

New features of the PCA-Seq method in time series analysis (on the example of COVID-19)

Efimov V.M.^{1,2,3,5*}, Polunin D.A.², Kovaleva V.Y.³, Efimov K.V.⁴

¹ Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia

³ Institute of Systematics and Ecology of Animals, SB RAS, Novosibirsk, Russia

⁴ Higher School of Economics, Moscow, Russia

⁵ Tomsk State University, Tomsk, Russia

* efimov@bionet.nsc.ru

Key words: Phase portraits, dynamic system, SSA

Motivation and Aim: A feature of the mass reproduction of living organisms (viruses, bacteria, lemmings, rabbits, humans, etc.) is the "explosive" exponential growth of the population at the initial stage. This occurs when population growth is proportional to the abundance already present. At the beginning of the coronavirus pandemic, the number of cases increases exponentially, and the number of deaths, obviously a more important indicator, is a relatively small proportion of cases and is late by about two weeks. And although the causal relationship of these processes is beyond doubt and, it would seem, the second indicator should be easily predicted two weeks in advance, the statistical correlation between them turns out to be rather weak. Quarantine, vaccination and re-infection contribute to the official statistics. There is a need for more powerful methods for analyzing such data. A 1D time series is traditionally represented as the sum of a trend, cyclical components, and noise. However, the typical behavior of Covid-pandemic indicators does not correspond to this perception: an almost imperceptible and difficult to fix beginning; exponential growth; plateau – fluctuations around a certain level; decline, slow or fast. Cyclicity manifests itself in the form of waves of different scale and duration with intervals of different lengths between them. On a logarithmic scale, a deviation from a straight line means that there are obstacles to the reproduction of the virus. The first obstacle is quarantine, the next is vaccination. The PCA-Seq method (a generalization of SSA) is suitable for the analysis of such processes [1]. It made it possible to obtain similar phase portraits of the dynamics of morbidity and mortality from coronavirus in Italy (the first country in which a pandemic began in Europe).

Methods and Algorithms: Time series for daily new cases and new deaths from Covid-19 in Italy obtained from the website https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/. After Log transformation, the PCA-Seq method (N = 823; Lag = 50) was applied to them [1, 2]. This method is a generalization of the well-known SSA (singular spectral analysis) [3], which is also used in research on Covid-19 [4]. PCA-Seq, unlike SSA, allows you to do any transformations of trajectory matrix rows and calculate any other Euclidean distances between the rows [5].

Results: The figure shows the phase portraits of the first two principal components of the dynamics of new cases of morbidity and mortality. Despite the orthogonality, the second PC is a derivative of the first one (this follows from the loadings of the PC2). This is a well-known property of the principal components of time series. In turn, the correlation coefficient of PC1 with the mean is practically equal to one. This means that when PC2

> 0 (above the red line) the average morbidity (or mortality) increases, otherwise it falls. Rotation is always clockwise. Taking into account the inertia of the process, this circumstance can be confidently used for forecasting at a qualitative level if the trajectory is sufficiently far from zero. For example, both morbidity and mortality should increase at the moment, so the trajectory definitely intends to cross the red line. But it is difficult to predict whether it will be a small loop or a large circle.

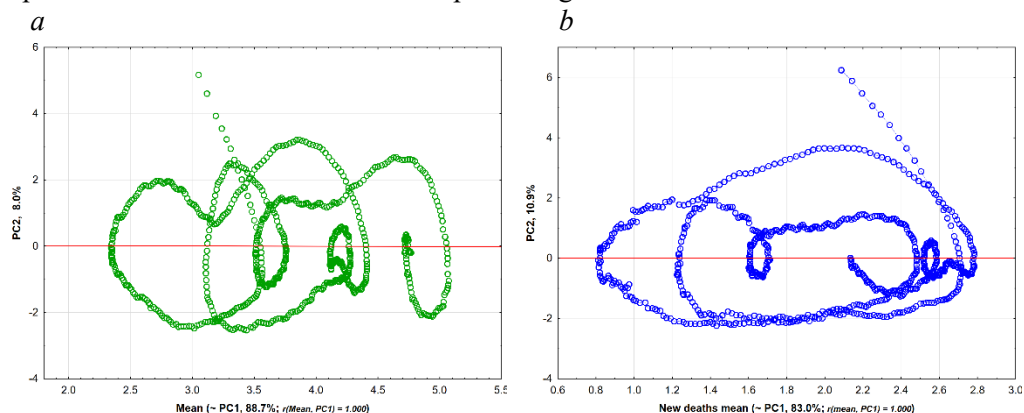


Fig. 1. Phase portraits of first two Covid-19 dynamic principal components: a) daily new cases; b) daily new deaths. In both cases PC1 dynamic is equivalent Mean one

Conclusion: As the figure shows, the behavior of the trajectories is similar, as it should be, since one process is generated by another. On the other hand, this behavior appears to be complex enough to hopefully be described by simple dynamic models. Somewhat reassuringly, something like differential equations emerges as a rather unexpected consequence of straight-forward statistical processing.

Acknowledgements: The study is supported by the Kurchatov Genomic Centre of the Institute of Cytology and Genetics, SB RAS (075-15-2019-1662).

References

1. Efimov V.M., Efimov K.V., Kovaleva V.Y. Principal component analysis and its generalizations for any type of sequence (PCA-Seq). *Vavilov J Genet Breed.* 2019;23(8):1032-1036.
2. Efimov V.M., Polunin D.A., Kovaleva V.Y., Efimov K.V. The PCA-seq method applied to analyze of the dynamics of COVID-19 epidemic indicators. *J Phys Conf Ser.* 2021;1715(1):012025.
3. Golyandina N. Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing. *Wiley Interdiscip Rev Comput Statist.* 2020;12(4):e1487.
4. Kalantari M. Forecasting COVID-19 pandemic using optimal singular spectrum analysis. *Chaos Solit Fractals.* 2021;142:110547.
5. Efimov V.M., Efimov K.V., Polunin D.A., Kovaleva V.Y. New possibilities of the PCA-Seq method in the analysis of time series (on the example of solar activity). *J Phys Conf Ser.* 2021;2099(1):012034.