Cross-correlation of publications dynamics and pandemic statistics

Dmitry Lande[1], Leonard Strashnoy[2,3]
[1]Institute for Information Recording of NASU (Ukraine)
[2,3]Amazon; UCLA, Infectious disease department (USA)

This work investigates the cross-correlation between information sources and factual data about infection and mortality rates. The source data (datasets) was obtained from the aggregator of factual information Our World In Data and the content monitoring system InfoStream. The results of this research allow for the conclusion that society is greatly capable of making prognoses where the trends from information sources precede actual results.

Cross-correlations (a model)

Studying different time series, connected to various events, a great deal of significance is attributed to cross-correlation. Cross-correlation determines the similarity of how processes develop though it does not always reflect a causal connection. Although when there is such a connection, a high correlation does signal this. A sequence of cross-correlations, implemented in the command ***xcorr*** в Matlab [1, 2], computed in the formula:

$$R_{XY}(m) = E\{x_{n+m}, y_n\}$$

where x_n and y_n – research subjects time series, $-\infty < n < \infty$, and $E\{\cdot\}$ – the operator the average (expected) value. (It should be noted that some researchers, in contrast to authors who prefer to use in the quality of values cross-correlations such a convolution: $\hat{R}_{XY}(m) = E\{x_n, y_{n+m}\}$, which is being implemented in

Matlab, a different function is used ***crosscorr***. Results execution options this function in practice it is a mirror image reflection results execution options ***xcorr*** functions, and not affects conclusions this work).

It is known that the cross-correlation process by yourself (i.e., autocorrelation) reaches a maximum when $m = 0$. Cross-correlation different processes can reach maximum when other values m.

Consider for example the simplest model sequences from 100 points, in which one is the first the sequence ($Series_1(i)$) "precedes" second ($Series_2(i)$) by 10 values the argument (Fig. 1):

$$Series_1(i) = \sin\left(\pi \frac{i}{100}\right);$$

$$Series_2(i) = \begin{cases} 0, & i \le 10; \\ Series_1(i-10), & i > 10. \end{cases}$$

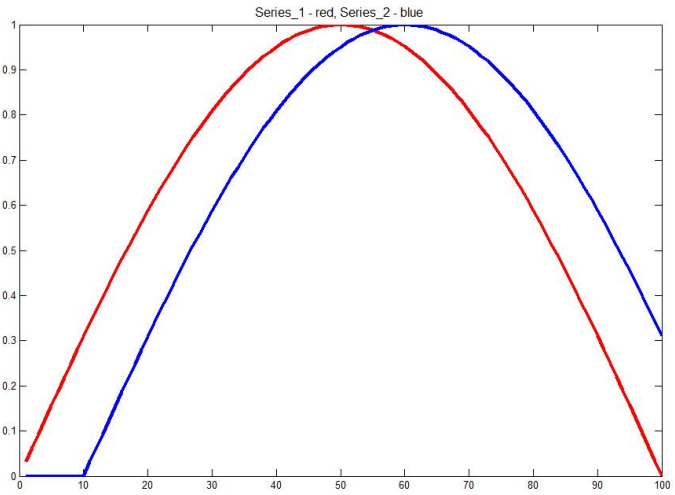


Fig. 1. Initial Data model

Cross-correlation in this case it has an explicit value expressed skewness relative points 0, reaching maximum when $m \approx 10$ (Fig. 2).

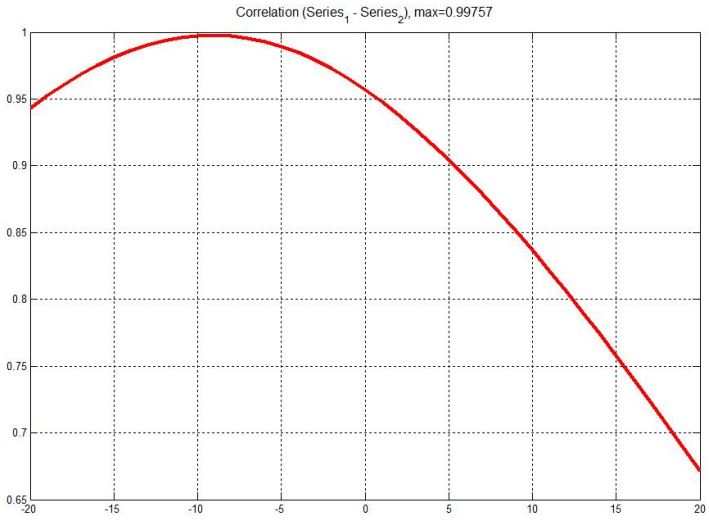


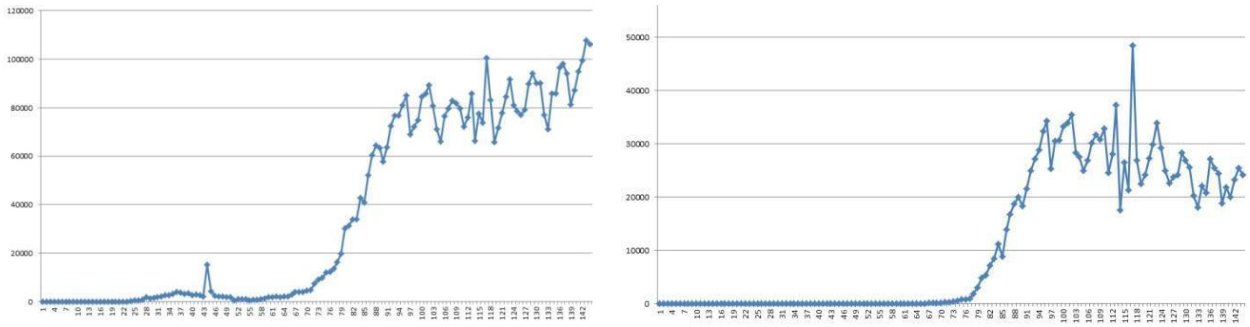
Fig. 2. Cross-correlation model coefficients

Sources of information and timelines

Information sources used in this work include data gathered from ***Our World In Data*** (is a project of the Global Change Data Lab). Within this source data about the pandemic (https://ourworldindata.org/coronavirus-source-data) is presented in a daily renewable "cleaned up" version, without the anomalous spikes connected to technical failures, and in highly integrated formats .xslx, .csv, .json.

The Ukrainian system InfoStream was used to gather data from social media (http://infostream.ua – that covers approximately 1000 of the world's leading English language publications government, press, and information agency web sites, as well as 10 social media networks) [3]. The InfoStream system provides a dynamic of daily announcements from websites and social media, partially in regards to the disjunction request (COVID/coronavirus).

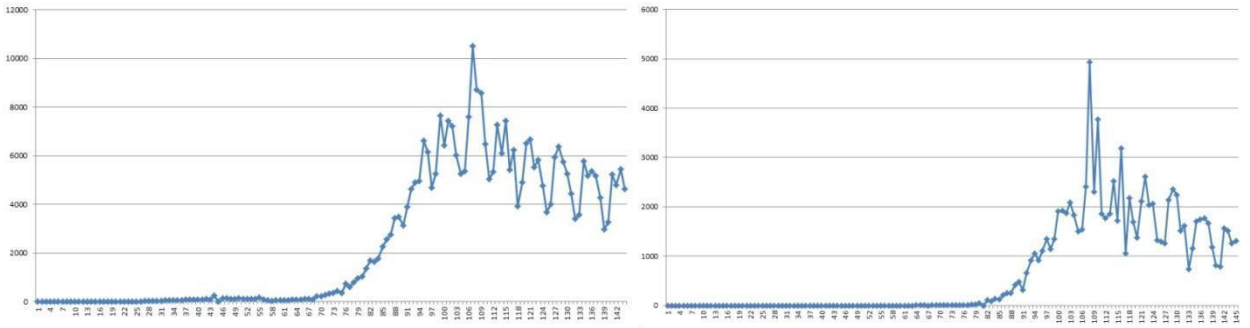
In Fig. 3. daily infection and mortality rates are presented ourworldindata.org (a) throughout the world and (b) the U.S., starting 01/01/2020 from the data set brought from the website Our World In Data.



a) b)

Fig. 3. Dynamics of infection by day (world/USA) ourworldindata.org

In Fig. 4. presented dynamics chart for mortality rate by day from ***ourworldindata.org*** by worldwide (a) and the United States (b) by dataset, to the given one on the website Our World In Data, starting from 01/01/2020.



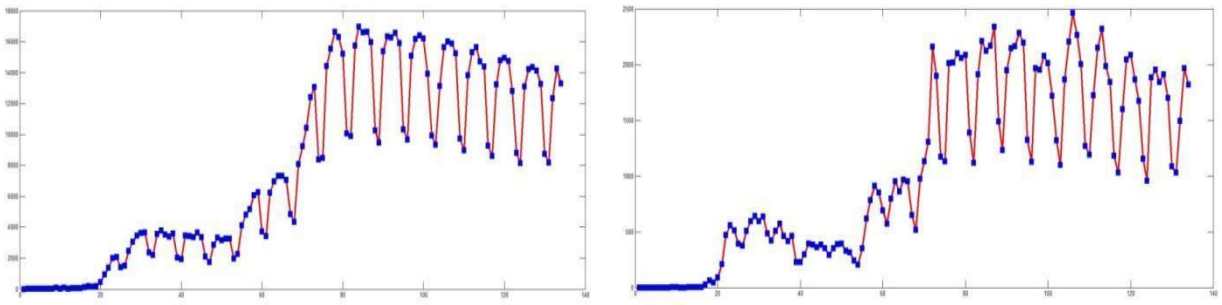
a) b)

Fig. 4. Dynamics of mortality by day (world/USA) ourworldindata.org

In Fig. 5. a dynamic of publication rates from different websites, gathered by the InfoStream system, is presented, relating to (a) the world and (b) the U.S., starting 01/01/2020. Queries input into the system were:

1. COVID|coronavirus

2. (COVID|coronavirus)&(USA|"United States")



a) b)

Fig. 5. Dynamics publications in web sites (world/USA)

In Fig. 6. a dynamic of publication rates from social media, gathered by the InfoStream system, is presented, starting 01/01/2020.The query input into the system was: "***COVID|coronavirus***".

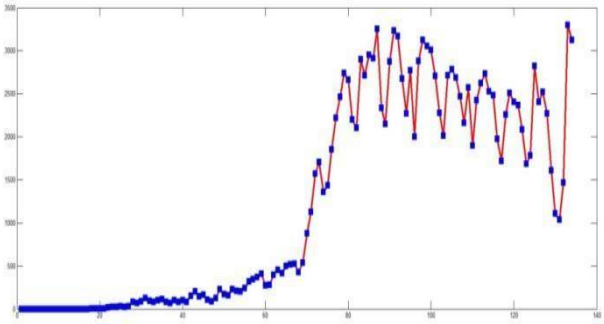
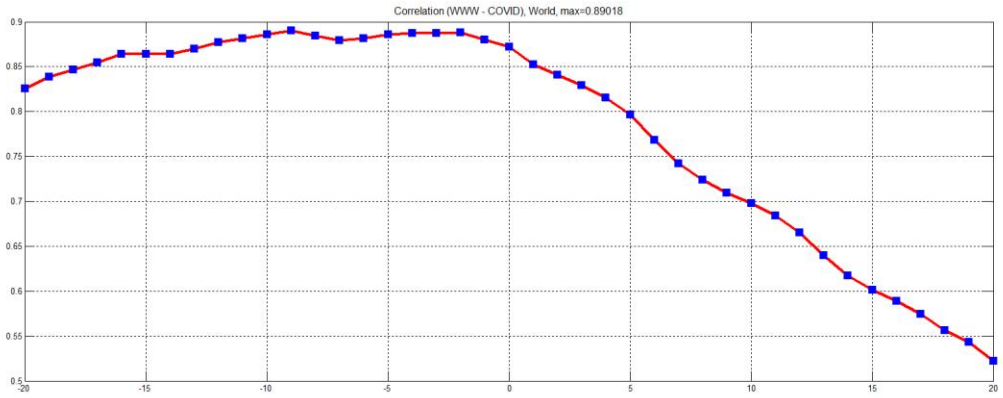


Fig. 6. Dynamics publications in social networks

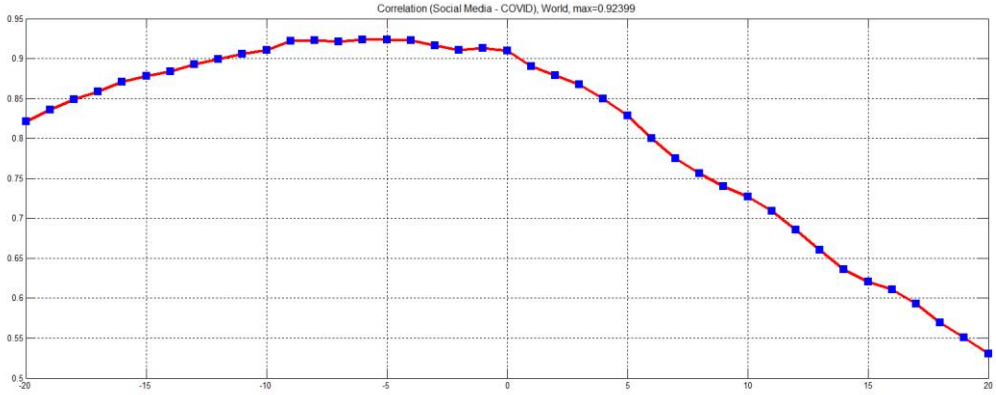
Cross-correlations real processes

As noted above, if the left side of the cross-correlation charts is higher than the right side, then it means that the first process "precedes" the second, correlating with the second with a negative shift. In this scenario, the cross-correlation charts of the processes of publication dynamics on the web (data obtained from the InfoStream system, the processes of infection or mortality rates (Fig. 7, 8.)) look paradoxical. Without direct proof of a causal relationship, we can conclude that

information processes on the internet and social media precede real processes of the pandemic.

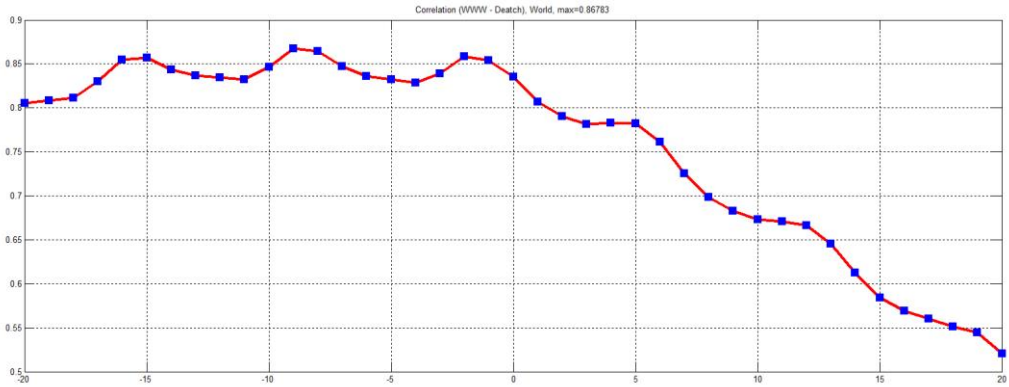


a)

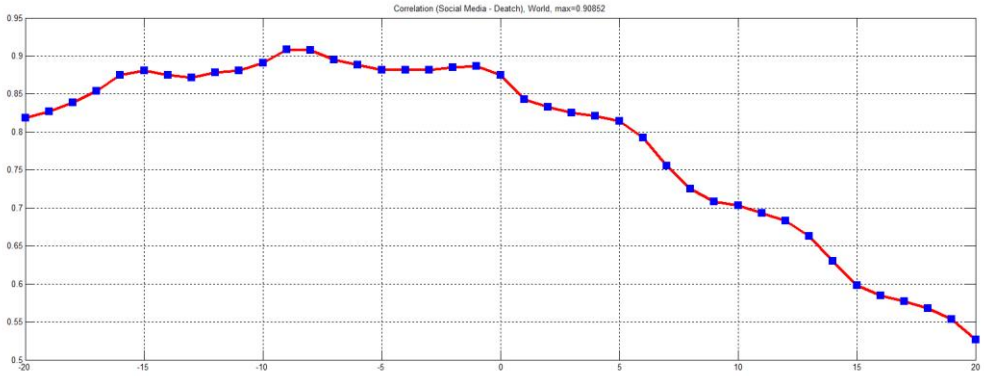


b)

Fig. 7. Cross-correlation coefficients of publication dynamics on the web (a) or social media (b) and dynamics of the processes of infection rates throughout the world.



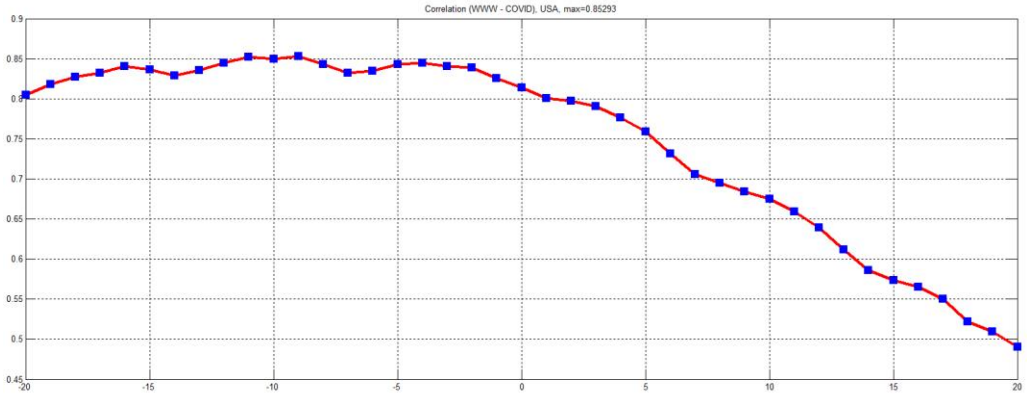
a)



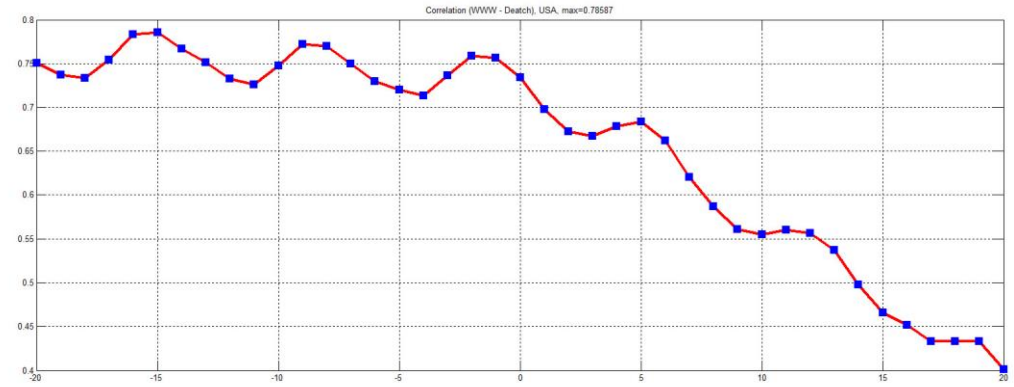
b)

Fig. 8. Cross-correlation of publication dynamics on the web (a) or social media (b) and dynamics of the processes of mortality rates throughout the world.

For a particular country (the U.S.) cross-correlation of the processes of publication dynamics on the web and the dynamics of infection and mortality rates are presented in Fig. 9. As can be seen, the charts are likewise asymmetrical. The fluctuating component corresponds to the periodic nature of publication volume in social media based on the days of the week.



a)



b)

Fig. 9. Cross-correlation coefficients of publication dynamics on the web and:
a) infectability by day; b) mortality by day
in the United States

Conclusions

Having analyzed the data of the daily infection and mortality rates of the processes of the COVID-19 Pandemic and their reflections in the indicated sources in social media a result was obtained which indicates that real processes of the daily infection and mortality rates were preceded by information processes in social media.

The explanation for this, possibly, is society's capacity for prognostication.

The volume of publications may indicate a coming problem. It appears, in fact, that all the media of the world were writing about COVID-19 when only a few dozen cases of infection were known in Wuhan. As we can see, trends in the information sources about the pandemic on average precede real results of the pandemic.

Literature

1. Steven T. Karris “Numerical Analysis Using MATLAB and Excel”. Orchard Publications (2007). 627 p.
2. Orfanidis, S.J., Optimum Signal Processing. An Introduction. 2nd Edition, Prentice-Hall, Englewood Cliffs, NJ (1996).
3. A. Dodonov, D. Lande, V. “Tsyganok, etc. Information Operations Recognition. From Nonlinear Analysis to Decision-Making”. LAP Lambert Academic Publishing (2019). 292 p. ISBN: 978-620-0-27697-1