Detecting malicious pdf documents using semi-supervised machine learning

Jianguo Jiang, Nan Song, Min Yu, Kam-Pui Chow, Gang Li, Chao Liu, Weiqing Huang

Information entropy contribution to COVID-19 waves analysis *

Iliyan Petrov[0000−0002−0062−6666]

Bulgarian Academy of Sciences (BAS); Institute of Information and Communication Technologies (IICT), akad. G. Bonchev str, bl. 2, 1113 Sofia, Bulgaria
iliyan.petrov@iict.bas.bg, petrovindex@gmail.com
https://www.iict.bas.bg/ipdss/i-petrov.html

Abstract. The beginning of COVID-19 pandemics was sudden and unexpected in terms of scale and symptoms, channels and territory of propagation in different countries. This article discusses the possible information theory contribution for analysing the waves of pandemics on the example of Bulgaria. Under conditions of uncertainty and non-sufficient statistics the simple and robust data-driven approach based on the concept of information entropy provides additional possibilities for analysing the dynamics of epidemic waves.

Keywords: covid-19 · covid-19 waves · SIR model · information theory · entropy

1 Introduction

System complexity and structural evolution are key issues for characterizing the specifics of dynamic systems in a large number of areas. The research of complex systems requires not only adequate methods and tools for treating and analysing the large volumes of data, but also a systematic approach for collecting and selecting of raw data. The sudden and unexpected outbreak of COVID-19 created unprecedented for many decades stress in all social sectors [8]. In this situation, it is a challenging task to develop models for long-term accurate prediction of pandemics evolution. The traditional forecasting approaches are usually based on deterministic extrapolation of general indicators and their average aggregated statistics often results in different prognostics about the dynamics of infections. Reliable statistics and consistent methods are critical for government policies and medical measures about "how" and "when" restrictions on mobility should be imposed, revised or lifted in all sectors, including education.

This study explores a research path based on information theory, entropy and agent-oriented analysis for complementing the general Susceptible-Infected-Removed (SIR) compartmental statistics [7] with additional macroscopic insight into the time-dependant and territory spread of pandemics [1].

* This research is supported by Bulgarian FNI fund though project "Modelling and Research of Intelligent Educational Systems and Sensor Networks (ISOSeM)", contract KP-06-H47/4 from 26.11.2020

2 Information entropy

2.1 Competition Evolutionary Model for Shannon Entropy

The assessment of information entropy is performed by specific indicators with relatively simple mathematical algorithms at two consecutive levels. Traditionally, natural sciences (physics, computer science, telecommunications) use the information theory concept for entropy to assess diversity, uncertainty, and chaos [5]. In our studies we use the small letter "e" and capital letter "E" to denote "entropy" at micro- and macro-level. Also, to distinguish different indicators we use the first letter from the family name(s) of their author(s) in front of the symbol of entropy - for example, "SE" for "Shannon Entropy" [9], defined as:

- Level-1: transformation of the values of components' relative parts "p_n" (probabilities or weights) into results which can be defined as "micro-level entropy (individual entropy)" in the basic function *"SE"*:

$$SE = -p_n.log_b p_n \tag{1}$$

- Level-2: summing of results of "micro-level Shannon entropy" "SE" for obtaining the cumulative indicator for nominal macro-entropy:

$$\Sigma SEnom = \Sigma_{w=1}^{n} SE(p_n) = -\Sigma_{w=1}^{n} p_w.log_b p_n \tag{2}$$

Shannon Entropy is usually defined in 3 basic formats: $log_2 p_n$ - with information measured in "bits"; lnp_n - with information measured in "nats"; and $log_{10} p_n$ - with information measured in "hartleys". To frame the systems' evolution in the entropy concept we introduce a novel Competition Evolutionary Model (CEM). In CEM, the values of maximal cumulative entropy for configurations with equal (symmetrical) weights of components are defined as:

$$SEmax(sym) = \Sigma_{i=1}^{n} SE(1/n) = -n{*}1/n.log_2(1/n) = -log_b(1/n) = log_b n \tag{3}$$

To define the values of sub-symmetric configurations SE(subsym) are introduced the following parameters:

- number of Transition Interaction Steps "TIS", and size of "TIS" defined as $\Delta p(TIS) = 1/TIS$;
- Part lost by leader - "$\Delta p(LL)$";
- Structural Phases (SP) - configuration wit "n" system components;

In CED, the competitive interactions between the consecutively increasing number of "n" competitors form separate consecutive Structural Phase (SP) within which at each TIS the leader loses the constantly defined part "$\Delta p(LL)$" of his initial resource "$1/(n-1)$" which is redistributed evenly between the other equal competitors until the full equalization of all weights in the system "$p_w = 1/n$)". To model a full evolutionary process, logically, we have to consider the following condition: $\Delta p(TIS) = 1/TIS \geq \Delta p(LL)$. For simplifying this presentation we consider $\Delta p(TIS) = 1/TIS = \Delta p(LL) = 1/100 = 1\%$. The starting values $p(start)$ of the leader and his "$n-1$" competitors are defined as:

$$p_{start(leader)} = \frac{1}{n-1} \tag{4}$$

$$p_{start(equalcomp)} = \frac{1 - \frac{1}{n-1}}{n-1} = \frac{n-2}{(n-1)^2} \tag{5}$$

In each SP_n the number "m_{SP_n}" of TIS is defined as:

$$m_{SP_n} = \frac{1}{n(n-1)\Delta p(TIS)} \tag{6}$$

The interim transition parts "$p_{transit}$" of the leader and hiss "$n-1$" competitors within each SP_n are defined as:

$$p_{transit(leader)} = \frac{1}{n-1} - m_{q(SP_n)}.\Delta p(LL) \tag{7}$$

$$p_{transit(equalcomp)} = \frac{1 - \left(\frac{1}{n-1} - m_{q(SP_n)}.\Delta p(LL)\right)}{n-1} \tag{8}$$

where "$q(SP_n)$" is the consecutive number of "m" in each SP_n.

Finally, all competitors are ending the serial of TIS in each SP_n with equal parts(weights) "$p_n = 1/n$". In this context, the symmetric and sub-symmetric macro-level configurations form a discrete path of the boundary of the maximum and sub-maximum levels of entropy in a scenario defined as "Equalization within each population" (Fig 1). The system's balance can be defined as:

$$\left(\frac{1}{n-1} - m_{q(SP_n)}.\Delta p(LL)\right) + (n-1)\left[\frac{1 - \left(\frac{1}{n-1} - m_{q(SP_n)}\Delta p(LL)\right)}{n-1}\right] = 1 \tag{9}$$

The Maximal Shannon Entropy $\Sigma SEmax(sym)$ for each population is reached in the unique and fully symmetric configuration [6] when all components have equal parts or weights ($p_1 = p_2 = ...p_n = 1/n$) as defined in in eq.(3) - $\Sigma SEmax(sym) = log_b n$.

2.2 Maximal and normalized of cumulative entropy

Maximal Entropy. The Shannon Entropy is used in several areas (telecommunications, computer science, biology, economics, etc.) for assessing the diversity and uncertainty in different systems with a large number of components [3, 4]. Therefore, the values of nominal cumulative Shannon Entropy $\Sigma SEnom$ for non-symmetrical configurations can be very different and have to be normalised if we need to compare them with other indicators - usually, the universal and dimensionless scale "0 to 1". Such normalization is achieved by comparing (dividing) the nominal cumulative entropy values $\Sigma SEnom$ to a selected level of maximal entropy $\Sigma SEmax(sym)$:

$$\Sigma SEnorm(sym) = \Sigma SEnom / \Sigma SEmax(sym) \tag{10}$$

Multi-system normalization. This approach consists to select some maximum level of entropy for a symmetrical system with higher number of equal components "n_{max}" and to apply it for normalizing all systems configurations (symmetric and non-symmetric) with a lower number of components "n" ($n_{max} > n$). It is useful for comparing systems with different number of components "n" - in the case of COVID-19, "n" could be the number of sub-national territorial units in different countries (states, departments, regions). The maximum level for such "multi-system normalization" has to be carefully selected - if it is to high the classification capacity of most of the normalized values will be substantially be reduced and the classification of capacity will be limited. This method is also useful for comparing different systems the dimensionless scale (0 to 1), but its major inconvenience is that so far in practice it is difficult to have a convention for such "n_{max}" and its universal level of "maximal entropy". In our opinion, a reasonable level for such normalization in the social sector could be a macro-state with 1024 equal components with results displayed in Table 1.

Table 1. Selected values of Shannon Entropy in the "log_2" format

n	1024	100	50	20	10	5	4	3	2	1
$p_w = 1/n$	0.0098	0.01	0.02	0.050	0.10	0.20	0.25	0.33	0.5	1
$se(log_2)$	0.01	0.0664	0.1128	0.216	0.332	0.464	0.5	0.528	0.5	-
$SEmax(nominal)$	10	6.64	5.64	4.32	3.32	2.32	2.00	1.58	1.00	-
$SEmax(normalized)$	1	0.664	0.564	0.432	0.332	0.232	0.200	0.158	0.100	-

Figure 1 includes graphical visualization of the continuous basic functions for individual entropies "se" in the three variants of logarithmic bases ($b = 2$, $b = 2.718$ and $b = 10$) and the respective functions of nominal cumulative, which after normalization produce identical numeric and graphical results. Logically, for all values of "b" the Shannon Entropy produces similar in shape and different in size structural spaces defined by convex parabolas.

Single-system normalization. To overcome some of the inconveniences of the "multi-system normalization" another popular method applies the opposite approach - the "single-level normalization" for each system with a certain number of components. Unfortunately, after such normalization all symmetric system configurations with equal relative weights ($p_w = 1/n$) but different number of components (n) produce identical maximal entropy $SEmax(sym) = 1$. By dividing the nominal (cumulative) entropy value for any asymmetrical system with unequal components weights ($p_w \neq 1/n$) to the maximal entropy value for this system we obtain the "normalized intra-system Shannon Entropy" $SEnorm$ as a convenient heterogeneity measure for systems with an equal number of components:

$$SEnorm = \frac{SEnom}{SEmax(sym)} = \frac{-\Sigma_{i=1}^{n} p_n . log_b p_n}{log_b n} \tag{11}$$

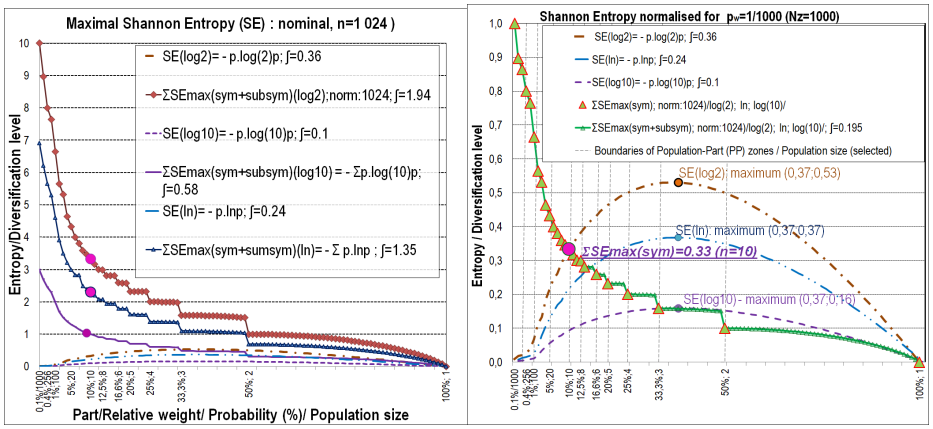

Fig. 1. Nominal and normalized cumulative Shannon Entropy

Although simple, such an approach is inconvenient for comparing systems with a different number of components but as it is frequently used in the majority of publications and for simplifying this presentation we apply it in this report.

3 Selecting data for COVID-19 entropy dynamics

The purpose of this paper is to explore the most general issues of COVID-19 dynamics focusing on the basic trends and waves of pandemics development. Table 2 reviews the main statistical indicators available publicly in the official site of the Ministry of Health Care [2]. We observe all available information but as it is not equally consistent on national and regional levels we form a panel of 5 main statistical parameters - new positive tests/day, active (infected) cases, hospitalised patients, intensively treated patients, and lethal/fatal outcomes/day.

For example, except for hospitalized patients, a large number of "Recovered cases" are qualified as such on the basis of a formal expiration of the 10-14 day quarantine periods for "Active cases". Different test results seem not be equally reliable, but after confirmation they are automatically added to the "Active (infected) cases". Nevertheless, we retain them as a "first alert signal" and a useful wave indicator for COVID-19 spreading. "Fatalities" are the most sensitive issue, but their analysis would require detailed medical information and consideration of specific demographic factors in different countries, regions and social groups.

A major technical inconvenience for "New positive tests" and "Fatalities" is their volatility with chronic peaks on Mondays/Tuesdays and minimums on Saturdays and Sundays. As in many other countries, this is due to the fact that most clinical laboratories are operational only 5 or 6 days/week, and for that reason, these parameters are smoothed with a "7 day moving averages":

$$MovAv7 = (v_{t-3} + ... + v_t + ... + v_{t+3})/7 \tag{12}$$

Table 2. Selected values in Shannon Entropy the $SE(log_2)$ format

Parameter	Reliable	Transparent	Spread risk	National data	Regional data
New positive tests/day	middle	high	middle	yes	yes
Active (infected)	high	high	high	yes	yes
Hospitalized cases	high	high	middle	yes	no
Intensive Care	high	high	middle	yes	no
Fatalities/day	high	low	low	yes	no
Hospital free beds	low	low	low	no	no
Recovered/day	middle	middle	middle	yes	no
Vaccinations	middle	high	middle	yes	no
Infected medic. staff	high	high	low	yes	no
Age of infected	high	high	middle	yes	no

The 7 day period reflects reliably the infection period in which COVID-19 can be detected with high probability. On one hand, a 3-4 day period is not sufficient the smooth the formal distortions of data over the weekends. On the other hand, a 14 day observation period can reduce the possibility to monitor the dynamics of spreading and implement adequate measures for preventing the effects of pandemics. This was the case in many small and big countries in the EU, but also in Brazil, India, Russia and the USA. The normalized to maximums results of panel indicators (new positive case/day - 3669/ 28 March 2021,active cases - 95442/ 7 Dec. 2020, hospitalizes case - 10649 / 5 April 2021, intensive care - 813 / 5 April 2021, fatalities - 140 / 27 Nov. 2020) are shown in Fig.2

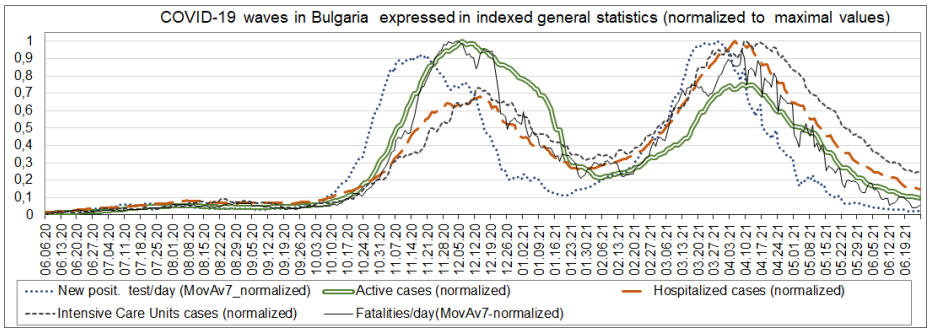

Fig. 2. COVID-19 general data panel statistics for Bulgaria (normalized to maximums)

A very important issue in the COVID-19 crisis is the vaccination process, but unfortunately, its volume and speed in many EU countries are very limited due to organizational weaknesses in national health care services and delivery delays from vaccine suppliers in the whole period until May 2021. In this context, the reported problems with some of the vaccines led to unexpected demotivation and

uncertainty among the population and health services. The information about the vaccination process was generalized mainly on a national level, and for that reason its entropy aspects could not be analysed adequately.

4 Nominal and normalized COVID-19 entropy

Under conditions of uncertainty and insufficient traditional statistics the entropy approach can provide additional possibilities for exploring the dynamics and spread of COVID-19. The signals from panel statistics can be supported with reliable information about the entropy level for the risky infection factor at sub-national, namely the "Active cases". Taking into account that in all countries the regional units have very different population in addition to the entropy assessment based on nominal raw data ("region by region") we introduce a second more objective indicator which is adjusted to the same size of population of 100,000 (100K) citizen, e.g. "Active cases per region of 100000(100K) persons". The results in the formats of nominal cumulative entropy and single-system normalized entropy are displayed in Figure 3.

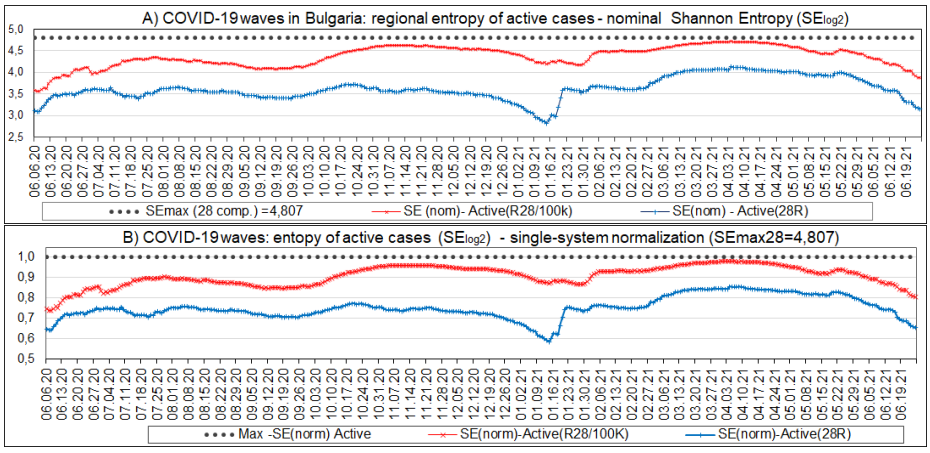

Fig. 3. Nominal and normalized regional entropy of COVID-19 spread in Bulgaria

The graphical analysis confirms, that the nominal and normalized formats produce similar profiles which values differ only in scale. In the two indicators, the profiles of entropies differ, since they are based on different sets of data. They provide a clear indication about three consecutive COVID-19 waves - two completed and one in the process of fading away in June 2021 when both indicators are slightly higher than the levels of June 2020. The first entropy indicator is based on official data about the 28 regions ($SEnom$ and $SEnorm$ "Actives per region") and produces lower levels of entropy, but at the same time with higher sensitivity and dynamics. This is easy to explain since the raw data set of

relative weights reflects the information about the distribution of "Active cases" in different territorial structures. Actually, the largest region in Bulgaria is the capital Sofia with a population of 1.33 Million (19.2% of total population), and the smallest is Vidin - with 0.082 Million, (1.2%).

In the indicator "Active(28R/100K)" the population of all regions is virtually equalized for a basis of 100000 citizens (100K) to ensure a better comparability of results. Thus, we create a possibility for monitoring the spread of infection in a virtually homogenized environment. Such data driven abstraction allows to receive another objective view about the dynamics of pandemics and on an equal basis to explore in future other territorial (population density, social mobility) or agent-related (age, income, gender, health status, etc.) parameters.

Logically, after such equalizing the $SEnom$ and $SEnorm$ formats for "Active(28R/100)" produce a higher level of entropy with less volatility and these results are more reliable for considering the overall effects of pandemics. At the same time, the higher volatility in "Active(28R)" is reflecting the nominal intra-regional dynamics which is potentially very helpful as "alert signalling" for new waves that usually start from bigger and more active regions or cities and later spreads in the remaining regional units. As a result, the combined pair of entropy indicators provides a deeper insight into the available data and a double macroscopic vision for framing the duration of COVID-19 waves.

5 Increased entropy in 2nd and 3rd COVID-19 waves

At the beginning of the crisis, doctors and data scientists did not have enough experience to explore the COVID-19 data but later several approaches, methods, and models were developed and applied in different countries. One possible path of research is to combine the analysis of traditional time series that reflects only one parameter (mono-statistics) with time series for structural entropy which reflect the aggregated results for all the components (agents) in the system. In our case, the data set contains 28 data series for each of the 28 regions, which can be regarded as agents interacting with the infection. In this sense, the mono-parametric and entropy time series would be more useful if considered in parallel for analysing the evolution of pandemics waves.

In this report we focus on the territorial dimension of entropy and the spreading of infection among 28 Bulgarian regions. Figure 4 displays the attempt to capture and frame the COVID-19 waves with the aid of the two regionally aggregated regional entropy indicators for "Active cases" (discussed above, Fig. 3) and two of the five panel statistics - "Active cases" and "New positive tests" which can be associated with the genesis and the main processing risks of the infection in the SIR/SEIR models. Due to the very limited number of tests the 1st wave was not as dramatically reflected in the Bulgarian official statistics as in other countries. As logically expected, during the 2nd and the 3rd waves the registration of "New positive tests" precedes the accumulation of "Active cases".

The two entropy indicators reflect correctly from the start the specifics of all three cycles. In addition, the up-turns of territorial diversification preceded

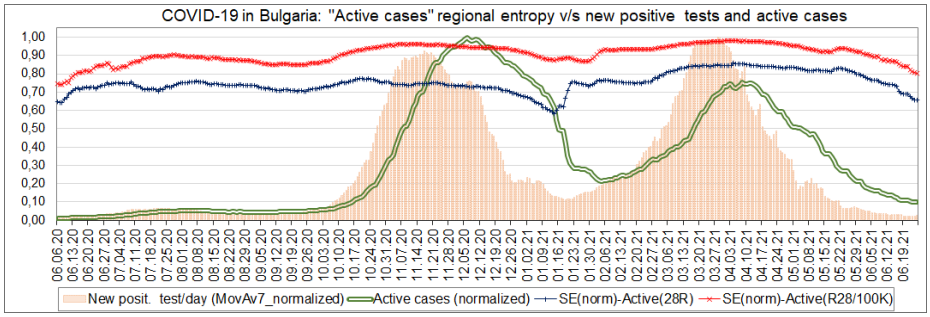

Fig. 4. COVID-19 waves in Bulgaria: entropy, new positive tests and active cases.

with 8-12 days the main traditional alarming signals for "New positive case". Unfortunately, in the spring and summer of 2020, this aspect did not receive due attention and analysis. Later, in each subsequent wave is observed a tendency for increasing of nominal and normalized entropy for "Active cases per region per 100K" above certain levels (in our case between 0.85 and 0.9). This is a reliable indication that the spread of infection is not linear or centralized and cannot be mitigated with liberal measures on a regional level. In this case, the entropy approach reflects the material information of virus spreading. Further, the rising of entropy levels above 0.90-0.95 reflects a diffusion type spreading and a danger of non-controllable chaos for a longer period. This is particularly true for the cases of intrusion of new virus variant that are characterized by accelerated spreading, longer recoveries, and more serious complications.

Regions and cities with open economies and active international and regional exchanges for tourism, work, business, and transport are particularly vulnerable to become the main gateways for more rapid spread of infections. In Bulgaria, this was the case for several regions and cities, whose population is more mobile in search of work or leisure recreations. In addition, the high entropy of "Active cases" can be regarded as an indirect indication about potential hidden channels of infections in areas with lower access and quality of health care services.

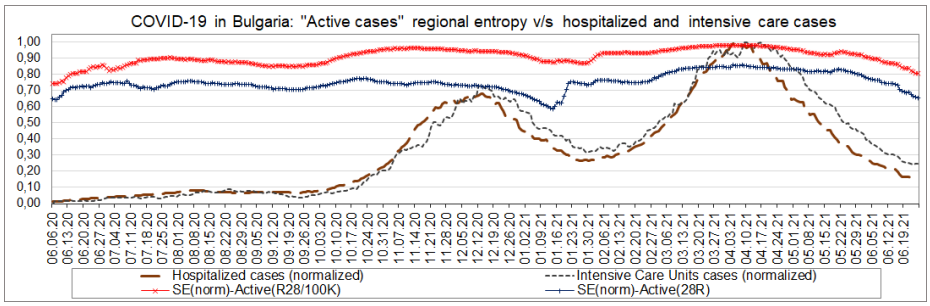

Fig. 5. COVID-19 waves in Bulgaria: entropy, hospitalized and intensive care cases.

As a next step, Figure 5 displays the attempt to frame the COVID-19 waves with two entropy indicators for "Active cases" aggregated on a national level and the other mono-statistical indicators from the panel statistics - "Hospitalized patients" and "Intensive care units cases", which can be linked with the results and back-end of SIR model. Here also, the first wave of hospitalization was not as clearly expressed as in other countries (Italy, Spain, UK). In the 3rd wave which started in February 2021 the "hospitalizations" and "fatalities" registered new record maximums in parallel with the highest entropy levels. These results can be interpreted as a confirmation, that the new mutations of the Coronavirus have more serious effects, taking into account that during the 2nd and the 3rd waves Bulgaria was dominated by the "British variant" of the virus. This observation is particularly valuable in the light of the possible rabid spreading of the Indian ("delta") variant of the virus in the summer of 2021.

The minimum and maximum values of the two entropy indicators and the five panel statistics allow to compare and analyse the duration periods of COVID-19 waves in Table 3:

Table 3. Selected values of Shannon Entropy in $SE(log_2)$ format

Wave	Indicators	Start	Min	Peak	Max	End	Min	Duration
	$SE(norm)/region$	6-6-20	0.644	7-8-20	0.759	25-9-20	0.706	99 days
	$SE(norm)/region100k$	6-6-20	0.745	28-7-20	0.857	15-9-20	0.848	91 days
	New positive tests/day	6-6-20	41	6-8-20	227	15-9-20	133	99 days
1st	Active cases	6-6-20	980	7-8-20	5205	15-9-20	4402	99 days
	Hospitalized patients	6-6-20	147	12-8-20	861	21-9-20	718	105 days
	Intensive care	6-6-20	12	24-8-20	74	24-9-20	28	109 days
	Fatalities/day	6-6-20	0	4-8-20	9	20-9-20	4	104 days
	$SE(norm)/region$	25-9-20	0.706	23-10-20	0.773	14-1-21	0.584	110 days
	$SE(norm)/region100k$	15-9-20	0.848	13-11-20	0.962	29-1-21	0.867	134 days
	New tests/day	15-9-20	143	9-12-20	3980	19-1-21	416	124 days
2nd	Active cases	18-9-20	4402	6-12-20	95442	4-2-21	20496	137 days
	Hospitalized patients	21-9-20	718	14-12-20	7244	27-1-21	2818	127 days
	Intensive care	24-9-20	28	13-12-20	595	29-1-21	257	124 days
	Fatalities/day	20-9-20	2	4-12-20	140	27-1-21	30	127 days
	$SE(norm)/region$	14-1-21	0.584	4-4-21	0.857	30-6-21	0.655	165+
	$SE(norm)/region100k$	15-1-21	0.867	4-4-21	0.980	25-6-21	0.804	160+
	New tests/day	20-1-21	416	28-3-21	3680	25-6-21	361	155+
3rd	Active cases	4-2-21	20496	2-4-21	70919	25-6-21	28490	150+
	Hospitalized patients	27-1-21	2818	5-4-21	10649	25-6-21	4201	160+
	Intensive care	29-1-21	257	29-3-21	773	25-6-21	460	160+
	Fatalities/day	27-1-21	30	2-4-21	126	25-6-21	4	160+

Both entropy indicators for regional entropy of "Active cases" registered increasing maximal values in the three consecutive waves. This is evident in the gradual entropy increase of "SE(norm)-Active/100k/region" in term of normal-

ized entropy values: 1st wave - 0.857; 2nd wave - 0.962; 3-rd wave - 0.98. At the same time, the duration of periods with maximum levels of entropy in each consecutive wave is also increasing: 1st wave - 2 weeks with maximum of ~ 0.85; 2nd wave - 2,5 months (mid. Oct. - end Dec.) at ≥ 0.94. During the current 3rd wave the maximum level of entropy of ≥ 0.97 was reached very quickly (within the 1st week of February,2021) and since than the up-turn phase duration was approximately 5.5 months. The down-turn phase in the 3rd wave was longer than in some other countries due to the low speed of the vaccination process and the more liberal social containment measures.

The duration periods of COVID-19 waves framed by the two entropy and five panel statistic indicators were between 91 and 104 days in the 1st wave, and between 110 and 137 days in the 2-wave. According to information until 21 May, 2021 the 3rd wave is still continuing and will be much longer than previous waves. In the 2nd and 3rd waves, the starting moments in the two entropy indicators differed by 10 days and they preceded the panel indicators with 3-10 days in the 1st wave and with 13-20 day in the 2nd wave. In the 3rd wave the entropy indicators preceded the general panel indicators with 10-15 days. The main reason for this is the fact that entropy as a structural parameter is able to reflect in real-time the dynamics of distribution. Although it may seem that the 3rd wave is about to end, the serious risks persist. The prediction capacity of entropy can be very useful for defining more adequate prevention policies and national an regional levels. The major advantage of entropy is due to fact, that it captures immediately the dynamics of diversification of the virus propagation instead of relying on the cumulative values in the standard indicators. In other words, the increase of entropy is a clear signal that the channels of propagation are getting larger which will result in more active spread of pandemics. In addition, we cane note that the larger spread of the two indicators between the 2nd and the 3rd wave was and indication about the more active propagation of the virus through the channels of major cities and regions which contributed for the worse results in the 3rd wave.

Within the 1st wave the maximum of $SE(norm)/region100k$ was reached in 71 days with an increase of 0.112; during the 2nd wave - in 88 days with an increase of 0,114; and during the 3rd wave in 62 days for an increase of 0.113. The increased entropy in the 2nd and 3rd waves can be explained by several factors: the winter period, penetration of new and more infectious variants of the virus, more liberal policies for social mobility. The record high levels of normalized entropy closer to the absolute maximum of "1" in Bulgaria during the 3rd wave can be explained by the domination of the "British variant" and the very liberal containment measures compared to other countries in the EU. The national responses to such challenging "natural evolution processes" as COVID-19 will depend for long periods on the specifics of socio-economic and psycho-cultural specifics in different countries, regions, and social groups. These specifics should be more seriously considered in the ways of reporting official statistics, defining containment measures and implementing vaccination plans.

6 Conclusions and further research

The logical combination of reliable research methods is a promising path for exploring more efficiently the dynamics and evolution of complex systems and processes such as COVID-19 and other pandemics. Under conditions of uncertainty, the entropy approach allows to add new dimensions in the exploration of insufficient public data and to obtain valuable macroscopic insight on the spread of infections on national and regional levels. Being based even on limited empirical data this study confirms that the data-driven entropy approach provides very useful holistic view and prediction capability for enhancing the analysis of COVID-19. Our findings confirm that at the end of the 3rd wave the infection is in temporary retreat in Bulgaria as in many other countries and continues to be a serious risk for public health, economic development, social stability, and national security. The robust entropy concept can be a valuable contribution for analysing complex epidemic processes and optimizing the activities in many areas, and especially in health care services, education, communications, energy, etc.

The improvement of publicly available statistics should contribute to developing more reliable models and justified policies. Future research can be enlarged on internal and international levels and will definitely include the analysis of vaccination results for improving the quality of comparative studies and the knowledge about COVID-19 and eventual similar and subsequent epidemics.

References

1. Bandt, C.: Entropy ratio and entropy concentration coefficient, with application to the covid-19 pandemic. Entropy (Basel) **22**(11) (2020). https://doi.org/10.3390/e22111315. PMID: 33287080; PMCID: PMC7712116
2. Bulgarian Ministry of health: Official web site for covid-19 - last accessed: 25 june 2021, https://coronavirus.bg
3. Golan, A., Judge, G., Miller, D.: Maximum Entropy Econometrics: Robust Estimation with Limited Data. John Wiley, New York (1996)
4. Harte, J., Newman, E.: Maximum information entropy: a foundation for ecological theory. Trends Ecol. Evol. **29**(7), 384–389 (2014)
5. Jaynes, E.: Information theory and statistical mechanics. Physical Review **106**, 620–630 (1957)
6. Jaynes, E.: On the rationale of maximum-entropy methods. In: CONFERENCE 1982, Proceedings of IEEE. vol. 70, pp. 939–952 (1982)
7. Kenah, E., Robins, J.: Network-based analysis of stochastic sir epidemic models with random and proportionate mixing. Journal of Theoretical Biology **249**(4), 706–722 (2007). https://doi.org/10.1016/j.jtbi.2007.09.011
8. Li, et al.: Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. New England Journal of Medicine **382**(13), 1199–1207 (2020). https://doi.org/10.1056/NEJMoa2001316
9. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal **27**(3), 379—423 (1948)