New possibilities of the PCA-Seq method in the analysis of time series (on the example of solar activity)

V M Efimov¹,²,³,⁴, K V Efimov⁵, D A Polunin² and V Y Kovaleva³

¹ Institute of Cytology and Genetics SB RAS, 10 Ac. Lavrentieva Ave., Novosibirsk 630090, Russia
² Novosibirsk State National Research University, 1 Pirogova St., Novosibirsk 630090, Russia
³ Institute of Systematics and Ecology SB RAS, 11 Frunze St., Novosibirsk 630091, Russia
⁴ National Research Tomsk State University, 36 Lenina Ave., Tomsk 634050, Russia
⁵ Higher School of Economics – National Research University, 20 Myasnitskaya St., Moscow 101000, Russia

E-mail: efimov@bionet.nsc.ru

Abstract. When analyzing a 1D time series, it is traditional to represent it as the sum of the trend, cyclical components and noise. The trend is seen as an external influence. However, the impact can be not only additive, but also multiplicative. In this case, not only the level changes, but also the amplitude of the cyclic components. In the PCA-Seq method, a generalization of SSA, it is possible to pre-standardize fragments of a time series to solve this problem. The algorithm is applied to the Anderson series – a sign alternating version of the well-known Wolf series, reflecting the 22-year Hale cycle. The existence of this cycle is not disputed at high solar activity, but there are doubts about the constancy of its period at this time, as well as its existence during the epoch of low solar activity. The processing of the series by the PCA-Seq method revealed clear oscillations fluctuations of almost constant amplitude with an average period of 21.9 years, and it was found that the correlation of these oscillations with the time axis for 300 years does not differ significantly from zero. This confirms the hypothesis of the existence of 22-year oscillations in solar activity even at its minima, like the Maunder minimum.

1. Introduction.
When processing a 1D time series by the principal component method (PCA) [1-11]), two problems usually arise. First: if there are cyclical components in the original series, then two PCs correspond to each of them. As a rule, one of them is derivative of the other [12]. It happens that there is a third component that modulates the first two in amplitude. In this case, all three PCs are mutually orthogonal, despite the obvious common cause that generates them.

The second problem is that even if some empirical patterns are found, it is not enough to recognize them as real. Naturally, the PCA cannot say anything about the nature of the found regularities. On the other hand, correlations between the PCs and the already known external factors (EFs) can say about it, if they are. If the correlations are large enough and reliable, this can serve as a hint for interpretation or even the interpretation itself; this is already the competence of a specialist in the relevant subject area.
We use traditional solar activity data as an example. The coordinates of the planets in the solar system are usually considered as EFs, although no convincing results on their effect on solar activity have yet been obtained [13]. However, the PC correlations can be considered not only with them, but also with the time axis, which was done much less frequently. In our opinion, a reliable zero correlation with the time axis may testify in favor of the hypothesis that the found regularity really exists and can be continued in both directions from the interval under study.

2. Material and methods.

The data on the annual Wolf numbers from 1700 to 2020 and the numbering of solar cycles were taken from the site [14]. Since double 22-year cycles have a physical meaning [15], Anderson's trick is used, when the Wolf numbers for odd cycles are multiplied by -1 [16]. Indeed, it looks more natural to study Anderson's series instead of Wolf's series (figure 1).

![Figure 1. Dynamics of Wolf numbers (SSN11) and Anderson numbers (SSN22).](image)

The resulting series (SSN22) was processed by the PCA-Seq method [17], which is a further development of the PCA for time series (PCA-TS). In PCA-TS, the original time series of length N is converted by sliding window of length L into a sequence of length N-L+1 [3], forming a matrix $X = (X_{ij})$ of size $(N-L+1) \times L$ [4], called the trajectory matrix (TM) [5]. Let $X = (x_1, x_2, \ldots, x_N)$ be a time series. Then the i-th row of TM $X$ is the vector $X_i = (x_i, x_{i+1}, \ldots, x_{i+L-1}) = (x_{i1}, x_{i2}, \ldots, x_{iL})$, $X_j = x_{i+j-1}$, $i = 1, \ldots, N-L+1$, $j = 1, \ldots, L$; $X$ is considered as an "object–trait" matrix and is processed by the PCA method. The processing consists in the fact that the covariance (correlation) matrix of traits through the eigendecomposition is represented as $X^TX = QDQ^T$, where $Q$ is an orthogonal eigenvector matrix, $D$ is
a diagonal matrix, and the sought PCs are found by the formula $P= XQ$. Obviously, $X= PQ^T$. For some reason, such as forecasting, we may only be interested in a certain subset of PCs. Obviously, they define a subspace of lower dimension. Then we can zero out (not exclude) the remaining PCs in the matrix $P$ and denote $P$ as $P^\wedge$. Then $X^\wedge = P^\wedge Q^T$ is the reconstructed TM.

In 1989, under the influence of physicists [3,5,6], the method was named SSA [5] and further developed under this name. Diagonal averaging turned out to be a very useful innovation [10,11]. Each element $x_i$ of the original series corresponds in TM to a series of its copies $(\ldots, x_{k-\alpha+1}, \ldots)$, $1\leq\alpha \leq L$ located along the TM diagonal. In the reconstructed $X^\wedge$, the values along the diagonals corresponding to one element of the original series are no longer necessarily equal to each other, but they can be averaged and thus obtain a reconstructed time series $X^\wedge$ of the same size as the original one corresponding to the selected subset of PCs.

Recently, we have proposed a new method for decomposing a sequence of elements of any type into PCs – PCA-Seq [17]. The main difference between PCA-Seq and SSA is that instead of the eigendecomposition of the covariance matrix of traits, the decomposition of the matrix of squares of Euclidean distances between objects is used – the Gower method of principal coordinates PCoA [18]. The objects are TM rows. PCs are obtained from TM via EDM without an eigenvector matrix $Q$. Since the distances between symbol rows have been known for a long time, it immediately became possible to obtain numeric PCs for non-numeric sequences, for example, molecular [17, 19].

In the numerical version, PCA-Seq and SSA are equivalent if the sum of the squared differences between TM rows is left as the squared distances between objects, as is implicitly and by default accepted in SSA. The eigenvector matrix $Q$ can be easily obtained from the matrix of correlations between PCs and TM. Therefore, diagonal averaging can also be done.

But in the numerical version, PCA-Seq, unlike SSA, allows you to do any transformations of TM rows and calculate any other Euclidean and close to them distances between the rows, if this is dictated by the logic of the problem being solved. For example, to “remove” the amplitude modulation, one can do standardization of TM rows [20], what is done in this work too (not to be confused with standardization by columns, as in SSA). (In fact, this is a sliding standardization of the original series, if we apply diagonal averaging to the TM standardized by rows.) And Euclidean distances in the numerical case can be obtained in very different ways [21]. Therefore, the capabilities of the PCA-Seq are much wider in the numerical case than those of the SSA.

As for the reconstruction from a subset of PCs, in the general numerical case TM is also numeric, and PCs are orthogonal. Therefore, it is always possible to calculate linear regression by PCs of interest under for each TM column and get its reconstruction. And for reconstructed TM, diagonal averaging is always applicable. Thus, the PCA-Seq in the numerical case always allows a reconstruction of the original series of the same length to be obtained. This means that SSA is a special case of PCA-Seq.

From the Anderson’s series (SSN22), a TM $X= (X_{ij})$ with a size of 300x22 is formed. Matrix $X^\prime$ is obtained by standardization of TM $X$ by rows (each row is standardized by subtracting the mean and dividing by the standard deviation). The means and standard deviations formed two new time series: the row mean $M= (M_i)$ and the row standard deviation $S= (S_i), i=1, \ldots, N-L+1$ (Fig. 2a).

$$X^\prime_{ij} = X_{i-L+j} \quad i=L, \ldots, N; \quad j=1, \ldots, L;$$

$$M_i = \frac{1}{L} \sum_j X_{ij}; \quad S_i^2 = \frac{1}{L} \sum_j (X_{ij} - M_i)^2; \quad X^\prime_{ij} = \frac{X_{ij} - M_i}{S_i},$$

where $X$ is the original series of length $N$; $L$ – lag; $X= (X_{ij})$ – TM of size $(N-L+1)\times L$, $M_i$ – $i$-th row mean, $S_i^2$ – $i$-th row variance, $X^\prime= (X^\prime_{ij})$ – TM X standardized by rows.

Between all the $X^\prime$ rows, the EDM was calculated, from which the PCs were obtained by the Gower’s method PCoA [18]. Fig. 2b, 2c shows the first two PCs, accounting for 89.2% of the total variance, and their phase portrait. Obviously, we are observing a clear oscillatory process. There is no doubt that the phase trajectory is practically a circle. However, we cannot call it strictly periodic, since a complete
rotation does not take place in equal time. This is clearly seen for the range 1800–1830 (years). Therefore, for each pair of neighboring fragments, that is, for one year, the angle of rotation $\omega_i$ and the radius relative to the center of the circle $r_i$ are calculated (Fig 2d). In other words, it is the instantaneous frequency and amplitude of the oscillations. Reconstruction of the Anderson's series by two first PCs (SSN22-rec) is represented in Fig. 3.

All calculations were performed in the STATISTICA12 [22], PAST4 [23] and Jacobi4 [19] packages.

3. Results
The dynamics of the 22-year moving mean, the moving standard deviation and the radius of the Anderson series are not stationary, since their nonzero correlations with the time axis, although not too great, are significant. The variance sum of the first two PCs of the standardized Anderson series is almost 90%, and their cyclicality is beyond doubt (figure 2b, 2c, 3.). The correlations with the time axis of the dynamics of the angular velocity, SSN22 and SSN22-rec do not differ significantly from zero. In figure 2d that, despite the rather significant deviations, the dynamics of the angular velocity still stubbornly returns to a certain constant level. The mean angular velocity of 0.287 over 300 years corresponds to a period of 21.9 years. In our opinion, all this may mean that it is a manifestation of some inner oscillatory process with a constant period, which is modulated in amplitude and possibly in phase by another, more irregular process, which, in turn, manifests itself in the dynamics of $S$ (Std) and $M$ (Mean) (figure 2a). The same inner process with an almost constant amplitude can be seen in figure 3. Perhaps this is a manifestation of the solar dynamo [13].

![Figure 2.](image-url)

**Figure 2.**

- a) Dynamics of the 22-year sliding mean $M$ and standard deviation $S$ of the Anderson series;
- b) Dynamics of the first two PCs of the standardized trajectory matrix;
c) Phase portrait PC1 – PC2 and biplot of the influence of lags (shifts by 1, 2, …, 22 years);  

d) Dynamics of the radius of rotation and angular velocity of the phase trajectory in Fig. 2c).

**Figure 3.** Dynamics of Anderson numbers (SSN22) and its reconstruction from two first PCs (SSN22-rec) by PCA-Seq method. Perhaps this is a manifestation of the solar dynamo [13].

4. Discussion

It is well known that before the beginning of the Wolf series there is a rather extended Maunder period, during which there was no solar activity at all. This circumstance would seem to exclude any constant periodicity. However, nothing prevents us from assuming that the modulating process is close to zero while maintaining the constancy of the oscillatory process. Indeed, in some works it was found that during the Maunder minimum the 11-year periodicity of solar activity was nevertheless, albeit very weak [13, 24].
Table 1. Means and correlation coefficients with the time axis of the Anderson series parameters.

|        | N=300       | Mean  | Correlation | p-value |
|--------|-------------|-------|-------------|---------|
| Mean   | -1.510      | -0.383| <1E-4       |         |
| Std    | 98.488      | 0.344 | <1E-4       |         |
| Radius | 0.944       | 0.251 | 1E-05       |         |
| Omega  | 0.287       | 0.038 | 0.517       |         |
| SSN22  | -1.172      | -0.041| 0.464       |         |
| SSN22-rec | -1.067     | -0.004| 0.941       |         |

As for the modulating one, it is also similar to a cyclical one with an oscillation period of about a hundred years, but since so far only three incomplete irregular oscillations have been observed, it is difficult to say something more definite about this. In the literature, it is known as the Gleisberg cycle [25].

5. Conclusion

In the PCA-Seq method, it is possible to pre-standardize fragments of a time series to eliminate the multiplicativity of the influence of an external factor. The algorithm is applied to the Anderson series—a sign alternating version of the well-known Wolf series, reflecting the 22-year Hale cycle. The processing of the series by the PCA-Seq method revealed clear oscillations of almost constant amplitude with an average period of 21.9 years, and it was found that the correlation of these oscillations with the time axis for 300 years does not differ significantly from zero. This confirms the hypothesis of the existence of 22-year oscillations in solar activity even at its minima, like the Maunder minimum.

Acknowledgements

This work was supported by the RFBR grant #19-07-00658-a and the Budget Project of the ICG SB RAS 0259-2021-0009.

References

[1] Karhunen K 1947 Ann. Acad. Sci. Fennicae Ser. A137
[2] Loève M 1948 In: Processus Stochastiques et Movement Brownien Paris: Hermann
[3] Takens F 1981 In: Dynamical Systems and Turbulence Berlin Heidelberg: Springer pp 366–81
[4] Efimov V M and Galaktionov Y K 1983 Zhurnal Obshchey Biologii 3 pp 343–52
[5] Broomhead D S and King G P 1986 Physica D 20 pp 217–36
[6] Fraedrich K 1986 Journal of Atmospheric Sciences 43 pp 419–32
[7] Efimov V M, Galaktionov Y K and Shushpanova N F 1988 Analysis and Prediction of Time Series by the Principal Component Method Novosibirsk: Nauka Publ p 71
[8] Vautard R and Ghil M 1989 Physica D: Nonlinear Phenomena 35 pp 395–424
[9] Elsner J B and Tsonis A A 1996 Singular Spectrum Analysis: A New Tool in Time Series Analysis New York: Plenum Press p 164
[10] Golyandina N, Nekrutkin V and Zhigljavsky A A 2001 Analysis of time series structure: SSA and related techniques CRC press p 320
[11] Golyandina N and Zhigljavsky A 2020 In Singular Spectrum Analysis for Time Series Berlin, Heidelberg: Springer pp 21–90
[12] Bartlett M S 1978 An introduction to stochastic processes: with special reference to methods and applications. 3rd Ed Cambridge: Cambridge University Press p 405
[13] Obridko V N and Nagovitsyn Yu A 2017 Solar activity, cyclicity and forecasting methods SPB: VVM p 446
[14]  http://www.sidc.be/silso/infosnytot. Sunspot Number | SILSO. www.sidc.be. Date of treatment: December 2, 2020
[15]  Nagovitsyn Yu A 2007 Astronomy Letters 33 pp 340–5
[16]  Anderson S N 1939 Bel. System Tech. J 18 p 292–9
[17]  Efimov V M, Efimov K V and Kovaleva V Y 2019 Vavilov J Gen and Breed 23 1032–6
[18]  Gower J C 1966 Biometrika 53 325–38
[19]  Polunin D, Shtaiger I and Efimov V 2019 bioRxiv 803684
[20]  Gower J C and Legendre P 1986 Journal of classification 3 pp 5–48
[21]  Takalo J 2021 Solar Physics 296 pp 1–15
[22]  Hill T, Lewicki P 2006 Statistics methods and applications a comprehensive reference for science industry and data mining Tulsa Okla UK: StatSoft Ltd p 719
[23]  Hammer Ø, Harper D A T and Ryan P D 2001 Palaeontologia Electronica 4 p 9
[24]  Vaquero J M, Kovaltsov G A, Usoskin I G, Carrasco V M S and Gallego M C 2015 Astronomy & Astrophysics 577 A71
[25]  Gleissberg W 1958 J Brit Astron Assoc 68 pp 148–52