A New Method to Detect Outliers in High-frequency Time Series

Ilaria Lucrezia Amerise[1], Agostino Tarsitano[1]

[1] Department of Economics, Statistics and Finance, University of Calabria, Cosenza, Italy

Correspondence: Ilaria Lucrezia Amerise, Department of Economics, Statistics and Finance, University of Calabria, Cosenza, Italy. E-mail: ilaria.amerise@unical.it

Received: September 29, 2018 Accepted: November 12, 2018 Online Published: November 19, 2018

doi:10.5539/ijsp.v8n1p16 URL: https://doi.org/10.5539/ijsp.v8n1p16

Abstract

The objective of this research is to develop a fast, simple method for detecting and replacing extreme spikes in high-frequency time series data. The method primarily consists of a nonparametric procedure that pursues a balance between fidelity to observed data and smoothness. Furthermore, through examination of the absolute difference between original and smoothed values, the technique is also able to detect and, where necessary, replace outliers with less extreme data. Unlike other filtering procedures found in the literature, our method does not require a model to be specified for the data. Additionally, the filter makes only a single pass through the time series. Experiments show that the new method can be validly used as a data preparation tool to ensure that time series modeling is supported by clean data, particularly in a complex context such as one with high-frequency data.

Keywords: SARIMA models, segmentation, Dice coefficient, simulations

1. Introduction

An important topic in time series analysis is how to deal with data that consist of on-the-minute, hourly, daily or weekly observations. Among the reasons for this interest is that high-frequency time series inevitably show unexpected spikes (peaks and troughs) that appear to be grossly inconsistent with neighboring values. Since occasional large disturbances may have serious consequences for model identification and parameter estimation in time series, it is important to attenuate their adverse effects before the data are used. This paper presents a new filter intended to remove or reduce potentially troublesome behavior in a time series, even though, in the preliminary stage, we ignore the specific model that is eventually to be applied to the data.

Let $p_t \geq 0$ be the observed values at period t and n be the length of the time series $p_t, t = 1, 2, \cdots, n$. We assume that, for each point of time, p_t is given by

$$p_t = \widehat{p}_t + a_t \tag{1}$$

where a_t is a random variable with zero mean and finite variance σ_a^2. The values $\widehat{p}_1, \widehat{p}_2, \cdots, \widehat{p}_n$ lie on a function (the reference curve) that should be flexible enough to represent a wide range of curvatures, but that should also be as smooth as possible. In this article, we detect extreme spikes (or outliers) by examining the absolute difference between observed values and the corresponding point in the reference curve. Therefore, peaks or troughs that are too high are considered anomalous and become candidates for an appropriate statistical treatment. However, only one pass is made through the data because of the length of the time series. The reference curve is obtained by solving the following problem: given a real $\lambda > 0$ and a positive integer m, find the values of $\widehat{\mathbf{p}} = (\widehat{p}_1, \cdots, \widehat{p}_n)$ that minimize the convex combination:

$$Q(\widehat{\mathbf{p}}, m, \lambda) = \frac{(1-\lambda)}{F_m} F(\widehat{\mathbf{p}}) + \frac{\lambda}{S_m} S(\widehat{\mathbf{p}}), \qquad 0 \leq \lambda \leq 1 \tag{2}$$

with

$$F(\widehat{\mathbf{p}}) = \sum_{t=1}^{n} (\widehat{p}_t - p_t)^2, \qquad S(\widehat{\mathbf{p}}) = \sum_{t=m+1}^{n} (\nabla^m \widehat{p}_t)^2 \tag{3}$$

where ∇ denotes the difference $\nabla \widehat{p}_t = \widehat{p}_t - \widehat{p}_{t-1}$. The function (2) has two terms: goodness of fit and smoothness. $F(\widehat{\mathbf{p}})$ measures fidelity to the data in terms of the squared deviations between smoothed and observed values. In particular, F_m is the maximum of $F(\widehat{\mathbf{p}})$, which occurs when all m-th differences are equal to zero. In this case, the reference curve is determined by fitting to $\mathbf{p}$ a polynomial of degree $(m-1)$ by the least squares. For example, if $m = 1$ then $F_m = n \cdot var(\mathbf{p})$, where $var(\mathbf{p})$ is the variance of the observed values. The term $S(\widehat{\mathbf{p}})$ expresses the smoothness as the sum of squares of m-th differences between smoothed values. The constant S_m is the maximum of $S(\widehat{\mathbf{p}})$, which occurs when $\widehat{p}_t = p_t, \forall t$, implying that $S_m = \sum_{t=m+1}^{n} (\nabla^m p_t)^2$. The constants F_m and S_m re-scale $Q(\widehat{\mathbf{p}}, m, \lambda)$ to the [0, 1] interval, so that smoothness

and goodness-of-fit are balanced consistently. In short, the normalized linear filter (2)-(3) can be considered a variant of the Whittaker-Henderson graduation. See Knorr (1984). Central to (2) is the trade-off between $F(\widehat{\mathbf{p}})$, which relates to the goodness-of-fit and $S(\widehat{\mathbf{p}})$, relating to the smoothness. The equilibrium between goodness of fit and smoothness is achieved by a reasoned choice of the smoothing constant λ. If $\lambda \to 0$, then the dominant component will be the squared Euclidean norm $\|\widehat{\mathbf{p}}-\mathbf{p}\|^2$ and $\widehat{\mathbf{p}}$ will increasingly resemble the original values, no matter how irregular $\mathbf{p}$ may be. As $\lambda \to 1$, smoothed prices approach the polynomial $\widehat{p}_t = \sum_{j=0}^{m-1} b_j t^j$ $t = 1, 2, \cdots, n$ regardless of the goodness-of-fit component. A very simple choice is $\lambda = 0.5$, which implies that fidelity and smoothness are equally balanced. Apart from these three cases, the solution of (2) is a serious concern because the degree to which we are justified in sacrificing fidelity in order to obtain smoothness varies greatly from one problem to another. See Whittaker (1923). The paper is organized as follows: in the next section, we present the normalized linear filter (NLF) together with computation of the thresholds beyond which outliers are detected. The effectiveness of the proposed method is assessed in section 3 by comparing the results of SARIMA models fitted to original time series with those of the same models fitted to filtered time series. The final section discusses our findings and points out some improvements for further applications.

2. Optimal Smoothing

The smoothness component of the NLF smoother can be rewritten as

$$\nabla^m \widehat{p}_t = \sum_{j=1}^{n} \mathbf{D}_{m,t,j} \widehat{p}_j, \quad t = 1, 2, \cdots, (n-m), \tag{4}$$

where $\mathbf{D}_m$ is an m-th differencing matrix with $(n-m)$ rows and n columns

$$\mathbf{D}_m = \begin{pmatrix} \mathbf{d}_m & 0 & \cdots, & 0 & 0 \\ 0 & \mathbf{d}_m & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{d}_m & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{d}_m \end{pmatrix}. \tag{5}$$

The matrix $\mathbf{D}_m$ transforms a column vector into the column of m-th differences of the elements in thew vector. A typical row of $\mathbf{D}_m$ contains $n-(m+1)$ zeros and the $1 \times (m+1)$ vector $\mathbf{d}_m = (d_0, d_1, \cdots, d_m)$ starting from column t and ending with column $t+m+1$. The elements $\mathbf{d}_m$ are the successive binomial coefficients of order m with alternating signs

$$d_j = (-1)^j \binom{m}{j}, \ j = 0, 1, 2, \cdots, m. \tag{6}$$

The nonzero elements of $\mathbf{D}_m$ form a diagonal band from the upper left to the lower right. Moreover, thanks to $\mathbf{D}_m$, (2) can be presented as

$$\mathbf{Q}(\widehat{\mathbf{p}}, m, \lambda) = \frac{(1-\lambda)}{F_m} (\widehat{\mathbf{p}} - \mathbf{p})^t (\widehat{\mathbf{p}} - \mathbf{p}) + \frac{\lambda}{S_m} (\mathbf{D}_m \widehat{\mathbf{p}})^t (\mathbf{D}_m \widehat{\mathbf{p}}) \ . \tag{7}$$

To minimize $\mathbf{Q}(\widehat{\mathbf{p}}, m, \lambda)$, its derivatives with respect to the $\widehat{\mathbf{p}}$ have to be equated to zero

$$\frac{\partial \mathbf{Q}(\widehat{\mathbf{p}})}{\partial \widehat{\mathbf{p}}} = 2\left[\frac{(1-\lambda)}{F_m} (\widehat{\mathbf{p}} - \mathbf{p}) + \frac{\lambda}{S_m} \mathbf{D}_m^t \mathbf{D}_m \widehat{\mathbf{p}}\right] = \mathbf{0}, \tag{8}$$

which leads to $[\mathbf{I}_n + \beta(\mathbf{D}_m^t \mathbf{D}_m)]\widehat{\mathbf{p}} = \mathbf{p}$ with $\beta = [\lambda(1-\lambda)^{-1}(F_m/S_m)]$ where $\mathbf{I}_n$ is the $(n \times n)$ identity matrix of order n. The second order condition for a minimum of (7) is

$$\frac{\partial^2 \mathbf{Q}(\widehat{\mathbf{p}})}{\partial \widehat{\mathbf{p}} \, \partial \widehat{\mathbf{p}}^t} = 2\frac{(1-\lambda)}{F_m} \left[\mathbf{I}_n + \beta\left(\mathbf{D}_m^t \mathbf{D}_m\right)\right] \text{being positive definite} \ . \tag{9}$$

It is easy to show that the matrix $\mathbf{A} = [\mathbf{I}_n + \beta(\mathbf{D}_m^t \mathbf{D}_m)]$ is a symmetrical and positive definite and therefore $\widehat{\mathbf{p}} = \mathbf{A}^{-1}\mathbf{p}$. The computation of $\widehat{\mathbf{p}}$ is simple as long as the scheme described above is applied to short time series, but the solution appears problematic for long time series. However, there are fewer difficulties than at first appear. Indeed, several authors, have devised very efficient computing software by exploiting the characteristics of the matrices involved. See, for example, Garcia (2010) and Cornea-Madeira (2017).

2.1 Choice of the Smoothing Constant

There are various techniques for choosing the smoothing constant. For example, Brooks *et al.* (1988) applied the general-

ized cross-validation (GCV) score suggested in Golub & Wahba (1979). Guerrero (2008) outlines a new formalization of the concept of trend to choose the smoothing constant. Frasso & Eilers (2015) proposes the use of two curves for finding the appropriate value of λ automatically. The problem that both methods have is that the optimal value leads to substantial under (over)-smoothing and the resulting long-term pattern tends to have too many (or too few) wiggles that often show up in the wrong places. Other techniques, such as Morozov's discrepancy principle (see, for example, O'Leary (2001)) often work well, but produce quite unexpected results when their outcome is inaccurate. These considerations lead us to a direct calculation of the smoothing constant. As in Knorr (1984), our point of departure is an appealing analogy between $\lambda \in [0, 1]$ and the confidence level of a statistical test where 0.50 or 0.60 are of little interest, but levels such as 0.95 or 0.99 may be important in terms of identifying the appropriate result of a test. To achieve a reasonable degree of smoothness, the recommended constant is the weighted average:

$$\hat{\lambda} = \frac{F_m * 0.95 + S_m}{100\,(F_m + S_m)} \,. \tag{10}$$

The rationale is that when a time series behaves like a polynomial of degree $(m-1)$, then F_m is near its maximum (and hence $S_m \to 0$), we need less smoothness; when the reference values are very similar to the observations (that is, $F_m \approx 0$), we need more smoothness. Therefore, strategy (10) can provide adequate smoothness without producing unnecessarily large deviations from the observed values. At first glance, the interval [0.95, 1] may appear to be narrow, but it is not, because of the extreme reactivity of the NLF toward λ.

2.2 Detection of Extreme Spikes

Let $\widehat{a}_t = p_t - \widehat{p}_t,\ t = 1, 2, \cdots, n$ be the absolute difference between observed values and points on the reference curve. We look for deviations $\widehat{a}_t$ that fall outside the following range

$$\tilde{\mu} - K\tilde{\sigma} < \widehat{a}_t < \tilde{\mu} + K\tilde{\sigma} \qquad t = 1, 2, \cdots, n \tag{11}$$

where $\tilde{\mu}$ is a measure of central tendency, $\tilde{\sigma}$ is a measure of scale and K is a positive multiplier. Since high-frequency time series often contain atypical values, the two statistics need to be robust to the presence of outliers. In this regard, in our experiments, we considered two well-known robust statistics. The measure of location is chosen to be the Sen rank weighted mean (Sen (1964))

$$S_\nu = \left[\binom{\nu}{2j+1}\right]^{-1} \sum_{i=1}^{\nu} \binom{i-1}{j}\binom{\nu-i}{j}\widehat{a}_{(i)} \,, \tag{12}$$

where $\widehat{a}_{(i)}$ is the i-order statistic with $0 < j < (\nu-1)/2$. Our choice is $j = 2$. As a robust statistic of scale we use the first quartile of the sorted pair-wise absolute differences:

$$Q_\nu = 2.21914 \left\{ |\,|\widehat{a}|_i - |\widehat{a}|_j\,|; i < j, |a|_i, |a|_j > 0 \right\}_{(q)} \,, \tag{13}$$

where $q = \binom{n}{2}/4$. See Rousseeuw & Croux (1993). If $\widehat{a}_t$ surpasses the warning limits in (11), then the corresponding value is considered an extreme spike. This, however, does not imply that the spike should automatically be eliminated. The presence of sharp peaks and/or narrow valleys is a rule rather than an exception in high-frequency time series. If too many of them are deleted and/or imputed, by using an average of the remaining data for example, then prediction modes may be applied to an unrealistic time series. It would be better to down-weight dubious observations rather than reject them. We recommend replacing a suspect spike p_t with a linear combination of observed and smoothed values

$$\tilde{p}_t = \gamma p_t + (1-\gamma)\,\widehat{p}_t \qquad 0 < \gamma < 1 \,. \tag{14}$$

In so doing, we preserve the peak or trough nature of the data point. In other words, we assume that the direction of the changes is compatible with the local behavior of the time series, but the magnitude is substantially larger than what is expected under standard conditions. The NLF has three parameters that need to be specified: K, m and γ. We note that the choice of m is not related to the degree of non-stationarity of a stochastic process or even to the representation of a trend by an algebraic curve. Rather, m is associated with the leveling necessary to correct the oscillations: the higher the degree of the polynomial in the smoothness component, the greater is the danger of getting misleading results when using smoothed values. The appropriate value of m, K and γ must be determined experimentally.

2.3 Segmentation

While NLF is a useful tool, it can be slow and inefficient if the time series is so long that the robust statistics of location and scale used in (12)-(13) completely loose their representativeness with respect to the various "local behavior" of time

data. A possible strategy that could be used is a breakdown of long time series into contiguous non-overlapping periods (or segments) and separate, indipendent application of the NLF filter to each segment. Let n_b be the number of segments and let $n_c = \lfloor n/n_b \rfloor$ be the common length of the segments. The pairs of integers $[b_i = 1 + (i-1)n_c, e_i = in_c], i = 1, 2, \cdots, n_b$ indicate the start and the end of each segment, respectively. Note that if $n(u_b n_c) < n$, then n_b is set equal to n. NLF can now be applied to each segment $p_{b_i}, p_{b_i+1}, \cdots, p_{e_i}, i = 1, 2, \cdots, n_b$. Clearly, the segments do not have common elements and together constitute the original time series. In many applications, the calculation of the number of segments and the determination of the boundaries is formulated as an optimization problem. There are various methods and systems that can be used to solve the problem of dividing a time series into periods of similar behavior. See, for example, Keogh *et al.* (2004). Given the variety of requirements that can be proposed in any segmentation procedure and the lack of specific esperiences we restrict our attention to a mere subdivision of the time series into a prefixed number of segments of equal size. The combination of optimal smoothing and extreme spikes detection described in this section has been implemented in an *R* script which is available from the authors upon request.

3. Monte Carlo Analysis

When a filter is applied to a time series, an obvious question arises: how effective is the filter? This section reports a simulation study on the performance of the NLF detection/correction method in a specific example of high-frequency time series (hourly electricity prices) analyzed in the framework of seasonal ARIMA models. Accuracy is assessed relative to the number and size of observations being classified, correctly or wrongly, as outliers. See Janczura *et al.* (2013).

3.1 SARIMA Models

We analyze hourly time series of electricity spot prices from 1am on Friday, 1 January 2016 to 12 pm on Sunday, 31 December 2017, one for each macro-region of the Italian electricity market GME, 2018. Every time series is $n = 17544$ hours long. The dominant seasonality is $s = 24$. For each time series we will identify and estimate a SARIMA$(p, d, q) \times (P, D, Q)_s$ model

$$p_t = \left[\phi^*(B)\right]^{-1} \left[\theta^*(B)\, a_t\right] \ , \tag{15}$$

where $a_t, t = 1, 2, \cdots$, are mutually uncorrelated random variables with zero mean and finite variance σ_a^2. The symbol B denotes the backward shift operator and $\phi^*(B)$ and $\theta^*(B)$ are polynomials

$$\begin{cases} \phi^*(B) &= 1 - \phi_1^* B - \phi_2^* B^2 - \cdots - \phi_{p^*}^* B^{p^*} \\ \theta^*(B) &= 1 - \theta_1^* B - \theta_2^* B^2 - \cdots - \theta_{q^*}^* B^{q^*} \end{cases} , \tag{16}$$

where $p^* = p + sP$, $q^* = q + sQ$ are the orders of the AR and MA polynomials, respectively. If all the roots of $\phi^*(B)$ are greater than one in absolute value, then the process is stationary. What is more, if all roots of $\theta^*(B)$ are greater than one in modulus, with no single root common to both polynomials, the process is invertible. We do not expect to obtain accurate results in terms of fitting ability. One reason for this is that , although hourly electricity prices exhibit numerous seasonalities ranging from daily to weekly to monthly, our study only considered $s = 24$. Moreover, we have ignored many factors that could act as external regressors: holiday effects, temperatures, alternative energy sources and heteroskedasticity. We also ignored the interconnections used for managing possible congestion occurring in the electricity market. Nonetheless, we think that, for the purposes of the present work, it is sufficient just to achieve an acceptable fit to the observed values. We have studied six time series $p_{t,j}, t = 1, 2, \cdots, n; N; j = 1, 2, \cdots, 6$, one for each zone of the Italian electricity market, by using the *auto.arima* function of the *R* package $forecast$ (Hyndman (2015)), which chooses whether to include autoregressive and moving average components (and how many terms to include for each one) by using the AICc criterion. In particular, we use $0 \le p, d, q, P, D, Q \le 2$ which include 729 distinct processes to be explored for each time series. The search for the best model is carried out in a non-stepwise automatic mode with parameters that are constrained to be stationary. The models reported in Table 1 satisfy the suggested criterion.

Table 1. Best SARIMA models for electricity zonal prices in Italy.

Parameter	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6
ϕ_1	0.9030	0.8853	0.8834	1.5789	0.883	1.5128
ϕ_2	–	–	–	-0.5963	–	-0.5315
θ_1	0.1097	-0.0159	0.0073	-0.6875	-0.1581	-0.7843
θ_2	-0.0150	-0.0592	-0.0786	-0.1324	-0.1310	-0.0540
Φ_1	0.2304	0.2044	0.2181	0.1493	0.1774	–
Θ_1	-0.9162	-0.9134	-0.9207	-0.9184	-0.9199	-0.7689
σ_a^2	15.950	24.646	21.342	17.092	50.203	49.710
AICc	98278	105907	103386	99498	118374	118202

3.2 Effects of Smoothing on Point Forecast Accuracy

In order to assess the effects of the normalized linear filter (NLF), we compare some accuracy measures before and after the filter usage. In this regard, the time series analyzed in Table 1 are considered to be the "training" period of our analysis. To this, we add the days from Monday, 1 January 2018 to Friday, 5 January 2018 inclusive (120 hours), which acts as the "validation" period.

The point forecast $\widehat{p}_{n,t,j}$ at origin n and lead time t of the j-th time series is obtained by identifying and estimating a SARIMA process for the training period. The search for the best model is conducted as described in the preceding paragraph. Forecast errors are obtained from the difference between actual values in the validation period and the corresponding forecast produced using the values in the training period: $e_{n,t,j} = p_{n+t,j} - \widehat{p}_{n,t,j}, t = 1, 2, \cdots, L$ where $L = 120$ is the forecast horizon

$$e_{n,l} = \sum_{j=0}^{l-1} \psi_j a_{L+l-j} \qquad \text{where } \sum_{i=0}^{\infty} |\psi_i| < \infty, \psi_0 = 1. \tag{17}$$

To evaluate the specific impact of NLF, we use the coefficient proposed by Hyndman & Koehler (2006)

$$g = \sum_{t=1}^{L} \frac{|e_{n,t}|}{\widehat{Q}} \qquad \widehat{Q} = (L-24)^{-1} \sum_{t=24+1}^{L} |\widehat{p}_{n,t} - \widehat{p}_{n,t-24}| \tag{18}$$

As the numerator and denominator both involve values on the scale of the original data, the quantity g is independent of the scale of the data.

For ease of comparison, we also report the more common forecasting criteria: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). See, for example, Samir & White (2017)

$$\begin{aligned} MAE &= \frac{\sum_{t=1}^{L} |p_{n+t,j} - \widehat{p}_{n,t,j}|}{L}, \quad RMSE = \sqrt{\frac{\sum_{t=1}^{L} \left(p_{n+t,j} - \widehat{p}_{n,t,j}\right)^2}{L}} \\ MAPE &= L^{-1} \sum_{t=1}^{L} \frac{|p_{n+t,j} - \widehat{p}_{n,t,j}|}{p_{n+t,j}} \end{aligned} \tag{19}$$

Table 2 shows the results for a filter with the following parameter' settings : $m = 2$, $\gamma = 0.25$, $K = 5.25, n_b = 4$.

Table 2. Accuracy of point forecasts before and after smoothing.

Stage	Index	Zone 1	Zone 2	Zone 3	Zone 4	Zonel 5	Zone 6
Before sm.	g	284.881	268.166	257.345	236.163	263.925	263.370
	MAE	9.890	9.745	9.542	7.965	14.985	9.941
	RNSE	11.642	11.578	11.419	10.381	19.060	11.759
	MAPE	0.314	0.311	0.305	0.274	0.420	0.315
After sm.	g	268.699	266.991	246.678	231.663	223.858	253.997
	MAE	8.768	8.971	8.445	7.086	12.718	8.854
	RNSE	10.205	10.444	9.995	9.127	16.199	10.316
	MAPE	0.276	0.282	0.269	0.242	0.355	0.278

It can immediately be noticed that the smoothing brings about an improvement in forecasting results in all of the six zones. This is confirmed if one looks at the values of all four accuracy measures before and after the use of the NLF method. The coefficients after smoothing are always lower than the same coefficients computed before smoothing. Although the findings in Table 2 do not appear striking, the fact that the progress, although small, is obtained at a small computational cost should not be ignored.

3.3 Simulation Design

The models in Table 1 are used to generate $N = 250$ time series $\widehat{p}_{t,i,j}, t = 1, 2, \cdots, n; i = 1, 2, \cdots, N;\ j = 1, 2, \cdots, 6$. In this regard, we used *simulate.Arima* of the *R* package $forecast$ (Hyndman, 2015). Successively, the time series are intentionally corrupted with gamma distributed noise, that is, the original data points are replaced with simulated outliers.

In view of the poor fitting, simulated time series may be affected by two types of systematic error. First, some of the $\widehat{p}_{t,i,j}$ may be negative, so contradicting the standard assumption in models of electricity prices. Second, although generated by

a stationary Gaussian process, unwanted spikes are always possible in a very long time series. A necessary consequence is the appearance of "endogenous" spikes, which usually give rise to false positive errors. The simulation design must cancel or attenuate both forms of bias. Our strategy is to do as follows:

- 0. Simulate a time series $\widehat{p}_{i,j}$ from one of the models in Table 1. Set $t = 1$ and $\nu_a = 0$, where ν_a is the number of artificial outliers.
- 1. Remove zero values. Let n^z be the number of simulated values less than or equal to zero and let P^z be the corresponding time points. Set $\widehat{p}_{t\in P^z,i,j} = p_{t\in P^z,i,j}$. Hence, negative values are replaced with the corresponding observed values. If, however, the negative values are more than 5 % of the total length, then reject the time series and return to point 0.
- 2. Modify "endogenous" outliers. Let $Q_{\theta_1} < Q_{\theta_2}$ be the quantiles of $\widehat{p}_{i,j}$ defining the thresholds outside which simulated values may be confused with "exogenous" outliers, but will not be taken into account for accuracy. Let P^L and P^U be the observations in $\widehat{p}_{i,j}$ less than Q_{θ_1} and greater than Q_{θ_2}, respectively. Change $\widehat{p}_{t\in P^L,i,j} = (1 - u_1)\widehat{\mu}_{i,j}$ and $\widehat{p}_{t\in P^U,i,j} = (1 + u_2)\widehat{\mu}_{i,j}$ where u_1 and u_2 are random numbers in the $[0, 0.25]$ interval and $\widehat{\mu}_{i,j} = E(\widehat{p}_{i,j})$. Values which are potentially too low (too high) are replaced with random values near to, but lower than (but greater than) $\widehat{\mu}_{i,j}$.
- 3. Insert an outlier. Set $t_1 = 1 + (t - 1), t_2 = r + (t - 1)$. Compute the mean $\widehat{\mu}_{i_1:i_2}$ and the standard deviation $\widehat{\sigma}_{i_1:i_2}$ of simulated values $\widehat{p}_{t_1,i,j}, \cdots, \widehat{p}_{t_2,i,j}$. Define the bounds $L_t = \widehat{\mu}_{i_1:i_2} - \eta\widehat{\sigma}_{i_1:i_2}$, $U_t = \widehat{\mu}_{i_1:i_2} + \eta\widehat{\sigma}_{i_1:i_2}$. The positive constant η, in practice, governs the range of values entitled to become outliers.
- 4. If $L_t < \widehat{p}_{t,i,j} < U_t$ then $\widehat{p}_{t,i,j}$ it is not a good candidate because the risk of generating a non-detectable outlier is too high. Increase t by one and repeat step 3, provided that $t \leq (n - r + 1)$. Otherwise stop.
- 5. Generate a random number u in the $[0, 1]$ interval. If $u \geq \tau$ then increase t by one and go to step 3. The constant τ controls the rarity of outliers.
- 6. Generate g_t from a gamma(a, β) distribution probability, where $a = \alpha\widehat{\mu}_{i,j}$, $E(g_t) = (\alpha/\beta)\widehat{\mu}_{i,j}$, $var(g_t) = (\alpha/\beta^2)\widehat{\mu}_{i,j}$. If $\widehat{p}_{t,i,j} < \widehat{\mu}_{i,j}$, then set $\widehat{p^*}_{t,i,j} = \widehat{p}_{t,i,j} - g_t$, otherwise set $\widehat{p^*}_{t,i,j} = p_{t,i,j} + g_t$. Increase t and ν by one and return to step 3.

Figure 1 shows an example simulated from the first model in Table 1.

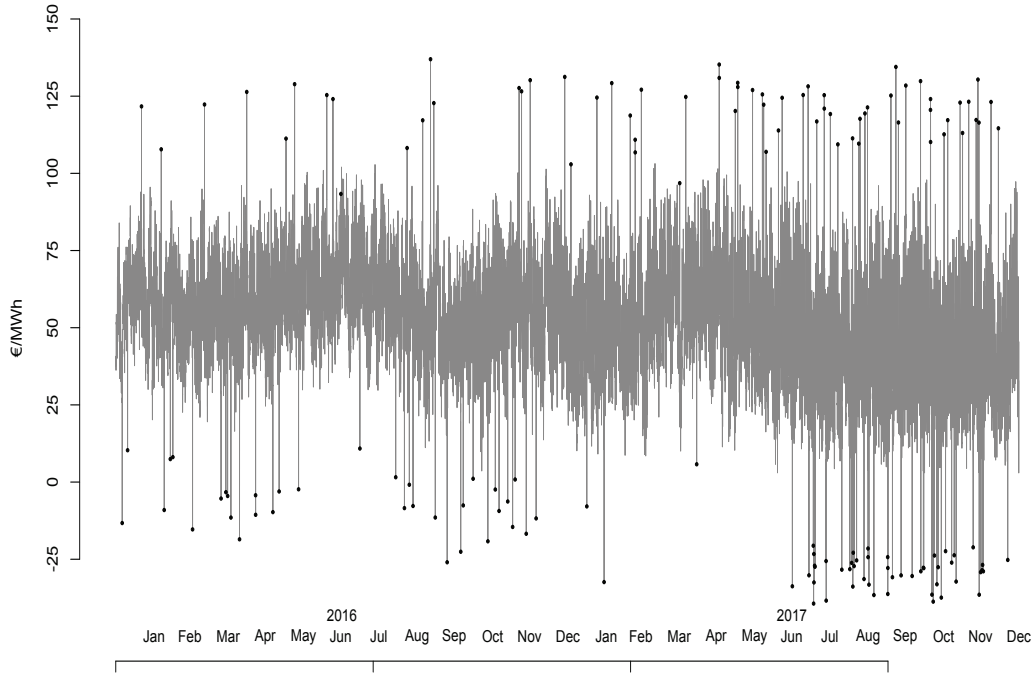

Figure 1. Simulated time series with $\nu_a = 152$ spikes marked in black.
Parameters: $\tau = 0.40$, $\alpha = 1.4$, $\beta = 2$, $\eta = 2.3$, $r = 24$, $\theta_1 = 0.001$, $\theta_2 = 0.999$

One aim of the simulation design is to keep error rates as low as possible, while remaining realistic. Despite the efforts made, not all the simulated spikes can be considered "extreme" in a meaningful sense and, thus, are likely to induce false negative errors. Additionally, false-positive errors remain a possibility when adhering strictly to the simulated time series. The occurrence of both false positive and false negative errors leads to incorrect decisions that could represent a limitation to the present work.

3.4 Performance of the Normalized Linear Filter (NLF)

The ability of NLF to identify outliers can be assessed by comparing the number of artificial outliers ν_a inserted into the time series with the number of outliers identified by the method ν_d. More particularly, we refer to the 2×2 table of the decisions taken

	Detected as outlier	Not detected as outlier
True outlier	True positives (A)	False negatives (B)
Non-outlier	False positives (C)	True negatives (D)

. (20)

A valid anomaly detection technique maximizes decisions of type A while, at the same time, keeping decisions of the types B and C at the lowest levels possible. Obvious measures of performance are

$$C_1 = \frac{A}{A+C}; \qquad C_2 = \frac{A}{A+B}\,. \tag{21}$$

The frequency A of values correctly considered as extreme spikes is central for both the coefficients. The frequency D of non-outliers, not detected as outliers is not included because outliers, by nature, are rare and, consequently, D is much larger than $A, BorC$ and its involvement would give a distorted picture of the degree of success. Coefficients (21) are plausible indices of performance, but have an evident drawback: they are not symmetrical with respect to B and C. It is therefore reasonable to choose some kind of mean of C_1 and C_2. We apply the harmonic mean of C_1 and C_2, known as the "coincidence index" (Dice, 1945).

$$C_3 = \frac{A}{\left[\dfrac{(A+C)+(A+B)}{2}\right]} = \frac{2A}{2A+B+C} = \frac{A}{A+0.5\,(B+C)}\,. \tag{22}$$

We have $0 \leq C_3 \leq 1$ where 0 implies that no outliers are detected and 1 indicates that all, and only, the outliers are detected. The larger the coefficient becomes, the greater the effectiveness of the detection method is. We have evaluated (22) for all the simulation runs (250 time series of $17,544$ time points for each zone). The parameter' setting of the NLF method is the same as in paragraph 3.2. In Table 3 we report mean and standard deviations of ν_s, ν_d, C_1, C_2 and C_3 (averaged over the repetitions). The column headed $\Delta\%$ refers to the relative variation $\Delta\% = (\nu_a - \nu_d)100/\nu_a$ of the true/detected outliers. With an initial general examination, we note the consistent behavior of the mean value of the coincidence index C_3. As expected, C_3 (as well as C_1 and C_2) increases as τ, the frequency of the simulated outliers, increases. Practically the same conclusions apply to the standard deviations of all of the coefficients. Regarding the number of simulated and detected extremes, we observe that the latter is slightly, but systematically higher than the former. This is presumably due to the parameter setting, which is the same throughout the SARIMA models whereas a more differentiated approach seems warranted. Coefficient C_2 is known as sensitivity, that is, the probability the proposed method will discover a real outlier if there is one. In all the repetitions, the average value of C_2 is around 99% (with a negligible standard deviation), so indicating that there are few false negatives. Coefficient C_1 is known as the positive predictive value, that is, the probability that a detected outlier is indeed an extreme spike. Table 3 shows that the average values of C_1 are relatively low (but remain at acceptable levels) only for $\tau = 0.10$, that is, in the case of time series that are contaminated by sporadic outliers. The findings in Table 3 show that the NLF method is capable of great accuracy in detecting extreme spikes in high-frequency time series. The values reached by the Dice coefficient C_3 are particularly remarkable and, hence, we can state that when NLF classifies an observed value as an outlier, it does so with a high degree of reliability.

4. Conclusions and Future Research

The filtering of high-frequency time series removes noise introduced by anomalous conditions, which are mostly self-explanatory and might not contribute significantly to modeling and forecasting. The NLF method described in this paper is a fast, easily applicable and versatile pre-processing treatment of sequences affected by a spike phenomenon. Experi-

Table 3. Comparison of average performance measures over 250 runs.

		Mean						St. Dev.				
τ	Z	ν_a	ν_d	$\Delta\%$	C_3	C_1	C_2	ν_a	ν_d	C_3	C_1	C_2
	2	27	33	18.2	0.901	0.851	0.984	12	16	0.123	0.166	0.046
	3	22	27	18.5	0.923	0.877	0.994	10	16	0.105	0.152	0.024
	4	23	27	14.8	0.912	0.860	0.990	10	14	0.105	0.155	0.035
	5	20	23	13.0	0.946	0.918	0.989	11	18	0.091	0.135	0.031
	6	21	26	19.2	0.912	0.861	0.994	11	17	0.120	0.171	0.022
0.20	1	52	58	10.3	0.928	0.888	0.982	22	25	0.074	0.115	0.043
	2	53	59	10.2	0.935	0.901	0.986	22	29	0.083	0.125	0.035
	3	42	46	8.7	0.947	0.915	0.991	18	20	0.075	0.112	0.031
	4	44	49	10.2	0.950	0.918	0.991	18	23	0.063	0.100	0.022
	5	41	44	6.8	0.965	0.948	0.989	23	33	0.061	0.093	0.027
	6	39	44	11.4	0.950	0.919	0.995	19	23	0.079	0.122	0.015
0.40	1	108	114	5.3	0.950	0.925	0.981	43	47	0.047	0.08	0.032
	2	107	110	2.7	0.954	0.944	0.971	44	45	0.052	0.069	0.061
	3	81	85	4.7	0.971	0.956	0.989	35	37	0.042	0.067	0.025
	4	88	91	3.3	0.967	0.951	0.987	42	43	0.045	0.072	0.028
	5	72	74	2.7	0.978	0.973	0.985	39	41	0.037	0.058	0.029
	6	79	82	3.7	0.977	0.962	0.994	37	41	0.034	0.057	0.014

mental findings show that the proposed methods can efficiently clean long time series and improve their quality. There are currently several ways in which a time series can be smoothed and filtered. One example is the Savitzky-Golay method, which is based on local least-squares polynomial approximation. See Barak (1995) and Shafer (2011) for more details. Another example is the moving weighted average discussed by Borgan (1979) in which each observation consists of a value determined by a local polynomial plus a random error term that satisfies the usual constant variance and uncorrelated assumptions of linear statistical models. A further possibility is the de-noising technique based on wavelets transform (Weron, 2006 [section 2.4.8]). Weron & Zator (2015) also show the validity of a smoother based on the Hodrick-Prescott filter. Finally, there is the *tsoutliers* function(of the R package *forecast*) for the automatic detection and replacement of outliers. We have not compared the aforementioned techniques with the NLS method. This is because of both the lack of software tools to assist application of these methods and the strict dependence of their algorithm on certain parameter settings, which have to be supplied by the user according to the desired strategy and which cannot be generalized beyond their own area of specialization. We plan in the future to compare our results with those obtained through the other techniques looking at a careful design of the experiment, which precludes one method from being preferred merely because data used for the comparison are more in accordance with the theory on which the method is based.

References

Barak, P. (1995). Smoothing and differentiation by and adaptive-degree polynomial filter. *Analytical Chemistry, 67*, 2758–2762. https://doi.org/10.1021/ac00113a006

Borgan, O. (1979). On the theory of moving average graduation. *Scandinavian Actuarial Journal, 3*, 83-105. https://doi.org/10.1080/03461238.1979.10413714

Brooks, R. J., Stone, M., Chan, F. Y., & Chan, L. K. (1988). Cross-validatory graduation. *Insurance: Mathematics and Economics, 7*, 59-66. https://doi.org/10.1016/0167-6687(88)90097-2

Cornea-Madeira, A. (2017). The explicit formula for the Hodrick-Prescott filter in a finite sample. *Review of Economics and Statistics, 99*, 314-318.

Dice, L. R. (1945). Measures of the amount of ecological association between species. *Ecology, 26,* 297–307. https://doi.org/10.2307/1932409

Frasso, G., & Eilers, P. H. C. (2015). L- and V-curves for optimal smoothing. *Statistical Modelling, 15*, 91-111. https://doi.org/10.1177/1471082X14549288

Garcia, D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational Statistics & Data Analysis, 54*, 1167-1178. https://doi.org/10.1016/j.csda.2009.09.020

Gestore dei Mercati Energetici (GME) (Italian National Board for Energy markets) http://www.mercatoelettrico.org/En/Tools/Accessodati.aspx?ReturnUrl=

Golub, G. H., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*, 215-223. https://doi.org/10.2307/1268518

Guerrero, V. M. (2008). Estimating trends with percentage of smoothness chosen by the user. *International Statistical Review/Revue Internationale de Statistique, 76,* 187-202. https://doi.org/10.111 l/j.l751-5823.2008.00047

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*, 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Hyndman, R. J. (2015). Forecast: forecasting functions for time series and linear models. Available online: http://rpackages.ianhowson.com/cran/forecast/

Janczura, J., Trück, S., Weron, R., & Wolff, R. (2013). Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling. *Energy Economics, 38,* 96-110. http://dx.doi.org/10.1016/j.eneco.2013.03.013

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2004). Segmenting time series: a survey and novel approach. In M. Last, A. Kandel, H. Bunke (Eds.), Data Mining in Time Series Databases, Singapore: World Scientific Publishing Co., 1-21.

Knorr, F. E. (1984). Multidimensional Whittaker-Henderson graduation. *Transactions of Society of Actuaries, 36*, 213-240

O'Leary, D. P. (2001). Near-optimal parameters for Tikhonov and other regularization methods. *SIAM Journal for Scientific Computation, 23*, 1161-1171 https://doi.org/10.1137/S1064827599354147

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88,* 1273-1283. https://doi.org/10.1080/01621459.1993.10476408

Samir, S., & White, A. (2017). Short and long-term forecasting using artificial neural networks for stock prices in Palestine: a comparative study. *Electronic Journal of Applied Statistical Analysis, 10,* 14-28. https://doi.org/10.1285/i20705948v10n1p14

Sen, P. K. (1964). On some properties of the rank-weighted means. *Journal Indian Society of Agricultural Statistics, 16*, 5-61

Schafer, W. W. (2011). What is a Savitzky-Golay Filter? *IEEE Signal Processing Magazine, 28,* 111-117. https://doi.org/10.1109/MSP.2011.941098

Weron, R. (2006). Modeling and Forecasting Electricity Loads and Prices. A Statistical Approach. John Wiley & Sons, Chichester, England

Weron, R., & Zator, M. (2015). A note on using the Hodrick-Prescott filter in electricity markets. *Energy Economics, 48* 1-6. https://doi.org/10.1016/j.eneco.2014.11.014

Whittaker, E. T. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society, 41,* 63-73. https://doi.org/10.1017/S0013091500077853