Outlier Identification and Adjustment for Time Series

UNECE Expert Meeting on Statistical Data Editing, Wien, 7-9 October.

Markus Fröhlich

Vienna, 8. October 2024

www.statistik.at

Independent statistics for evidence-based decision making

STATISTICS AUSTRIA

The Information Manager

Overview

- Motivation
- Outlier Adjustment for Time Series
  - Requirements
  - Standard procedures
  - Robust Model

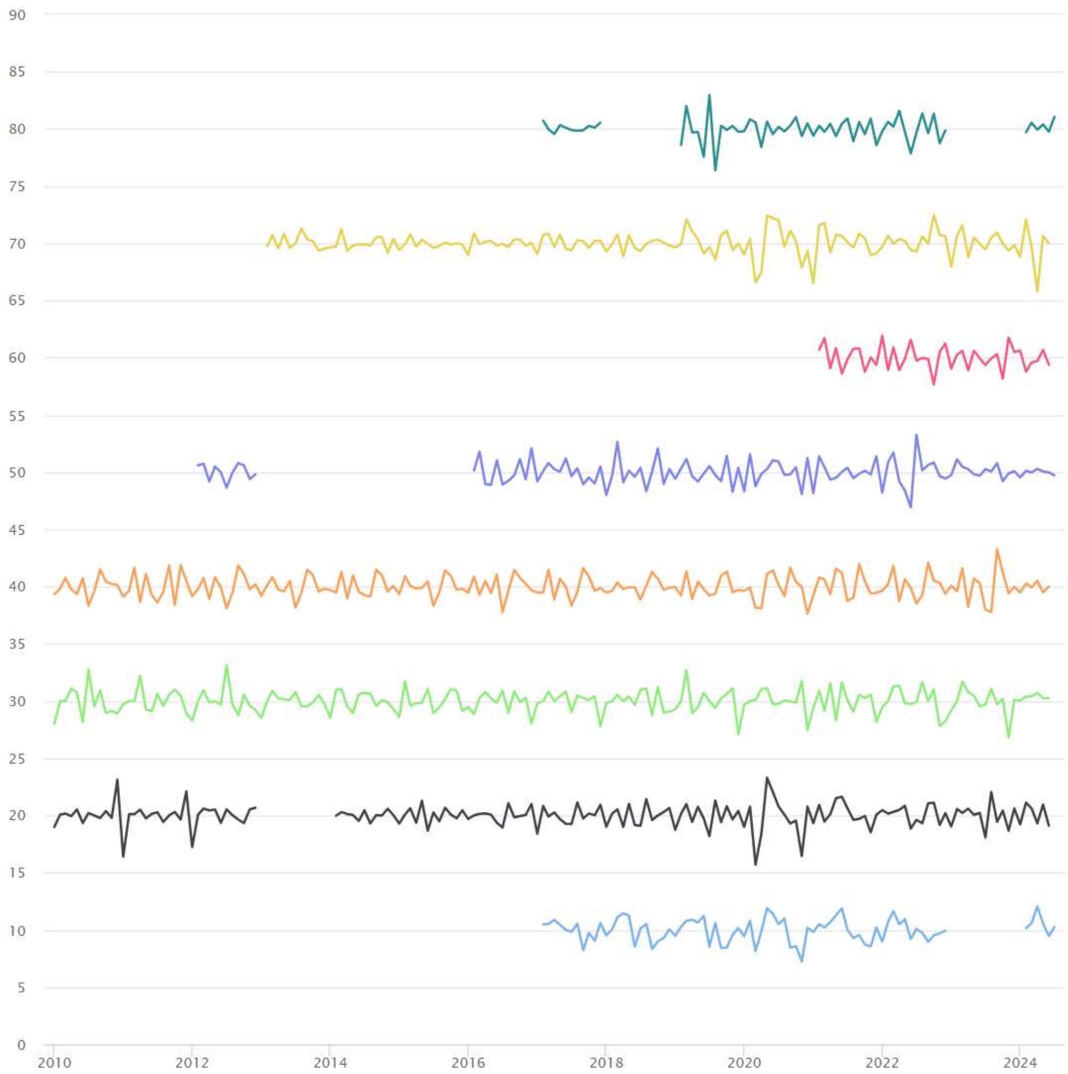
Short Term Statistics (NACE Sections B-F)

- Organized as cut-off survey
  - Total of 65 000 enterprises and establishments
  - Survey: about 10 000 enterprises
    - ~ 17% in terms of the number of enterprises
    - ~ 95% in terms of total turnover
  - Mandatory participation

Short Term Statistics

- Publication
  - Early Estimates t+30
  - Indices (Production, Turnover, …) at t+40
  - Absolute Numbers at t+70
  - Absolute Numbers at t+1Y (12 months)

- Correction of „Data Errors“ for Early Estimates
- Automatic procedure
- Time Series Approach



Outlier Adjustment - Requirements

- „Assist“ data editing process
- Number of outliers should not exceed 20%, rather much less
- Focus on outliers at the very end of time series
- Identification procedure shoud be conservative

Outlier Adjustment – Standard Procedures

- Seasonal Adjustment Methods
  - X13-Arima/Seats, Tramo/Seats
    - Time series with more than 35 Observations (monthly)
    - Sensitivity of Outlier Identification

- R Procedures
  - tsoutliers from Forecast Package
  - tsoutliers Package
  - for time series of any length

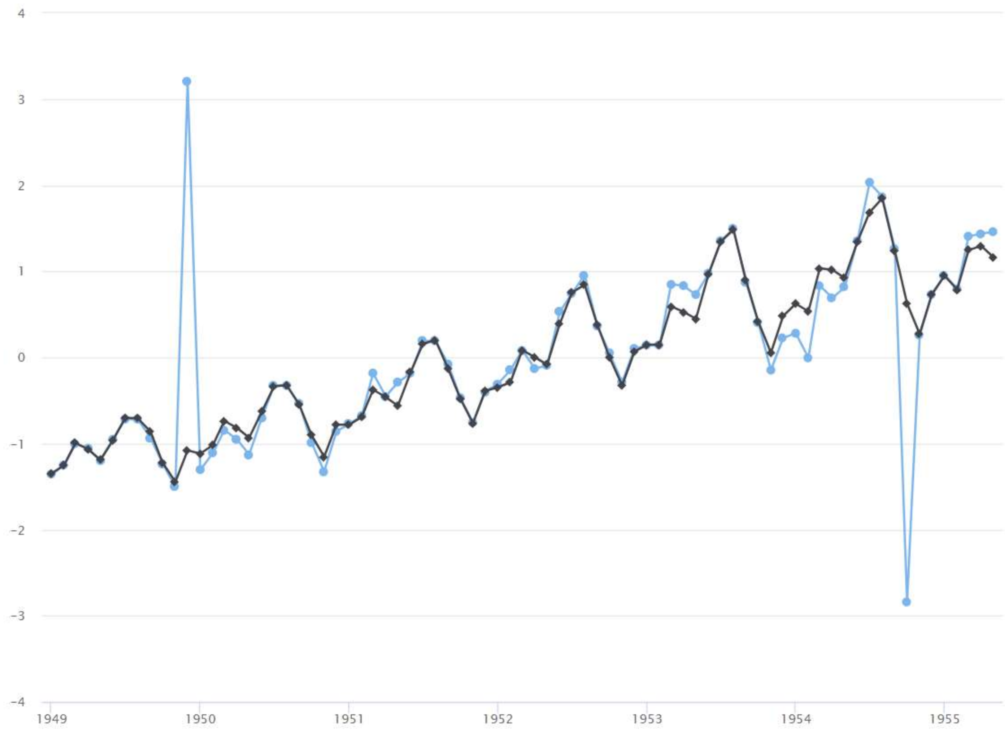
Outlier Adjustment and Replacement

- Robust time-series approach based on Rousseeuw et al.

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

Outlier Adjustment and Replacement

- Estimation with Support Vector Regression
  - applicable for nonlinear models
  - robust approach



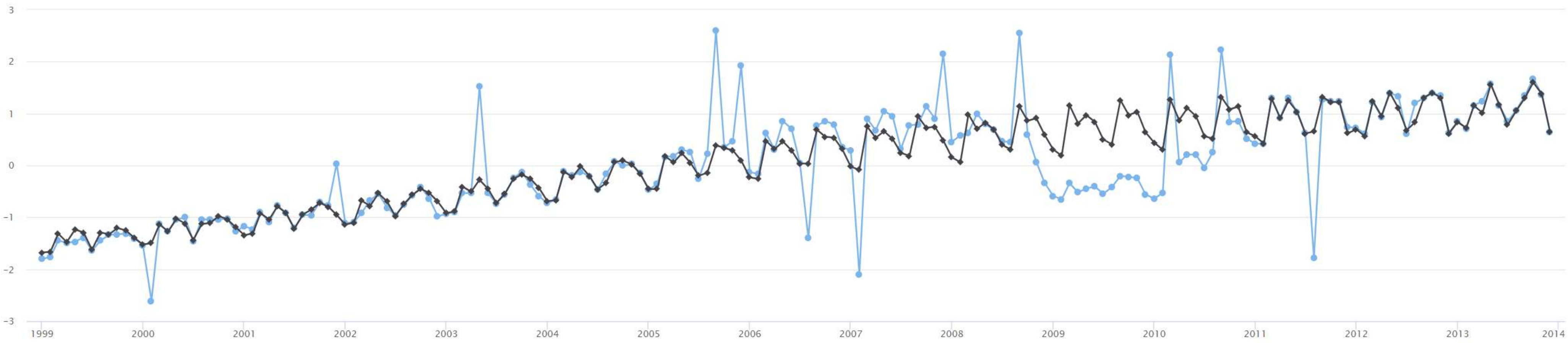
Identification of Level Shifts

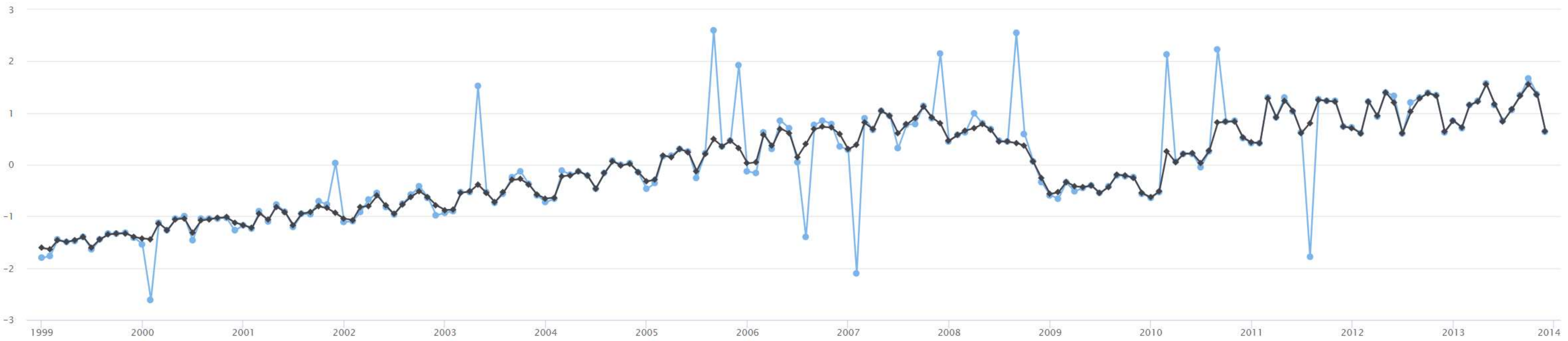
- Sequentially:

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

  – OLS regression with level shift position between 4,…,N-3
  – Fixing level shift for regression with minimal residual variance AND significant size of level shift

Identification of Level Shifts



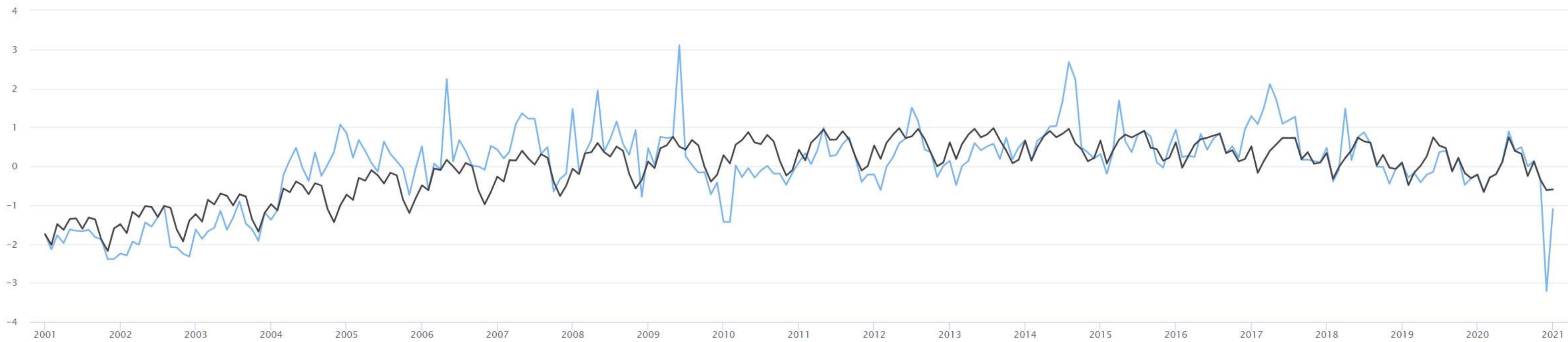


Long Time Series

- Changing seasonality
  - Split time series into smaller parts
  - Select segments (or windows) of the time series randomly
  - Fit the model for each segment with SVR
  - Repeat b times

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

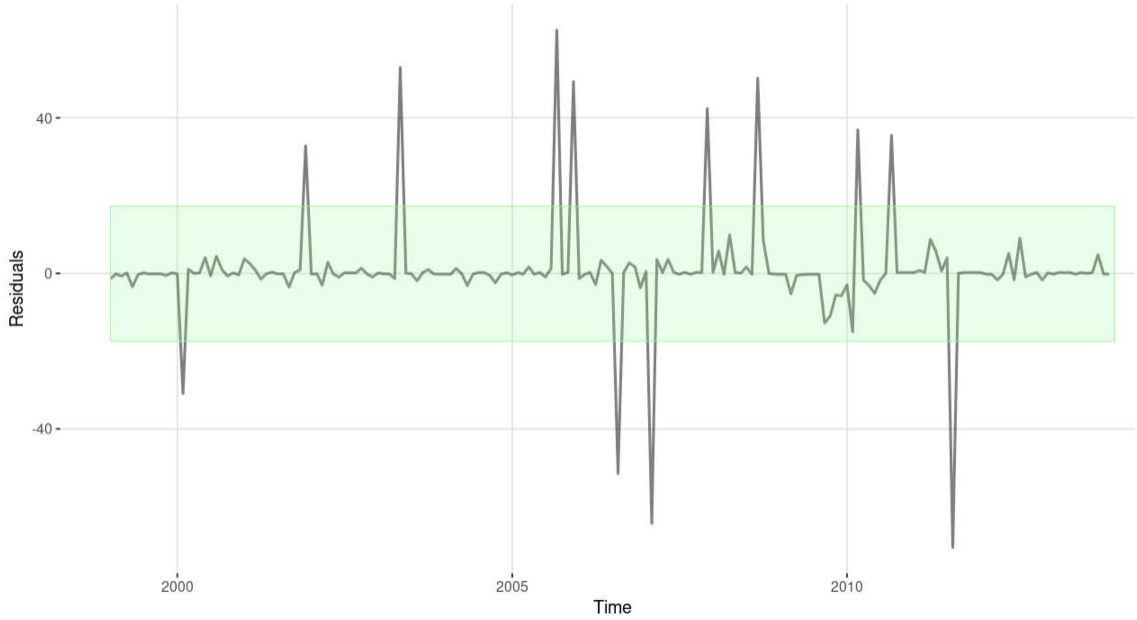
Long Time Series



Identification of Outliers

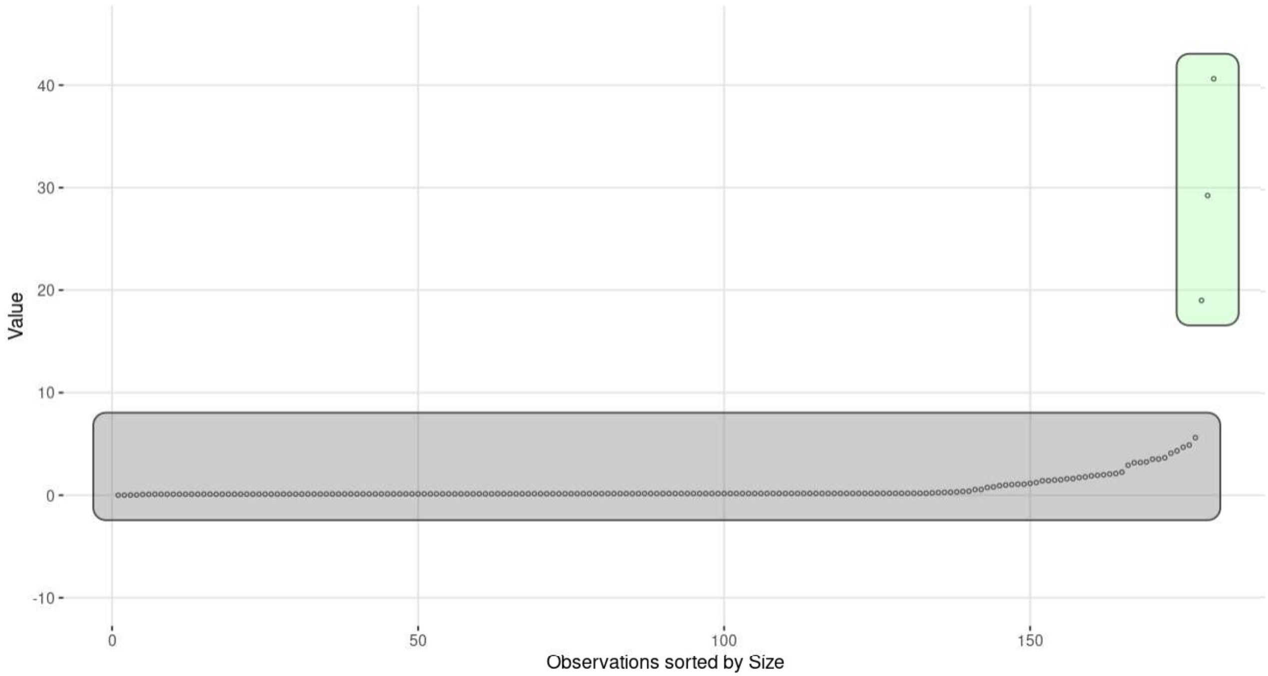
- Outlier-identification based on Residuals

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



Identification of Outliers

- Outlier-identification based on Residuals
- With K-Means Clustering



Results – Long Time Series (168-180 Obs)

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

Results – Long Time Series (168-180 Obs)

Method	Outliers	Low	Zero	Exact	High
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

Results – Long Time Series (168-180 Obs)

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

Results – Short Series (16, 8 Obs)

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
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

Bibliography

[1] 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: http://dx.doi.org/10.2307/1392565.

[2] 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.

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

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

[5] Fröhlich M. Outlier Identification and Adjustment for Time Series. Statistical Journal of the IAOS, vol. 40, Issue 2, pp. 389-402, 2024.

Please address queries to
Markus Fröhlich
Mail: markus.froehlich@statistik.gv.at

STATISTIK AUSTRIA
Guglgasse 13, 1110 Wien

Independent statistics for evidence-based decision making

STATISTICS AUSTRIA
The Information Manager