2 Example of Time Series Forecasting by SSA[1]

Let us illustrate the 'Caterpillar'-SSA technique [1] by the example of the time series forecast. We consider the time series FORT (monthly volumes of fortified wine sales in Australia from January 1984 till June 1994, Fig. 1). Certainly, time series forecasting should be based on the preliminary time series investigation. In this case we rely on the time series analysis performed in [2] (see also http://www.gistatgroup.com/cat/). We examine both the initial time series of length equal to 174 and its subseries consisting of the first 120 points. We name the former FORT174 whereas the latter FORT120. Presented graphs serves to illustrate concepts and theoretical aspects of the method.

Generally, forecasting by the Caterpillar-SSA method should be applied to time series governed (may be approximately) by linear recurrent formulae (LRFs). That's why we start with the study of the series FORT from this point of view.

2.1 Linear recurrent formula governing the time series

As we found out in [2], FORT time series can be decomposed into a sum of a signal and a noise. We performed this decomposition quite well using $L = 84$ (w-correlation between the signal component and the noise component is small enough; it is equal to 0.004). Thus, we know that the signal produces 11 eigentriples, so its trajectory space should be rather well approximated by a subspace of dimension 11. In turn, the subspace generates a LRF of dimension $83 = L - 1$ (we call such formulae *full LRFs*), such that the reconstructed time series (the signal) would approximated by a time series governed by this LRF.

Table 1 presents the information for 19 (of 83) leading roots of the characteristic polynomial corresponding to this LRF. The roots (recall that they are complex numbers) are ordered in descending of their moduli. The label "compl."for the "Type"column of Table 1 notes that this line relates to two conjugate complex roots. In this case, the absolute value of the root imaginary part is listed in the table. The first six rows can be interpreted easily: the rows 1-3 and 5-6 correspond to conjugate complex roots, which produce harmonics with periods 6, 4, 2.4, 12, and 3. Moduli greater than one correspond to harmonics with increasing amplitudes, a modulus less than one reflects decreasing amplitude. These results are in accordance with [2, Fig. 9]. The forth row of the table corresponds to the real-valued root with modulus equal to 0.997. There are no more main roots, since their amount should be equal to 11. Other roots are extraneous; furthermore, their moduli are less than one. Fig. 2 represents all the roots in complex coordinates (Re,Im).

Let us check whether the time series is well fitted by the full LRF. Maximum error of total approximation is equal to 132 and is smaller than 10% of the time series values. Note that we use the first 83 points as initial data for the LRF and so the approximation error is calculated starting from the 84th point.

Let us consider a new minimal LRF of dimension 11 generated by the eleven main polynomial roots. If we took the points 73-83 as the initial data for this LRF, it would better approximate the time series (maximum error is equal to 94). No wonder that if we took the first 11 points as the initial data, then the maximum error estimated on the points 12-174 would increase to 495. Thus we conclude that the time series is well approximated by the time series governed by the minimal 11-dimensional LRF with roots represented in Fig. 3. So we know the analytical form of the approximating time series up to constants before addends. Relation between results of local approximation is the same (magnitudes of errors are smaller).

[1] This is the example section of [3]. See also http://www.gistatgroup.com/cat/

Таблица 1: Time series FORT: roots of the characteristic polynomial for the full LRF

N	Re	Im	Modulus	Frequency	Period	Type
1	0.497	0.871	1.003	1.053	5.969	compl.
2	-0.002	1.000	1.000	1.573	3.994	compl.
3	-0.870	0.489	0.998	2.630	2.389	compl.
4	0.997	0.000	0.997	0.000	no	real
5	0.861	0.497	0.994	0.524	12.002	compl.
6	-0.478	0.866	0.989	2.075	3.028	compl.
7	-0.094	0.972	0.976	1.667	3.768	compl.
8	-0.391	0.894	0.975	1.983	3.168	compl.
9	0.796	0.563	0.975	0.615	10.212	compl.
10	0.401	0.888	0.975	1.147	5.480	compl.

Since we know the exact period of the time series periodical component (due to its seasonal behavior), we can adjust the LRF by changing the roots so that they correspond to periods 6, 4, 2.4, 12 and 3. (Recall that there is a one-to-one correspondence between linear recurrent formulae and roots of the characteristic polynomials.) This 11-dimensional formula is called an *adjusted minimal LRF*. The local approximation errors corresponding to the adjusted minimal LRF even decrease here.

Let us give an analytical form of the time series governed by the adjusted minimal LRF:

$$\begin{aligned} f_n = C_1 0.997^n + C_2 0.994^n \sin(2\pi n/12 + \phi_2) + \\ + C_3 \sin(2\pi n/4 + \phi_3) + C_4 1.003^n \sin(2\pi n/6 + \phi_4) + \\ + C_5 0.998^n \sin(2\pi n/2.4 + \phi_5) + C_6 0.989^n \sin(2\pi n/3 + \phi_6). \end{aligned} \tag{1}$$

The coefficients C_i and ϕ_i are determined by initial data. The addends are ordered by decreasing of their contributions to their sum for terms from 1 to 174 (i.e., by eigenvalue shares). Recall that ordering by roots moduli generally differs from ordering by eigenvalues, since roots moduli are related to rates of increasing/decreasing of time series components and thereby influence a future behavior of the time series governed by the corresponding LRF.

Thus, the preliminary investigation implies that the time series FORT well fits to the required model, so we can start the forecasting.

2.2 Forecast and confidence intervals

Robust long-term forecast. As we demonstrated above, the time series FORT is quite appropriate for forecasting: it consists of the signal perfectly approximated by the time series which is governed by the LRF and of the white noise. Fig. 4 depicts the first 24 points of the recurrent 60-step forecast. This forecast is performed using the full LRF produced by the subspace, which is generated by the leading 11 eigentriples. The results of the vector forecast as well as of the recurrent forecast that uses the minimal adjusted LRF 1 are very similar to the results depicted in Fig. 4 so we don't present them. Such coincidence just confirms robustness of the performed forecast.

Fig. 5 represents the bootstrap confidence intervals with confidence level equal to 0.95. Certainly, we should check whether the residuals satisfy necessary conditions. Thus we should check whether the residual time series $\widetilde{F}^{(2)} = F - \widetilde{F}^{(1)}$, where $\widetilde{F}^{(1)}$ is reconstructed using the leading 11 eigentriples signal, is a realization of independent normally distributed random variables. Testing of residuals has already been carried out during the analysis of the time series. Hypothesis of independence is not rejected whereas the criterion for testing of the normality hypothesis yields the p-level equal to 0.01. This slight non-normality in the residual distribution can lead to a small modification of the confidence intervals but we can rely on confidence limits anyway. You see that size of the confidence intervals increases during five years very slow.

Thus, we have constructed the robust 60-step forecast of the FORT time series. We can rely on it under the assumption that the structure of the time series will not change in future.

Short-term forecast. It is impossible for some data to do a proper long-term forecast. It can be caused by: 1) small length of a time series as compared to a noise magnitude and/or to a signal complexity; 2) non-linear structure of a signal. We consider structure of a signal as non-linear if the signal is poorly approximated by a time series governed by an LRF of relatively small dimension. Below we demonstrate only the first case.

Let us truncate FORT174 and consider the subseries consisting of the first 120 points of the initial time series, as well as we did for the time series analysis in [2]. We use designations FORT120 for this subseries, whereas FORT174 for the initial time series. The time series analysis performed in [2, Appendix B] yields that for the window length $L = 60$ the signal is produced by the leading 11 eigentriples. Recall that we observed mixing of two harmonics arising at the singular value decomposition stage. However, this mixture is insignificant for forecasting if we include these eigentriples 8–11 together.

Let us construct 60-step forecast for the FORT120 time series using the leading 11 eigentriples, as well as we did for FORT174. The structure of the eigenspace produced by the leading 11 eigentriples is the same both for the initial time series and for the subseries. However, since FORT120 is shorter, the quality of approximate separability of a signal from a noise is not so good. Error of global approximation confirms it; its value is equal to 312 (compare with 132 for FORT174).

When we forecast the truncated time series, we can compare the forecast values with the known time series values (that is, we have so called validation period). It helps us to estimate the forecast quality. Fig. 6 depicts the recurrent forecast of FORT120 (the bold solid line) as well as the FORT174 time series. It can be seen that the forecast describes the future behavior for the first 24 points very well, but then discrepancy begins to increase. Table 2 contains information about the leading 19 of 59 roots of characteristic polynomial corresponding to the full LRF used for the forecast of FORT120. The comparison with Table 1 makes clear that such behavior of FORT120 forecast is caused by a large modulus of complex conjugate roots corresponding to amplitude-modulated harmonic with 6-months period (the first line of Table 2). Fig. 6 confirms the supposition that this half-year harmonic with increasing amplitude is responsible for the discrepancy of the forecast with the time series data.

To demonstrate less robustness of the FORT120 time series forecast in comparison with reliability of the FORT184 forecast, we present confidence intervals. Fig. 7 shows that sizes of the bootstrap confidence intervals increase falling even into negative values area.

Let us make certain that the lack of separability does account for the poor quality of the long-term FORT120 forecast rather than changing of the time series structure after the 120th point. We apply to FORT120 the adjusted minimal LRF produced by the FORT174 time series and given by the formula (1). The result is presented in Fig. 8 and demonstrates a quite good quality of the forecast (the bold solid line connects the forecasted points while the real values are marked by squares).

By means of forecasting truncated time series we can estimate forecast errors, for example, the mean-square deviation of forecasted values from real values. The recurrent forecast using the

Таблица 2: Time series FORT120: roots of the characteristic polynomial of the full LRF

N	Re	Im	Modulus	Frequency	Period	Type
1	0.505	0.878	1.013	1.049	5.990	compl.
2	-0.885	0.480	1.007	2.644	2.376	compl.
3	0.000	1.000	1.000	1.571	4.001	compl.
4	0.997	0.000	0.997	0.000	no	compl.
5	0.862	0.496	0.994	0.522	12.033	compl.
6	-0.490	0.851	0.982	2.093	3.002	compl.
7	0.366	0.896	0.968	1.183	5.311	compl.
8	0.768	0.587	0.966	0.652	9.635	compl.
9	-0.128	0.957	0.966	1.704	3.688	compl.
10	-0.899	0.350	0.965	2.771	2.268	compl.

full forecasting LRF produces mean-square error equal to 549, the vector forecast yields to the smaller error 535, whereas the forecast using the minimal adjusted LRF diminishes the error up to 314.

Thus, the considered example of the time series analysis/forecasting confirms theoretical results about relation between forecast quality and separability accuracy.

Список литературы

[1] Golyandina N., Nekrutkin V., and Zhigljavsky A. *Analysis of Time Series Structure: SSA and Related Techniques*, London: Chapman & Hall/CRC, 2001, 305 P.

[2] Golyandina N. *The 'Caterpillar'-SSA Method for Time Series Analysis: Training Aids*, St.Petersburg, St.Petersburg State University, 2004, 87 P. (In Russian).

[3] Golyandina N. *The 'Caterpillar'-SSA Method for Time Series Forecasting: Training Aids*, St.Petersburg, St.Petersburg State University, 2004, 53 P. (In Russian).

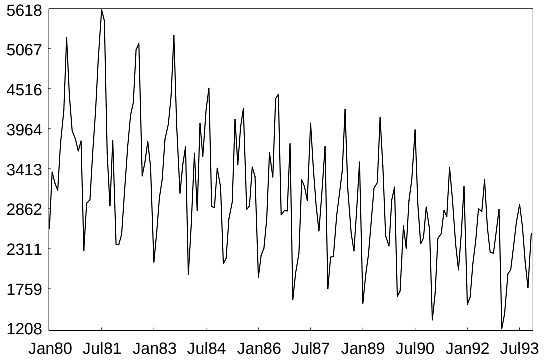

Рис. 1: FORT174: initial time series of monthly volumes of fortified wine sales in Australia

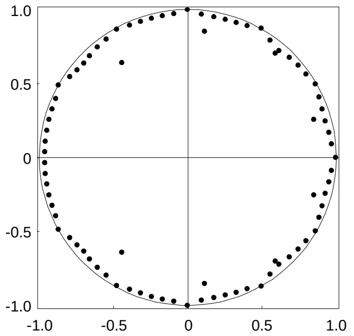

Рис. 2: FORT174: main and extraneous roots of characteristic polynomial for the full LRF

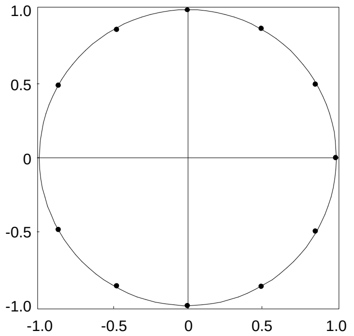

Рис. 3: FORT174: roots of characteristic polynomial for the minimal LRF

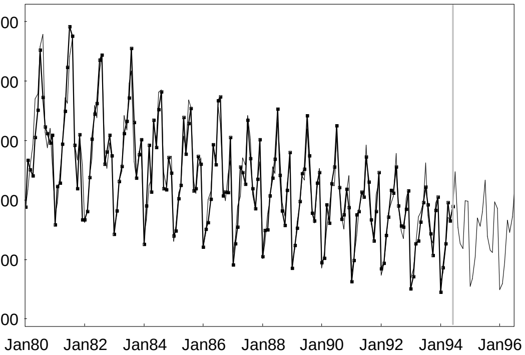

Рис. 4: FORT174: recurrent 24-step forecast

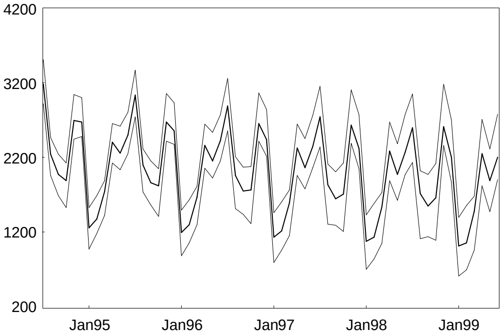

Рис. 5: FORT174: recurrent 60-step forecast and bootstrap confidence intervals

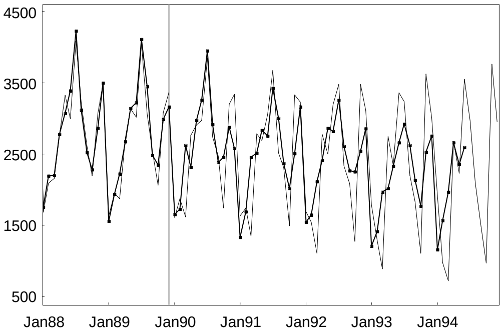

Рис. 6: FORT120: recurrent 60-step forecast using the full LRF on the background of the real values

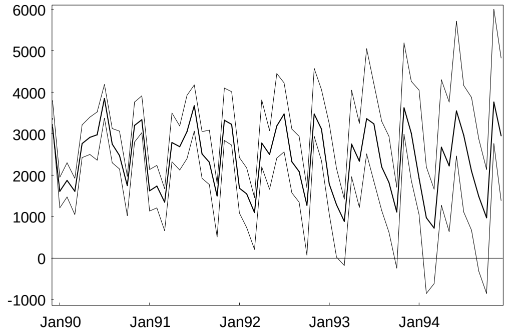

Рис. 7: FORT120: recurrent 60-step forecast and bootstrap confidence intervals

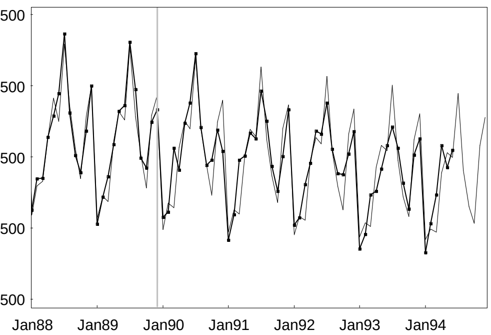

Рис. 8: FORT120: Recurrent 60-step forecast using the minimal adjusted LRF (FORT174) on the background of the real values