Journal of Statistical and Econometric Methods, Vol.9, No.1, 2020, 47-66
ISSN: 1792-6602 (print), 1792-6939 (online)
Scientific Press International Limited, 2020

# An $L_1$ smoother for outlier cleaning of time series

**Ilaria Lucrezia Amerise[1] and Agostino Tarsitano[2]**

### Abstract

This paper introduces a new robust outlier cleaner specific for high-frequency time series data and provides guidelines for researchers who wish to use this procedure before the analysis process starts . The essence of the method is a fully automatic , data-driven procedure based on fit- ting , by least absolute deviations, a reference function to the actual time series. Once the reference curve has been defined, it can be used to establish bands such that all observations that deviate from the reference curve by more than a prefixed amount will be replaced. Properties of the new screening tool are investigated through the accuracy of simultaneous prediction intervals produced by Box-Jenkins models applied to real data, before and after the outlier cleaner usage. It is shown that the new method can be validly used as a data preparation technique to ensure that statistical analysis is supported by clear-cut data.

**Mathematics Subject Classification:** 90C05; 62M20; 37M10
**Keywords:** Linear programming; simultaneous prediction intervals; electricity prices, pre-processing time series.

[1] Department of Economics, Statistics and Finance; University of Calabria, via P. Bucci, cubo 1/c, Rende, CS. E-mail: ilaria.amerise@unical.it

[2] Department of Economics, Statistics and Finance; University of Calabria, via P. Bucci, cubo 1/c, Rende, CS. E-mail: agostino.tarsitano@unical.it

Article Info: *Received* : July 10, 2019. *Revised* : September 14, 2019.
*Published online* : January 5, 2020

# 1 Introduction

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 on model identification and parameter estimation, it is important to attenuate their adverse effects before data are used. This paper presents a robust smoother intended to detect and then remove or reduce potentially troublesome behaviors in a time series even if, at a preliminary stage, we ignore the specific technique that would be applied to it. The line of reasoning we follow is that a large proportion of spikes are caused by recurrent patterns of change that we cannot afford to ignore or to modify. However, a small numbers of peaks and troughs have such large magnitudes that any use of these observations in model fitting, without an adequate theoretical and empirical knowledge base, would be meaningless. The spirit of such a formulation is that spikes have to be dealt with in a pre-processing stage and not as an integral part of the time series modelling. This paper introduces a new robust outlier cleaner specific for high-frequency time series data and provides guidelines for researchers who wish to use this procedure before the analysis process starts. For this purpose, we assume that a time series simply consists of spikes and other not better-specified aspects that are superimposed upon a reference curve.

$$p_t = \widehat{p}_t + u_t, \qquad t = 1, 2, \cdots, n \tag{1}$$

where $p_t \geq 0$ is the value observed at period $t$ and $n$ is the length of the time series. The reference values $\widehat{\mathbf{p}} = (\widehat{p}_1, \cdots, \widehat{p}_n)$ belong to the reference curve and the residuals $u_t$ are assumed to have zero mean, finite variance and to be not necessarily uncorrelated. Once the reference curve has been defined, it can be used to establish bands such that all observations that deviate from the reference curve by more than a fixed amount will be replaced. The tacit idea is that identification of extreme fluctuations has to be carried out before any forecast technique is implemented. The first statistical task is the choice of the reference values $\widehat{\mathbf{p}}$ that approximate the observed values as well as possible while, at the same time, penalizing curvilinearity of the smoothing behavior. The next step is the construction of two thresholds one from above and one from below, which enables outliers to be detected and replaced. To carry out those tasks, the paper is organized as follows: in Section 2, we present the least absolute deviations smoothing (LADS) and show how a valid filtering be carried out by using linear programming. The method is fully automatic and very robust. Section 3 deals with the detection and mitigation of outliers in time series. Section 4 examines the construction of simultaneous prediction intervals derived from Box-Jenkins models. Section 5, A subsection is devoted to analyze hourly time series of electricity prices on the Italian market. The results are compared to ascertain whether (and, if so, how much) data cleaning affects the accuracy of simultaneous prediction limits (here used for the first time in the literature of filtering time series). The final section discusses our findings and points out some extensions and improvements for further applications of the proposed method.

## 2 Least absolute deviation smoothing

The reference curve is unknown and must be estimated. The primary objective of our study is to develop a function representation that has a smooth nonlinearity and it is close to the observed values. These requirements are antithetical: a curve forced to pass through all the observed values will not be free of any irregularity while a very smooth curve can rarely capture all of the important features of a time series and, therefore, the eventual choice is necessarily a compromise solution. In this paper, we suggest a reference curve, which solves the following problem: given a real $\lambda$ and a positive integer $m$, find the reference values $\widehat{\mathbf{p}}=(\widehat{p}_1,\cdots,\widehat{p}_n)$ that minimize the linear combination

$$Q\left(\widehat{\mathbf{p}}\right)=\lambda F(\widehat{\mathbf{p}})+(1-\lambda)S(\widehat{\mathbf{p}}), \quad 0\leq\lambda\leq 1 \tag{2}$$

with

$$F\left(\widehat{\mathbf{p}}\right)=\sum_{t=1}^{n} w_{1,t}\left|\widehat{p}_t-p_t\right|, \quad S\left(\widehat{\mathbf{p}}\right)=\sum_{t=m+1}^{n} w_{2,t}\left|\Delta^m\widehat{p}_t\right|. \tag{3}$$

where $\Delta$ denotes the difference $\Delta\widehat{p}_t=\widehat{p}_t-\widehat{p}_{t-1}$. The use of the least absolute deviations instead of the conventional least squares is appropriate because the former are less sensitive to big fluctuations in values (outliers) than the latter. See [17], [5] and [4]. The weights $w_{1,t}=w_t/F_{max}, t=1,\cdots,n$ are known non-negative numbers (with at lest one greater than zero) representing the importance given to the distance between reference and actual values associated to a particular period. A zero weight is given to an observations not being considered in the smoothing procedure, and positive weights characterize terms actually used. When one does not have sufficient information to determine a coherent weighting system, then a simple set of equal weights can be enough. For example, [7] found equal weighting to be more valuable for his data.

The constant $F_{max}$ is the maximum value of $F(\mathbf{p})$, which occurs when all $m$-th differences are equal to zero. This implies that $\widehat{\mathbf{p}}$ is determined by fitting a polynomial of degree $(m-1)$ to $\mathbf{p}$ by using the least absolute deviations. The constant $S_{max}$ is the maximum possible value of $S(\mathbf{p})$, which occurs when reference values are equal to observed values and hence $S_{max} = \sum_{t=m+1}^{n} |\Delta^m p_t|$. The constants $F_{max}$ and $S_{max}$ re-scale the objective function $Q_{m,\lambda}(\widehat{\mathbf{p}})$ to the $[0,1]$ interval, so that $\lambda$ consistently balances smoothness against the goodness-of-fit for different time series.

The rationale of (2) is the trade-off between $F(\widehat{\mathbf{p}}_\lambda)$ that is inversely related to goodness-of-fit and $S(\widehat{\mathbf{p}}_\lambda)$ that is inversely related to the smoothness of the reference curve. If $\lambda \rightarrow 1$, then the dominant component will be the normalized city block metric of the residuals and $\widehat{\mathbf{p}}_\lambda$ will increasingly resemble the observed values more closely, no matter how irregular they may be. As $\lambda \rightarrow 0$, $\widehat{\mathbf{p}}_\lambda$ approaches the polynomial $\widehat{p}_{t,m} = \sum_{j=0}^{m-1} b_j t^j \;\; t = 1, 2, \cdots, n$ regardless of the fit component. Apart from these extreme cases, however, the solution of (2) is a serious concern because there does not appear to be an easy, efficient method to account for the two conflicting components.

## 2.1 A cost-parametric linear programming solution

In order to simplify the solution of problem (2), we can replace the differences between smoothed and observed values in the $F$ component at period $t$, say $f_t = \widehat{p}_t - p_t$, with the sum of two non-negative variables:

$$|f_t| = |\widehat{p}_t - p_t| = f_t^+ + f_t^-, \;\; f_t^+ = \begin{cases} f_t & \text{if } \widehat{p}_t \geq p_t \\ 0 & otherwise \end{cases}, \quad f_t^- = \begin{cases} -f_t & \text{if } \widehat{p}_t < p_t \\ 0 & otherwise \end{cases} \tag{4}$$

The same can be done for the component $S$, that is, $|\Delta^m \widehat{p}_t| = |s_t| = s_t^+ + s_t^-$. Given the order of difference $m$, we can formulate (2) as a cost-parametric linear programming problem

$$\begin{aligned} &\min_{\widehat{\mathbf{p}}_m \in R^n} \left(\mathbf{c} + \lambda \mathbf{a}\right)^t \mathbf{x} \\ &\text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \qquad \mathbf{x} \geq \mathbf{0}_{2(2n-m)}, \qquad 0 \leq \lambda \leq 1. \end{aligned} \tag{5}$$

where $\mathbf{x}^t = [(\mathbf{f}^+)^t|(\mathbf{f}^-)^t|(\mathbf{s}^+)^t|(\mathbf{s}^-)^t]$ is a $2(2n-m)$ row vector of “decision variables” and $\mathbf{c}$ and $\mathbf{a}$ are $2(2n-m)$ partitioned vectors of “costs” such that

$$\mathbf{c}^t = \left[\mathbf{w}_1^t \;|\mathbf{w}_1^t \;|\mathbf{0}_{n-m}^t \;|\mathbf{0}_{n-m}^t\right], \qquad \mathbf{a}^t = \left[-\mathbf{w}_1^t \;| - \mathbf{w}_1^t \;|\mathbf{w}_2^t \;|\mathbf{w}_2^t\right] \tag{6}$$

The symbols $\mathbf{f}^+$ and $\mathbf{f}^-$ denote $n \times 1$ vectors and $\mathbf{s}^+$ and $\mathbf{s}^-$ are $(n-m) \times 1$ vectors. The weights $\mathbf{w}_1$ are given by $w_{1,t} = w_t/F_{max}, t = 1, 2, \cdots, n$ and the weights $\mathbf{w}_2$ are given by $w_{2,t} = 1/S_{max}, t = m+1, \cdots, n$. The symbol $\mathbf{0}_{n-m}$ represents an $(n-m) \times 1$ column vector with all components equal to zero. The matrix $\mathbf{A}$ is an $(n-m) \times 2(2n-m)$ partitioned matrix

$$\mathbf{A} = \left[\; \mathbf{D} \;\middle|\; -\mathbf{D} \;\middle|\; \mathbf{I}_{n-m} \;\middle|\; -\mathbf{I}_{n-m} \;\right] \tag{7}$$

where $\mathbf{I}_{n-m}$ denotes the $(n-m)$ identity matrix and $\mathbf{D}$ is a $(n-m) \times n$ banded matrix, *i.e.* the non-zero elements are in a band centered on the main diagonal

$$d_{i,j} = \begin{cases} (-1)^{m+j-i} \binom{m}{j-i} & i = 1, 2, \cdots, m;\; j = i, i+1, \cdots, i+m \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The right-hand side of the equality constraints in (5) is given by $\mathbf{b} = \mathbf{D}\mathbf{p}$ where $\mathbf{D}$ is such that the elements of the vector $\mathbf{b}$

$$b_i = \sum_{j=i}^{i+m} (-1)^{m+j-i} \binom{m}{j-i} p_j \quad i = 1, 2, \cdots, n-m \tag{9}$$

are the differences of order $m$ of the observed values. The matrix $\mathbf{A}$ is assumed to be of full row rank. Smoothed values can be then obtained from the decision variables of the optimal solution as: $\widehat{p}_t = p_t + (f_t^+ - f_t^-)$.

[16] show that the set of admissible values of $\lambda$ can be partitioned into a finite number $\nu$ of subintervals.

$$\widehat{p}_t\left(\lambda\right) = \begin{cases} p_t, \quad t = 1, \cdots, n \text{ if } \lambda = \lambda_0 = 0 \\ (1-\lambda_i)\,\mathbf{w}_1^t\left(\mathbf{f}^+ + \mathbf{f}^-\right) + \lambda_i \mathbf{w}_2^t\left(\mathbf{s}^+ + \mathbf{s}^-\right) \text{ if } \lambda \in [\lambda_{i-1}, \lambda_i), i = 1, \cdots, \nu \\ \sum_{j=0}^{m-1} b_j t^j, \; t = 1, \cdots, n \text{ if } \lambda = \lambda_\nu = 1 \end{cases} \tag{10}$$

The central relationship in (10) means that an optimal basic index set for some fixed value of $\lambda$ would remain optimal for a range of $\lambda$. The parametric linear programming procedure is not difficult to implement (see, for example, [12], [21]). All the calculations are executed by the statistical language R and the scripts are available by the authors upon request.

## 2.2 Choice of the smoothing constant

The choice of $\lambda$ is as important as it is arbitrary. One way to proceed is to solve problem (5) for various values of $\lambda$ and then deciding which value constitutes a good choice on the basis of visual comparisons. A less subjective method is to compute the objective function (2) for a fixed set of values such as $\lambda \in L = (0.01, 0.05, 0.10, \cdots, 0.90, 0.95, 0.99)$. It can be shown that $Q[\mathbf{p}(\lambda)]$ is a positive and concave function of $\lambda$. In Figure 1, we report the relationship between $\lambda$ and $Q[\widehat{\mathbf{p}}(\lambda)]$, which is obtained by evaluating (5) for each $\lambda$ in $L$. The example we consider consists of the hourly spot price on the Italian day-ahead energy market from 01:00 on Monday, 1st January 2018 to 24:00 on Wednesday, 31st January 2018 for a total of $n = 744$ observations.

The curves reveal an inverted $U$-shaped relationship between the minimum of $Q(\widehat{\mathbf{p}}_{m,\lambda})$ and $\lambda$ for each order of differencing. This behavior indicates that, as $\lambda$ growths, the minimum of $Q(\widehat{\mathbf{p}}_{m,\lambda})$ increases, and after reaching a turning point, will diminish. As an operative strategy, we consider the element of the grid where the turning point is located as the best $\lambda$ because, at this level, smoothness turns from a positive into a negative influence on goodness-of-fit. For example, the curve appear to be almost symmetrical only for $m = 1$; thus, only in this case $\lambda = 0.5$ is the optimal choice for the smoothing constant. Needless to say, it is a rather cumbersome and expensive way of finding the smoothing constant, but it has the merit of being completely automatic and data-driven.

## 2.3 Order of differencing

The order of difference is another parameter influencing the smoothing, but we also note that establishing the correct order of differencing is not crucial

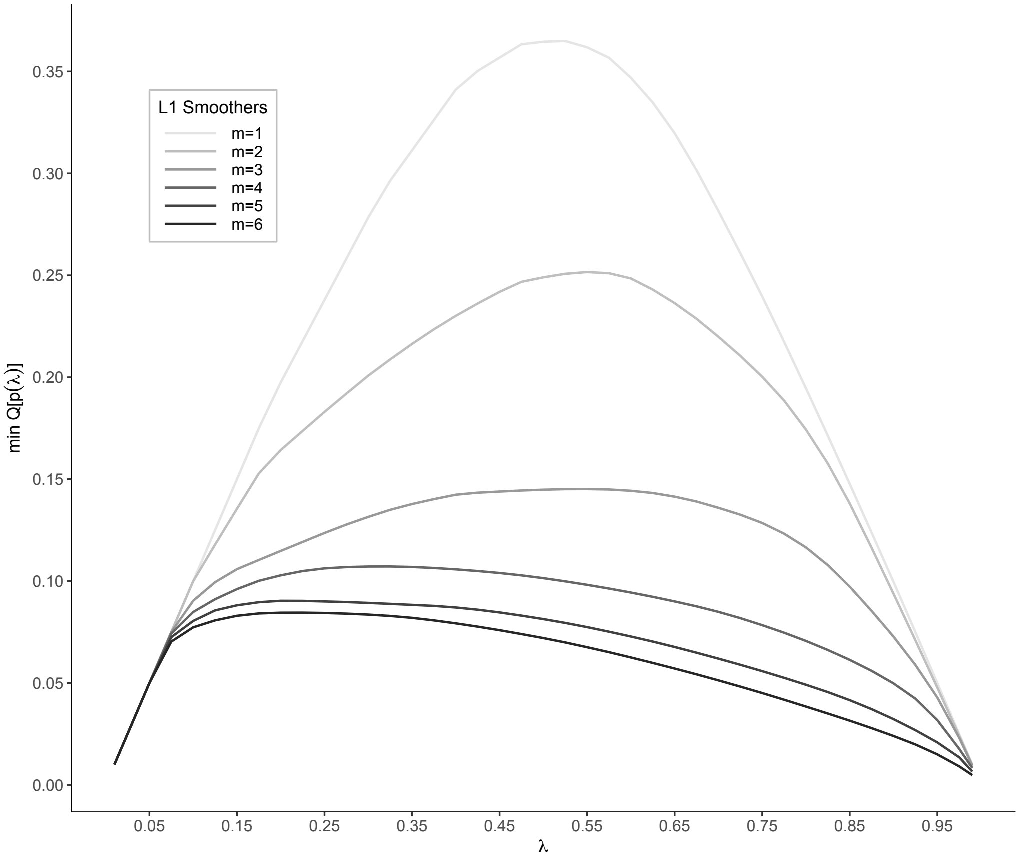

Figure 1: Relationship between the loss function and the smoothing parameter.

to achieve smoothness. In fact, only two special cases: $m = 2$ or $m = 3$, are worthy to be considered. See, for example, [13], [20], [1]. On the basis of several experiments with time series of different complexity, we can state that $m = 3$ is adequate for many problems. In the sequel, we assume $m = 3$, although the computer routine that implements our algorithm will have $m$ as an external parameter.

# 3 Detection and mitigation of outliers

The least absolute deviation smoothing (LADS) method proposed in the preceding section can be particularly useful in the treatment of outliers when there is no a priori information that can be used to identify and remove abnormal measurements. Outliers are detected via the difference between original and reference values $\widehat{u}_t = \widehat{p}_t - p_t$ $t = 1, 2, \cdots, n$ by looking for points that are poorly predicted by the reference curve. In this regard, it is necessary to define lower and upper thresholds for the residuals $\widehat{u}_t$ which delimit what we accept as common cause of variations.

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

where $K$ is a positive multiplier, $\tilde{\mu}$ is a robust measure of central tendency, $\tilde{\sigma}$ is a robust measure of dispersion (robustness is required because the mean and the variance are vulnerable to the influence of outliers). In our experiments, we use the Sen rank weighted mean ([18])

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

where $\widehat{a}_{(i)}$ is the $i$-order statistic with $0 < j < (\nu - 1)/2$. Our choice is $j = 2$ if $\nu > 5$, otherwise $S_\nu = median(|\widehat{u}_t|), |u|_t > 0$. The integer $\nu \leq n$ is the number of residuals that are different from zero (in absolute value). This restriction is necessary because the residuals arising from the solution to the linear programming problem discussed in Section 2.1 contains a certain number of zero values, which if fully included in the computation of the two statistics, would reduce their robustness.

The statistic used as a robust scale estimator is the first quartile of the sorted pair-wise differences between all residuals.

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

where $q = \binom{n}{2}/4$. See [15].

The factor $K$ appearing in the bands (11) establishes the aggressiveness of the LADS in rejecting/replacing outliers. Tolerant choices (large values) of $K$ effectively turn the filter off since no modifications are suggested by warnings. Conversely, aggressive choices (small values) of $K$ lead to the refusal of most of the observed values. We have adopted the traditionally four-sigma rule, that is, $K = 4$, which, in a random sample from a standard Gaussian distribution, would reject the 0.0063% of the units. This means that only 6 observations out of $100'000$, may be expected to lie beyond a distance of $\pm 4\sigma_p$ from the reference curve.

subsectionReplacement of outliers

If a residual $\widehat{u}_t$ surpasses the warning bands, then the corresponding value is considered an outlier. This does not imply that the outlier should automatically be eliminated. The sharp decision of whether to keep or reject an observation is, to some degree, wasteful. While the removal of aberrant values may improve the performance of forecasting models, it may end up suppressing some important feature of the time series. The presence of sharp peaks and/or narrow valleys is a rule rather than an exception in most high-frequency time series. If too many of them are deleted and imputed, for example, using an average of the remaining data, the forecasting tecnique may be adapted to an unrealistic time series without sufficient information about peaks and valleys. We argue that, it would be better to down-weight dubious observations rather than reject them.

Indeed, when considered the relevance of the spikes in time series, we do not want to drastically smooth out such maxima. On the other hand, local minima should be, at least partially, preserved because they could represent particular conditions that need to be accounted for in time series setting. In summary, we propose the replacing of suspect outliers with a weighted average of observed and reference values

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

where $t$ runs over all the periods with values falling outside the fences (11). The greater is $\gamma$, the closer is the averaged value $\tilde{p}_t$ to the observed outlier $p_t$ and the smaller is the contribution of the reference value $\widehat{p}_t$. As $\gamma$ decreases, the strategy (14) yields average values which lie more and more closer to the reference time series, thus exaggerating the role of the smoothing procedure.

Let $r_{p,\widehat{p}}$ denote the Pearson correlation coefficient between $p$ and $\widehat{p}$ and let $\sigma^2_p$ be the variances of observed outliers and $\sigma^2_{\widehat{p}}$ the variance of the corresponding values of the reference curve $\widehat{p}$. We have

$$\sigma^2_{\tilde{p}} = \gamma^2\sigma^2_p + (1-\gamma)^2\,\sigma^2_{\widehat{p}} + 2\gamma\,(1-\gamma)\,\sigma_p\sigma_{\widehat{p}}r_{p,\widehat{p}} \tag{15}$$

Without loss of generality, we set $\sigma^2_p = \theta^2\sigma^2_{\widehat{p}}$, where $\theta > 0$ is a proportionality factor. We obtain

$$\begin{aligned} \sigma^2_{\tilde{p}} &= \gamma^2\theta^2\sigma^2_{\widehat{p}} + (1-\gamma)^2\,\sigma^2_{\widehat{p}} + 2\left(\gamma-\gamma^2\right)\sigma^2_{\widehat{p}}\theta r_{p,\widehat{p}} \\ &= \gamma^2\sigma^2_{\widehat{p}}\left[\theta^2 - 2r_{p,\widehat{p}}\theta + 1\right] - 2\gamma\sigma^2_{\widehat{p}}\,(1-\theta r_{p,\widehat{p}}) + \theta^2\sigma^2_{\widehat{p}} \end{aligned} \tag{16}$$

Differentiating with respect to $\gamma$, and equating to zero we get the minimum of $\sigma^2_{\tilde{p}}$ occurring when

$$\gamma = \frac{1 - r_{p,\widehat{p}}\theta}{\theta^2 - 2r_{p,\widehat{p}}\theta + 1} \ . \tag{17}$$

In fact, the second derivative

$$\frac{d^2\sigma^2_{\tilde{p}}}{dx^2} = 2\sigma^2_{\widehat{p}}\left[\theta^2 - 2r_{p,\widehat{p}}\theta + 1\right] \tag{18}$$

is positive because the discriminant $2\sqrt{r^2_{p,\widehat{p}} - 1}$ of the quadratic equation on the right-hand side of (18) has two complex roots unless $r_{p,\widehat{p}} = 1$, that is, unless the reference values are a proportional transformation of the observed values. It follows that any non trivial average always improves the reference values. Additionally, since $\gamma$ must be positive, then $\theta$ must be less than one, that is, the variance of reference values must be less than the variance of actual values. This is guaranteed by the fact that some of the extreme observations in $p$ have been brought closer to the mean level in $\widehat{p}$.

If the correlation between actual values and reference values $r_{p,\widehat{p}}$ is zero and $\sigma^2_{\widehat{p}}$ is an exact estimate of $\sigma^2_p$ with $(\theta = 1)$ then the appropriate value of $\gamma$ is 0.5 i.e. $\tilde{p}_t$ is the simple mean between $p_t$ and $\widehat{p}_t$. In general, however, $r_{p,\widehat{p}}$ is greater than zero and the optimum value of $\gamma$ is not that obvious. The empirical work undertaken so far has been rather limited. While awaiting additional studies, we suggest $\gamma = 0.25$, which enables the smoothed time series to maintain the shape, up to some degree, if not the magnitude, of maxima (with surrounding peaks) and minima (with surrounding valleys). It must be pointed out that, after performing a vast amount of computation to evaluate (14), we concluded that the question of how to combine observed and reference outliers does not seem to be critical, within certain limits, to forecast accuracy.

In Figure 1, aberrant spikes appearing in the original time series ($20/744 \approx 2.7\%$), are replaced by the values obtained by the LADS method, while preserving peak and valley data points.

# 4 Outlier cleaning in a prediction framework

There are various indicators that can be used to quantify the impact of outlier cleaning. In this section, we evaluate the LADS method by analyzing

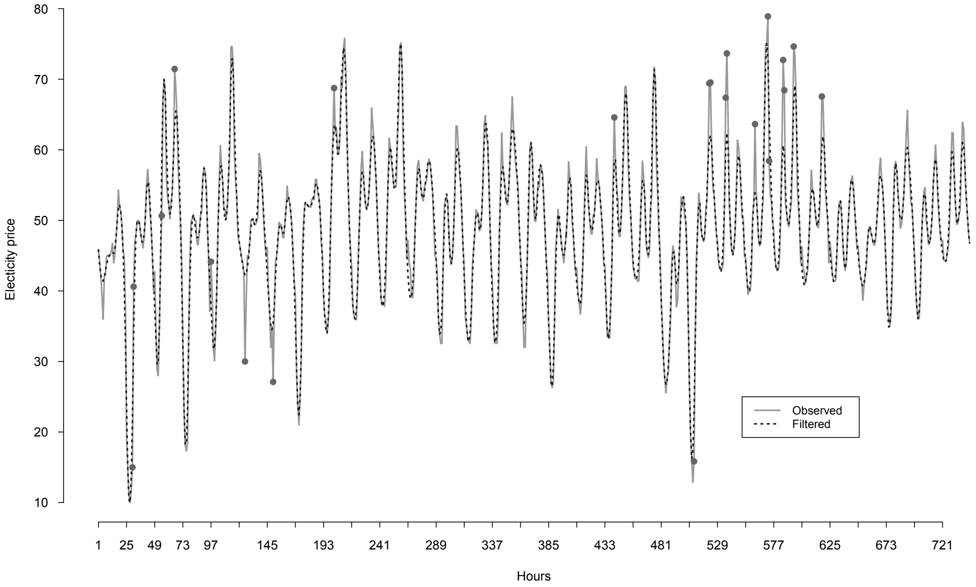

Figure 2: Detection of outliers in hourly single national price; $m = 3, K = 4, \beta = 0.25$

its effects on the accuracy of simultaneous prediction intervals (PIs) derived from Box-Jenkins seasonal models.[11] shows that point predictions are largely unaffected by outliers, provided that the outliers do not occur too close to the origin of the forecasts. However, outliers always affect the estimate of the variance of the residuals, thus impacting the width prediction intervals, as these intervals are proportional to the standard deviation of the residuals.

In this section, we will evaluate the consequences of leaving outliers in the data by splitting a time series into two parts: the “training” period, which ignores a number of the most recent time points, and the “validation” period, which is comprised only the ignored time points and constitutes a separate part of the time series. The training period is used to identify and estimate the model. The validation period is used to test the effects of our smoothing procedure with respect of the effectiveness of simultaneous prediction intervals.

## 4.1 Point predictions

Here, we assume that the time series $p_t, t = 1, 2, \cdots, n$ is adequately represented by a multiplicative seasonal autoregressive moving average with external regressors process (SARMAX)

$$p_t - \left(\beta_0 + \sum_{j=1}^{m} \beta_j X_{t,j}\right) = \left[\phi^*\left(B\right)\right]^{-1} \theta^*\left(B\right) a_t \tag{19}$$

where $a_t, t = 1, 2, \cdots$ , are independent and identically distributed random variables with mean zero and finite variance $\sigma_a^2$, $B$ is the backward shift operator and $\phi^*\left(B\right)$ and $\theta^*\left(B\right)$ are polynomials

$$\begin{cases} \phi^*(B) & = 1 - \phi_1^* B - \phi_2^* B^2 - \cdots - \phi_{p^*}^* B^{p^*} \\ \theta^*\left(B\right) & = 1 - \theta_1^* B - \theta_2^* B^2 - \cdots - \theta_{q^*}^* B^{q^*} \end{cases} \tag{20}$$

where $p^*$ and $q^*$ are the orders of the AR and MA polynomials, respectively. For stationarity and invertibility, it is assumed that the roots of $\phi^*(B)$ and $\theta^*(B)$ lie outside the unit circle, with no single root common to both polynomials. The $X_{t,j}, j = 1, 2, \cdots, k$ are $k$ variables observed on day $t$ influencing the dependent variables; $\beta_j$ is a parameter measuring how the price $p_t$ is related to the $j$-th variable $X_{t,j}$. In order to keep the estimation problem tractable, the exogenous variables are all deterministic functions of time, *e.g.* calendar variables or orthogonal polynomials in time. Of course, in the case of binary variables one of the categories must be omitted to prevent complete collinearity. In each case, the use of binary variables precludes using the difference operators in (19).

An equivalent form of (19) is the infinite moving-average representation

$$e_t = p_t - \left(\beta_0 + \sum_{j=1}^{m} \beta_j X_{t,j}\right) = \sum_{i=0}^{\infty} \psi_i a_{t-i} \qquad \text{where } \sum_{i=0}^{\infty} |\psi_i| < \infty \tag{21}$$

with $\psi_0 = 1$. Coefficients $\psi$s are functions of the parameters $\phi$s and $\theta$s and can be easily obtained through recursive equations (see, for example, [6]). If the parameters of the process are known, then we can compute the minimum mean square error forecast $p_{n,l}$ of the future value $p_{n+l}$ starting from time $n$. The consequent forecast errors are

$$e_{n,l} = \sum_{j=0}^{l-1} \psi_j a_{n+l-j} \tag{22}$$

with $E[e_{n,l}] = 0$ and

$$Cov\left[e_{n,i}, e_{n,j}\right] = \sigma_{i,j} = \sigma_a^2 \left[\sum_{l=0}^{i-1} \psi_l \psi_{j-i+l}\right] \quad i, j = 1, \cdots, H. \tag{23}$$

where $H$ is the time horizon of the forecasts. Note that forecasting the regression term in (19) does not present particular difficulties because we have assumed the perfectly predictable nature of the regressors.

## 4.2 Simultaneous prediction intervals

In order to assess the effectiveness of our robust smoother, we compare the capacity of prediction intervals (PIs) to contain all $H$ future values, both in the presence and absence of smoothing. The scope is to determine two bands

$$P\left[\bigcap_{l=1}^{H} \left(C^1_{l,\alpha} \leq p_{n+l} \leq C^2_{l,\alpha}\right)\right] = 1 - \alpha \ . \tag{24}$$

such that the probability of consecutive future values $p_{n+l}, l = 1, 2, \cdots, H$ lying simultaneously within their respective range is (1–$\alpha$). The limits in (24) are $C^1_{l,\alpha} = p_{n,l} - c_\alpha \sigma_l$ and $C^2_{l,\alpha} = p_{n,l} + c_\alpha \sigma_l$. The quantity $\sigma_l^2$ is the residual variance at the $l$-th lead time

$$\sigma_l^2 = \sigma_a^2 \sum_{l=0}^{l-1} \psi_l^2 \qquad l = 1, 2, \cdots, H \ . \tag{25}$$

$c_\alpha$ is a quantile of the joint probability distribution of forecast errors

$$G^{-1}(c_\alpha) = 1 - \alpha \qquad \text{with} \quad G\left(c_\alpha\right) = Pr\left(|z_l| \leq c_\alpha, \ l = 1, 2, \cdots, L\right) \tag{26}$$

where $z_l = e_{n,l}/\sqrt{\sigma_l^2}, l = 1, \cdots, H$ are the standardized forecast errors. The computation of $c_\alpha$ requires an explicit hypothesis about the distribution of the forecast errors. More specifically, we assume that $G$ is the $H$-variate Gaussian distribution.

$$Pr\left(|z_l| \leq c_\alpha, \ l = 1, 2, \cdots, H\right) = \int_{-c_\alpha}^{+c_\alpha} \cdots \int_{-c_\alpha}^{+c_\alpha} f\left(z_1, \cdots, z_L\right) dz_1 \cdots dz_H \ . \tag{27}$$

Simultaneous PIs guarantee that the $H$ individual intervals include the respective expected value with a confidence level of (1–$\alpha$). See [6]. If $a_t$ is a Gaussian process and the $\phi$, $\theta$ and $\sigma_a^2$ coefficients are known, then $(z_1, \cdots, z_H)$ have an $H$-variate Gaussian distribution with mean vector $\mathbf{0}_H$ and correlation matrix

$$\boldsymbol{\Sigma} = (\rho_{i,j}) = \frac{\sum_{l=0}^{i-1} \psi_j \psi_{j-i+l}}{\sqrt{\sum_{l=0}^{i-1} \psi_l^2}\sqrt{\sum_{l=0}^{j-1} \psi_l^2}} \qquad i < j \ . \tag{28}$$

If the forecast errors are independent and identically distributed, then $\boldsymbol{\Sigma} = \sigma_a^2 \mathbf{I}_H$. In this case it is legitimate to use of the marginal prediction intervals

$$p_{n,l} \pm z_{\alpha/2}\sigma_l, \qquad l = 1, \cdots, H. \tag{29}$$

where $z_{\alpha/2}$ is the upper $\alpha$-th quantile of the univariate standard Gaussian distribution [3, sec. 5.2.4]. However, the hypothesis of independent or even uncorrelated forecast errors is illusory and has no validity in practical situations. Therefore, unless the observed values $p_{n+l}, l = 1, \cdots, H$ develop according to a known pattern, the probability that a given sequence lies completely inside all $H$ marginal PIs would be less than $100(1-\alpha)$, especially if $H$ is large. This is the reason why we have focused our efforts on simultaneous PIs.

It is evident that $c_\alpha$ in place of $z_{\alpha/2}$ is the only difference between the two types of limits, but marginal PIs do not vary with the forecast horizon. In this sense, [19] noted that, for example, at $\alpha = 0.05$ (i.e., at a coverage probability of 95%) the $\alpha/2$-quantile is 1.96 for all leading times. On the other hand, for simultaneous PIs, $c_{0.05}$ increases with $H$. For $H = 20$, we have $c_{0.05} = 2.8004$ which makes the extent of the spurious narrowness of marginal PIs clear. See [14].

## 4.3 Evaluation of PIs

The most important characteristic of PIs is their actual coverage probability (PIAC). We measure PIAC by the proportion of true values of the validation period enclosed in the bounds

$$PIAC_\alpha = 100H^{-1}\sum_{l=1}^{H} c_{l,\alpha} \quad \text{where } c_{l,\alpha} = \begin{cases} 1 & \text{if } p_{n+k} \in \left[C_{l,\alpha}^1, \quad C_{l,\alpha}^2\right] \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

If $PIAC_\alpha \geq (1-\alpha)$ then future values tend to be covered by the constructed bands, but this may also imply that the estimates of the variances in the forecast errors are positively biased. A $PIAC_\alpha < (1-\alpha)$ indicates under-dispersed forecast errors with overly narrow prediction intervals and unsatisfactory coverage behavior.

All other things being equal, narrow PIs are desirable as they reduce the uncertainty associated with forecast-based decision-making. However, high accuracy can be easily obtained by widening PIs. A complementary measure that quantifies the sharpness of PIs might be useful in this context. Here, we use the score function.

$$R_{l,\alpha} = \left(\frac{1-\alpha}{2}\right)\frac{\left(C^2_{l,\alpha} - C^1_{l,\alpha}\right)}{p_{n+l}}, \quad l = 1, 2, \cdots, H. \tag{31}$$

This expression reflects a penalty proportional to the narrowness of the intervals that encompass the true values at the nominal rate. The penalty increases as $\alpha$ decreases, to compensate for the tendency of prediction bands to be broader as the confidence level increases. Of course, the lower $S_{l,\alpha}$ is, the more accurate PI will be. The average value of the score width across time points

$$ASW_\alpha = \frac{1}{H}\sum_{l=1}^{L} R_{l,\alpha} \tag{32}$$

can provide general indications of PIs performance.

# 5 Empirical Analysis

In this section, we perform an experimental evaluation of our method. In particular, we examine $144 = 24\times 6$ daily time series of prices, one for each hour of the day and each zone of the Italian electricity market. Due to transmission capacity constraints, Italy is partitioned into six zones: North, Center-North, Center-South, South, Sardinia and Sicily with a separate price for each zone. When there is no transmission congestion arbitrage opportunities force the prices in each zone to be equal. See [9]. All the time series are long $1'976$ days long, but the last three weeks ($H = 21$) are reserved for assessing the accuracy of PIs. Thus, only the first $1,895$ days are used for estimation and validation of the models.

Parameters can be estimated by optimizing the log-likelihood function of (19), provided that $p$, $q$, $P$, $Q$ are known and errors are Gaussian random variables. Since we ignore the order of the polynomials, the estimation is repeated for different values of $p,\ q,\ P\ and\ Q$. The search of the best SARMAX model is conducted within the bounds $0 \leq p, q, P, Q \leq 3$ which include 256 distinct processes to be explored for each time series. Note that $p^* = p + sP$, $q^* = q + sQ$. The search for the best model is carried out in non-stepwise automatic mode using the *auto.arima* function of the $R$ package $forecast$ ([10]) with parameters constrained to be stationary. It should be pointed out that PIs tend to perform poorly when the residuals are not Gaussian. In consequence, even under the most favorable conditions, the PIs in (24) are de facto "approximate" PIs.

To compute (27), we apply the method proposed by [8]. Table 1 shows the results at the confidence levels $(80, 85, 90, 95)$. Columns labeled "none" display the actual prediction interval coverage rate (PIAC) and the average width (ASW) when time series have not undergone a pre-processing stage. Columns labeled "LADS", display the analogous results obtained after applying our robust smoother. In the initial general examination, we note the consistency of the behavior of PIAC and ASW with the latter decreasing as the former increases, for each zone, either in the presence or absence of filtering. Naturally, this is a confirmation of the expected behavior of the score function (32). What appears immediately clear are the notable differences between the various zones, reflecting the climatic diversity of geographical zones, different size of the zones and price differentials (see, e.g., [2]). It is no coincidence that the most negatively affected zones are the problematic large islands of Sicily and Sardinia, which suffer from poor interconnections and frequent congestion.

To have an idea of the effects of LADS in reducing the impact of price spikes on forecasting, we compare the narrowness of the prediction intervals reported in the columns headed $\mathrm{ASW}_\alpha$ at the first level. Figures shown in column "LADS" are systematically and significantly lower than those shown in column "none". Additionally, more precise forecasts are obtained without appreciably reducing the coverage rate. The performances of SARMAX models, combined with the LADS, appear to be moderately satisfactory with respect of the improved accuracy and efficiency of the prediction intervals. The main result is that, in the absence of smoothing, SARMAX models consistently yield PIs with greater than nominal coverage rates. The robust smoother corrects

Table 1: Improvement in the accuracy of prediction intervals.

| | | $PIAC_\alpha$ | | $ASW_\alpha$ | |
|---|---|---|---|---|---|
| Zone | $(1-\alpha)$% | None | LADS | None | LADS |
| 1 | 80 | 87.66 | 82.16 | 13.65 | 10.27 |
| | 85 | 91.41 | 87.01 | 11.29 | 8.62 |
| | 90 | 94.37 | 93.05 | 8.42 | 6.51 |
| | 95 | 96.93 | 95.46 | 4.87 | 3.76 |
| 2 | 80 | 89.83 | 83.17 | 15.33 | 11.04 |
| | 85 | 92.00 | 87.60 | 12.61 | 9.33 |
| | 90 | 95.15 | 93.44 | 9.43 | 7.01 |
| | 95 | 97.92 | 96.06 | 5.45 | 4.04 |
| 3 | 80 | 93.58 | 90.22 | 17.73 | 11.48 |
| | 85 | 94.96 | 93.05 | 14.63 | 9.56 |
| | 90 | 96.73 | 96.26 | 10.87 | 7.17 |
| | 95 | 98.32 | 97.67 | 6.35 | 4.13 |
| 4 | 80 | 95.55 | 90.82 | 17.48 | 11.85 |
| | 85 | 96.35 | 93.05 | 14.54 | 9.91 |
| | 90 | 97.13 | 95.26 | 10.84 | 7.46 |
| | 95 | 98.12 | 97.06 | 6.27 | 4.30 |
| 5 | 80 | 99.10 | 97.22 | 28.26 | 19.27 |
| | 85 | 99.10 | 98.61 | 23.37 | 15.80 |
| | 90 | 99.10 | 99.21 | 17.39 | 11.72 |
| | 95 | 99.10 | 99.60 | 10.11 | 6.75 |
| 6 | 80 | 98.71 | 97.22 | 22.85 | 13.65 |
| | 85 | 98.71 | 98.61 | 18.91 | 11.15 |
| | 90 | 98.90 | 98.81 | 14.04 | 8.37 |
| | 95 | 98.90 | 99.21 | 8.09 | 4.86 |

the coverage rates, but not in a way to alter the impression of over-dispersed forecast errors. This is an unwanted conservatism, primarily due to inflated estimates of the forecast error variances, which, in turn, can be attributed either to unsuspected behavior of the time series in the validation period, or to the length of the forecast horizon or, ultimately, to the weakness of the usual Box-Jenkins approach, when applied to electricity price time series.

# 6 Conclusions and future research

If one fits a model to a time series that has not been properly filtered, then important temporal patterns could remain buried in the noise and outliers might have detrimental effects on forecast accuracy. As we have seen, the LADS method is not only effective in favoring the best conditions for the application of forecasting models, but also have neutral or inhibiting effects when the tuning constants are appropriately chosen. Naturally, we do not claim that our method achieves the best, or even a satisfactory, result under all circumstances, or even under most. Nonetheless, it does have the advantage that it reduces the width of simultaneous prediction intervals deriving from SARMAX models while keeping the coverage rate close to the nominal level. As such, our robust smoother adds a very promising new methodology to the data analysis toolbox within the area of statistical data cleaning.

In the future, we intend to apply the LADS method to other forecasting techniques such as Holt-Winters, trend-seasonal decomposition, artificial neural networks. In addition, we plan to compare our results with those obtained with other smoothers looking at a careful design of the experiment which allows to overcome the main challenge posed to smoothing methods comparisons. Most specifically, it must be ensured that none of the methods should have an unfair advantage merely because the selected tuning constants and/or the data used for the applications are more in accordance with the model on which the method is based.

# References

[1] I.L. Amerise and A. Tarsitano, A new method to detect outliers in high-frequency time series, *International Journal of Statistics and Probability*, **8**, (2019), 16–24.

[2] S. Bigerna and C.A. Bollino, Ramsey prices in the Italian electricity market, *Energy Policy*, **88**, (2016), 603–612.

[3] G.E.P. Box and G.M. Jenkins, *Time Series Analysis: Forecasting and Control*, San Francisco, Holden-Day, 1976.

[4] F.Y. Chan, L.K. Chan, J. Falkenberg and M.H. Yu, Applications of linear and quadratic programmings to some cases of the Whittaker-Henderson graduation method, *Scandinavian Actuarial Journal*, (1986), 141–153.

[5] F.Y. Chan, L.K. Chan and M.H. Yu, A generalization of Whittaker-Henderson graduation, *Transactions of Society of Actuaries*, **36**, (1984), 183–211.

[6] S.H. Cheung, K.H. Wu and W.S. Chart, Simultaneous prediction intervals for autoregressive-integrated moving-average models: A comparative study, *Computational Statistics* & *Data Analysis*, **28**, (1998), 297–306.

[7] M. Chiodi, A partition type method for clustering mixed data, *Rivista di Statistica Applicata*, **2**, (1990), 135–147.

[8] A. Genz, Numerical computation of multivariate normal probabilities, *Journal of Computational and Graphical Statistics*, **1**, (1992), 141–149.

[9] A. Gianfreda and L. Grossi, Forecasting Italian electricity zonal prices with exogenous variables, *Energy Economics*, **34**, (2012), 2228–2239.

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

[11] J. Ledolter, The effect of additive outliers on the forecasts from ARIMA models, *International Journal of Forecasting*, **5**, (1989), 231–240.

[12] T.L. Magnanti and J.B. Orlin, Parametric linear programming and anticycling pivoting rules, *Mathematical Programming*, **41**, (1988), 317–325.

[13] A.S. Nocon and W.F. Scott, An extension of the Whittaker-Henderson method of graduation, *Scandinavian Actuarial Journal*, **1**,(2012), 70–79.

[14] N. Ravinshanker, Shao-Yen Wu and L. Glaz, Multiple prediction intervals for time series: comparison of simultaneous and marginal intervals, *Journal of Forecasting*, **10**, (1991), 445–463.

[15] P.J. Rousseeuw and C. Croux, Alternatives to the median absolute deviation, *Journal of the American Statistical Association*, **88**, (1993), 1273–1283.

[16] T. Saaty and S. Gass, Objective function (part 1), *Journal of the Operations Research Society of America*, **2**, (1954), 316–319.

[17] D.A. Schuette, A linear programming approach to graduation, *Transactions of Society of Actuaries*, **30**, (1978), 407–431.

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

[19] J. Siu-Hang Li and W-S. Chan, Time-simultaneous prediction bands: A new look at the uncertainty involved in forecasting mortality, *Insurance: Mathematics and Economics*, **49**, (2011), 81–88.

[20] R. Weron and M. Zator, A note on using the HodrickPrescott filter in electricity markets, *Energy Economics*, **48**, (2015), 1–6.

[21] Y. Yao and Y. Lee, Another look at linear programming for feature selection via methods of regularization, *Statistical computing*, **24**, (2014), 885–905.