing parameters and the appropriate choice of the auxiliary robust scale estimator.

The classical exponential and Holt–Winters methods are briefly reviewed in the next section. Furthermore, the next section presents robust exponential and Holt–Winters smoothing. The method we propose boils down to applying a robust data pre-cleaning procedure, after which the traditional method is applied. We provide guidelines for the choice of smoothing parameters, starting values and scale estimation procedures. The third section presents a simulation study comparing the forecast performance of the classical and the robust smoothing methods. An illustration using seasonal real data can be found in the fourth section. The fifth section summarizes and gives concluding remarks.

EXPONENTIAL AND HOLT–WINTERS SMOOTHING

Suppose we have a time series y_t, which is observed at $t = 1, \ldots, T$. The exponential smoothing method computes the smoothed series $\tilde{y}_t$ according to the following recursive scheme:

$$\tilde{y}_t = \lambda y_t + (1-\lambda)\,\tilde{y}_{t-1} \tag{1}$$

DOI: 10.1002/for

where λ is a smoothing parameter taking values between zero and one and regulating the degree of smoothing. A starting value $\tilde{y}_m$ for equation (1) is obtained by averaging over the first m observed values $y_1, \ldots, y_m$, where m is the length of the startup period. An h-step-ahead prediction at time t is denoted by $\hat{y}_{t+h|t}$ and given by the most recent smoothed value: $\hat{y}_{t+h|t} = \tilde{y}_t$.

For trended time series, Holt (1959) proposed to include a local trend variable F_t:

$$\begin{aligned} \tilde{y}_t &= \lambda_1 y_t + (1-\lambda_1)(\tilde{y}_{t-1} + F_{t-1}) \\ F_t &= \lambda_2(\tilde{y}_t - \tilde{y}_{t-1}) + (1-\lambda_2)F_{t-1} \end{aligned} \tag{2}$$

As for exponential smoothing, the smoothing parameters λ_1 and λ_2 take values between zero and one. Guidelines on how to select the smoothing parameters are presented below. For Holt's method (2), an h-step-ahead forecast of y_t is obtained as

$$\hat{y}_{t+h|t} = \tilde{y}_t + hF_t \tag{3}$$

The starting values $\tilde{y}_m$ and F_m of the recursive equations in (2) can be obtained by a linear ordinary least squares fit in a startup period, as described in Bowerman *et al.* (2005). More specifically, regressing y_t versus the time t, for $t = 1 \ldots m$, yields an intercept $\hat{\alpha}_0$ and a slope $\hat{\beta}_0$ resulting in

$$\begin{aligned} \tilde{y}_m &= \hat{\alpha}_0 + \hat{\beta}_0 m \\ F_m &= \hat{\beta}_0 \end{aligned} \tag{4}$$

Holt's method in equation (2) can easily be extended to time series with seasonality, in which case the method is referred to as the Holt–Winters method (see Winters, 1960). We address seasonality in the fourth section. For convenience, we use in this paper the name 'Holt–Winters method' for both the setting with and without seasonality.

The exponential and Holt–Winters forecasting methods in equations (1) and (2) are not robust with respect to outlying observations. One extremely large or small observation results in high or low values of the smoothed series over some period and deteriorates the forecast performance. To obtain better forecast accuracy in the presence of outliers, robust versions of this technique are needed.

Cipra (1992) proposes a robust exponential and Holt–Winters smoothing method based on M-estimation. In the case of ordinary exponential smoothing, the value of the smoothed series at time t, $\tilde{y}_t$, is the solution of the following minimization problem:

$$\tilde{y}_t = \arg\min_{\theta} \sum_{i=1}^{t} (1-\lambda)^{t-i}(y_i - \theta)^2 \tag{5}$$

To obtain robustness according to the idea of M-estimation, the square in equation (5) is replaced by a suitable function ρ applied to standardized data:

$$\tilde{y}_t = \arg\min_{\theta} \sum_{i=1}^{t} (1-\lambda)^{t-i} \rho\left(\frac{y_i - \theta}{\sigma_t}\right) \tag{6}$$

where $\hat{\sigma}_t$ estimates the scale of the residuals $r_t = y_t - \hat{y}_{t|t-1}$. By taking a function ρ which increases more slowly than the quadratic function, the effect of outlying values is reduced. Cipra (1992) uses

DOI: 10.1002/for

the ρ-function with derivative $\rho' = \psi$ and the Huber-ψ function with $k = 1.645$ (the Huber-ψ function is given in equation (11)). For the scale estimates, Cipra (1992) proposes to use

$$\hat{\sigma}_t = 1.25\lambda_\sigma |r_t| + (1-\lambda_\sigma)\hat{\sigma}_{t-1} \tag{7}$$

where the parameter λ_σ is another smoothing parameter taking values between zero and one. After algebraic manipulations and making use of an iterated weighted least squares algorithm, recursive formulae are obtained for the smoothing methods proposed by Cipra (1992). In the sequel of the paper we refer to this approach based on *M*-estimation as the weighted regression (WR) method.

A new robust smoothing method

The approach we recommend in this paper is different from that of Cipra (1992), and turns out to yield a better performance than the WR method, as will be shown in the third section. The idea is to replace the observed y_t in equation (1) by a 'cleaned' version y_t^*. The robust exponential smoothing recursive equation is then simply given by

$$\tilde{y}_t = \lambda y_t^* + (1-\lambda)\tilde{y}_{t-1} \tag{8}$$

and similarly for the Holt smoothing system of equations (2)

$$\begin{aligned} \tilde{y}_t &= \lambda_1 y_t^* + (1-\lambda_1)(\tilde{y}_{t-1} + F_{t-1}) \\ F_t &= \lambda_2(\tilde{y}_t - \tilde{y}_{t-1}) + (1-\lambda_2)F_{t-1} \end{aligned} \tag{9}$$

For both the exponential and Holt–Winters smoothing, the cleaned series y_t^* is obtained as

$$y_t^* = \psi\left(\frac{y_t - \hat{y}_{t|t-1}}{\hat{\sigma}_t}\right)\hat{\sigma}_t + \hat{y}_{t|t-1} \tag{10}$$

where the ψ-function is applied to standardized one-step-ahead forecast errors. As before, forecasts are computed by equation (3). The scale of these forecast errors is estimated by $\hat{\sigma}_t$. The ψ-function reduces the influence of outlying observations. In this paper we take the Huber ψ-function

$$\psi(x) = \begin{cases} x & \text{if } |x| < k \\ \text{sign}(x)k & \text{otherwise} \end{cases} \tag{11}$$

The pre-cleaning based on the Huber ψ-function can be interpreted as replacing unexpected high or low values by a more likely value. More specifically, if the difference between the observed y_t and its predicted value at $t-1$ is small, the cleaned value y_t^* simply equals the observed y_t. On the other hand, if the deviation between the predicted and the observed value is too large, the observation is considered an outlier and gets replaced by a boundary value dependent on k. This positive constant k hence regulates the identification of outliers. A common choice of k is two, implicitly assuming a normal distribution of the one-step-ahead forecast errors $r_t = y_t - \hat{y}_{t|t-1}$. If k is set to infinity, the robust smoothing procedure reduces to the classical method.

DOI: 10.1002/for

The pre-cleaning process in equation (10) makes use of an estimated scale $\hat{\sigma}_t$ of the one-step-ahead forecast errors. This scale can simply be estimated by applying any robust scale estimate on all previous values of r_t. For example, one could take the median absolute deviation (MAD), defined as

$$\hat{\sigma}_t = \underset{1\le s\le t}{\mathrm{MAD}}(r_s) = \underset{1\le s\le t}{\mathrm{med}}\left|r_s - \underset{1\le s\le t}{\mathrm{med}}\, r_s\right|$$

However, this implicitly assumes that the scale remains constant. An option allowing for (slowly) varying scale is to estimate σ_t by the following update equation:

$$\hat{\sigma}_t^2 = \lambda_\sigma \rho\left(\frac{r_t}{\hat{\sigma}_{t-1}}\right)\hat{\sigma}_{t-1}^2 + (1-\lambda_\sigma)\hat{\sigma}_{t-1}^2 \tag{12}$$

where we make use of a *bounded* loss-function ρ. Here, we choose the biweight ρ-function defined by

$$\rho(x) = \begin{cases} c_k\left(1-\left(1-(x/k)^2\right)^3\right) & \text{if } |x| \le k \\ c_k & \text{otherwise} \end{cases} \tag{13}$$

where c_k is a constant to achieve consistency of the scale parameter for a normal error distribution. For the common choice of $k = 2$ we have $c_k = 2.52$. The scale estimator in (12) corresponds to a τ^2-scale estimator as in Yohai and Zamar (1988), now computed in a recursive way. This τ^2-scale estimator is highly robust, but at the same time almost as efficient as the standard deviation. Equation (12) yields a much more robust scale estimation procedure as compared to (7). Indeed, the estimated scale in (7) can be made arbitrarily large by including extreme outliers in the data, since the impact of $|r_t|$ is still unbounded.

Combining the robust Holt system of equations (9) and the pre-cleaning process in equation (10) yields the following recursive scheme:

$$\begin{aligned} \tilde{y}_t &= \lambda_1 \psi\left(\frac{y_t - (\tilde{y}_{t-1} + F_{t-1})}{\hat{\sigma}_t}\right)\hat{\sigma}_t + \tilde{y}_{t-1} + F_{t-1} \\ F_t &= \lambda_2(\tilde{y}_t - \tilde{y}_{t-1}) + (1-\lambda_2)F_{t-1} \end{aligned} \tag{14}$$

Predictions of y_t can be obtained by straightforward application of the forecast formula (3), as for the classical method. Outlying observations will not attract the smoothed values as much as in the classical method, resulting in more accurate predictions in the presence of outliers. We refer to the above equations, together with equation (12), as the robust Holt–Winters method, abbreviated RHW. It is important to note that this procedure still yields an easy-to-implement forecasting method. The choices we made for the ψ-function, the ρ-function, and the scale estimator are standard in the modern literature on robust statistics (e.g. Maronna *et al.*, 2006).

Not only do the update equations need to be robust to outliers, but also the starting values. For the starting value in exponential smoothing, we suggest using the median of the first m observations instead of the mean. For Holt–Winters smoothing, we replace the ordinary least squares regression in equation (4) by the repeated median estimator (Siegel, 1982), where $\hat{\alpha}_0$ and $\hat{\beta}_0$ in (4) are given by

DOI: 10.1002/for

$$\hat{\beta}_0 = \text{med}_i\left(\text{med}_{i\neq j}\frac{y_i - y_j}{i-j}\right) \quad \text{and} \quad \hat{\alpha}_0 = \text{med}_i\left(y_i - \hat{\beta}_0 i\right)$$

for $i, j = 1, \ldots, m$. The repeated median regression has been shown to have good properties for time series smoothing; see, for example, Davies *et al.* (2004) and Fried (2004). A starting value of the recursive scale estimation in (12) is obtained by the MAD of the regression residuals in this startup period.

Selecting the smoothing parameters

In this subsection, we give guidelines for the choice of the smoothing parameters λ_1 and λ_2. A first approach is rather subjective and based on a prior belief of how much weight should be given to the current versus the past observations. If one believes that current values should get high weights, large values of the smoothing parameters are chosen.

Another possibility for selecting the smoothing parameters is to use a data-driven procedure optimizing a certain criterion. For every λ (in the case of exponential smoothing) or for every possible combination of values for λ_1 and λ_2 (Holt smoothing), the one-step-ahead forecast error series r_t is computed. The best smoothing parameters are those resulting in the smallest one-step-ahead forecast errors. A common way to evaluate the magnitude of the one-step-ahead forecast error series is by means of the mean squared forecast error (MSFE):

$$\text{MSFE}\,(r_1, \ldots, r_T) = \sum_{t=1}^{T} r_t^2 \tag{15}$$

However, since we deal with contaminated time series, the MSFE criterion has a drawback. One very large forecast error causes an explosion of the MSFE, which typically leads to smoothing parameters being biased towards zero. For this reason, we propose to use a robustified version of the MSFE, based on a τ-estimator of scale:

$$\tau^2(r_1, \ldots, r_T) = s_T^2 \frac{1}{T}\sum_{t=1}^{T} \rho\left(\frac{r_t}{s_T}\right) \tag{16}$$

where $s_T = 1.48\ \text{Med}_t|r_t|$. The τ^2-estimator downweights large forecast errors r_t. It requires a bounded ρ-function for which we choose again the biweight ρ-function as in (13). The effect of extremely large forecast errors on criterion (16) is limited. If the ρ-function in (16) is replaced by the square function, this procedure reduces to the usual minimization of the MSFE. Other criteria than the MSFE or τ^2-measure could also be considered, as, for example, mean or median absolute prediction error. More details can be found in the review article Gardner (2006).

COMPARISON OF THE FORECAST PERFORMANCE

In this section we conduct a simulation study to compare the forecast performance of the classical and the robust Holt–Winters methods for trended data. For the classical method we use the MSFE criterion for selecting the smoothing parameters. Optimization with respect to λ_1 and λ_2 is carried out by a simple grid search. The classical Holt method, denoted by HW, is compared to the two

DOI: 10.1002/for

robust schemes presented above. The first is the weighted regression approach of Cipra (1992) (WR), and the second the robust Holt method (RHW) that we recommend. To illustrate the importance of the use of an appropriate robust scale estimator in the recursive equations, we also apply the proposed RHW method with the non-robust scale estimator (7) instead of (12). This variant will be abbreviated RHW′. For the robust methods, the smoothing parameters are chosen optimally by a simple grid search according to the τ^2 criterion in (16). Furthermore, the smoothing parameter λ_σ, needed in (7) and (12), is fixed at 0.2 and the length of the startup period is fixed at $m = 10$.

As mentioned before, the robust Holt method proposed in (9) applies the classic update equations to a pre-cleaned time series. Instead of using the proposed cleaning rule (10), other simple pre-cleaning schemes where outliers are replaced by fitted values can be used; see, for example, Gardner (1999). Some systems routinely reject extreme values and replace them with averages, where a value is declared as 'extreme' when it has a distance of more than 2 standard deviations from the average. An obvious drawback of this approach is that it cannot be applied to trending time series. Hence, as an alternative to computing the average, a regression line needs to be fitted. To allow for stochastic trends, the regression fit will be computed locally within a window. More precisely, for every observation y_t, we fit by ordinary least squares a linear trend from the data $y_t, y_{t-1}, \ldots, y_{t-n}$ in a window of length $n = 20$, and compute the fitted value $\hat{y}_t$. Then we define an upper and lower bound as

$$\mathrm{UB}_t = \hat{y}_t + 2\,\hat{\sigma}_{\mathrm{OLS}}$$

$$\mathrm{LB}_t = \hat{y}_t - 2\,\hat{\sigma}_{\mathrm{OLS}}$$

where $\hat{\sigma}_{\mathrm{OLS}}$ is the standard deviation of the regression residuals. Whenever the observed time series exceeds the bounds, i.e. $y_t > \mathrm{UB}_t$ or $y_t < \mathrm{LB}_t$, it is replaced by its fitted value. Identifying outliers as those observations with large residuals in a linear regression is standard and described in most statistics textbooks; see, for example, Kutner *et al.* (2005). After the pre-cleaning phase, the classic Holt's smoothing method in equation (2) can be applied to the cleaned series. This 2-sigma rule for pre-cleaning is further referred to as HWc and is included in the simulation study.

We first compare the forecast performance when the data are generated according to a local linear trend model. For this model, it is known that Holt's smoothing method is optimal. However, exponential smoothing methods generally work very well for many different types of models. To study whether this property also holds for the robust smoothing methods, we also compare the forecast performance of the different methods under model misspecification.

Local linear trend model

The model from which we simulate the time series y_t is the local linear trend model. This model specifies that the observed series y_t is composed of an unobserved level α_t and trend β_t:

$$y_t = \alpha_t + e_t, \qquad e_t \sim N(0, \sigma^2) \tag{17}$$

where α_t and β_t are given by

$$\begin{aligned} \alpha_t &= \alpha_{t-1} + \beta_{t-1} + \eta_t, & \eta_t \sim N(0, \sigma_\eta^2) \\ \beta_t &= \beta_{t-1} + \nu_t, & \nu_t \sim N(0, \sigma_\nu^2) \end{aligned} \tag{18}$$

We take the error terms e_t in equation (17) to follow a standard normal distribution (or a certain contaminated version of it) and set both σ_η and σ_ν in equation (18) equal to 0.1, so that the variation

DOI: 10.1002/for

around the local level and trend is large as compared to the variation of the level and trend themselves.

Our main interest is the forecast performance of the smoothing methods. Therefore we evaluate them with respect to their prediction abilities. We simulate 1000 time series of length 105, for which we let the smoothing algorithm run up to observation 100 and then forecast observations 101 and 105. We compare the one- and five-step-ahead forecast errors under four different simulation schemes. These schemes correspond to four possibilities for the innovation terms e_t in equation (17) and are given in Table I. For the out-of sample period, $t = 101, \ldots, 105$, we do not allow for outliers, since these are unpredictable by nature. As such, 1000 forecast errors are computed for each method and for every simulation scheme at horizons $h = 1$ and $h = 5$. The outcomes are presented in a boxplot in Figure 1 and summarized by their mean squared forecast error[1] in Table II. We only present the results for $h = 1$, since the relative performance of the different procedures is similar for $h = 5$. Results for $h = 5$ are available upon request. The τ^2-scale estimate of the forecast errors is also presented in Table II, measuring the magnitude of the forecast errors without interference of extreme errors that blow up the MSFE.

The first simulation setting is the reference setting, where the simulated time series follow a non-contaminated local linear trend model. This setting is referred to as the clean data setting, and indicated by CD. The top left panel of Figure 1 presents a boxplot of the one- and five-step-ahead forecast errors for the HW, HWc, WR, RHW′ and RHW methods. Their median forecast error is not significantly different from zero, as is confirmed by a sign test. This property is called median unbiasedness. The dispersion of the forecast errors is about the same for all methods, with the exception of the HWc approach. Standard HW combined with cleaning based a 2-sigma rule yields an important increase in the forecast error. The reasons for this bad performance of HWc is that the local regression fits assume local linearity, which does not hold in the model we simulate from. Hence, clean observations will be falsely declared as outliers, leading to an efficiency loss. Table II shows that the classical method has the smallest MSFE and τ^2-scale, although the difference from the robust methods WR, RHW′ and RHW is rather small. We conclude that in the clean data case the method we advocate, RHW, performs only slightly worse than the non-robust standard HW.

In the second setting, a small fraction of the clean observations is replaced by symmetric outliers (SO), i.e. equal probability of observing an extremely large or small value. More specifically, we replace on average 5% of the error terms e_t in (17) by drawings from a normal distribution with mean zero and standard deviation 20. One-step-ahead forecast errors are presented in the top right panel of Figure 1. Similarly as for the clean data, the forecast methods are median unbiased. The RHW methods show a smaller dispersion of the forecast errors than all other methods. This can also be seen from the MSFE and τ^2 figures in Table II. The MSFE is the smallest for the RHW method,

Table I. Simulation schemes

CD	Clean data	$e_t \overset{\text{i.i.d.}}{\sim} N(0,1)$
SO	Symmetric outliers	$e_t \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)N(0,1)+\varepsilon N(0,20)$, where the outlier probability ε equals 0.05
AO	Asymmetric outliers	$e_t \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)N(0,1)+\varepsilon N(20,1)$, where the outlier probability ε equals 0.05
FT	Fat-tailed data	$e_t \overset{\text{i.i.d.}}{\sim} t_3$

[1]All reported MSFE values in Table II have a standard errors below 0.05, except for the setting ‘Fat-tailed’ (FT), where the standard deviation is about 0.5.

DOI: 10.1002/for

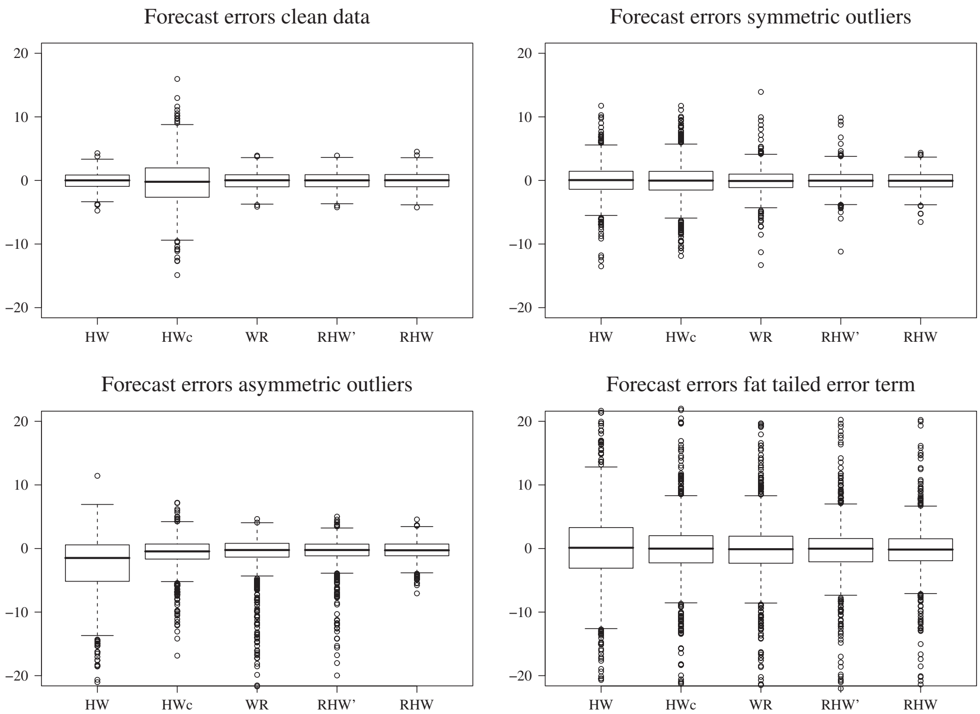

Figure 1. One-step-ahead forecast errors of 1000 replications of a time series of length 100, simulated according to sampling schemes CD (clean data, top left), SO (symmetric outliers, top right), AO (asymmetric outliers, bottom left) and FT (student-t_3 error terms, bottom right). We consider standard Holt–Winters (HW), HW with standard pre-cleaning (HWc), weighted regression (WR), robust HW without robust scale (RHW′) and with robust scale (RHW)

Table II. The MSFE and τ^2-scale for one-step-ahead forecast errors for four different simulation schemes under the local linear trend model

Setting		HW	HWc	WR	RHW′	RHW
CD	MSFE	1.76	1.84	1.79	1.83	1.86
	τ^2	1.77	1.88	1.90	1.91	1.93
SO	MSFE	7.00	2.43	3.92	2.65	2.10
	τ^2	4.68	2.07	2.47	2.01	2.04
AO	MSFE	32.26	10.34	29.85	16.96	2.21
	τ^2	13.85	3.05	2.65	2.03	1.92
FT	MSFE	3475.37	2656.47	2706.17	2643.62	2638.70
	τ^2	25.31	10.37	11.36	7.91	7.23

while the τ^2 for the RHW and RHW′ methods are comparable. The reason for this is that the RHW and RHW′ methods show similar forecast performance in general, but occasionally the RHW′ methods have much larger forecast errors. This illustrates the advantage of the robust pre-cleaning and scale estimation of the RHW method.

DOI: 10.1002/for

Table III. Average value of the selected smoothing parameters for four different simulation schemes, and using different Holt–Winters smoothing methods

	HW		HWc		WR	RHW′		RHW	
	λ_1	λ_2	λ_1	λ_2	λ	λ_1	λ_2	λ_1	λ_2
CD	0.36	0.25	0.37	0.27	0.29	0.36	0.36	0.39	0.34
SO	0.20	0.17	0.31	0.23	0.27	0.43	0.24	0.37	0.32
AO	0.14	0.12	0.23	0.18	0.29	0.66	0.15	0.37	0.32
FT	0.15	0.12	0.19	0.16	0.21	0.44	0.17	0.30	0.26

The third setting is similar to the previous setting, but here the data contain asymmetric outliers (AO) instead of symmetric ones. We replace on average a small fraction $\varepsilon = 0.05$ of the clean data by large positive values. The boxplot in the bottom left panel of Figure 1 shows that the median prediction error is now below zero, since adding large positive outliers makes the forecasts upward biased. A sign test indicates that this median bias is significant at the 1% level, for all methods. However, the bias is much less prevalent for the robust methods than for the classical HW. Note that in this setting with large asymmetric outliers, the HWc method does a much better job than before and improves now on the standard HW. However, it is still not competitive with the recommended RHW approach. As can be seen from Table II, the RHW method is the most accurate one. The advantage of using RHW instead of WR and RHW′ is much more pronounced than before, as is also confirmed by the boxplot.

In the two previous settings (SO and AO), the outliers occur with probability ε. In the next simulation setting, every observation follows a fat-tailed (FT) distribution, namely a Student-t with three degrees of freedom. The boxplot in the bottom right panel of Figure 1 illustrates that for this FT distribution the robust methods are much better than standard HW. Hence, not only in the presence of outliers, but also for FT time series, there is a gain in terms of forecast accuracy by using a robust approach. Table II indicates that in case of the Student-t errors the MSFE is the smallest for the RHW method. The τ^2 criterion also indicates better forecast performance of the RHW compared to the other methods.

In the simulation results presented here, the selection of the tuning parameters might be influenced by the presence of outliers. Table III presents the average value of the smoothing parameters under the four simulation settings.[2] One sees readily from Table III that, when using the classical HW procedure, the selected values for the smoothing parameters decrease strongly in presence of contamination of the SO and SA type.[3] For example, λ_1 decreases from an average value of 0.36 to 0.14 under AO. Also the HWc and RHW′ methods lack stability here. This is in contrast to RHW, showing that it is not only important to use a robust pre-cleaning algorithm as in (10), but that also robust selection of the smoothing parameter is important.

Model misspecification

One of the reasons exponential smoothing is often used in practice is its good performance under a variety of models. Forecasts based on the Holt–Winters method are optimal at the local linear trend model considered above. Obviously, the true data-generating process may deviate from this model, causing model misspecification. Therefore we compare the forecast performance of the different

[2] All values reported in Table III have a standard error of about 0.002.

[3] The results in Table III for the FT distribution are not directly comparable to those of the (contaminated) normal ones, since also in the absence of sampling error and outliers, their optimal values are different.

DOI: 10.1002/for

methods for other models. It will turn out, however, that the robust smoothing method RHW is still to be preferred.

In the first setting, we simulate according to an ARIMA (1, 1, 1) model given by

$$(1-0.2L)(1-L)\,y_t = (1-0.5L)\,e_t$$

where L denotes the lag operator and the innovations $e_t \sim N(0, 1)$. Secondly, we consider series following a deterministic trend model with correlated errors:

$$y_t = 0.5t + e_t \qquad \text{with} \qquad (1-0.2L)\,e_t = (1-0.5L)\,\delta_t$$

and $\delta_t \sim N(0, 1)$. In both settings, we study clean data (CD) and series contaminated with symmetric outliers (SO). For the latter, we add realizations of an $N(0, 20)$ to on average 5% of the observations. The one-step-ahead forecast errors, with and without outliers, are presented in Figure 2 and the accompanying MSFE and τ^2-values can be found in Table IV. Similar figures are available upon request for the five-step-ahead forecast errors.

It is striking how poor the HWc method performs in the ARIMA clean data setting. This can be explained by the fact that the cleaning process involves fitting a linear trend, while the data contain a stochastic trend. At the deterministic trend model, however, the HWc method performs as expected: a small efficiency loss with respect to HW in the absence of outliers, and an improved performance under contamination. But even in this setting the HWc approach cannot outperform our recommended robust RHW. The four other methods (HW, WR, RHW′ and RHW) perform comparably for the two models if no outliers are present. When we add 5% contamination, the RHW is to be preferred under both forms of model misspecification, as can be seen from the MSFE values in Table IV and the boxplots in Figure 2.

In this subsection we studied the performance of the different procedures when we deviate from the local linear trend model. We can conclude again that RHW performs comparably to standard HW in the absence of outliers, and better than the other robust smoothing methods in the presence of outliers. Obviously, we cannot guarantee that this will happen for all possible time series. In particular, when a level shift is present, there will be some loss of efficiency for the robust approach, since the first few observations after a level shift will be detected as outliers.

Table IV. The MSFE and τ^2-scale for one-step-ahead forecast errors for series generated under an ARIMA model and a deterministic trend model, without (CD) and with (AO) outliers

			HW	HWc	WR	RHW′	RHW
ARIMA(1, 1, 1)	CD	MSFE	1.58	4.25	1.61	1.44	1.67
		τ^2	1.59	2.19	1.48	1.39	1.55
	SO	MSFE	20.03	7.25	11.74	19.71	5.98
		τ^2	9.70	5.22	4.92	5.25	4.94
Deterministic trend	CD	MSFE	1.55	1.61	1.59	1.61	1.61
		τ^2	1.51	1.49	1.55	1.52	1.52
	SO	MSFE	9.95	3.44	5.45	3.81	3.36
		τ^2	5.27	3.44	3.73	3.05	3.09

DOI: 10.1002/for

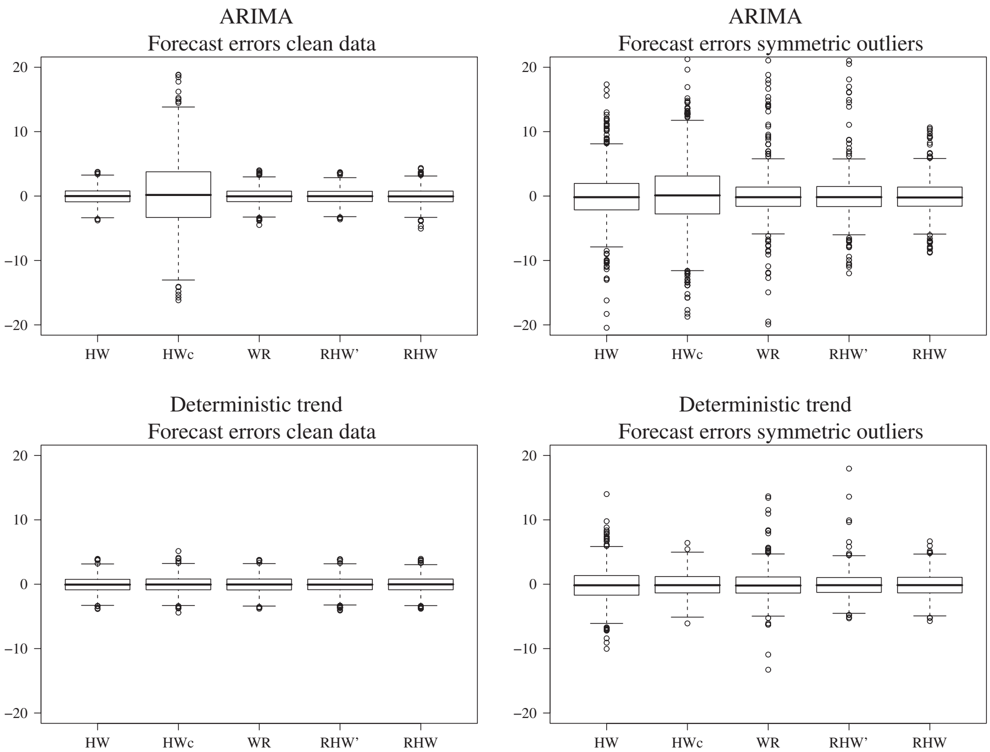

Figure 2. One-step-ahead forecast errors of 1000 replications of a time series of length 100, simulated according to an ARIMA model and a deterministic trend model, with and without outliers. We consider standard Holt–Winters (HW), Holt–Winters with the 2-sigma rule pre-cleaning (HWc), weighted regression (WR), robust Holt–Winters without robust scale (RHW′) and with robust scale (RHW)

REAL DATA EXAMPLE: SEASONALITY

In this section, we apply Holt–Winters smoothing to a real time series with seasonality. The Holt smoothing method given in (2) can easily be generalized to allow for seasonality, as originally described in Winters (1960). A recursive equation for smoothing the seasonal component is then included. The update equations can easily be made robust in the same manner as before, i.e. by replacing y_t by its cleaned version y_t^*, as defined in (10). The resulting recursive scheme is given by

$$\begin{aligned} a_t &= \lambda_1(y_t^* - S_{t-s}) + (1-\lambda_1)(a_{t-1} + F_{t-1}) \\ F_t &= \lambda_2(a_t - a_{t-1}) + (1-\lambda_2)F_{t-1} \\ S_t &= \lambda_3(y_t^* - a_t) + (1-\lambda_3)S_{t-s} \end{aligned} \tag{19}$$

where s is the order of the seasonality. The extra smoothing parameter λ_3 takes values between zero and one, and is fixed at 0.1 throughout the section. In the first equation, a_t is interpreted as the

DOI: 10.1002/for

smoothed series corrected for seasonality. The smoothed series $\tilde{y}_t$ is then defined as the sum of a_t and the smoothed seasonality component S_t:

$$\tilde{y}_t = a_t + S_t$$

and forecasts are obtained by means of

$$\hat{y}_{t+h|t} = a_t + hF_t + S_{t+h-qs}, \qquad q = \lceil h/s \rceil$$

where $\lceil z \rceil$ denotes the smallest integer larger than or equal to z.

To obtain robust starting values for the update equations (19), we run a repeated median regression in the startup period. We obtain a robust regression fit $\hat{y}_t = \hat{\alpha} + \hat{\beta}t$, for $t = 1, \ldots, m$, resulting in starting values

$$a_m = \hat{\alpha} + \hat{\beta}m, \quad F_m = \hat{\beta}, \quad \text{and}$$

$$S_{i+(p-1)s} = \text{Med}\left(y_i - \hat{y}_i, y_{i+s} - \hat{y}_{i+s}, \ldots, y_{i+(p-1)s} - \hat{y}_{i+(p-1)s}\right) \text{ for } i = 1, \ldots, s$$

The length of the startup period m is a multiple p of the seasonal period s: $m = ps$. Here, p equals 3, such that the median in the above equation is computed over three observations, allowing for one outlier among them while still keeping the robustness. Also the simple pre-cleaning method HWc based on ordinary least squares can easily be adjusted by including seasonal dummies in the regression stage.

The Holt–Winters smoothing technique with seasonality is applied to the resex time series (see, for example, Maronna *et al.*, 2006). This series presents sales figures of inward telephone extensions, measured on a monthly basis from January 1966 to May 1973 in a certain area in Canada. The resex time series is plotted in Figure 3. We see two large outliers near the end of the time series,

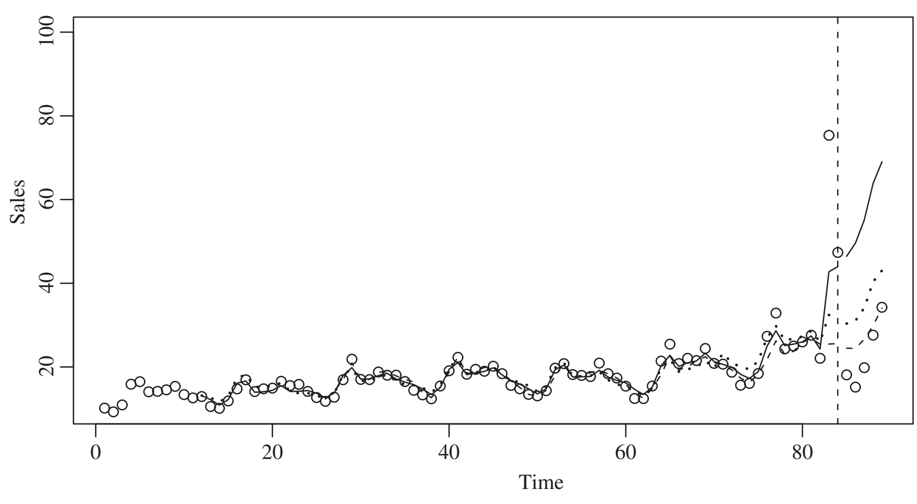

Figure 3. Resex time series(large dots). The Holt–Winters smoothed series up to observation 84 is given, with respective forecasts for observations 85–89, using classical HW (solid line), HWc (dotted line), and the robust RHW method (dashed line)

DOI: 10.1002/for

observations 83 and 84, as the result of a price promotion in period 83 with a spillover effect to period 84. These are expected to strongly affect the forecasts using the classical method. We apply the classical HW and the robust Holt–Winters (RHW) smoothing algorithms to the series up to observation 84 and predict the last five observations. For comparison, we also apply the pre-cleaning method HWc discussed above, now with seasonal dummies included in the regression fit.

The smoothing parameters to be used in the recursive schemes are chosen according to a grid search, where we let λ_1 and λ_2 vary from zero to one in steps of 0.1. The HW method selects $\lambda_1 = 0.3$ and $\lambda_2 = 0.2$, HWc takes $\lambda_1 = 0.1$ and $\lambda_2 = 0.2$, and the RHW method chooses $\lambda_1 = 0.7$ and $\lambda_2 = 0.1$. The robust method takes a much larger value for λ_1, and will smooth the series less than HW and HWc. The standard approach tends to overestimate the persistence in series, due to the presence of the outliers.

The smoothed series and forecasts of the last 5 weeks, according to the HW, HWc and RHW methods, are depicted in Figure 3. Up to the outlying values, the classical and robust smoothed series are very similar. In the forecast period, right after the occurrence of the outliers, the performance of the classical method deteriorates, while the robust smoothing still performs reasonably well. The mean squared forecast error over time horizon five equals 2423 for the classical method. This mean squared forecast error is reduced to 174 for HWc, thanks to the pre-cleaning. However, a further substantial improvement is obtained when the proposed robust Holt–Winters procedure is used, since it achieves a mean squared error of only 37. This illustrates that in real data applications a robust method can considerably improve forecast performance.

CONCLUSION

Exponential and Holt–Winters smoothing are frequently used methods for smoothing and forecasting univariate time series. The main reason of the popularity of these methods is that they are simple and easy to put into practice, while at the same time quite competitive with respect to more complicated forecasting models. In the presence of outlying observations, however, the performance of the Holt–Winters forecasting performance deteriorates considerably. In this paper a robust version of exponential and Holt–Winters smoothing techniques is presented. By downweighting observations with unexpectedly high or low one-step-ahead prediction errors, robustness with respect to outlying values is attained. Moreover, as shown in the simulation study, the robust Holt–Winters method also yields better forecasts when the error distribution is fat-tailed. There is only a small price to pay for this protection against outliers, namely a slightly increased mean squared forecast error at a normal error distribution.

It is important to note that the proposed robust Holt–Winters method preserves the recursive formulation of the standard method. Hence the forecasts are still computed online, without computational effort, using explicit expressions. The choices we made for the loss-function ρ, the tuning constant k, and the auxiliary scale estimate are all standard in the robustness literature. This paper also explains how robust selection of the smoothing parameters, starting values, and auxiliary scale estimates can be dealt with. These choices are crucial for a good overall performance of the forecast methods. Earlier proposals, such as Cipra (1992) and Cipra and Romera (1997), did not treat these issues. The robust method then applies standard Holt–Winters smoothing to the cleaned time series. We also made a comparison with another pre-cleaning method, based on a standard 2-sigma outlier detection rule, and showed that such rules do not provide a satisfactory performance. Applying robust outlier detection rules for cleaning time series (see Pearson, 2005) would yield better results, but

DOI: 10.1002/for

such rules are not applicable to non-stationary series, and do not lead to a simple recursive computing scheme. We consider the latter to be a key property of a Holt–Winters smoothing scheme. The advantages of the cleaning method (10) proposed in this paper are that (i) it can cope with correlated observations, (ii) it can be applied to non-stationary data, also with stochastic trend, (iii) it is robust to large outliers, (iv) it is simple to compute, and (v) its recursive scheme allows for online cleaning of the series. We are not aware of other cleaning methods sharing these properties.

Both the simulation study and the real data example illustrate the advantages of the robust forecasting procedure. The presented robust Holt–Winters method is robust with respect to outliers, results in accurate and stable predictions, and works well for different types of data. This is of great importance for practical business forecasting, where the underlying data-generating process is unknown and outliers frequently occur. Moreover, thanks to the simplicity of the procedure, it can be implemented without much effort with the appropriate software tools.

There are some issues left open for future research. For example, we did not develop formulas for prediction intervals (as in Hyndman *et al.*, 2005). Another issue is the behavior of the method in the presence of level shifts or other forms of structural break. In this case, one would need to allow the smoothing parameters to vary over time, as described in Taylor (2004). Alternatively, more ad hoc detection rules for structural breaks, as in Kirkendall (2006), can be applied.

ACKNOWLEDGEMENTS

This research has been supported by the Research Fund KU Leuven and the Fonds voor Wetenschappelijk Onderzoek (contract number G.0594.05). The second author gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft (SFB 475, 'Reduction of complexity in multivariate data structures').

REFERENCES

Bowerman B, O'Connell R, Koehler A. 2005. *Forecasting, Time Series, and Regression*. Thomson Books Cole: Duxbury Press, Pacific Grove, California.

Cipra T. 1992. Robust exponential smoothing. *Journal of Forecasting* **11**: 57–69.

Cipra T, Romera R. 1997. Kalman filter with outliers and missing observations. *Test* **6**: 379–395.

Davies P, Fried R, Gather U. 2004. Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference* **122**: 65–78.

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

Gardner E. 1999. Note: rule-based forecasting vs. damped-trend exponential smoothing. *Management Science* **45**: 1169–1176.

Gardner E. 2006. Exponential smoothing: the state of the art. Part II. *International Journal of Forecasting* **22**: 637–666.

Holt C. 1959. Forecasting seasonals and trends by exponentially weighted moving averages. *ONR Research Memorandum 52*.

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

Kirkendall N. 2006. Monitoring for outliers and level shifts in kalman filter implementations of exponential smoothing. *Journal of Forecasting* **11**: 543–560.

Kotsialos A, Papageorgiou M, Poulimenos A. 2005. Long-term sales forecasting using Holt–Winters and neural network methods. *Journal of Forecasting* **24**: 353–368.

Kutner M, Nachtsheim C, Neter J, Li W. 2005. *Applied Linear Statistical Models* (5th edn). McGraw-Hill Education: New York.

DOI: 10.1002/for

Makridakis S, Wheelwright S, Hyndman R. 1998. *Forecasting, Methods and Applications* (3rd edn). Wiley: Hoboken, NJ.
Maronna R, Martin R, Yohai V. 2006. *Robust Statistics, Theory and Methods*. Wiley: Hoboken, NJ.
Martin R, Thomson D. 1982. Robust-resistant spectrum estimation. *Proceedings of the IEEE* **70**: 1097–1115.
Masreliez C, Martin R. 1977. Robust Bayesian estimation for the linear model and robustifying the kalman filter. *IEEE Transactions on Automatic Control* **22**: 361–371.
Pearson R. 2005. *Mining Imperfect Data: Dealing with Contamination and Incomplete Records*. Siam: Philadelphia.
Romera R, Cipra T. 1995. On practical implementation of robust kalman filtering. *Communications in Satistics* **24**: 461–488.
Ruckdeschel P. 2001. *Ansatze zur Robustifizierung des Kalman-Filters*. Dissertation, Bayreuther Mathematische Schriften, Bayreuth.
Siegel A. 1982. Robust regression using repeated medians. *Biometrika* **68**: 242–244.
Taylor J. 2004. Smooth transition exponential smoothing. *Journal of Forecasting* **23**: 385–404.
Taylor J. 2007. Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research* **178**: 154–167.
Winters P. 1960. Forecasting sales by exponentially weighted moving averages. *Management Science* **6**: 324–342.
Yohai V, Zamar R. 1988. High breakdown estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* **83**: 406–413.

Author's biographies:
Sarah Gelper is an Assistant Professor of Marketing at the Erasmus School of Economics (Rotterdam, the Netherlands). This article is a result of her research as a PhD student at Katholieke Universiteit Leuven.

Roland Fried is an Associate Professor for Biostatistics in the Department of Statistics at TU Dortmund University (Dortmund, Germany). Before this he was a visiting professor in the Department of Statistics at University Carlos III (Madrid, Spain). His research interests include time series and spatial processes, robustness, change-point detection, and computational statistics.

Christophe Croux is a Professor of Statistics and Econometrics at the Katholieke Universiteit Leuven (Leuven, Belgium). His research interests are robust statistics, multivariate data analysis, computational statistics, predictive marketing modeling, and applied time series analysis. His work has been published in the Biometrika, Journal of the American Statistical Association, and The Review of Economics and Statistics, among others. He serves on the editorial board of Journal of Multivariate Analysis.

Author's addresses:
Sarah Gelper and **Christophe Croux**, Faculty of Business and Economics, Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium.

Roland Fried, Department of Statistics, University of Dortmund, 44221 Dortmund, Germany.

DOI: 10.1002/for