On the usage of the Singular Spectrum Analysis for precision estimation and editing of total atmospheric delay time series

M. S. Vasyuta[1], *V. Ya. Choliy*[1]

[1]Taras Shevchenko National University of Kyiv, Glushkova ave., 4, 03127, Kyiv, Ukraine
ecratos12@gmail.com

We use Singular Spectrum Analysis (**SSA**) for precision estimation of time series of total zenith atmospheric delay for a list of European GNSS data stations proceed in Main Astronomical Observatory GNSS processing center. The series are downloadable at [1]. Analysis of the principal components of the series allowed us to clean the series by removing noise out of them. With the capabilities of SSA some gaps in the data were filled out.
Key words: spectrum analysis, data quality, time series forecast.

INTRODUCTION

Singular Spectrum Analysis (SSA) and its two-dimensional extension (2dSSA) are frequently used in time series analysis. Author' experience along with some useful references were presented in [3], [4]. Here we investigate one promising feature of SSA: its forecasting capabilities. There are two variants of forecasts with SSA, but here we use linear recurrent formula (LRF). Our purpose - to fill the gaps in time series of total atmospheric delay keeping the statistical properties of the series. As a byproduct the estimation of noise part of the series will be made using Principal Component Analysis (PCA) as a key step of SSA.

Let outline here the properties of LRF. The explanation of the SSA itself can be found in [1], [2]. Given: $X_N = (x_1, x_2, \dots, x_N), N > 2$ - the initial series, $L : 1 < L < N$ - window length; $L^r \subset R^L, r < L$ - some linear space. Given: ort vector $e_L = (0, 0, \dots, 1) \in R^L, \notin L^r$. M - amount of the forecasted points to be found.

There is the algorithm for recurrent forecast:

1. $\mathbb{X} = (\vec{\mathbf{X}}_1 : \vec{\mathbf{X}}_2 : \cdots : \vec{\mathbf{X}}_K), K = N - L + 1$ - trajectory matrix for X_N;
2. $\vec{\mathbf{P}}_1, \dots, \vec{\mathbf{P}}_r$ - ortonormalised basis L_r;
3. $\hat{\mathbb{X}} = (\hat{\mathbb{X}}_1 : \hat{\mathbb{X}}_2 : \cdots : \hat{\mathbb{X}}_K) = \sum\limits_{i=1}^{r} \vec{\mathbf{P}}_i \vec{\mathbf{P}}_i^T \cdot \mathbb{X}$ - ortogonal projection of the vectors $\hat{\mathbf{X}}_i$ on L_r;
4. $\tilde{\mathbb{X}} = (\tilde{\mathbf{X}}_1 : \tilde{\mathbf{X}}_2 : \cdots : \tilde{\mathbf{X}}_K)$ - hankelised matrix $\hat{\mathbb{X}}$.
   Than $\tilde{\mathbb{X}}$ is trajectory matrix of some series $\tilde{X}_N = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_N)$;
5. $\forall \vec{\mathbf{Y}} \in R^L$ sign as $\vec{\mathbf{Y}}_\triangle$, the vector, consist of the last $L-1$, and $\vec{\mathbf{Y}}^\triangledown$ - consist of the first $L-1$ its own components;
6. $\nu^2 = \pi_1^2 + \pi_2^2 + \cdots + \pi_r^2$, where π_i - latter component of the vector $\vec{\mathbf{P}}_i$.

It might be shown that the last component of any vector $\vec{\mathbf{Y}} = (y_1, y_2, \dots, y_L)^T \in L_r$ is an linear combination of its first components:

$$y_L = a_1 y_{L-1} + a_2 y_{L-2} + \cdots + a_{L-1} y_1,$$

and parameters vector is:

$$\vec{R} = (a_{L-1}, a_{L-2}, \dots, a_2, a_1)^T = \frac{1}{1-\nu^2} \sum_{i=1}^{r} \pi_i \vec{\mathbf{P}}_i^\triangledown,$$

[1]ftp://ftp.mao.kiev.ua/pub/gnss/products/IGS05/

and it doesn't depend upon the preselected basis. Recurrently forecasted series $G_{N+M} = (g_1, g_2, \ldots, g_{N+M})$ can be build by recurrently extended the initial series:

$$g_i = \begin{cases} \tilde{x}_i, & i = 1, N, \\ \sum\limits_{j=1}^{L-1} a_j g_{i-j}, & i = N+1, N+M. \end{cases} \cdot$$

MODEL

Time series of editable is represented as tropospheric total delay values at step of one hour (here and below designated as TROTOT). Producing of the series is a part of activities of Main Astronomical Observatory GNSS processing center [5]. For further processing the arithmetic mean in all series were removed due to requirements of the SSA. Window size was preselected equal to one solar day, i.e $L = 24$ – the least obvious period of the series.

Numerical experiment have taken place to estimate the number of principal components necessary to represent total delay signal in time series. It was done with the largest solid part of european data - 10400 values ($433^d 8^h$) from 2002-01-17 00:30:00 to 2003-03-26 08:30:00 at the BOR1 station. Figure 1 gives the outline of total dispersion of the series represented by principal components. In author' previous work [4] it was shown that the series principal components, determined with SSA [2] might be subdivided in "noisy" and "deterministic" ones. In Figure 1 the horizontal part of graph includes a lot of "noisy" principal components. That is why we subdivide the series principal components by the position of turning point. It can be seen from Figure 1 that three, or maybe four principal components are sufficient for explanation of "deterministic" part of the series. Adding extra components adds nothing but noise to the result.

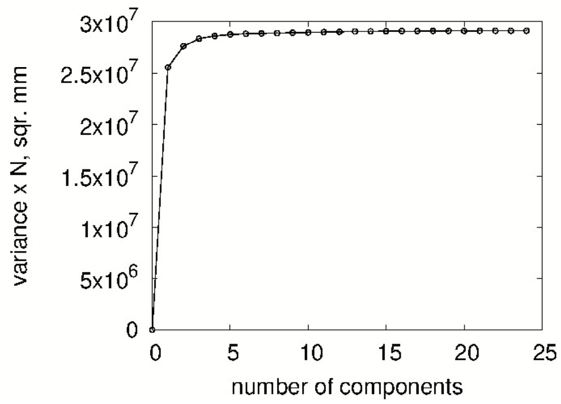

Fig. 1: Summary variance of first components

It should be noted that choosing different window sizes in SSA does not give any fundamentally new results for our testing – the SSA provides the same variance behavior of this kind. The table below shows the estimated signal and noise variances. It is typical for all processed series.

Table 1: Variance analysis result of the example series.

Selection	Total variance, mm^2
Σ^2	2801
A	2729
B	2753
Ratio	Value
A/Σ^2	0.9743
B/Σ^2	0.9829

In Table1 we present Σ^2 is total variation of the series, A - variation of the first three principal components, B - the same for four components. It is clearly seen that three components are responsible for

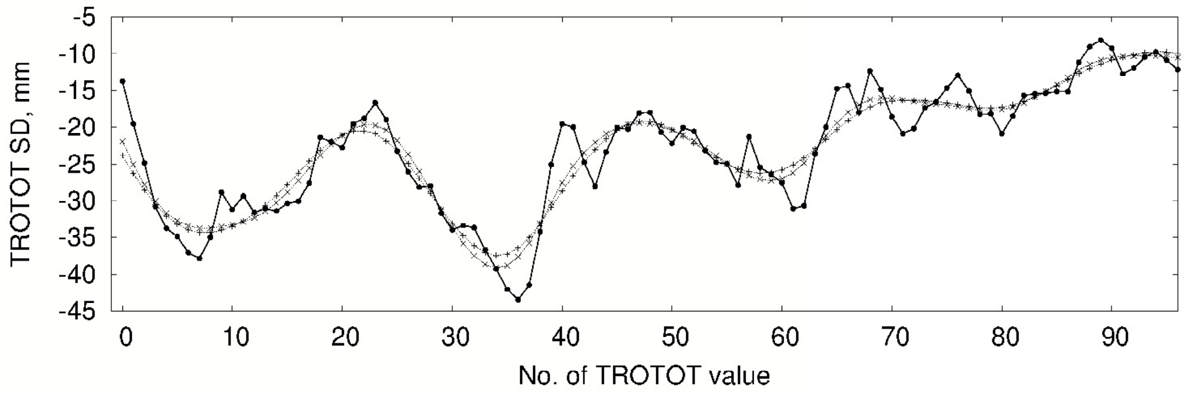

Fig. 2: 3 & 4 components visualization

0.9743 of total series variance. The remaining part of total variance is variance of the noise. Variation of the noise should be estimated as total variance minus given values. A and B used on the next pages in the same context. Figure 2 visualize all series with solid, the first three (+) and the first four selections (×) with dotted line. SD means standard deviation, i.e. value. This selections imitates the behavior of the series decently. It's not possible to clearly define 4-th component as principal one, as it seen in figure 2 - selections behave very similar. Adding it does not change the behavior of the principals.

GAP FILLING

Every gap in the data were interpreted as a place for internal forecasts. Left end of the gap is the starting point for forward forecast. The right end of the gap is the place where the backward forecast is applied. Those forecasts are moving to each other point by point and after all close the gap. It is obvious that the gap should not be wider than twise the size of the SSA window L. But sometimes the forecast works much better.

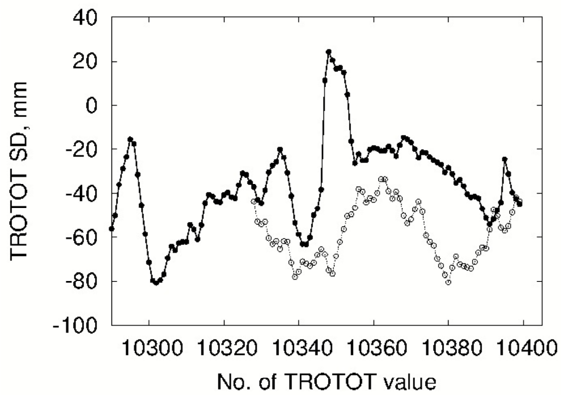

Fig. 3: Test gap filling demo, BOR1

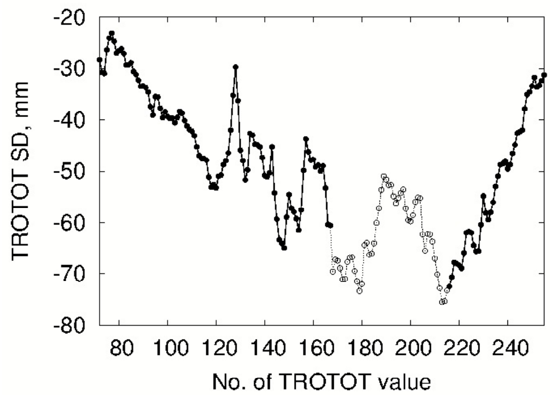

Fig. 4: Real gap filling demonstration, ALCI

We have used common way in testing forecasting capabilities, replacing real data by predicted ones. The performing on example data from model section with gap size equals 72 ($3L$) is shown in figure 3. The prediction decently mimics the appearance of oscillations, but can't handle with abrupt changes or outliers. These issues are discussed below.

The figure 4 shows a gap in ALCI series 48 points long (2 days) – from 2003-12-07 00:30:00 to 2003-12-09 00:30:00 – perfectly filled in.

PREEDITING THE SERIES

Unfortunately sometimes is happens that the gap is filled wrong.

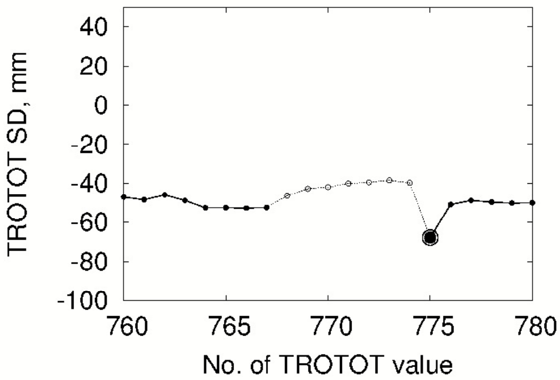

Fig. 5: Another gap filling demonstration, ALCI

Marked dot (date: 2004-01-01 07:30:00) in figure 5 is outlier. The LRF processing may lead to series' divergence. For this example forward forecast from left end of the gap looks OK, but backward one from right gap end diverges to abnormal negative values (not shown).

It may be proposed some explanation, say, measuring device was broken, or after repairing it gave wrong results. Such values not make practical sense because the probability of them is very small. That is why we applied sequential analysis of the original series near the gaps to throw away outliers [6]. The results are presented in table 2 with 3σ rejection criterion.

FINAL RESULTS

Summarizing all the remarks made it came up with the final results shown on the Table 2. Location column separates the Ukrainian and European stations.

We assume SSA works great in case of predicting of the series behavior. It can be seen in figure 2. The sequential analysis makes good help to make prediction smoother - so it helps to minimize station's equipment random error. However, if TROTOT suffers abrupt changes, this changes reproduced weakly in modeling, as it can be seen in figure 3. Theoretically, recurrent $T - th$ order forecast can be processed with sufficient precision $2T$ points forward. Practically, it works alright to up to $10T$-sized gaps (SULP). Anyway if existing gaps are very wide (like in EVPA, GRAZ, ISTA, RIGA) the filling of the gaps with LRF fail.

CONCLUSIONS

We have used Singular Spectrum Analysis for TROTOT series processing. We find out that if $L = 24$, the first three or four principal components produce decent modeling of "deterministic" part of the series. On this basis we can fill gaps in series confidently. Also we assume total dispersion of "noisy" principal components evaluates quantity of noise in series decently.

The computation time of processing the SSA takes from several minutes up to 3-4 hours on the common laptop (Eigen3 library (open-source, MPL2 license)[2], C++, Intel Core i3-5005, 4 cores), depending on series size ($\approx O(N^2)$).

REFERENCES

[1] Golyandina N., Nekrutkin V., Zhigljavsky A. Analysis of time series structure. Chapman, New York, 2001.- 310 p.

[2] Golyandina N., Zhigljavsky A. Singular spectrum analysis for time series. Springer, Berlin, 2013.- 125 p.

[2] http://eigen.tuxfamily.org

Table 2: Final Results.

name	location	points in total	outliers	A, mm^2	B, mm^2	Σ^2, mm^2	$1 - A/\Sigma^2$	$1 - B/\Sigma^2$
ALCI	ukr	25704	13	2548	2571	2609	0.023	0.015
BACA	euro	6949	0	2896	2917	2949	0.018	0.011
BAIA	euro	6960	0	2895	2915	2950	0.019	0.012
BOR1	euro	78120	9	2525	2553	2608	0.032	0.021
BUCU	euro	67623	25	64100	73060	147500	0.565	0.505
CNIV	ukr	10227	3	2468	2491	2529	0.024	0.015
COST	euro	7392	8	58770	73630	149100	0.606	0.506
CRAO	ukr	57181	6	1978	2031	2106	0.061	0.036
DNMU	ukr	11509	12	2362	2385	2422	0.025	0.015
EVPA	ukr	61776	13	1.044e+07	1.055e+07	1.259e+07	0.171	0.162
GLSV	ukr	78040	6	3736	3803	3915	0.046	0.029
GRAZ	euro	78120	7	1.385e+19	1.811e+19	3.072e+19	0.549	0.411
ISTA	euro	60000	6	7.538e+07	1.086e+08	7.750e+08	0.903	0.860
KHAR	ukr	34609	6	2529	2568	2629	0.038	0.024
KLPD	euro	15144	12	2159	2183	2215	0.025	0.014
LAMA	euro	78120	14	66840	85430	135700	0.507	0.370
MDVJ	euro	27552	22	2589	2612	2658	0.026	0.017
MIKL	ukr	39768	2	2699	2813	3527	0.235	0.202
MOBN	euro	27552	9	2585	2605	2647	0.024	0.016
PENC	euro	78120	8	5126	5823	7390	0.306	0.212
POLV	ukr	47753	11	2651	2697	2830	0.063	0.047
RIGA	euro	78120	15	1.314e+07	1.411e+07	1.661e+07	0.209	0.151
SHAZ	ukr	21144	21	6302	6729	7842	0.196	0.142
SULP	ukr	45228	8	2908	2955	3045	0.045	0.030
TRAB	euro	59983	24	4556	4925	7398	0.384	0.334
UZHL	ukr	64762	18	14860	15650	17030	0.127	0.081
VLNS	euro	67336	32	2062	2079	2107	0.022	0.013
ZECK	euro	78120	4	20850	24200	32140	0.351	0.247

[3] Choliy V., Formal estimation of the random component in global maps of total electron content // Advances in Astronomy and Space Physics.- 2016.- v.6, N.1.- P.56-60.

[4] Choliy V., Precision estimation and forecasting of the Earth orientation parameters with the SSA method // Kinematics and Physics of Celestial Bodies.- 2015.- v.31, N.4.- P.72-80.

[5] Ishchenko M., Reprocessing GPS-observations of permanent stations GPS-regional network for the GPS-weeks 935-1399 // Space science and technology.- 2014.- v.20, N.3.- P.41-48.

[6] Mytropolsky A., Statistical calculations technique. Science, Moscow, 1971.- 379 p.