Statistical Journal of the IAOS 40 (2024) 389–402
DOI 10.3233/SJI-230109
IOS Press



# Outlier identification and adjustment for time series

Markus Fröhlich*
*Statistics Austria, Vienna, Austria*

**Abstract.** Identification and replacement of erroneous data is of fundamental importance for the quality of statistical surveys. If statistical units are continuously sampled over an extended period, time series methods can facilitate this task. Numerous outlier identification and replacement procedures are accessible for this particular purpose, like RegArima Approaches within the seasonal adjustment procedures in X13-Arima or Tramo/Seats. These algorithms can be used to identify different types of outliers, like additive outliers, level shifts or transitory changes. In this paper an alternative outlier identification procedure is proposed which is based on a nonlinear model estimated with support vector regressions. The focus of this procedure is on the identification of additive outliers and on the applicability for short time series with less than 3 years of observations.

Keywords: Time series, X13 arima/seats, tramo/seats, outlier adjustment, support vector regression

## 1. Introduction

Statistical Institutions are requested to provide results of their surveys in increasingly shorter time intervals. Yet, their standard data editing procedures are not always conceived to process the data at these early time limits. However, early available data which is used for grossing up or for econometric modeling has to be checked for data errors in advance. An automatic outlier identification and replacement procedure would be important for this purpose.

The outlier identification and replacement procedure proposed in this paper was motivated by the nowcasting procedure (see [1] for a definition of nowcasting) for short term statistics implemented at Statistics Austria. This method is based on thousands of unedited original records provided by early responding enterprises and therefore eliminating data errors or at least minimizing their effect influences the quality of nowcasting results. Time series methods were employed for the identification of outliers because in this specific case the population remains constant over time and is surveyed at regular intervals. The following requirements were considered specifically important in this respect:

- fully automated data processing because of the large number of time series,
- applicability to long and short time series,
- considering seasonality because data is sampled monthly (quarterly),
- focus on point outliers.

Practically, automatic detection of erroneous data is not feasible. If suspicious data points were identified through an automatic procedure the next step would be a validity check of these data. Extreme data points can readily arise from exceptional circumstances, and prove correct after verification. These data should be kept unchanged and not be edited. In order to maximize rates of correctly adjusted data and to minimize false alarm ratios critical values for outlier identification can be calibrated heuristically. The focus in this paper is on the detection of data errors at specifically early stages of the data editing process. Therefore, the procedure has to run fully automatically without any possibility to intervene manually. Moreover, this approach is limited to point outliers, because after some time-lag, longer series of consecutive extreme values should be detected by standard data editing procedures. If the position of outliers were known in advance they could easily be considered in a time series model. However, because this is not the case in the intended scenario, a model

*Corresponding author: E-mail: markus.froehlich@statistic.gv.at.

which is robust to outliers was implemented. A robust model would not be influenced by outlying observations and therefore, outliers could be identified via the residuals, which should exhibit large peaks or valleys at their positions.

### *1.1. Methods for outlier identification and replacement*

Numerous time series methods are available for outlier treatment, like those based on [2] or extensions. The seasonal adjustment programs X13 Arima and Tramo/Seats offer automatic outlier detection and – replacement, based on the work of [2] with extensions proposed in [3]. Different kinds of outliers like point outliers, transitory changes or level shifts can be identified with these procedures. However, these methods are limited to seasonal time series with a minimum length of 36 monthly observations and 16 observations for quarterly data respectively. An alternative outlier identification and replacement procedure based on [4] is provided by the R-Package `tsoutliers`. Like X13 or Tramo, this package offers the identification of point outliers, transitory changes and level shifts, and can be applied to shorter time series with less than 12 monthly observations as well. An outlier adjustment function, `tsoutliers` (not to confuse with the R-Package `tsoutliers` from above), implemented in the R-Package `forecast`, was provided by [5]. This procedure decomposes the time series into seasonal, trend and a remainder component by multiple seasonal trend decomposition (MSTL, see [6]). For nonseasonal series only a trend and remainder component are estimated. Outliers are identified by analyzing the distribution of the remainder component. Alternatively, [7] developed an approach based on the discrete wavelet transform of a time series. The method is capable to identify additive outliers and innovational outliers. [8] proposed a robust method for the detection of outliers in time series. Contrary to X13 or Tramo/Seats, they proposed a purely deterministic model, which was developed for fraud detection and it is applicable for very short time series as well. In the next subsection an introductory example, which should demonstrate the principal idea of the proposed method is provided. The basic model used in this approach is based on the method of [8].

### *1.2. Principal approach for outlier identification*

The procedure discussed in this paper addresses economic time series which are sampled monthly or quarterly. In this respect, a time series model including a trend- and seasonal component was considered to be appropriate. Moreover, the model should be robust to the occurrence of outliers, whose position and magnitude is unknown in advance. It was expected, that a robust model would fit actual economic time series well with the exception of outlying observations. Larger deviations between actual data and fitted data will occur at the position of outliers. Extreme values are easy to separate from the rest of the data.

For illustration of this approach, the famous airline data set (see [9]), "enriched" with 3 randomly added outliers is depicted in the top panel of Fig. 1a. The model described in the following sections was fit to this time series. The time series of the original data (solid line) was well captured by the model (dashed line) with the exception of the 3 outliers, which did not influence the fitted line. Plotting the residuals with adequately chosen thresholds,[1] outliers could be identified correctly (see Fig. 1a lower panel).

The rest of this paper is structured as follows: the specification of the basic model for outlier identification is presented in Section 2.1. The treatment of specific issues like changing seasonal patterns or the occurrence of level shifts in the time series are discussed in Sections 2.3 and 2.4 respectively. Section 3 discusses the practical identification of extreme values from the residuals of the model. Two different approaches are presented in Subsections 3.1 and 3.2. Section 4 reflects on the setup of the simulation study (Section 4.1), the competing benchmark methods (Section 4.2), the settings for the outlier identification method presented in this paper (Section 4.1.1) and the results of a simulation study (Section 4.3). Concluding remarks summerize the findings of this paper (Section 5).

## 2. Modelling time series for outlier detection

As discussed above, the approach proposed in this paper relies on identifying outliers from the residuals of a fitted (robust) model. The basic model, introduced in Section 2.1, is a purely deterministic model which is rather restrictive as both, trend and a seasonal pattern are assumed to be fixed for the whole time span. Hence, in Sections 2.2.1 and 2.3 extensions of this model are discussed, that allow for a seasonal pattern that changes over time and the inclusion of level shifts.

[1] See Section 3 for details on the calculation of threshold values.

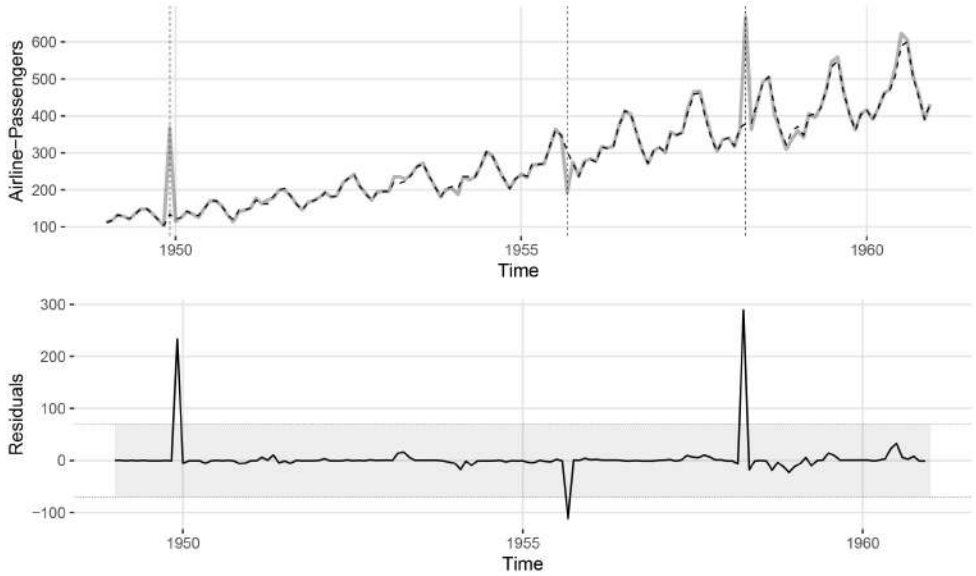


Fig. 1. Top Panel: Airline passenger data with 3 randomly added point outliers. The solid line represents the original time series with added outliers, the positions of outliers are marked with horizontal dotted lines. The dashed line represents the fitted values. The fit for this time series was very close to the original values with the exception of the three outliers which influenced the fitted values only marginally. Therefore, outliers could be identified easily from the residuals in this case. Lower Panel: Residuals of the fit with the model described in this paper. Horizontal dotted lines represent critical values for outlier identification. The outliers exceed these limits considerably and could therefore be identified correctly.

### 2.1. Specification of the basic model

The outlier identification procedure presented in this paper is based on the model proposed in [8]:

$$y_t = \sum_{a=0}^{A} \alpha_a t^a \tag{1}$$
$$+ \left[\sum_{b=1}^{B} \left(\beta_{b,1} \cos\left(\frac{2\pi b}{12} t\right) + \beta_{b,2} \sin\left(\frac{2\pi b}{12} t\right)\right)\right]$$
$$\left(1 + \sum_{g=1}^{G} \gamma_g t^g\right) + \delta_1 I(t \geqslant \delta_2) + e_t,$$

where the first line represents a trend component of order $A$, the second line represents a seasonal component with possible non constant amplitudes and finally the third line represents a level shift of intensity $\delta_1$ at position $\delta_2$ and an irregular component $e_t$ which is assumed to be stationary. This model was essentially adopted for the procedure presented in this paper. In the first analysis of a time series, the level shift term $\delta_1 I(t \geqslant \delta_2)$ is omitted from the model and included only in later steps once it has been identified, see Section 2.4 for details.

### 2.2. Estimation

For the estimation of the parameters of the non-linear model in (1), [8] used a combination of the FastLTS algorithm for robust linear regression ([10]) and the alternating least squares method (see [11]). Differing from that, the estimation of the model for the approach presented in this paper was performed with support vector regressions (SVR). SVR were selected because of two reasons: first, they capture the nonlinear nature of Eq. (1) and secondly, SVR represent an approach which is robust to outliers (see e.g. [12]). Thus, outliers could be identified as large peaks or valleys in the residuals of the model. The basic idea of the proposed method therefore was to estimate model (1) with SVR, which fits the majority of the data points but not the outlying observations well. These extreme points could then be identified via the distribution of the residuals.

#### 2.2.1. Support vector regression

For the linear regression problem $f(x) = \beta_0 + \mathbf{x}^T\boldsymbol{\beta}$ least squares regression minimizes the sum of squared residuals, $SSR = \sum_{i=1}^{n}(y_i - f(x_i))^2$. Thus, a single extreme observation can influence the parameter estimates considerably. In order to limit the influence of outliers, robust regression models have been developed, where small residuals enter the objective function quadratically and large residuals only in their absolute value (see [13]). Differently, with support vector regression (see [14] or [15]) the parameters $\boldsymbol{\beta}$ and $\beta_0$ are calculated minimizing the following loss function (see [16]):

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V_\varepsilon(y_i - f(x_i)) + \frac{\lambda}{2}||\beta||^2 \qquad (2)$$

where

$$V_\varepsilon(r) = \begin{cases} 0 & \text{if } |r| < \varepsilon \\ |r| - \varepsilon & \text{otherwise} \end{cases}. \qquad (3)$$

Thus, residuals which are smaller than a predefined constant $\varepsilon$ do not contribute to the calculation of the loss function, the contribution of all observations larger than this constant is linear. Therefore, the influence of outliers is reduced compared to OLS-Regression. The term $\frac{\lambda}{2}||\beta||^2$ is a penalization term for large $\beta$ values similar to the penalization term in ridge regression (see [17]). The parameter $\lambda$ can be estimated by cross-validation. Moreover, SVR can deal with nonlinear relationships by employing so-called kernel functions (see [12]). Different kernels can be used like linear, gaussian or radial kernels, depending on the concrete regression problem. Radial kernels were used for the estimation of mode (1).

### 2.3. Changing seasonal pattern

In a time series observed over longer horizons, seasonal and cyclical patterns could change over time. Therefore, the basic model (1) will not be adequate for the whole length of the series.

Figure 2 shows a time series with monthly observations over 20 years and a changing seasonal pattern. Fitted values obtained from estimating the model with SVR over the entire time series length were plotted in the same panel (top panel). Obviously, the fitted line does not closely align with the original time series.[2] Instead of estimating one single model for the whole time-series $\mathbf{y} = (y_1, \ldots, y_N)$ (global method), separate models for shorter time periods of length $n < N$ could improve the fit (local method). To achieve local estimates, $b$ samples $\mathbf{y}_i^{(j)} = (y_i^{(j)}, y_{i+1}^{(j)}, \ldots, y_{i+n-1}^{(j)})$ of consecutive observations of the original time series of length $n$ are selected, $j \in \{1, \ldots, b\}$. The algorithm starts with a random number $i$, selected from the interval $\{1, \ldots, (N-n)\}$. This random number selects the first time series sample $(y_i^1, y_{i+1}^1, \ldots, y_{i+n-1}^1)$. Then, this first sample is estimated with a SVR and the fitted values $(\hat{y}_i, \hat{y}_{i+1}, \ldots, \hat{y}_{i+n-1})$ are entered to an estimation matrix $A$ (see Eq. (4)). Next, the second starting value is randomly selected from the interval $\{1, \ldots, (N-n)\}$ and again $n$ fitted values are calculated via SVR. This procedure is repeated until $b$ samples are selected and fitted values $(\hat{y}_1, \hat{y}_2, \ldots)$ are generated for each sample. The fitted values are entered into a $(b \times N)$ matrix $\mathbf{A}$ as follows:

$$\mathbf{A}_{b\times N} = \qquad (4)$$

$$\begin{bmatrix} - & \hat{y}_{1,2} & \hat{y}_{1,3} & \cdots & \hat{y}_{1,n} & \hat{y}_{1,n+1} & - & \cdots & \cdots & \cdots & - \\ - & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & - & \hat{y}_{2,N-n} & \cdots \hat{y}_{2,N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \hat{y}_{b,1} & \hat{y}_{b,2} & \cdots & \cdots & \hat{y}_{b,n} & - & \cdots & \cdots & \cdots & \cdots & - \end{bmatrix}$$

Each row of the matrix $\mathbf{A}$ contains $n$ estimated values and $N - n$ missing values. In the first row of the matrix the first entry is missing, i.e. the first random number $i$ was 2. Therefore, the first time window for which values were estimated was $(y_2, ..., y_{n+1})$. The number $b$ (= number of rows) has to be selected such that every column of matrix $\mathbf{A}$ contains at least 3 non-missing values. Finally, the $N$ fitted values, $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_N)$ are calculated as medians of each column of matrix $\mathbf{A}$, i.e. $\hat{\mathbf{y}} = (med(\mathbf{A}_{.1}), med(\mathbf{A}_{.2}), ..., med(\mathbf{A}_{.N}))$.

The "optimal" length of the time window $n$ was determined depending on the total number of observations $N$. On the one hand $n$ should be long enough to capture several full seasons (years) in order to have enough observations for the estimation of the seasonal components. On the other hand, $n$ should not be too large in order to avoid averaging changing seasonal fluctuations. Finally, all series with more than 48 observations were uniformly assigned a value of $n = 48$. For shorter series, $n$ was gradually reduced down to a minimum of 24 observations, which was supposed to be a minimum length for the estimation of the seasonal factors. For all time series with $N < 25$ observations the global method was applied. For quarterly time series the limits for the global and local method were not tested. However, four or five years of observations for the local method, with a minimum of $N = 8$ or $N = 12$ observations for the global method could be a reasonable starting point for the selection of $n$. When applying this algorithm to the turnover series of Fig. 2 the fitted line aligns much closer with the original time series path than with the global method (compare upper and lower panel of Fig. 2).

### 2.4. Level shifts

The level-shift term in Eq. (1) constitutes a problem for this procedure because the timing of a possible level-shift is initially unknown. Fitting the SVR-model

[2] The inclusion of a possible level shift around the year 2005 and/or 2010 could improve the fit. However, at this point no level shift was included.

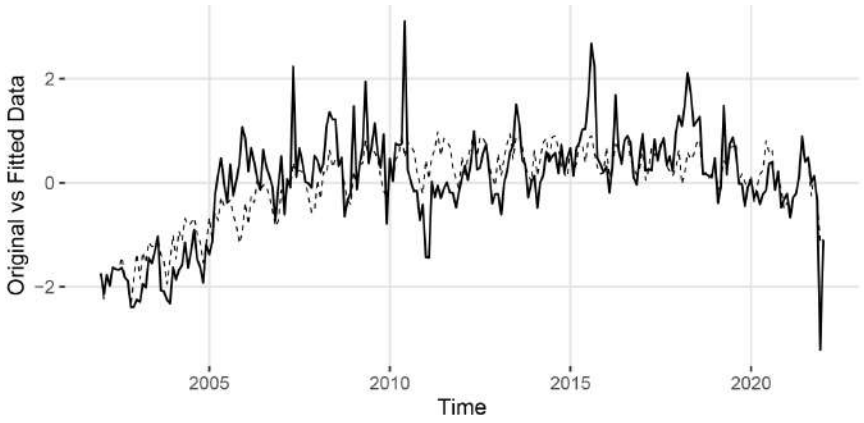


Fig. 2. Top Panel: Scaled turnover series (solid line) of one single enterprise with fitted values calculated for the whole span of the series (dashed line). Lower Panel: Scaled turnover series (solid line) with fitted values calculated for short time windows (dashed line).

to time series ignoring existing level shifts leads to large residuals before and after the level shift. However, adding dummy-variables for each observation could lead to estimation problems (overspecification).

Level shifts in a time series were handled with two different strategies which are presented in the next two Subsections. Firstly, a sequential procedure was implemented adding dummy variables to the model one after another (see Subsection 2.4.1). Secondly, differencing the data changes the shape of outliers in the original data. If level shifts could be observed in the original series, they will not occur as level shifts in the differenced data. Therefore, a model without any level shift component can be fitted to the differenced data (see Subsection 2.4.2).

### 2.4.1. *Sequential Identification of possible Level Shifts*

In an initial step, before fitting the SVR-model, possible level-shift positions have to be identified. The level-shift identification was done by OLS regression. For this purpose, the equation

$$log(y_t) = \sum_{a=0}^{A} \alpha_a t^a \tag{5}$$
$$+ \left[ \sum_{b=1}^{B} \left( \beta_{b,1} cos\left(\frac{2\pi b}{12} t\right) + \beta_{b,2} sin\left(\frac{2\pi b}{12} t\right) \right) \right]$$
$$+ \delta_1 I(t \geqslant \delta_2) + e_t,$$

was estimated sequentially with different level-shift positions $\delta_2$.[3] The first regression was fitted with $\delta_2 = 4$, the last with $\delta_2 = N - 3$, assuming that a level shift would not occur before the fourth observation and not after the last but 4th observation. Thus, $N - 6$ regressions were calculated in the first step. The first level-shift, if any, was fixed for the regression with minimal residual variance of all $N - 6$ regressions if the size of the level shift $\delta_1$ was significant as well.[4] Thereafter, the level shift position $\delta_2$ was fixed and the procedure was repeated until no further level shift could be identified.

Figure 3 shows a time-series with a level-shift in 2008. The level-shift was identified correctly (vertical line) and the fit of the model improved considerably around the level-shift. The quality of the level-shift identification procedure depends on the amplitude of the level-shift and whether there are other point outliers close to the level-shift (see Fig. 3).

### 2.4.2. *Differencing*

Applying the procedure for identification of level shifts described above, results could be improved considerably. However, if level shifts are not correctly identified, the subsequent identification of point outliers is affected negatively. Therefore, an alternative approach for this procedure was considered. If the differenced time-series $\Delta y_t = y_t - y_{t-d}$ (with $d = 1$ or $d = 12$ for monthly data) is analysed instead of the original data, level-shifts would not be an issue. Potential level-shifts or transitory changes in the original series are transformed to point outliers by differencing. However, the pattern for point outliers and level-shifts for differenced data differs from the original pattern (see Fig. 4 for

[3] Differing from Eq. (1) the term $\left(1 + \sum_{g=1}^{G} \gamma_g t^g\right)$ was omitted. Non-constant amplitudes of the seasonal fluctuations were considered by taking logs on the left hand side of Eq. (5).

[4] If the $p$-value of the estimated parameter was smaller than 0.05.

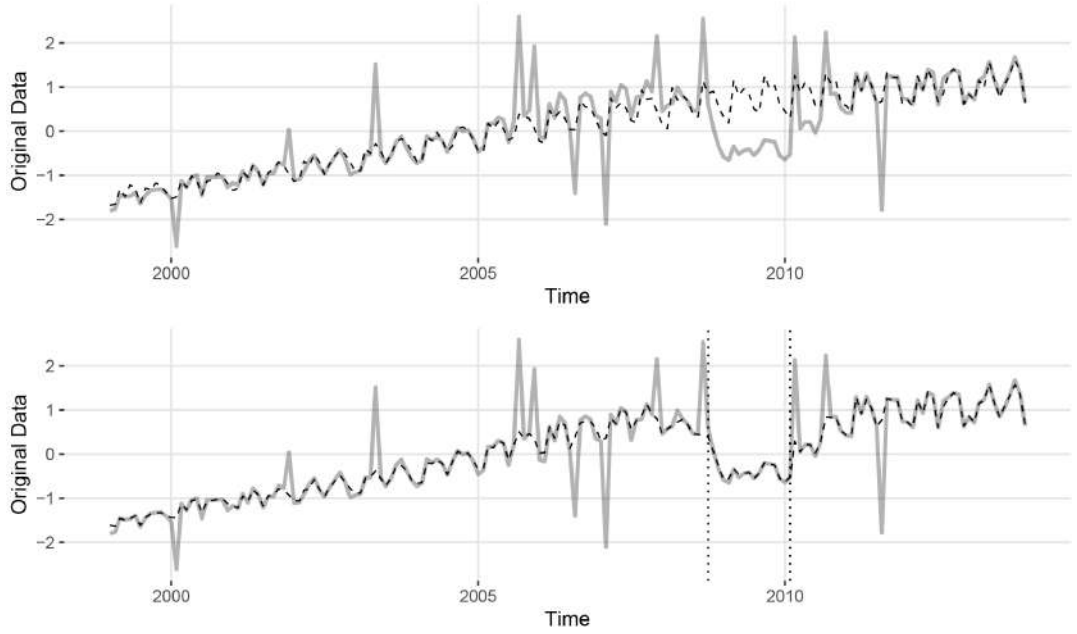


Fig. 3. Time-series with a level-shift in the second half of 2008. A second level shift was identified early 2010, possibly caused by the point outlier in this period. The solid line shows the original path of the time series with twelve randomly included point outliers. The first level-shift was correctly identified in this case (vertical grey line in the lower panel of the plot). The upper panel shows the fit with SVR ignoring the level-shift (dashed line). The lower panel shows the fit with SVR including two dummy variables, considering the prior identified level-shifts. The fit in the surrounding of the level-shift could be improved considerably in this case. In both cases the global method was applied omitting the replication algorithm.

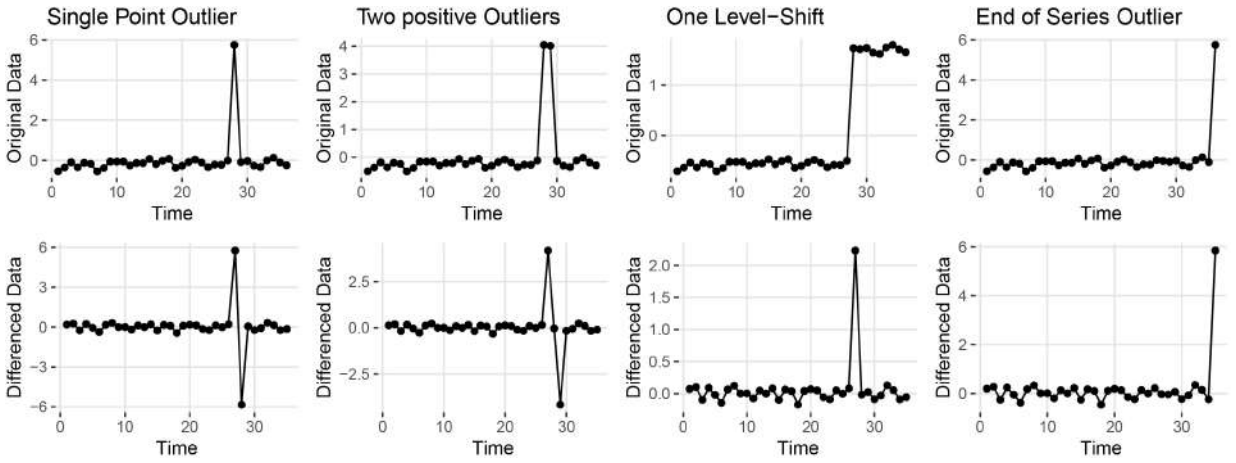


Fig. 4. Examples of the behaviour of differenced data for different patterns of outliers. The original time-series are plotted in the top panel, corresponding differenced series in the lower panel.

illustration of the outlier pattern of first differences). A single point outlier in the original data leads to two consecutive additive outliers with opposite sign in the differenced data. The only exception are single outliers in the very beginning or the very end of the series which are single outliers in the differenced data. However, signs of these single outliers are different depending on whether they occur at the beginning of the time-series or at the end. On the other hand, transitory changes are converted to single outliers. Sequences of consecutive outliers (sometimes called ramps) could be interpreted as level-shifts in one direction followed by a level-shift in the opposite direction after several observations (depending on the length of the ramp).

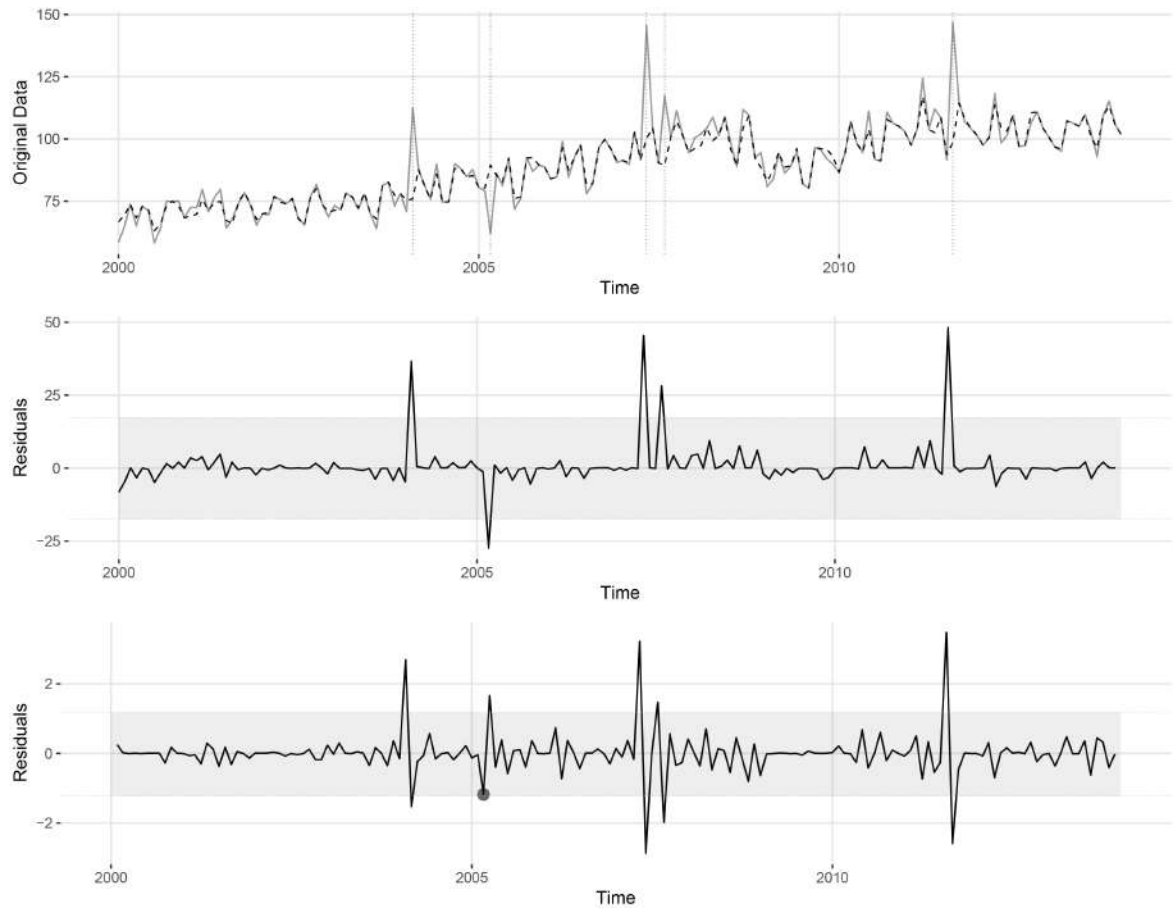


Fig. 5. Top Panel: Time Series with 5 randomly added outliers (marked with vertical dotted lines) plus fitted line to original data generated with SVR (dashed line). Middle: Residuals from estimation with SVR with critical thresholds added (shaded area). All outliers could be identified correctly in this case. Lower Panel: Residuals from estimation of differenced data with SVR. For each outlier-position a peak followed by a valley (or vice versa) can be observed. However, while the peak in April 2005 exceeds the threshold, the immediately preceding value in March 2005 (marked with the dot) does not. Therefore, in this case the second outlier was not correctly identified, whereas the other 4 outliers were identified correctly.

For outlier identification, the same procedure as described for un-differenced data can be applied to the first differences of the original series. However, the preliminary estimation of level shift positions could be omitted in this case. SVR were applied to the first differences of the data and subsequently the residuals of the differenced data and the fitted line were analysed. Point outliers were identified if two subsequent values of the residuals exceeded the critical value in opposite direction or if single outliers were identified in the very beginning or the very end of the time series (see Fig. 5).

## 3. Practical outlier identification

After model estimation, possible outliers have to be identified. The practical outlier identification procedure depends on the method. Parametric methods, like X13-Arima or Tramo/Seats, employ RegArima-models with dummy variables for every single observation. Outliers are identified as the values for which the effect of the corresponding dummy variable is significant. Nonparametric approaches, as implemented in the `tsoutlier` function of the R-Package `forecast`, focus on the distribution of the residuals.

### *3.1. Identification based on the distribution of the residuals*

The aforementioned R-Package `forecast` identifies outliers if residuals are beyond the interval

$$[Q_1 - 3 * \text{IQR}; \quad Q_3 + 3 * \text{IQR}], \tag{6}$$

Table 1
Parameter settings used for simulation of outlier identification

| Length of time-series | Regressors | | | | Local/Global Method |
|---|---|---|---|---|---|
| | Trend polynomial | Harmonics | Varying amplitudes | Level-shift | |
| 8 Obs | $A = 3$ | None | None | Not considered | Global Method |
| 16 Obs | $A = 3$ | $B = 2$ | None | Not considered | Global Method |
| 32 Obs | $A = 3$ | $B = 6$ | None | Yes | Global Method |
| 168–180 Obs | $A = 3$ | $B = 6$ | $G = 2$ | Yes | Local Method |

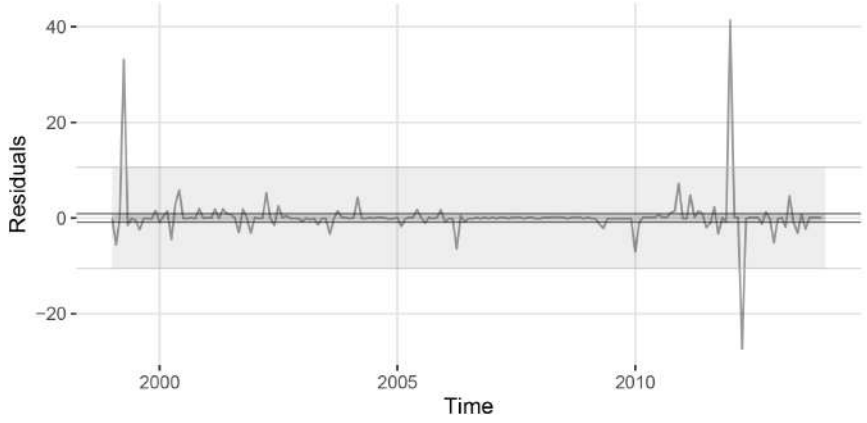


Fig. 6. Residuals of a time series with 3 added outliers. The shaded areas represent the limits specified in (6) (dark grey area) and (7) (bright grey area) respectively.

where $Q_1$ and $Q_3$ are the first and third quartile and IQR is the interquartile-range of the distribution of the residuals.

Figure 6 shows the plot of the residuals of a time series with 3 outliers and a fitted SVR-model. The dark grey area indicates the limits given in Expression (6). These limits are too narrow for outlier identification with the method proposed in this paper. The same problem could be observed for other time series with differing numbers of outliers as well. The SVR-model closely follows the path of the time series, resulting in residuals that are highly concentrated around zero. As an alternative for the method proposed in this paper, outliers were identified as observations with residuals outside the interval

$$[P_{10} - 3 * (P_{90} - P_{10}); \quad P_{90} + 3 * (P_{90} - P_{10})], \quad (7)$$

where $P_{10}$ and $P_{90}$ are the 10th and 90th percentile of the residuals respectively.[5]

The light grey area in Fig. 6 was plotted using these limits. Adjusting the sensitivity of outlier identification for the simulation study (see Section 4.3) in this manner resulted in high proportions of correctly identified outliers. However, the variance of the residuals as well as the range of $(P_{90} - P_{10})$ depends on the number of outliers in the series – the quality of the model fit improves as the number of outliers in the time series decreases. Increasing the number of outliers yielded an increasing $10 - 90$ percentile range and therefore the identification rate of outliers (with unchanged multiplier 3) decreases with the number of outliers. However, the same issue could be observed for other benchmark methods as well (see Tables 2 and 3 in Section 4.3). In the concrete application of data cleaning for short term statistics, the number of outliers should be limited because of implemented editing procedures for historical data.

In Fig. 7 the residuals of the model fitted to the same series are plotted – top panel: original series with one added outlier, mid-panel: original series with twelve added outliers, lower panel: residuals of first plot (solid line) and second plot(dotted line). Obviously, the residuals from the series with 12 outliers are noisier than the residuals from the series with one outlier.

### 3.2. Outlier identification by clustering residuals

In addition to identifying outliers based on percentiles and the variability of residuals, an alternative

[5]It was assumed that the number of outliers would not exceed 20% of observations. The factor 3 in Eq. (7) was changed to 1 for differenced data and also for short time series with less than 36 observations.

Table 2
Performance of different Outlier Identification Procedures: Relative number of identified outliers for 17 × 100 time series with 168–180 observations

| Method | Outliers | Low | Zero | Exact | High |
|---|---|---|---|---|---|
| SVR | 1 | 0.04 | 0.00 | 0.80 | 0.16 |
| SVR kmeansH | 1 | 0.00 | 0.00 | 0.81 | 0.19 |
| SVR kmeans | 1 | 0.00 | 0.00 | 0.62 | 0.38 |
| SVRD | 1 | 0.01 | 0.00 | 0.63 | 0.36 |
| SVRD kmeansH | 1 | 0.01 | 0.00 | 0.88 | 0.11 |
| SVRD kmeans | 1 | 0.01 | 0.00 | 0.87 | 0.12 |
| tsoutliers | 1 | 0.01 | 0.00 | 0.49 | 0.49 |
| Tramo/Seats | 1 | 0.00 | 0.00 | 0.44 | 0.56 |
| Tramo/Seats AL | 1 | 0.00 | 0.00 | 0.81 | 0.19 |
| X13 Arima | 1 | 0.00 | 0.00 | 0.74 | 0.26 |
| tsoutliers FCT | 1 | 0.06 | 0.00 | 0.66 | 0.28 |
| SVR | 3 | 0.08 | 0.01 | 0.70 | 0.21 |
| SVR kmeansH | 3 | 0.16 | 0.00 | 0.81 | 0.03 |
| SVR kmeans | 3 | 0.12 | 0.00 | 0.85 | 0.03 |
| SVRD | 3 | 0.04 | 0.00 | 0.74 | 0.21 |
| SVRD kmeansH | 3 | 0.14 | 0.01 | 0.74 | 0.11 |
| SVRD kmeans | 3 | 0.14 | 0.01 | 0.74 | 0.11 |
| tsoutliers | 3 | 0.05 | 0.02 | 0.65 | 0.28 |
| Tramo/Seats | 3 | 0.00 | 0.00 | 0.46 | 0.53 |
| Tramo/Seats AL | 3 | 0.01 | 0.00 | 0.80 | 0.19 |
| X13 Arima | 3 | 0.00 | 0.00 | 0.73 | 0.27 |
| tsoutliers FCT | 3 | 0.18 | 0.00 | 0.58 | 0.24 |

Table 3
Performance of different Outlier Identification Procedures: Relative number of identified outliers for 17 × 100 time series with 168–180 observations

| Method | Outliers | Low | Zero | Exact | High |
|---|---|---|---|---|---|
| SVR | 7 | 0.34 | 0.01 | 0.60 | 0.05 |
| SVR kmeansH | 7 | 0.36 | 0.01 | 0.61 | 0.02 |
| SVR kmeans | 7 | 0.36 | 0.00 | 0.62 | 0.02 |
| SVRD | 7 | 0.30 | 0.06 | 0.49 | 0.16 |
| SVRD kmeansH | 7 | 0.56 | 0.05 | 0.31 | 0.08 |
| SVRD kmeans | 7 | 0.56 | 0.05 | 0.31 | 0.08 |
| tsoutliers | 7 | 0.23 | 0.05 | 0.53 | 0.19 |
| Tramo/Seats | 7 | 0.01 | 0.01 | 0.46 | 0.52 |
| Tramo/Seats AL | 7 | 0.02 | 0.01 | 0.78 | 0.19 |
| X13 Arima | 7 | 0.01 | 0.00 | 0.67 | 0.31 |
| tsoutliers FCT | 7 | 0.46 | 0.00 | 0.38 | 0.16 |
| SVR | 12 | 0.70 | 0.01 | 0.29 | 0.00 |
| SVR kmeansH | 12 | 0.53 | 0.01 | 0.45 | 0.01 |
| SVR kmeans | 12 | 0.53 | 0.01 | 0.45 | 0.01 |
| SVRD | 12 | 0.70 | 0.08 | 0.13 | 0.09 |
| SVRD kmeansH | 12 | 0.80 | 0.08 | 0.06 | 0.06 |
| SVRD kmeans | 12 | 0.80 | 0.08 | 0.06 | 0.06 |
| tsoutliers | 12 | 0.62 | 0.03 | 0.26 | 0.09 |
| Tramo/Seats | 12 | 0.03 | 0.01 | 0.45 | 0.51 |
| Tramo/Seats AL | 12 | 0.05 | 0.01 | 0.73 | 0.21 |
| X13 Arima | 12 | 0.02 | 0.01 | 0.59 | 0.39 |
| tsoutliers FCT | 12 | 0.70 | 0.01 | 0.21 | 0.08 |

approach was pursed. The intention was to perform a cluster analysis on the model residuals. The basic idea is that extreme residuals should be separated in one outlier-cluster and all other observations should be classified in a second cluster of non-outlying observations. As an initial step, the absolute values of the residuals were computed and sorted according to size. The number of clusters to identify was set to two. All observations in the cluster with the largest absolute residuals were then classified as outliers. Figure 8 shows an example with ordered absolute residuals from an SVR-estimation of a time series with 3 added outliers. The marked areas represent the two identified clusters. The three largest residuals were assigned to the second cluster – in this case the separation between outliers and non-outliers was performed correctly.

Yet, this procedure was not successful when dealing with time series without any outliers. In this case, the number of clusters to identify was again fixed to two. According to the procedure described above, all observations in the second cluster would be misclassified as outlier. To overcome this problem, a first attempt was to test for a significant difference between the means of the two clusters – in case of a significant result it can be concluded that the second cluster consisted of outliers. Otherwise, no outliers were assumed in the series. However, even for time series without any outliers this procedure gave significant results. As an alternative attempt, one artificial outlier was introduced into the residuals before conducting cluster analysis. In this way, the second cluster would include at least this one artificial outlier.[6] The intention was that in case of no outliers in the original residuals, the second cluster would only include the one artificial outlier. This artificial outlier was subsequently eliminated from the list of identified outliers.[7] In order to avoid introducing outliers at a position which was already extreme, the outlier was introduced at the position of the median value of the residuals.

### *3.2.1. Kmeans and hierarchical clustering*

K-means clustering is a common method for cluster analysis. The algorithm is executed through the following steps:

- The number $k$ of clusters is determined by the user.
- $k$ random points are fixed by the program as means ("centroids") for each cluster.
- The distances from each observation to each centroid are calculated.

[6] Moreover, with the inclusion of an artificial outlier in the residuals, the sensitivity of outlier adjustment could be controled by the magnitude of this outlier.

[7] In case of the analysis of differenced data, two consecutive additional outliers had to be introduced to the residuals – one positive and one negative.

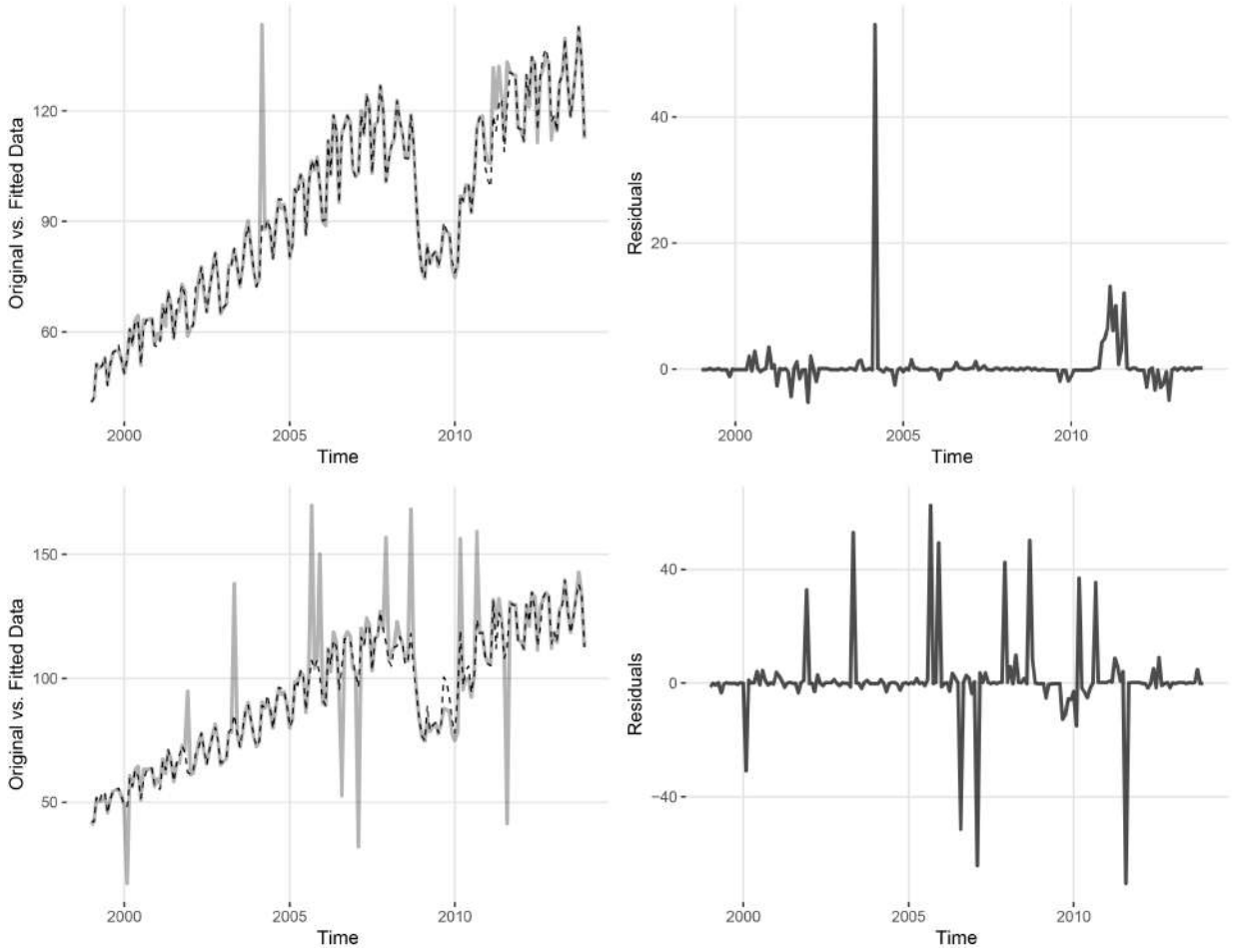


Fig. 7. Time series with one added outlier (upper left panel) and twelve added outliers (lower left panel) with corresponding residuals on the right side. The lower residual plot exhibits higher variability between the peaks for outliers than the upper plot (mind the scale).

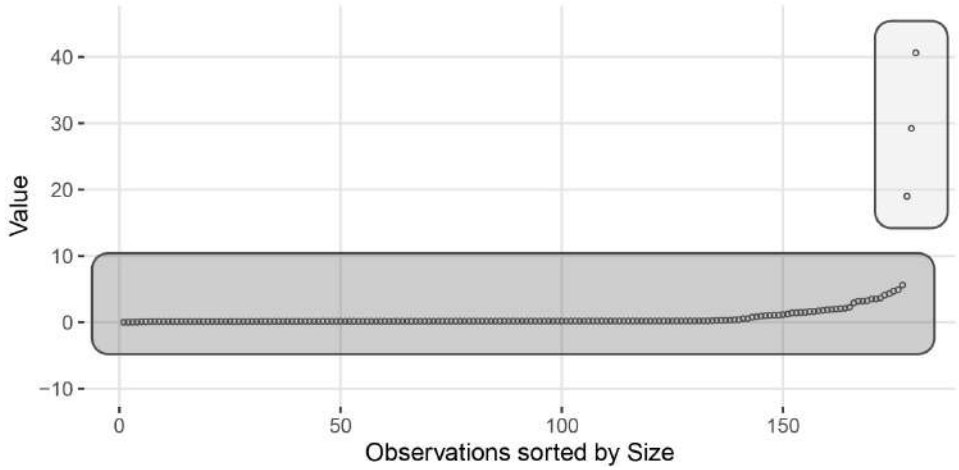


Fig. 8. Ordered absolute Residuals from estimation of a time series with 3 added outliers. The Estimation of the model was performed with SVR. K-means clustering was applied to this data. The cluster on the right hand side (largest absolute residuals) was correctly identified in this case.

- Each observation is assigned to its nearest centroid.
- Centroids are recalculated as means of each cluster.
- The algorithm terminates if no observation changes its assignment to a centroid.

K-means clustering is very sensitive to the initial selection of random centroids. Hence, the result of this clustering algorithm might be different with every new run of the clustering algorithm. In order to circumvent this problem, a hybrid of k-means- and hierarchical clustering was suggested by [18]. In this approach, hierarchical clustering is performed in a first step to identify the clusters. Thereafter, k-means clustering is performed with the centroids of the initially determined clusters as starting values. K-means clustering and the hybrid method are available with the function `kmeans` from the `stats` Package and the function `hkmeans` from the `factoextra`-Package (see [19]). Both methods are compared in the results section below (see Section 4.3).

## 4. Comparison of outlier detection methods

The objective of the outlier detection and replacement procedure discussed in this paper was to identify errors which are most influential in the original data and to eliminate or at least minimize their effect. The quality of subsequent nowcasting procedures is highly dependent on this procedure. The intended scenario for outlier identification was as follows:

- The outlier-identification procedure was intended to assist the data-editing process
- The number of ouliers should not exceed 20% of data, but realistically it should be much smaller
- The process should possibly focus on outliers at the very end of the time series because historically all outliers have already been corrected
- The adjustment of outliers should be conservative, prioritizing the correction of too few rather than too many outliers. This approach is preferred because suspicious values that are initially considered as outliers may later be validated to be correct.

### *4.1. Setup*

To mimic the planned application, the performance of the outlier identification procedure was evaluated based on 17 real-world time series taken from the R-Package `tsoutliers`, representing Industrial production indices in the manufacturing sector of European monetary union countries, with a length of 168 and 180 monthly observations. For the evaluation of the performance of the different outlier identification procedures, the series were contaminated with varying numbers of point outliers at random position and of random size. The size of the outliers was selected randomly between $(0.3x - 0.7x)$ for negative outliers and $(1.3x - 1.7x)$ for positive outliers, with $x$ the actual value of the time series. For time series with shorter time horizons (less than 48 observations), these factors were adjusted to $(0.1x - 0.5x)$ and $(1.6x - 2.0x)$. Thus, 100 contaminated series were generated for each country, resulting in 1700 series to adjust.

#### *4.1.1. Settings*

Automatic detection of erroneous data has to be applicable for time series of any length because even if only a few observations are available for an enterprise or if the size of the enterprise is small, influential errors could occur with negative consequences for data quality of the whole survey. Clearly, the time series approach has limits concerning the length of the series. If the time series to analyse is very short, say 1–5 observations a different approach could be more appropriate.[8] Model (1) can be applied to time series of any length if the regressors are selected adequately. The settings which were employed for the simulation study presented in Section 4.3 are shown in Table 1: the first column shows the different time series lengths, columns 2–4 exhibit the corresponding specifications of the parameters $A, B, G$ of the model Eq. (1), column 5 indicates whether a level shift was considered or not and column 6 shows if the global- or local method (see 2.3) was applied.

### *4.2. Methods*

All calculations were performed with well documented R-Packages ([20]). For the support vector regression we used the `train` function implemented in the `caret` package ([21]) as well as the function `ksvm` from the package `e1071` ([22]). Parameter tuning was done by cross-validation – the two procedures offer respective functions for this purpose. However, parameter tuning is also a time-consuming issue and therefore the parameter grids were limited. For `train` ten-fold

[8] This could be the specification of absolute thresholds which are unlikely to be exceeded.

cross-validation was performed to determine the cost parameter and the scaling parameter of the kernel function. Moreover, $\varepsilon$ was fixed to 0.01. A radial kernel function was used for the estimation of model (1) in order to capture its nonlinear structure.

For comparing the performance of outlier identification the following benchmark models were considered (The terms used here are kept in the tables of the Results Section):

- X13 Arima: Seasonal Adjustment Procedure X13-Arima ([23])
- Tramo/Seats: Seasonal Adjustment Procedure Tramo/Seats ([24]) (with automatic model identification)
- Tramo/Seats AL: Tramo/Seats (with fixed Airline-Model (AL))[9]
- tsoutliers: Outlier identification by decomposition into components and smoothing the rest component implemented in the R-Package `forecast` ([5])
- tsoutliers FCT: Outlier identification algorithm proposed by [4] implemented in the R-Package `tsoutliers` ([25]).

The R-Package `persephone` ([26]) was used for calculations with the seasonal adjustment procedures X13-Arima and Tramo/Seats. The identification of outliers was limited to additive outliers for X13-Arima, Tramo/Seats and `tso`, because most of these series exhibit some kind of level-shift/transitory change during the financial crisis in 2008 and the intention was to identify point outliers only. Moreover, data errors in the original time series were considered unlikely because the data represents aggregated series officially released by Eurostat. Therefore, the outliers identified by the different procedures should coincide with the number of outliers implemented artificially.

### *4.3. Results*

The following tables report the number of outliers introduced into the series (Outliers), the percentage of time series with too few identified outliers (Low), the percentage of time series with correct numbers of identified outliers (Zero), the percentage of time series with correctly identified numbers and positions of outliers (Exact) and the percentage of time series with too many identified outliers (High) for each method.

[9]It was found, that Tramo identified too many outliers when the model selection was performed automatically. The performance of Tramo improved considerably when outliers were identified based on the AL-Model.

Table 4
Performance of different Outlier Identification Procedures: Relative number of identified outliers for 17 × 100 time series with 32 observations

| Method | Outliers | Low | Zero | Exact | High |
|---|---|---|---|---|---|
| SVR | 1 | 0.31 | 0.01 | 0.62 | 0.06 |
| SVR kmeansH | 1 | 0.02 | 0.01 | 0.79 | 0.18 |
| SVR kmeans | 1 | 0.02 | 0.01 | 0.79 | 0.18 |
| tsoutliers | 1 | 0.08 | 0.26 | 0.60 | 0.07 |
| tsoutliers FCT | 1 | 0.18 | 0.05 | 0.24 | 0.54 |
| SVR | 2 | 0.52 | 0.01 | 0.44 | 0.04 |
| SVR kmeansH | 2 | 0.08 | 0.02 | 0.80 | 0.11 |
| SVR kmeans | 2 | 0.08 | 0.02 | 0.80 | 0.11 |
| tsoutliers | 2 | 0.22 | 0.12 | 0.58 | 0.09 |
| tsoutliers FCT | 2 | 0.13 | 0.05 | 0.26 | 0.55 |

Table 5
Performance of different Outlier Identification Procedures: Relative number of identified outliers for 17 × 100 time series with 16 observations

| Method | Outliers | Low | Zero | Exact | High |
|---|---|---|---|---|---|
| SVR | 1 | 0.27 | 0.00 | 0.69 | 0.03 |
| SVR kmeansH | 1 | 0.03 | 0.01 | 0.88 | 0.08 |
| SVR kmeans | 1 | 0.03 | 0.01 | 0.88 | 0.08 |
| tsoutliers | 1 | 0.15 | 0.00 | 0.63 | 0.22 |
| tsoutliers FCT | 1 | 0.12 | 0.00 | 0.73 | 0.14 |
| SVR | 2 | 0.69 | 0.01 | 0.30 | 0.00 |
| SVR kmeansH | 2 | 0.19 | 0.02 | 0.77 | 0.02 |
| SVR kmeans | 2 | 0.19 | 0.02 | 0.77 | 0.02 |
| tsoutliers | 2 | 0.26 | 0.01 | 0.57 | 0.16 |
| tsoutliers FCT | 2 | 0.26 | 0.01 | 0.62 | 0.11 |

Table 6
Performance of different Outlier Identification Procedures: Relative number of identified outliers for 17 × 100 time series with 8 observations

| Method | Outliers | Low | Zero | Exact | High |
|---|---|---|---|---|---|
| SVR | 1 | 0.17 | 0.04 | 0.79 | 0.00 |
| SVR kmeansH | 1 | 0.11 | 0.04 | 0.85 | 0.00 |
| SVR kmeans | 1 | 0.11 | 0.04 | 0.85 | 0.00 |
| tsoutliers | 1 | 0.29 | 0.00 | 0.58 | 0.13 |
| tsoutliers FCT | 1 | 0.44 | 0.00 | 0.49 | 0.06 |

The SVR-model achieved very high correct outlier identification rates for long time series with 1–3 outliers. For all scenarios identification rates of 80%–85% were accomplished (see Tables 2 and 3). The rate was slightly lower for the approach with differenced data. Tramo/Seats AL and X13 yielded similar results with correct identification rates between 73% and 81%. The performance of the remaining models slightly fell behind in this respect, with rates of 44% to 66%.

Nearly all methods had problems when the number of outliers were increased. This observation can potentially be attributed to the fact that outliers of small magnitude will occur more often with a growing number of

outliers.[10] Therefore, correctly identifying a large number of outliers in one series may become more difficult. Also, the increasing variability of the series influences the identification process. Increasing the number of outliers to 7 and 12 respectively, reduced correct identification proportions considerably. In case of 7 outliers the SVR-model identified around 60% of outliers correctly, in case of 12 outliers this value decreased to 29% to 45%. However, the identification proportions for the model with differenced data was much lower.

Again, Tramo/Seats AL delivered high correct identification rates with increasing numbers of outliers. The identification rate for 7 and 12 outliers was still 78% to 73%. Such high rates could not be achieved with the other benchmark models.

The performance evaluation of various outlier identification procedures should encompass not only long time series but also series with shorter time horizons. Therefore, shorter time windows were extracted from the Industrial production index series. Three different scenarios were tested, time series with 32, 16 and 8 observations (data tuning in ksvm is only possible with at least 10 observations). The starting point for the time-windows was selected randomly between 1 and $N - n$, with $n$ representing the length of the desired time window, and $N$ the length of the original time series. The results of this simulation study are presented in Tables 4 to 6. As Tramo/Seats and X13-Arima require a minimum of 36 observations these methods were not included in this comparison.

For medium to short time series, the performance of the SVR-model was quite good with correct identification rates of over 60% to around 80%. Thus, the performance was clearly better than that achieved with the benchmark models. Some series could not be processed with the R-package `tsoutliers` as no appropriate Arima-Model could be identified.

## 5. Conclusions

The SVR-method introduced in this paper was tested on several macro-economic time series from European countries. These series have been officially released by Eurostat and therefore it can be expected that these data do not contain any errors. The time span of the series was 1999/2000 to 2013. Thus, several series exhibited a structural break in late 2008 because of the financial crisis. For the evaluation of the different outlier adjustment procedures various aspects were tested such as different numbers of outliers, amplitudes of outliers and time series lengths. The outliers were generated at random positions and with random amplitudes.[11] The identification of outliers was restricted to point outliers – thus, methods which could identify different types of outliers were restricted to point outlier search only.

The results achieved with the new outlier-identification and replacement method were promising. It is applicable to long time series but also for very short series. The main issue for this method was the run-time, which is considerably higher than that of Tramo/Seats or the function `tsoutliers` from the `forecast` Package. However, for shorter series the difference in run-time became smaller, because the time-consuming replication algorithm was only implemented for series with more than three years of observations.

The performance of several common benchmark methods was evaluated, yielding varying results. The methods X13-Arima and Tramo/Seats performed very well both in terms of correct identification rates as well as in terms of run-time. For Tramo/Seats the default settings had to be slightly adapted in order to improve the results. However, these methods require a length of 36 observations for monthly data at a minimum. In contrast, the outlier identification methods integrated in the `forecast` Package and the `tsoutliers` Package could be used for long and short time series. Moreover, with respect to the run-time the algorithm implemented in the `tsoutliers` function from the `forecast` Package outperformed all competing methods by far. However, the rate of correctly identified outliers was lower than for other competitors.

The outlined procedure will be further tested on enterprise level data for variables like industrial production, turnover or hours worked. It is expected that these series exhibit stronger irregular variations which could negatively influence the performance of the procedure. Furthermore, specific emphasis will be placed on the optimization of the algorithm in terms of runtime, which was not the primary intention of this first analysis.

[10] Outliers were introduced randomly and with random magnitude. Therefore, the probability of introducing outliers with small magnitude is higher with increasing number of outliers. Then it becomes difficult to classify all outliers correctly.

[11] The intended scenario was to edit data sets which are contaminated with errors like typos or misplacement errors, which occur randomly. Ideally, real data should not be corrected by the procedure, even if observations are of exceptional size.

## References


[1] Mazzi GL, Cannata RR. Ladiray D, Mazzi GL, editors. Handbook on Rapid Estimates. European Commission, Eurostat; 2017; doi: 10.2785/488740.

[2] Chang I, Tiao G, Chen C. Estimation of Time Series Parameters in the Presence of Outliers. Technometrics. 1988; 30: 193-204. doi: 10.1080/00401706.1988.10488367.

[3] Otto MC, Bell WR. Two issues in time series outlier detection using indicator variables. In: Proceedings of the American Statistical Association, Business and Economic Statistics Section; 1990; pp. 182-7.

[4] Chen C, Liu LM. Joint Estimation of Model Parameters and Outlier Effects in Time Series. Journal of the American Statistical Association. 1993; 88: 284-97. doi: 10.2307/2290724.

[5] Hyndman RJ, Khandakar Y. "Automatic Time Series Forecasting: The forecast Package for R". Journal of Statistical Software. 2008; 27. doi: 10.18637/jss.v027.i03.

[6] Bandara K, Hyndman R, Bergmeir C. MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns. International Journal of Operational Research. 2022; 1(1): 1. doi: 10.1504/ijor.2022.10048281.

[7] Bilen C, Huzurbazar S. Wavelet-Based Detection of Outliers in Time Series. Journal of Computational and Graphical Statistcs. 2002; 11: 311-27. doi: 10.1198/106186002760180536.

[8] Rousseeuw PJ, Perrotta DC, Riani M, Hubert M. Robust Monitoring of Time Series with Application to Fraud Detection. Econometrics and Statistics. 2019; doi: 10.1016/j.ecosta.2018.05.001.

[9] Box GEP, Jenkins GM. Time Series Analysis: Forecasting and Control. Holden-Day series in time series analysis and digital processing. Holden-Day; 1976; Available from: https//books.google.at/books?id=1WVHAAAAMAAJ.

[10] Rousseeuw PJ, Driessen KV. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics. 1999 08; 41: 212-23. doi: 10.1080/00401706.1999.10485670.

[11] Wold H. Research Papers in Statistics: Festschrift for Jerzy Neyman (ed FN David). 1966.

[12] Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Articles. 2008; 28(5): 1-26. doi: 10.18637/jss.v028.i05. Available from: https//www.jstatsoft.org/v028/i05.

[13] Huber PJ. Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. 1964; 35(1): 73 101. doi: 10.1214/aoms/1177703732. Available from: 10.1214/aoms/1177703732.

[14] Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support Vector Regression Machines. Advances in Neural Information Processing Systems. 2003; 11; 9.

[15] Smola A, Burges CJC, Drucker H, Golowich S, Hemmen L, Müller KR, et al. Regression Estimation with Support Vector Learning Machines. 2003; 11.

[16] Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. 2001; doi: 10.1007/978-0-387-84858-7.

[17] Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970; 12(1): 55-67. doi: 10.1080/00401706.1970.10488634. Available from: https//www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

[18] Chen B, Tai PC, Harrison RW, Pan Y. Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis. In: 2005 IEEE Computational Systems Bioinformatics Conference – Workshops (CSBW'05); 2005. pp. 105-8. doi: 10.1109/CSBW.2005.98.

[19] Kassambara A, Mundt F. Extract and Visualize the Results of Multivariate Data Analyses; 2020. R package version 1.0.7. Available from: https//CRAN.Rproject.org/package=factoextra.

[20] R Core Team. R: A Language and Environment for Statistical Computing. R. 2010. Available from: http//www.R-project.org/.

[21] Kuhn M. Caret: Classification and Regression Training; 2020. R package version 6.0-86. Available from: https//CRAN.R-project.org/package=caret.

[22] Meyer D. Support Vector Machines. The Interface to libsvm in package e1071. Online-Documentation of the package e1071 for R. 2001.

[23] Findley DF, Monsell BC, Bell WR, Otto MC, Chen BC. New Capabilities and Methods of the X12-ARIMA Seasonal-Adjustment Program. Journal of Business and Economic Statistics. 1998; 16: 127-52. doi: 10.2307/1392565.

[24] Gomez V, Maravall A. Seasonal Adjustment and Signal Extractin in Economic Time Series. Banco de Espana Working Papers 9809, Banco de Espana. 1998. doi: 10.1002/9781118032978.ch8.

[25] de Lacalle JL. Detection of Outliers in Time Series. https://cranrprojectorg/web/packages/tsoutliers/indexhtml.2015.

[26] Kowarik A, de Cillia G, Meraner A, Fröhlich M. Persephone, Production-Ready Seasonal Adjustment in R with RJDemetra. In: Conference on New Techniques and Technologies for Official Statistics (NTTS) [Internet]; 2021.