MARITIME TECHNICAL JOURNAL

2022 1 (224)

sciendo

DOI:10.2478/sjpna-2022-0001

IMPACT OF STARTING OUTLIER REMOVAL ON ACCURACY OF TIME SERIES FORECASTING

Vadim Romanuke

Polish Naval Academy, Faculty of Mechanical and Electrical Engineering, Śmidowicza 69 Str., 81-127 Gdynia, Poland; e-mail: v.romanuke@amw.gdynia.pl; ORCID ID: 0000-0003-3543-3087

ABSTRACT

The presence of an outlier at the starting point of a univariate time series negatively influences the forecasting accuracy. The starting outlier is effectively removed only by making it equal to the second time point value. The forecasting accuracy is significantly improved after the removal. The favorable impact of the starting outlier removal on the time series forecasting accuracy is strong. It is the least favorable for time series with exponential rising. In the worst case of a time series, on average only 7 % to 11 % forecasts after the starting outlier removal are worse than they would be without the removal.

Key words:

time series forecasting, outlier, ARIMA, forecasting accuracy, RMSE, MaxAE.

Research article

INTRODUCTION

In time series analysis and forecasting, data preparation and preprocessing is a very important phase before obtaining factual forecasts. Raw time series are usually are subject to low-pass filtering that removes random fluctuations and outliers [9, 10, 28]. In addition, this sometimes may help in restoring missed or misinterpreted data [2, 16]. Thus the time series is smoothed [12, 26, 27].

Surely, smoothing is not always perfect. Moreover, if the starting point in the time series is an outlier, the result of smoothing may be unsatisfactory. How much does it negatively influence the forecasting accuracy? Can the accuracy be improved by removing the starting outlier in the time series? These questions are to be studied and answered for further ascertaining the methods of improving the time series analysis quality.

GOAL

The goal is to determine the impact of the starting outlier removal on the time series forecasting accuracy. To achieve the goal, the following four tasks are to be completed:

1. To ascertain the most probable cause of the starting outlier in a preprocessed (smoothed) time series.
2. To suggest a method to remove it (the removal implies an appropriate modification).
3. To define a set of benchmark time series for testing the forecasting accuracy before and after the starting outlier removal.
4. To discuss and conclude on the obtained results.

THE PROBABLE CAUSE OF THE STARTING OUTLIER

Time series are not fitted by curves because then a fitting curve either approximates the time series average (expected value as a function of time) or leads to overfitting [20, 21]. In both cases, forecasting by curve-fitting is badly inaccurate. A time series can be only smoothed with a purpose to eliminate high-frequency fluctuations which are most probably consequences of true randomness [8, 15, 17,

27]. There are six basic smoothing methods [3, 4, 7, 13, 26]: moving average (MA), local regression using weighted linear least squares (LRWLS) and a 1st degree polynomial model (LRWLS-1), LRWLS and a 2nd degree polynomial model (LRWLS-2), Savitzky — Golay filter, the robust versions of LRWLS-1 and LRWLS-2. Every method does have its merits and demerits, but the starting outlier is not properly removed. An example to this is presented in fig. 1, where the starting outlier is clearly seen. Although the time series without the starting outlier is smoothed well enough (except for the robust LRWLS-2 method, which is not shown), no one of these methods removes the outlier appropriately. In the example, the robust LRWLS-1 method is the closest to solve this problem, but its modification of the outlier is followed by the changed value at the second time point resulting in the difference between the starting value and the second value becomes positive (whereas it is expected to be negative due to the starting outlier is likely to be less than the second value but just somehow badly "dropped" down). Therefore, even the robust LRWLS-1 method cannot be generally accepted for removing the starting outlier as it may generate an "updated" version of the outlier.

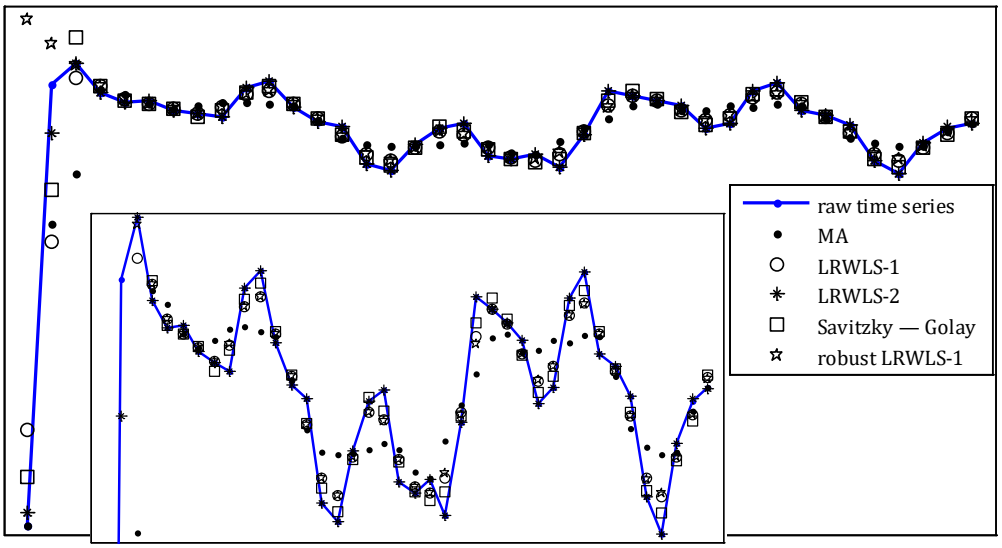

Fig. 1. An example of the starting outlier in a raw time series and time series after smoothing

So, smoothing either does not remove the starting outlier or modifies it inappropriately. Thus, the most probable cause of the starting outlier in a preprocessed (smoothed) time series is the obvious limitation of the smoothing methods. This should be rectified by suggesting a simple approach which would deal only

with the outlier removing it via satisfactory modification of its value. The remaining time series is left as it is (maybe, for smoothing by the considered methods or other manipulations).

THE STARTING OUTLIER REMOVAL

Denote by T the amount of a time series data, which are formally denoted by

$$\left\{y\left(t_i\right)\right\}_{i=1}^{T}, \tag{1}$$

where, without losing generality, $t_i = i$. Data (1) can be also referred to as the time series. If $y\left(t_1\right)$ is the outlier, then it might be removed by just setting

$$y\left(t_1\right) = \max_{k=\overline{2,T}} y\left(t_k\right) \text{ for } y\left(t_1\right) > y\left(t_2\right) \tag{2}$$

or

$$y\left(t_1\right) = \min_{k=\overline{2,T}} y\left(t_k\right) \text{ for } y\left(t_1\right) < y\left(t_2\right). \tag{3}$$

However, if the time series has a trend, either of modifications (2) and (3) is improper. Examples of this are shown in fig. 2.

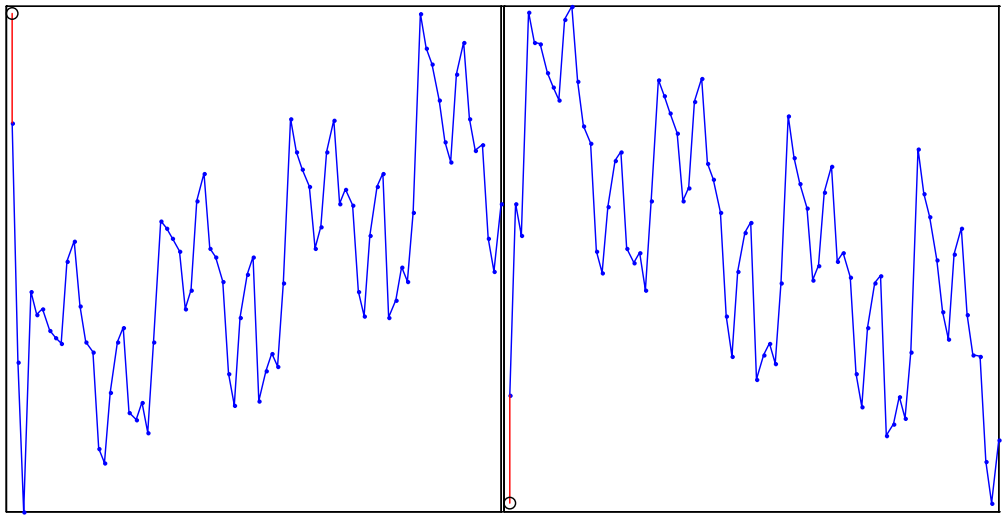

Fig. 2. Examples of the improper removal of the starting outlier by (2) and (3) due to a trend

Other types of the starting outlier modification involving subsequent time point values are improper as well unless the time series is detrended. In general case, determining the trend is influenced by the starting outlier also, so detrending may cause additional distortions of time series (1). Therefore, the starting outlier can be removed by simply making it equal to the second time point value:

$$y(t_1) = y(t_2). \tag{4}$$

Although the starting outlier modification by (4) is quite naive, it does not depend on a trend and thus remaining data

$$\left\{ y(t_i) \right\}_{i=2}^{T} \tag{5}$$

are left untouched.

BENCHMARK TIME SERIES

The benchmark time series are based on 12 random-like sequences (12 patterns) with repeatability, where every sequence is a stack of 6, 7, or 8 identical randomly-structured subsequences. These sequences are denoted by $\left\{ r_g(t) \right\}_{g=1}^{12}$, where every sequence is generated by using pseudorandom numbers drawn from the standard normal distribution (with zero mean and unit variance) [8, 15, 22, 23] by $t = \overline{1, T}$. In addition, vectors $\left\{ \Theta_l(T) \right\}_{l=1}^{30}$ of T pseudorandom numbers (these vectors are used to simulate noise and volatility), a set $\left\{ a_h > 0 \right\}_{h=1}^{6}$ of adjustable coefficients, and factor $\upsilon > 0$ indicating an oscillation frequency are used to form an initial set of benchmark time series.

Thus, a time series pattern without additional properties is

$$y_1(t) = \left[a_1 + 0.25\Theta_1(T) \right] r_1(t) + a_2\Theta_2(T). \tag{6}$$

A time series pattern with a linear trend is

$$y_2(t) = \left[a_1 + 0.25\Theta_3(T) \right] r_2(t) + a_2\Theta_4(T) + a_3 t, \tag{7}$$

and a time series pattern with seasonality is

$$y_3(t) = \left[a_1 + 0.25\Theta_5(T) \right] r_3(t) + a_2\Theta_6(T) + \left[a_4 + 0.25\Theta_7(T) \right] a_5 \cos(\upsilon t). \tag{8}$$

These three first patterns are then used in various combinations to form the remaining nine patterns including exponential extinction and rising properties. Thus, the nine patterns are formed as follows:

$$y_4(t)=\left[a_1+0.25\Theta_8(T)\right]r_4(t)+a_2\Theta_9(T)+a_3t+\left[a_4+0.25\Theta_{10}(T)\right]a_5\cos(\upsilon t), \quad (9)$$

$$y_5(t)=\left[a_1+0.25\Theta_{11}(T)\right]r_5(t)e^{-a_6t}+a_2\Theta_{12}(T), \quad (10)$$

$$y_6(t)=\left[a_1+0.25\Theta_{13}(T)\right]r_6(t)e^{a_6t}+a_2\Theta_{14}(T), \quad (11)$$

$$y_7(t)=\left[a_1+0.25\Theta_{15}(T)\right]r_7(t)e^{-a_6t}+a_2\Theta_{16}(T)+a_3t, \quad (12)$$

$$y_8(t)=\left[a_1+0.25\Theta_{17}(T)\right]r_8(t)e^{-a_6t}+$$
$$+a_2\Theta_{18}(T)+\left[a_4+0.25\Theta_{19}(T)\right]a_5\cos(\upsilon t)e^{-a_6t}, \quad (13)$$

$$y_9(t)=\left[a_1+0.25\Theta_{20}(T)\right]r_9(t)e^{-a_6t}+$$
$$+a_2\Theta_{21}(T)+a_3t+\left[a_4+0.25\Theta_{22}(T)\right]a_5\cos(\upsilon t)e^{-a_6t}, \quad (14)$$

$$y_{10}(t)=\left[a_1+0.25\Theta_{23}(T)\right]r_{10}(t)e^{a_6t}+a_2\Theta_{24}(T)+a_3t, \quad (15)$$

$$y_{11}(t)=\left[a_1+0.25\Theta_{25}(T)\right]r_{11}(t)e^{a_6t}+$$
$$+a_2\Theta_{26}(T)+\left[a_4+0.25\Theta_{27}(T)\right]a_5\cos(\upsilon t)e^{a_6t}, \quad (16)$$

$$y_{12}(t)=\left[a_1+0.25\Theta_{28}(T)\right]r_{12}(t)e^{a_6t}+$$
$$+a_2\Theta_{29}(T)+a_3t+\left[a_4+0.25\Theta_{30}(T)\right]a_5\cos(\upsilon t)e^{a_6t}. \quad (17)$$

Initially, a time series is generated by

$$a_1=2,\ a_2=0.175,\ a_3=0.01,\ a_4=5,\ a_5=0.18,\ \upsilon=0.02,\ a_6=0.0005,\ T=1680.$$

Then the time series is equidistantly downsampled so that 168 time points remain. These points are smoothed producing thus the benchmark time series. For each of patterns (6) — (17), 200 series are generated. For each of those 2400 series, ARIMA forecasts [2, 14] are made at $t=\overline{113,168}$ (i. e, the forecast length is one third of the available data). The forecasting accuracy is estimated by the corresponding root-mean-square error (RMSE) and the maximum absolute error (MaxAE) [6, 9, 11, 25] as follows. If

$$\left\{\tilde{y}(t)\right\}_{t=113}^{168} \tag{18}$$

are forecasted data, they are normalized with respect to the initial data:

$$\tilde{u}(t)=\frac{\tilde{y}(t)-\min\limits_{k=\overline{113,168}} y(k)}{\max\limits_{k=\overline{113,168}} y(k)-\min\limits_{k=\overline{113,168}} y(k)} \quad \text{by} \quad t=\overline{113,168}. \tag{19}$$

Test data

$$\left\{y(t)\right\}_{t=113}^{168} \tag{20}$$

are normalized as well:

$$u(t)=\frac{y(t)-\min\limits_{k=\overline{113,168}} y(k)}{\max\limits_{k=\overline{113,168}} y(k)-\min\limits_{k=\overline{113,168}} y(k)} \quad \text{by} \quad t=\overline{113,168}. \tag{21}$$

Then the RMSE is calculated as

$$\rho_{\text{RMSE}}=\sqrt{\frac{1}{56}\sum_{t=113}^{168}\left[u(t)-\tilde{u}(t)\right]^2} \tag{22}$$

and the MaxAE is calculated as

$$\rho_{\text{MaxAE}}=\max_{t=\overline{113,168}}\left|u(t)-\tilde{u}(t)\right|. \tag{23}$$

Obviously, MaxAE (23) registers information about the worst outlier [6, 11, 17]. Therefore, RMSE (22) and MaxAE (23) are used to see the averaged and worst errors in forecasting.

The 50 series which are forecasted the worst are extracted for each pattern. Their respective RMSEs are sorted in descending order, so each series is tagged to its number $z=\overline{1,50}$ ($z=1$ corresponds to the maximal RMSE). Finally, $y_g^{(\text{obs})}(1)=y_g(1)$, and the starting outlier is intensified for benchmarking as follows:

$$y_g(1)=y_g^{(\text{obs})}(1)-\sqrt{z}\left(y_g(2)-y_g^{(\text{obs})}(1)\right) \quad \text{by} \quad y_g(2)>y_g^{(\text{obs})}(1) \tag{24}$$

and

$$y_g(1)=y_g^{(\text{obs})}(1)+\sqrt{z}\left(y_g^{(\text{obs})}(1)-y_g(2)\right) \quad \text{by} \quad y_g(2)<y_g^{(\text{obs})}(1). \tag{25}$$

Graphical examples of three benchmark time series per pattern forecasted the worst are presented in fig. 3.

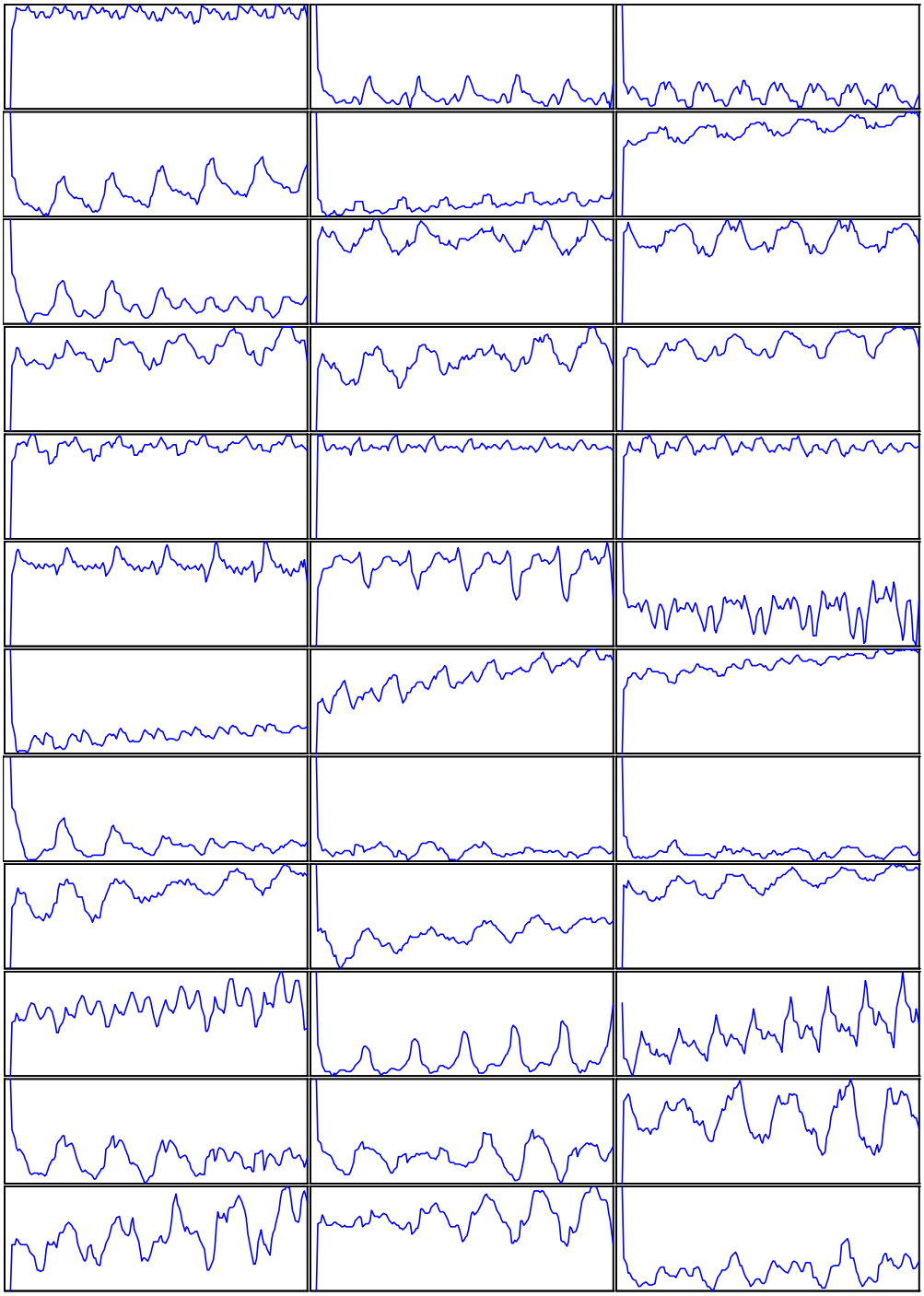

Fig. 3. Three benchmark time series per pattern forecasted the worst

THE FORECASTING ACCURACY IMPROVEMENT

For each of the extracted 600 time series, ARIMA forecasts are made prior to the starting outlier removal and after it. The difference between the respective RMSEs (22) of the forecasts is shown in fig. 4 as a polyline. The difference between the respective MaxAEs (23) is shown in fig. 5. Both plots confirm that the removal significantly improves the forecasting accuracy. The worst and best forecasts prior to the removal (squares) and after it (circles) are shown in fig. 6. However, it is

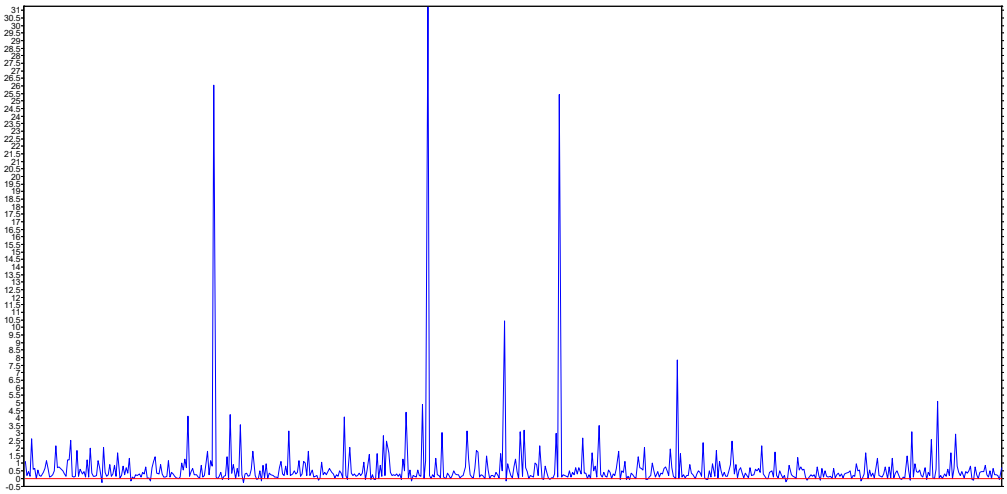

Fig. 4. The difference between RMSEs (22) of the 600 forecasts prior to the removal and after it

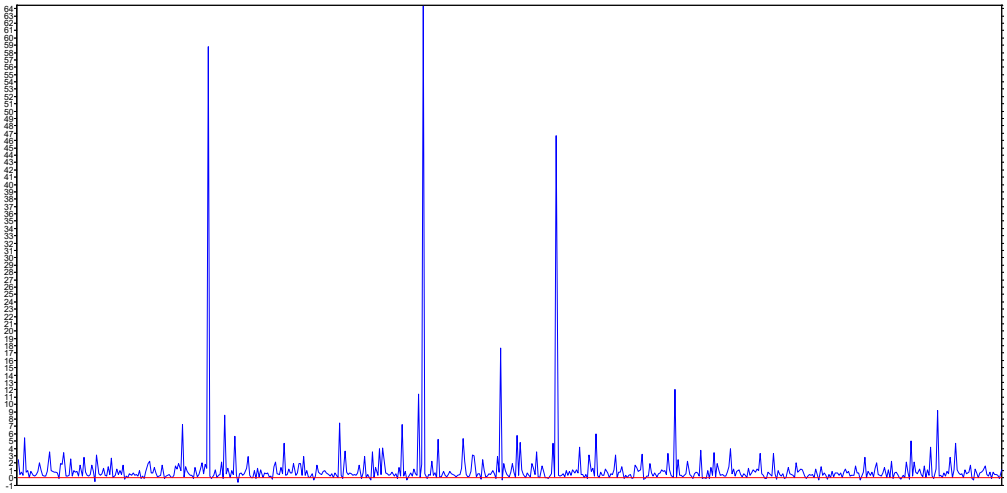

Fig. 5. The difference between MaxAEs (23) of the 600 forecasts prior to the removal and after it

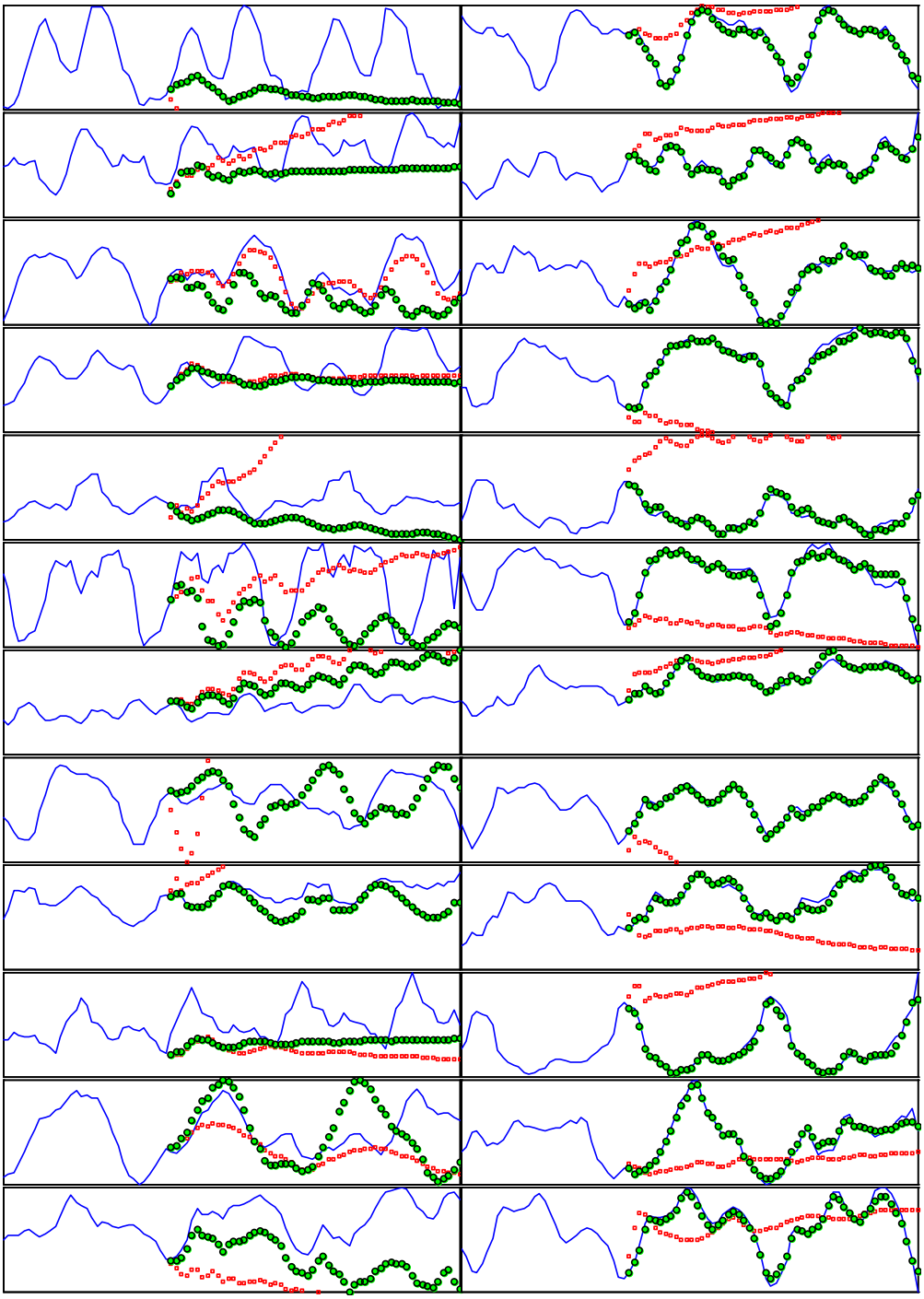

Fig. 6. The worst (left) and best (right) forecasts prior to the removal and after it

worth noting that not all the 1200 differences in fig. 4 and fig. 5 are positive (i. e., not every time series is forecasted more accurately after the starting outlier removal). Although the polylines do rarely drop below the horizontal zero level (difference) lines, there are 56 time series forecasted worse by the RMSE after the removal. This is 9.33 % of the benchmark volume. Besides, there are 65 time series forecasted worse by the MaxAE after the removal, which is 10.83 % of the benchmark volume. Obviously, a worse RMSE does not imply a worse MaxAE (with respect to forecasts prior to the removal) and vice versa (see fig. 7). If to consider the RMSE and MaxAE simultaneously, there are 44 time series forecasted worse after the removal, which is 7.33 % of the benchmark volume (see fig. 8). It is noticeable

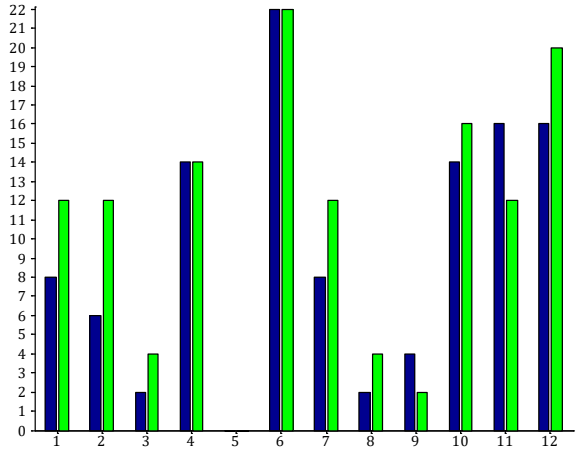

Fig. 7. The percentage of worse RMSEs (darker bars) and worse MaxAEs (lighter bars) after the starting outlier removal per pattern

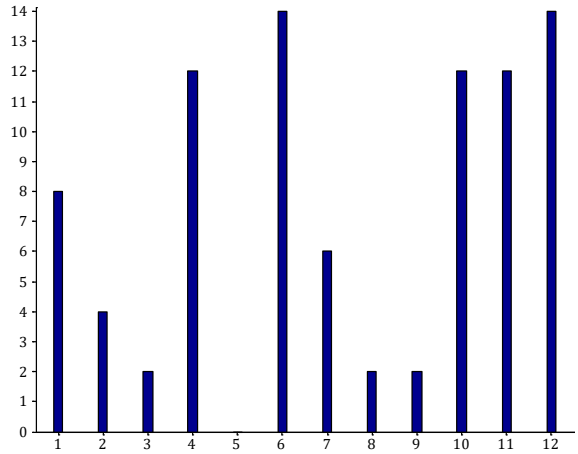

Fig. 8. The percentage of simultaneously worse RMSEs and MaxAEs after the removal per pattern

that the starting outlier removal has not worsened the accuracy for the 50 times series with exponential extinction by (10). This is seen in both fig. 7 and fig. 8, where the fifth bar place is empty (i. e., it is zero). On the contrary, the bar plots allow concluding on that the removal has the weakest favorable impact on the time series with exponential rising by (11).

DISCUSSION

The results presented in fig. 4 — 8, are obtained under roughly the worst conditions as the 50 worst-to-forecast time series out of 200 series have been studied per each pattern. This study approach reminds the maximin method, where an object or system is improved ("maximized") under the worst ("minimized") conditions [17, 18, 19, 24]. Thus, the best-under-worst-conditions behavior of the system is guaranteed. The results hereinabove obtained can be thought of as an approximately guaranteed "behavior" of the time series forecasts prior to the starting outlier removal and after it.

Based on studying the 600 time series divided into 12 patterns, both fig. 4 and fig. 5 (with the plots of the forecasting accuracy criteria) confirm that the presence of the starting outlier does negatively influence the forecasting accuracy. These plots also confirm that the forecasting accuracy is significantly improved by removing the starting outlier. However, the improvement is "guaranteed" only on average. Moreover, even if the accuracy is improved after the removal, it still may be unacceptable. For instance, eight of the worst-of-the-worst forecasts in fig. 6 (left column) have less both RMSE and MaxAE after the removal, but the accuracy is visibly very poor. The eighth subplot (the fifth from the bottom), by the way, corresponds to that case with the highest peaks of the RMSE (fig. 4) and MaxAE (fig. 5) differences (i. e., in this case of a time series having seasonality with exponential extinction the starting outlier removal has improved the accuracy best of all, but the result is quite unacceptable).

CONCLUSION

The starting outlier in a time series is effectively removed only by making it equal to the second time point value. In the worst case of a time series, on average

only 7 % to 11 % forecasts after the starting outlier removal are worse than they would be without the removal. Therefore, the favorable impact of the starting outlier removal on the time series forecasting accuracy is indeed strong. Nevertheless, the impact on time series with exponential rising is the least favorable. Their rate of the post-outlier-removal-accuracy drop is about 12 % to 20 %. Roughly speaking, the same percentage of time series with linear trend and seasonality, after the removal, are forecasted poorer also. Time series without additional properties (trend, seasonality, exponential, or other) have the post-outlier-removal-accuracy drop at about 8 % to 12 %.

The research might be furthered by considering a possibility to automatically detect the occurrence of the starting outlier removal. Such a possibility, for instance, can be based on using the Hampel filtering [1, 5]. The presence of a trend or other properties, however, may be an obstacle which will require a trickier Hampel filtering to detect outliers in univariate time series.

REFERENCES

[1] Astola J, Kuosmanen P., *Fundamentals of Nonlinear Digital Filtering*, CRC Press, 1997.

[2] Box G., Jenkins G., *Time Series Analysis: Forecasting and Control*, Holden-day, San Francisco, 1970.

[3] Cleveland W. S., Devlin S. J., *Locally-weighted regression: an approach to regression analysis by local fitting*, 'Journal of the American Statistical Association', 1988, Vol. 83, Iss. 403, pp. 596 — 610.

[4] Cleveland W. S., *Robust locally weighted regression and smoothing scatterplots*, 'Journal of the American Statistical Association', 1979, Vol. 74, Iss. 368, pp. 829 — 836.

[5] Davies L, Gather U., *The identification of multiple outliers*, 'Journal of the American Statistical Association', 1993, Vol. 88, Iss. 423, 782 — 792.

[6] Edwards R. E., *Functional Analysis. Theory and Applications*, Hold, Rinehart and Winston, 1965.

[7] Fox J., Weisberg S., *An R Companion to Applied Regression (3rd ed.)*, SAGE, 2018.

[8] Gubner J., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, 2006.

[9] Hamilton J. D., *Time Series Analysis*, Princeton University Press, Princeton, NJ, 1994.

[10] Han J., Kamber M., Pei J., 12. Outlier detection, in: *Data Mining: Concepts and Techniques (Third Edition)*, Morgan Kaufmann, 2012, pp. 543 — 584.

[11] Hyndman R., Koehler A., *Another look at measures of forecast accuracy*, 'International Journal of Forecasting', 2006, Vol. 22, Iss. 4, pp. 679 — 688.

[12] Kotu V., Deshpande B., *Data Science (Second Edition)*, Morgan Kaufmann, 2019.

[13] Mills T. C., Chapter 8. Unobserved Component Models, Signal Extraction, and Filters, in: *Applied Time Series Analysis: A Practical Guide to Modeling and Forecasting*, Academic Press, 2019, pp. 131 — 144.

[14] Pankratz A., *Forecasting with Univariate Box — Jenkins Models: Concepts and Cases*, John Wiley & Sons, 1983.

[15] Papoulis A., *Probability, Random variables and Stochastic processes*, McGraw-Hill, 1991.

[16] Randel W. J., *Filtering and Data Preprocessing for Time Series Analysis*, 'Methods in Experimental Physics', 1994, Vol. 28, pp. 283 — 311.

[17] Romanuke V. V., *Theoretic-game methods of identification of models for multistage technical control and run-in under multivariate uncertainties*, Mathematical Modeling and Computational Methods, Vinnytsia National Technical University, Vinnytsia, Ukraine, 2014.

[18] Romanuke V. V., *Identification of the machining tool wear model via minimax combining and weighting subsequently specific models*, 'Information processing systems', 2015, Iss. 12 (137), pp. 106 — 111.

[19] Romanuke V. V., *Meta-minimax approach for optimal alternatives subset regarding the change of the risk matrix in ensuring industrial and manufacturing labor safety*, 'Herald of Khmelnytskyi national university. Technical sciences', 2015, No. 6, pp. 97 — 99.

[20] Romanuke V. V., *Appropriateness of DropOut layers and allocation of their 0.5 rates across convolutional neural networks for CIFAR-10, EEACL26, and NORB datasets*, 'Applied Computer Systems', 2017, Vol. 22, pp. 54 — 63.

[21] Romanuke V. V., *An attempt of finding an appropriate number of convolutional layers in CNNs based on benchmarks of heterogeneous datasets*, 'Electrical, Control and Communication Engineering', 2018, Vol. 14, No. 1,pp. 51 — 57.

[22] Romanuke V. V., *Decision making criteria hybridization for finding optimal decisions' subset regarding changes of the decision function*, 'Journal of Uncertain Systems', 2018, Vol. 12, No. 4, pp. 279 — 291.

[23] Romanuke V. V., *Minimal total weighted tardiness in tight-tardy single machine preemptive idling-free scheduling*, 'Applied Computer Systems', 2019, Vol. 24, No. 2, pp. 150 — 160.

[24] Romanuke V. V., *A minimax approach to mapping partial interval uncertainties into point estimates*, 'Journal of Mathematics and Applications', 2019, Vol. 42, pp. 147 — 185.

[25] Romanuke V. V., *Wind speed distribution direct approximation by accumulative statistics of measurements and root-mean-square deviation control*, 'Electrical, Control and Communication Engineering', 2020, Vol. 16, No. 2, pp. 65 — 71.

[26] Savitzky A., Golay M. J. E., *Smoothing and differentiation of data by simplified least squares procedures*, 'Analytical Chemistry', 1964, Vol. 36, Iss. 8, pp. 1627 — 1639.

[27] Schelter B., Winterhalder M., Timmer J., *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, Wiley, 2006.

[28] Zhao Y., Chapter 7. Outlier detection, in: *R and Data Mining: Examples and Case Studies*, Academic Press, 2013, pp. 63 — 73.

WPŁYW USUNIĘCIA POCZĄTKOWEJ WARTOŚCI ODSTAJĄCEJ NA DOKŁADNOŚĆ PROGNOZOWANIA SZEREGÓW CZASOWYCH

STRESZCZENIE

Wartość odstająca w punkcie początkowym jednowymiarowego szeregu czasowego negatywnie wpływa na dokładność prognozowania. W ramach przeprowadzonych badań dokonano analizy wpływu usunięcia wartości odstającej poprzez zrównanie jej z wartością drugiego punktu czasowego. Uzyskane wyniki wskazują, że przyjęta metoda znacznie poprawia dokładność prognozowania dla większości szeregów czasowych. Jednak w przypadku szeregów czasowych z wykładniczym wzrostem, metoda okazała się mniej skuteczna. Minimalny wzrost dokładności prognozowania wynosił w tym przypadku od 7 % do 11 %.

Słowa kluczowe:

prognozowanie szeregów czasowych, wartość odstająca, ARIMA, dokładność prognozowania, RMSE, MaxAE.