Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cjas20

On regime changes of COVID-19 outbreak

A. Tchorbadjieff, L. P. Tomov, V. Velev, G. Dezhov, V. Manev & P. Mayster

To cite this article: A. Tchorbadjieff, L. P. Tomov, V. Velev, G. Dezhov, V. Manev & P. Mayster (2023) On regime changes of COVID-19 outbreak, Journal of Applied Statistics, 50:11-12, 2343-2359, DOI: 10.1080/02664763.2023.2177625

To link to this article: https://doi.org/10.1080/02664763.2023.2177625

JOURNAL OF APPLIED STATISTICS
2023, VOL. 50, NOS. 11–12, 2343–2359
https://doi.org/10.1080/02664763.2023.2177625

APPLICATION NOTE

On regime changes of COVID-19 outbreak

A. Tchorbadjieff [a], L. P. Tomov [b], V. Velev [c], G. Dezhov [d], V. Manev [e] and P. Mayster [a]

[a] Institute of Mathematics and Informatics Bulgarian Academy of Sciences, Sofia, Bulgaria; [b] Department of Informatics, New Bulgarian University, Sofia, Bulgaria; [c] Department of Infectious Diseases, Parasitology and Tropical Medicine, Medical University of Sofia, Sofia, Bulgaria; [d] Faculty of Mathematics and Informatics at Sofia University, Sofia, Bulgaria; [e] Fakultät für Mathematik und Informatik, Ruprecht-Karls University Heidelberg, Heidelberg, Germany

ABSTRACT

The COVID-19 pandemic has had a very serious impact on societies and caused large-scale economic changes and death toll worldwide. The first cases were detected in China, but soon the virus spread quickly worldwide and the intensity of newly reported infections grew high during this initial period almost everywhere. Later, despite all imposed measures, the intensity shifted abruptly multiple times during the two-year period between 2020 and 2022 causing waves of too high infection rates in almost every part of the world. To target this problem, we assume the data heterogeneity as multiple consecutive regime changes. The research study includes the development of a model based on automatic regime change detection and their combination with the linear birth-death process for long-run data fits. The results are empirically verified on data for 38 countries and US states for the period from February 2020 to April 2022. Finally, the initial phase (conditions) properties of infection development are studied.

ARTICLE HISTORY
Received 14 December 2020
Accepted 2 February 2023

KEYWORDS
COVID-19; linear birth–death processes; change point analysis; statistical inference for branching processes

MATHS
60M20; 60J85; 62-07

1. Introduction

The initial information about COVID-19 was scarce and intimidating in the beginning of 2020. Not surprisingly, the public interest was focused on almost real-time infection spread trackers and prediction tools. Most of them were based on already existing epidemic modelling theories, incorporating separately or in combinations specific parameters, such as spatial details or realistic population mixing structures, individual-based network models, and simple SIR-type models that incorporate the effects of reactive behaviour changes or inhomogeneous mixing [11].

There is also a long tradition of using stochastic epidemic models to simulate and predict the transmission dynamics of infectious diseases. They are important when the number of infectious individuals is small or the transmission and recovery rates vary with time due

CONTACT A. Tchorbadjieff atchorbadjieff@math.bas.bg Institute of Mathematics and Informatics Bulgarian Academy of Sciencies, Acad. Georgi Bonchev Str., Block 8, Sofia 1113, Bulgaria

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02664763.2023.2177625

to various reasons. Very popular stochastic models in epidemiology are continuous time Markov chains (CTMCs), stochastic differential equations (SDEs) and branching process approximations. Detailed review on stochastic epidemic modelling is available in [2,14].

In aim to implement a real-data stochastic application, we we use the theory of branching processes for modelling COVID-19 outbreak [5,16,30]. Estimates for the probability of extinction based on them have been applied frequently to populations, genetics, cellular processes, and to epidemics on networks. Some methods for the calculation of disease extinction thresholds in deterministic and stochastic models are summarised in [3]. They estimate the probability of a major outbreak for the susceptible – infectious – recovered (SIR) model when the population size is large and a small number of infectious individuals are considered.

Important assumptions when using branching processes are that each infected individual (usually noted as a particle in branching theory) spreads disease (gives birth) independently to others and everyone has an equal probability of getting infected. For a small number of infectious individuals and a large population size, these assumptions are realistic and similar approximations are very good in many cases. Another important advantage of selecting a stochastic solution is the lack of knowledge about population immunity similar to early assumptions [28].

Another possible approach is based on the multi-type branching processes with inhomogeneous Poisson immigration [26]. The advantage of this model is that the random arrival of infected persons makes a serious contribution to infection spread. However, the branching process is reduced to a single type with immigration when modelling COVID-19 outbreaks during the initial period outside Wuhan. The branching reproduction starts after multiple independent arrivals of infected patients zero. Then, local clusters emerge from them after a time delay dependent on the incubation period, estimated within the range of two to nine days with 95% confidence [22].

This dominance of immigration continues until the infection gains local dynamics and surpasses the imported cases. In order to describe this inhomogeneous dynamic, caused by different social factors, the data is split in consecutive intervals. The initial conditions for every interval can be considered altogether with the immigration. They cannot change the critical parameter of the branching reproduction, but impact the extinction probabilities [24,32].

Finally, the public reactions and measures against spread of contagious diseases are an important part of the response to them. The different *actions* may include quarantine at a local and national level, border closing, physical distances, etc. According to data models, all actions for prevention are assumed as constraints on *free infection spread*. The estimate of intervention effect could be obtained by stochastic SIR network epidemic model with preventive dropping of edges [7].

In accordance with all these assumptions, we designed and implemented a computational model for automatic fit and short time prediction of COVID-19 infection incorporating all significant changes in ambient conditions and natural trends. It is based on the linear birth-death processes $X(t), t > 0$, starting from one particle $X(0) = 1$, [32]. It is combined with change-point theory to implement the dynamic detection of the changes in infection spread rates. The model is calibrated and verified on data from 38 countries. Finally, the empirical explanations for detected changes in initial conditions and infection dynamics is looked for.

2. Data modelling

In general, the branching process $Z(t), t \geq 0, Z(0) = Z_0$, with probability generating function (p.g.f.) $F(t,s) = E[s^{Z(t)}|Z(0) = Z_0], |s| < 1, F(t,1) = 1$, is defined by the initial condition, the lifetime of particles and the offspring number with p.g.f. $h(s), |s| \leq 1$. Suppose, the p.g.f. $U_0(s)$ describes the number of particles $Z_0 = Z(0)$ at the initial moment, (say today). The branching intensity depends on the lifetime of particles. Knowing the independence of particle evolution, the assumption that the lifetime is a random exponentially distributed variable with parameter K guarantees the Markov property of this branching process $Z(t)$. It means that the entire past information of the process until the time moment $t > 0$ is presented in its current state $Z(t)$. Thus, the probabilities of future events are completely determined by the present state of the process – if its current state is known, its past behaviour provides no additional information in determining the probabilities of future events.

The linear birth-death process $X(t), t > 0, X(0) = 1$, can provide locally the best explanation of the branching evolution starting with a single particle. The probability of a birth (successful transmission) of an infectious individual is denoted by p, where $0 < p < 1$, describes the outcome of either two daughter particles' birth or parent demise. The offspring number is defined by a random variable η with p.g.f. $h(s) = E[s^\eta] = ps^2 + 1 - p$, $E[\eta] := h'(1) = 2p$.

The ultimate extinction probability, traditionally denoted by $q := \lim_{t\to\infty} P(X(t) = 0)$ is the smallest nonnegative solution of the equation $h(s) = s$. The infinitesimal p.g.f. is $f(s) := K(h(s) - s) = K(ps^2 - s + 1 - p)$ and the mean of the infinitesimal offspring number is denoted by $m = f'(1) = K(2p - 1)$. Then the mathematical expectation is $E[X(t)] = e^{mt}$. A branching process $X(t)$ is classified as supercritical if $m > 0, \frac{1}{2} < p < 1$, critical when $m = 0, p = 1/2$, or subcritical for $m < 0, 0 < p < \frac{1}{2}$. The extinction probability of the linear birth-death process to the finite time $0 < t < \infty$ is given by the value $P(X(t) = 0)$ in its explicit form [32].

Due to the Markov property, the information on the first time interval of approximation $t \in [0, t_1)$ is provided by (U_0, h_0, K_0). Thus, for inhomogeneous branching process $Z(t), t > 0, Z(0) > 1$, the information on the first time interval depends on the epidemics evolution, i.e. the probability of a successful transmission $p = p_0$ is in the time interval $0 \leq t < t_1$ and $K = K_0$. Then the medical research specialist can recalculate the new parameters (U_1, h_1, K_1) and the new transmission probability for the next interval of approximation, where $U_1(s) = E[s^{Z(t_1)}]$ describes the real data at the time moment $t_1 > 0$. The p.g.f. $U_1(s)$ and the random variable $Z(t_1)$ defines the initial conditions for the next interval of approximation, $t \in [t_1, t_2)$. Also, the intensity of reproduction and the offspring's number can be different from one interval of approximation to another, respectively (h_1, K_1) on the interval $t_1 \leq t < t_2$ and so on.

An important part of any pandemic onset is its development during the first days of contagion. In the case of pure imported infection, the local branching process begins only with immigration from outside and it is dependent on the incubation period. During the first days, the pandemic evolution is dominated by arrival intensity and define the initial conditions of the newborn branching process. Thus, this initial situation must be considered because the first COVID-19 virus variant has a quite long incubation period and all initial cases around the world except China are imported.

Appropriate probability distributions for modelling the initial conditions are Negative-Binomial (N-B) and Poisson (Po) distributions. Conveniently, the following development in the case of the linear birth-death process is defined in [24,32]. The N-B distribution may explain better the stochastic fluctuations. mainly due to the dominance of super-speeding contagion [23]. However, because the initial immigration of infected from China was limited event compared to overall population and passenger traffic until the end of February 2020, the Poisson distribution is selected as default.

The mainly used empirical estimator of epidemics dynamics is the basic reproduction number, R_0, known also as a threshold in deterministic epidemic theory. It is equivalent to the expected value of the newborn particles in branching processes. This parameter predicts a disease outbreak if $R_0 > 1$. Epidemiologists calculate R_0 tracing data for individual-level contacts at the onset of the epidemic. It is computed by averaging over the number of confirmed by tests secondary cases of many diagnosed individuals. But not surprisingly, this deterministic approach requires a complete history and it is not a reliable real-time measure for the development of the outbreak due to inevitable sampling bias. It could occur due to different factors, such as the pattern of contact underestimation, regional and local factors, methods and errors in testing, etc. A possible solution could be found in the search and implementation of more innovative and complex modelling like in [25,31].

In addition, we remark that the basic reproduction factor, R_0, is a deterministic constant and it does not reflect to changes in behaviour and social restrictions. For this reason, it is more appropriate to use the effective reproduction factor R_t at any moment $t > 0$ to model inhomogeneous development. However, the population growth rate r_t per unit in time is selected as a more convenient empirical measure. It is widely used in demographics and it is already known in the theory of branching processes [16]. The relation to productivity knowing the infection occurred in the first interval of approximation of the linear birth-death process with mean offspring number m is $R_t = e^{mt} = 1 + r_t$. Empirically, r_t is obtained from time series data by B1.

With a prolonged pandemic, the daily growth rate series $(r_1, r_2, \ldots)$ aggregates the typical for time series analysis trend and seasonality. A possible local fit of similar inhomogeneous birth-death process with linear time-dependent birth and death rates can be obtained by application with a generalised Gompertz growth model [4]. Another possible approach is to use Autoregressive Integrated Moving Average (ARIMA), similar to [1,13]. However, for inhomogeneous data, regime changes occurred by time cause temporal changes of ARIMA model parameters, such as in [33].

In aim to construct an automatic computational tool, the occurred regime changes in growth rate series have to be detected at the first step. They are identified by applying change point analysis. Initially, the cumulative sum control chart (CUSUM) method has been used at the early stage of the process. It is very efficient non-parametric test to detect small shifts in the mean of a process, introduced by E. S. Page in 1953–1955 [27]. Due to its simplicity, it is widely used in many different change point applications. In epidemics it is successfully used in early stages of infections [36]. It works as established continuous inspection scheme to detect unknown location parameter k where the mean value μ of i.i.d. $x_i, i = 1, \ldots, n$ changes significantly. In statistical terminology, this means to test the null hypothesis H_0 against the alternative H_A if exists $1 \leq k \leq n$

such that

$$H_0 : \mu_k = \mu$$
$$H_A : \mu_1 = \mu_2 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n.$$

The standard methods are based on properly standardised cumulative sums (CUSUM)s $\sum_{i=1}^{k}(x_i - \hat{x}_n), k = 1, \ldots, n$, where $\hat{x}_n = (1/n)\sum_{1\le i\le n} x_i$. The null hypothesis is denied if some derivative statistics W_n from the cumulative sum is too large.

There are different computational method modifications for estimation of W_n, generally following the original definition.The software,selected in this work, relies on the use the likelihood style ratio of the CUSUM functionals instead of the differences for estimation of magnitude for the computation of W_n. It is proposed and the required important limit theorems are proved in [12]. The computational implementation relies on library *changepoint* in *R statistics environment* [20].

The CUSUM method was applied on the stationary time series of daily growth rate changes, $r_i - r_{i-1}, i = 1, \ldots, k, \ldots n$, without any distributional assumptions [20]. It works well at the very beginning of the outbreak when data is very variate, small in size and hence with clear non-normal behaviour. However, in the later stages of process development, when the number of segments grew, the CUSUM began to fail. In this case, the CUSUM method produced bias towards the early stage of the outbreak, neglecting the later development. The empirical attempts of any other change point test for mean value did not produce any improvements.

For this reason, the conceptual hypothesis is changed – the regime changes are assumed as simultaneous differences of mean and variances of independent normally distributed changes. The dedicated computation again relies on likelihood-ratio procedure penalised by information criteria [10]. The software implementation is based on function *cpt.meanvar* from the *changepoint* package in *R* [20]. The advantage of this computational library is the implementation of several methods for optimisation of multiple segmentation process for large data. The available methods are Pruned Exact Linear Time (PELT), Binary Segmentation and Segment Neighbourhood [6,18,21,29]. The At Most One Change (AMOC) methods is also implemented, but it is mainly associated with CUSUM. However, the previous three ones provide a useful tool to deal with large data sets with many change points.

After all regime changes are detected, the data in segments between two consecutive change points are fitted with the new process parameters. Because the length of segments varies randomly, starting from a few elements, the precise real-time application of ARIMA is not feasible. However, the Markov property of the linear birth-death process $Z(t), t > 0, Z(0) > 1$, enables a quick regime switch, allowing real-time implementation in the segment time window $T = T_i, i = \overline{1:n}$. The linear birth-death process in selected segment is determined by the average probability estimate $\langle p_T \rangle$. It is obtained directly from the local mean rate parameter $\langle r_T \rangle$ in studied interval T:

$$\langle p_T \rangle = 1/(1 + e^{-\langle r_T \rangle L}). \tag{1}$$

The formula is derived from ultimate extinction probability q in the critical and supercritical process. In the later case, the probability is centred at $1 + r_t = 2$, requiring $L = 2$ to obtain $e^{-\langle r_T \rangle L} = q$ in Equation (1). Finally, an additional correction of L is introduced to adjust probability estimate to an optimal value. It is a country dependent small constant c that is computed by iterative optimisations based on minimisation of mean square error (MSE), B2. Another possible option for data adjustment is correction of intensity branching parameter $K = K_t, t = t_i$, after every regime change. Both corrections are simple and computationally cheap substitute for adopted in demographics method of blending – passing from one curve to other [8].

When the initial condition is either Poisson or Negative Binomial, the solutions for the linear birth–death process can be obtained analytically, as it is shown in [32]. However, the results obtained from the available numerical experiment are preferred for automated computations when the explicit determination of intermediate initial conditions is not critical. The simulator generates trajectories after 10,000 independent realisations as it is explained in the provided Supplement material, extending the realisation in [32]. The integrated results over all paths are obtained from averaged values. The computational procedure can be easily adjusted linearly by changing together or separately the scalars K_t and L.

3. Results

The performance of the combination between the linear birth–death process and change point analysis is tested on real data from a cumulative number of COVID infected patients by regional separation. Those included in this research data consist of reported daily new cases in 38 different entities. They were selected intentionally to represent geographical, political and cultural diversity. The main focus is set on data from countries assumed as open and relatively effective in COVID-19 response. The other criteria were variety in policy measures like the very different approaches such as Sweden and Kuwait, population density – North Dakota in comparison to Singapore and New York or total population over large territory coverage in countries like the USA, Canada, Australia, Russia, and India against cities like Hong Kong, Singapore and New York or densely populated countries like Japan and Netherlands. The expectations for common policy of the European Union are also considered. In general, the geographical distribution is as follows:

- The European Union (EU): Germany, France, Italy, Spain, Netherlands, Denmark, Sweden, Austria, Greece, Czechia, Croatia and Bulgaria.
- Outside of EU common policy: UK, Norway and Russia (incl. parts of Caucasus and Siberia)
- North America: Canada, USA plus Texas, Florida, California, South Dakota and New York.
- South America: Argentina, Brazil and Chile
- Far East Asia: Hong Kong, Japan, South Korea, Singapore and South Korea
- South Pacific and India: Australia, New Zealand and India.
- Rest of the World: South Africa, United Arab Emirates (UAE), Kenya, Kuwait, Israel and Turkey.

The major data source is the dedicated public John Hopkins' database and maps [14]. The data are preprocessed for irregularities and growth rate values are computed. The detailed explanations are available in provided Supplemental material.

The COVID-19 outbreak at every country began as independent arrivals of newly infected persons. Their number and intensity varied due to multiple reasons. But, the probabilistic distribution of daily infections is expected to follow either Negative Binomial or Poisson ones. If the process remains completely exhaustive, i.e. the infection dies without transition and the transmission probabilities remain infinitesimally small, but not in deterministic zeroes. Thus, having permanently single count infected persons the new super-spreader cluster could emerge, triggering a new branching process. Another possible source of renewed infection could be immigration.

By assumption, we have to distinguish the secondary pandemic outbursts from the initial conditions at the first infected person arrival. However, in the terms of modelling, every new wave of infection outburst is separated from the previous and following calm periods by clear regime changes. For data fit software, the moment t_i when change point occurs is considered as the end of the previous time window and the initial condition for linear birth–death process with new parameters.

3.1. Model initialisation

The early days after the first infected particles arrival are of high importance for further pandemic development. Due to the long incubation period and without registered local infection yet, the new daily counts are assumed as a Poisson counting process. Then, when the first local branching clusters are available, the rate of newly infected persons accelerates. At this moment neither branching nor immigration dominates. In case of local transmission collapse for any reason, the process $N(t), t > 0$, giving the number of particles alive remains homogeneous Poisson (HP) with constant parameter $\lambda > 0$ and mean λt, such as

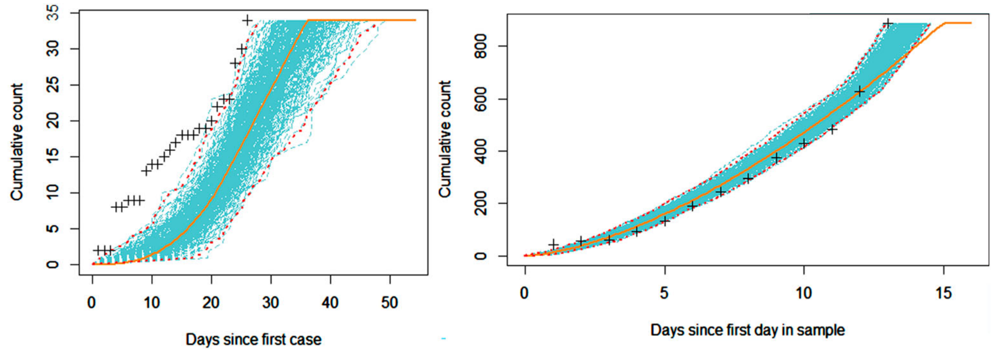
$$P(N(t) = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \quad n = 0, 1, 2, \dots. \tag{2}$$

However, with the prolonged process, two possible scenarios emerge separately or combined. The first is that immigration evolves due to inhomogeneous Poisson process similar to that reported in [17]. Another possibility is when the branching continues to expand causing growing domination of locally transmitted infection. The daily numbers $N(t), t > 0$ are time dependent. For generalisation of both cases, the considered distribution is non-homogeneous Poisson (NHP) with mean

$$\Lambda(t) = E(N(t)) = \int_0^t \lambda(s)\, \mathrm{d}s.$$

The location and estimation of the state before the first regime change are important tasks for initial conditions' definition and computation initialisation. Because any detailed specific rule does not exist, the first change point following either the 10th day after the first infection or after the first 50 cases are reported is expected to include completely initial conditions. As the data show, it is a time when exponential growth has just begun in most of the observed data series.

However, the hypothesis of homogeneous Poisson distribution of daily data before the first change point is not confirmed by empirical data from all studied 38 areas. The maximum likelihood estimation (MLE) shows that the mean parameter λ of the homogeneous



(a) (b)

Figure 1. A comparison between real IC data and simulated estimates from 300 repetitions for 3σ confidence intervals for HP and NHP distributed in UK. The HP trajectories are computed for λ equal to 1.37 and shown in (a). The NHP process is generated by Power Law model time function at^b, *where* $a = 7.2$ *and* $b = 0.56$. The real data are marked with $+$. (a) HP distributed data and (b) NHP distributed data.

Poisson distribution is a biased estimator. Thus, despite $n\lambda$ yielding a good approximation for cumulative counts at the end of the period with length n, the predictions for previous days are wrong. The MLE mean values for homogeneous Poisson distribution are computed by the function *fitdistr* in the packet *MASS* in *R* [35]. The obtained results are shown in Table A1.

For further analysis, this initial time window is split again on only two parts by the already adopted change point model, splitting homogeneous and non-homogeneous processes. In general, as a thumb rule, the first time ordered sample and the remaining part are usually HP and NHP distributed. We also note that in cases when there is not significant immigration, the second period could be assumed as a linear birth–death process. To distinguish them more clearly, in the following text we will denote the firstly occurred in time part, which is HP distributed, as *Initial condition*. The following part will be notified as a *mixing stage*, referring to still serious impact of arriving from abroad infection. The prediction fits for both intervals are significantly improved using the library *poisson* in [9]. Results with simulated trajectories for United Kingdom (UK) are shown in Figure 1.

From samples obtained in this way, several additional conclusions on the initial phase of the pandemic can be made. At first, a conclusion can be drawn of possible undocumented infections during the first weeks of pandemic due to low data quality. This conclusion is confirmed by examples like time series from the Netherlands, where subsection of HP distributed part is missing. Another hypothesis for data inconsistency can be yielded from the difference between projected trajectory of the HP process with optimal λ and the real data. In the UK, the empirical data are constantly outside the 3σ confidence intervals of the modelled HP process with λ directly obtained from MLE, see Figure 1(a). Note that this is the period exactly before lockdown when chaotic travels from and to the country occurred.

The next property noticed from the initial condition (IC) data is that the parameter λ for initial condition does not depend on population size, but λ weighted on population density depends on geographical location and the proportion of IC time length in days and the number of days when IC is detected after 1st February 2020 process. The conclusions are obtained from the Log-gamma Generalised Linear Model (GLM) with 34 degrees of freedom [15]. The optimal model is obtained after investigation of different scenarios. The selection is based on a stepwise procedure by minimising the Akaike Information Criterion (*AIC*), 10-fold cross-validation procedure and residuals heteroscedasticity. The final model is with mean square error (RMSE) of 165.33 and bias of −5.37. The details for computational process are available in Appendix B.

3.2. Data fit performance

The data modelling begins with those obtained from the initial change point split period, including initial condition and *mixing stage* periods. If followed technically strictly [32], the initialising values have to be computed by the estimate of HP distribution for the *initial condition*. However, due to data differences, the initial point is set at the first change point. The starting value for every trajectory is computed of the approximate value of $n\lambda_{whole}$, where n is the size of the whole period until the first detected regime change. This homogeneous Poisson parameter λ_{whole} is *used for simplicity* despite that it is biased and does not explain the data trajectories from the initial moment $t = 0$ to any $t < n$. However, it is good enough and easy to compute at the moment n, where all possible trajectories are aggregated. Computed values are shown in Table A1.

The fits are repeated regularly multiple times with different time series length during the period 2020–2022. At first, the computational algorithm relies on automatic non-inferred one step detection of all possible change points. Then, the local average probabilities p_T for every segment are computed from previously precomputed growth rates for all t. At the next step, all 10,000 trajectories are executed consecutively ordered by time. This splits the cumulative stochastic error of the process between all available trajectories (even zeros, if it exists). Thus, the expected number of infected people $E(X_t)$ and variation $V(X_t)$ are, respectively, the empirical mean and dispersion over all resulting trajectories at the moment t in the interval $t_i < t < t_{i+1}$.

This approach provides an easy to implement and effective automatic algorithm for obtaining a good estimate for the expected value at long distant time moment and short time future predictions. In addition, it is easy to adjust the final result by applying the correction value of c. The value can be computed by selecting the value with minimal overall MSE error. It is executed autonomously and without supervising modelling for all 38 countries. When the change points from this first iterations are obtained by the BinSeg segmentation algorithm, the fit is in good agreement with the overall trend (see Figure 2(a)). However, at some local time intervals, mainly during non-linear expansion due to two overlapping waves, this first iteration shows very low flexibility due to the large size of determined change point segmentation. For these cases, a local re-computations with a more sensitive segmentation algorithm as *PELT* improves significantly the goodness-of-fit of model (see Figure 2(b)). The errors are proved non-autocorrelated by ACF diagnostic and normally distributed (Kolmogorov–Smirnoff test for standarised values yields p-value of 0.8467).

Another major empirical observation from regime change detection in studied cases is that the frequency of detected change points usually is higher at the beginning and at the end of the period when infection rate is either strongly increasing or in decreasing mode

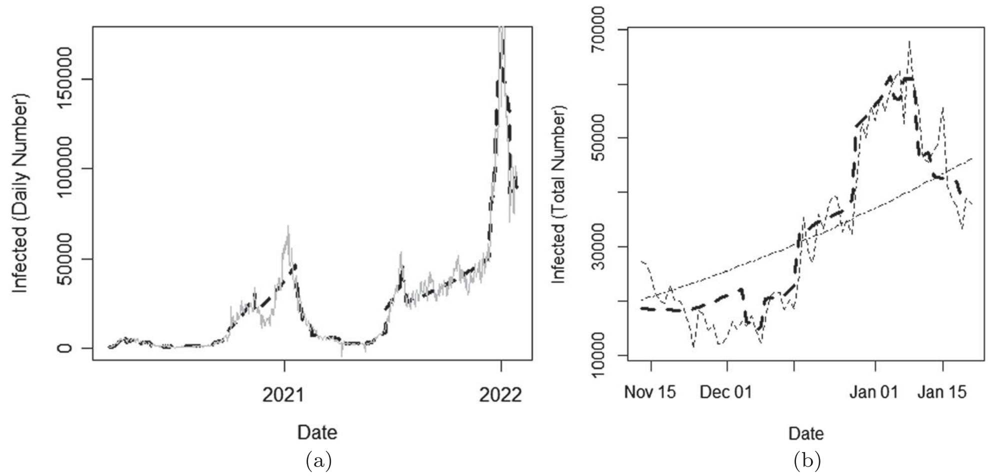


Figure 2. Examples of daily COVID-19 outbreak modelling for UK. The fit without correction shows observed (grey line) and predicted (dark) values for the period until 22nd January 2022. The second iteration fit shows empirical (grey) and predicted in first iteration (dotted) values for the period between 13 November 2020 and 21 January 2021. The second iteration fit is shown with dark dashed line. (a) UK first iteration and (b) UK second iteration.

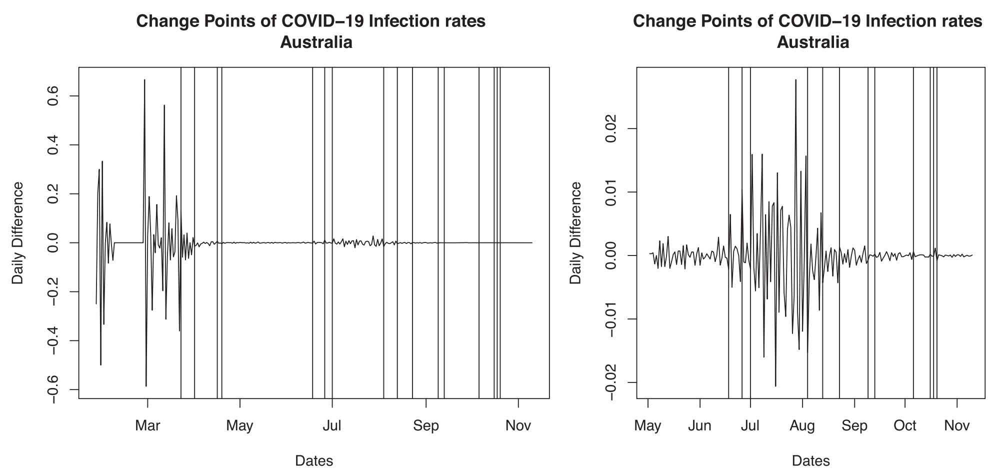


Figure 3. Detection of regime changes for daily differences in growth rate for Australia. The change points are shown with vertical lines. The whole process from the last day of IC to the end of November is shown in (a). The reduced process showing only the summer burst is shown in (b). (a) Complete period and (b) Reduced period.

(Figure 3). This is in direct relation to the long unsegmented periods during the non-linear expansion of daily cases. Having knowledge of this trend, the frequent change points can be used as time location predictors of wave arrival or extinction. Visually it is easily detectable for Australia data (see Figure 3).

After multiple differences in time repetitions of the model, change point split for the first days confirmed stable and precise for the countries with strong initial wave. The variations in detected locations of change points are negligible. However, the location of regime

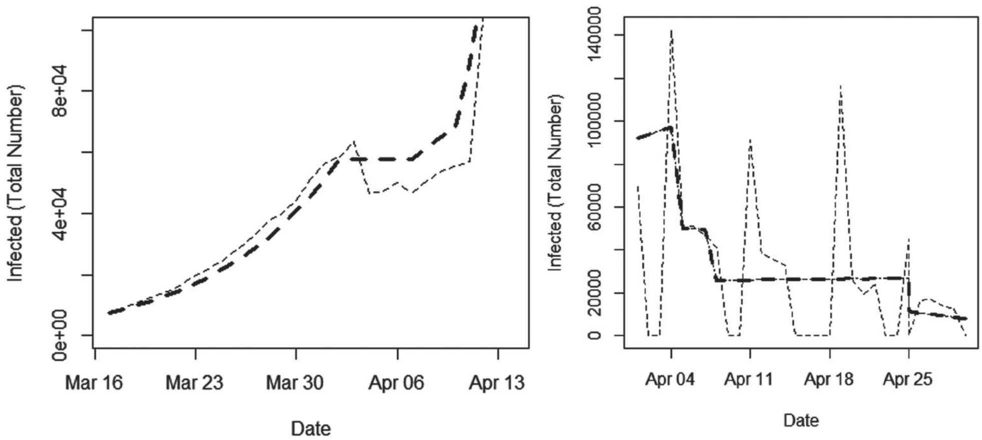


Figure 4. The figures when dynamic model recalibration is required. Case when negative cumulative level is detected in data from France in the middle of April 2020 (dotted line) (a). The periodic data effect (dotted line) appeared in UK data April 2022 (b). The first iteration results are shown in tick dashed lines. (a) Data from France with correction and (b) UK data periodicity.

changes in the early days is changing for some countries. The common observation for these cases is their delayed great shock. Such countries are Germany and Japan. Their infection rate shifted extremely to 2022 due to a strong reduction of the initial wave in 2020. Thus, when data from 2022 are included, the change points in early 2020 are changed.

There are cases when the model requires enforced recalibration. It is technically easy to apply dynamic level correction by resetting the process with new values due to the Markov property. Such case arises in France when data are updated by a reduction of the cumulative infection numbers in a single day (see Figure 4(a)). Because it is a one step negative correction, the change point analysis detects only occurrence of the correction, not the amplitude. Thus, the fit continues to follow the trend of average local probability requiring intensity resetting. The following recalibration is only a technical operation and its effect on results is only limited to the magnitude of overall intensity. Another, more complicated correction of the model is required when the measurement regime is changed to clear weekly periodicity (see Figure 4(b)).

4. Conclusion

The COVID-19 outbreak was a multiple wave pandemic driven by different viral variants. The previous sections successfully described the pandemic modelling fitted to the overall number of infected persons without distinction of data acquisition and report. However, focussing mainly on modelling, we missed some precise discussion about limitations. Possibly, distributions can be oversimplified, e.g. overdispersion can affect Poisson and negative binomial in real data scenarios. We also do not comment on different viral variants and that the instability of reproduction factor can play a role. However, these additional considerations could be modelled with appropriate data. The simplicity of software implementation allows concurrent computation of any sub-variant models. Further inference conclusions could be derived from a comparison and analysis of aggregated data.

For this reason, we share the complete software implementation with special supplements part in aim to make possible any further independent implementations. They may require additional development not considered neither in this work nor in the software. Another possible extended use is the implementation of geospatial regional modelling to study local differences and their impact on aggregated global level. This could be easily implemented, using separate modelling of regional data and following aggregation. This will improve the precision of predicted overall values by including lagging in time differences. The simulator source code can be downloaded from here.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Assen Tchorbadjieff acknowledges the support by the Bulgarian National Science Fund, grant No KP-06-H22/3. The authors acknowledge the provided access to the e-infrastructure of the NCHDC – part of the Bulgarian National Roadmap on RIs, with the financial support by the Grant No DO1-168/28.07.2022.

ORCID

A. Tchorbadjieff http://orcid.org/0000-0001-9322-262X
L. P. Tomov http://orcid.org/0000-0003-1902-6473
V. Velev http://orcid.org/0000-0003-0161-6993

References

[1] H. Alabdulrazzaq, M.N. Alenezi, Y. Rawajfih, B.A. Alghannam, A.A. Al-Hassan, and F.S. Al-Anzi, *On the accuracy of Arima based prediction of COVID-19 spread*, Results Phys. 27 (2021), p. 104509.
[2] L.J.S. Allen, *A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis*, Infect. Dis. Model. 2 (2017), pp. 128–142.
[3] L.J.S. Allen and G.L. Jr, *Extinction thresholds in deterministic and stochastic epidemic models*, J. Biol. Dyn. 6 (2012), pp. 590–611.
[4] M. Asadi, A. Di Crescenzo, F.A. Sajadi, and S. Spina, *A generalized Gompertz growth model with applications and related birth-death processes*, Ric. Mat. (2020), pp. 1–36.
[5] K. Athreya and P. Nay, *Branching Processes*, Springer-Verlag, Berlin, 1972.
[6] I. Auger and C.E. Lawrence, *Algorithms for the optimal identification of segment neighborhoods*, Bull. Math. Biol. 51 (1989), pp. 39–54.
[7] F. Ball, T. Britton, K.Y. Leung, and D. Sirl, *A stochastic sir network epidemic model with preventive dropping of edges*, J. Math. Biol. 78 (2019), pp. 1875–1951.
[8] B. Benjamin and J. Pollard, *The Analysis of Mortality and Other Actuarial Statistics*, Butterworth-Heinemann Ltd, Oxford, 1980.
[9] K. Brock and D. Slade, *Poisson (Simulating homogenous & non-homogenous poisson processes. r package version 1.0)*. Available at https://CRAN.R-project.org/package=poisson.
[10] J. Chen and A. Gupta, *On change point detection and estimation*, Commun. Stat. Simul. Comput. 30 (2013), pp. 665–697.
[11] G. Chowell, L. Sattenspiel, S. Bansal, and C. Viboud, *Mathematical models to characterize early epidemic growth: A review*, Phys. Life Rev. 18 (2016), pp. 66–97.
[12] M. Csorgo and L. Horváth (eds.), *Limit Theorems in Change-Point Analysis*, John Wiley & Sons Ltd, England, 1997.

[13] S.K. Das and B. Sujit, *A study on geospatially assessing the impact of COVID-19 in Maharashtra, India*, Egypt. J. Remote. Sens. Space Sci. 25 (2022), pp. 221–232.

[14] E. Dong, H. Du, and L. Gardner, *An interactive web-based dashboard to track COVID-19 in real time*, Lancet Infect. Dis. 20 (2020), pp. 533–534.

[15] E.W. Frees, *Regression Modelling with Actuarial and Financial Applications*, 1st ed., Cambridge University Press, Cambridge, 2010.

[16] T. Harris, *The Theory of Branching Processes*, 2nd ed., Springer-Verlag, Berlin, 1963.

[17] O. Hyrien and N.M. Yanev, *Branching stochastic evolutionary models of cell populations*, Biomath. Commun. 6 (2019), pp. 78–95.

[18] B. Jackson, J.D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, and T.T. Tsai, *An algorithm for optimal partitioning of data on an interval*, IEEE Signal Process. Lett. 12 (2005), pp. 105–108.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 1st ed., Springer Texts in Statistics, Springer New York, New York, 2013.

[20] R. Killick and I. Eckley, *Changepoint: An R package for changepoint analysis*, J. Stat. Softw. 58 (2014), pp. 1–19.

[21] R. Killick, P. Fearnhead, and I.A. Eckley, *Optimal detection of changepoints with a linear computational cost*, J. Am. Stat. Assoc. 107 (2012), pp. 1590–1598.

[22] N.M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A.R. Akhmetzhanov, S.M. Jung, and H. Nishiura, *Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data*, J. Clin. Med. 9 (2020), p. 538.

[23] J.O. Lloyd-Smith, S.J. Schreiber, P.E. Kopp, and W.M. Getz, *Superspreading and the effect of individual variation on disease emergence*, Nature 438 (2005), pp. 355–359.

[24] P. Mayster and A. Tchorbadjieff, *Supercritical Markov branching process with random initial condition*, C. R. Acad. Bulg. Sci. 72 (2019), pp. 21–28.

[25] L. Meyers, *Contact network epidemiology: Bond percolation applied to infectious disease prediction and control*, Bull. Am. Math. Soc. 44 (2007), pp. 63–86.

[26] K. Mitov, N. Yanev, and O. Hyrien, *Multitype branching processes with inhomogeneous poisson immigration*, Adv. Appl. Probab. 50 (2018), pp. 211–228.

[27] E.S. Page, *A test for a change in a parameter occurring at an unknown point*, Biometrika 42 (1955), pp. 523–527.

[28] E. Prompetchara, C. Ketloy, and T. Palaga, *Immune responses in COVID-19 and potential vaccines: Lessons learned from sars and mers epidemic*, Asian. Pac. J. Allergy Immunol. 38 (2020), pp. 1–9.

[29] A.J. Scott and M. Knott, *A cluster analysis method for grouping means in the analysis of variance*, Biometrics 30 (1974), pp. 507–512.

[30] B. Sevastyanov, *Branching Processes (in Russian)*, Nauka, Moscow, 1971.

[31] M. Stehlík, J. Kisel'ák, A. Dinamarca, E. Alvarado, F. Plaza, F. Medina, S. Stehlíková, J. Marek, B. Venegas, A. Gajdoš, and Y. Li, *Redacs: Regional emergency-driven adaptive cluster sampling for effective COVID-19 management*, Stoch. Anal. Appl. (2022), pp. 1–35.

[32] A. Tchorbadjieff and P. Mayster, *Models induced from critical birth-death process with random initial conditions*, J. Appl. Stat. 47 (2020), pp. 2862–2878.

[33] L. Tomov, S. Angelov, and A. Tchorbadjieff, *Age-specific mortality risk from COVID-19 in Bulgaria*, in *Computer Science and Education in Comput. Sci*, T. Zlateva and R. Goleva, eds., New Bulgarian University, Sofia, 2021, pp. 1–24.

[34] J.H. University, *COVID-19 data repository by the center for systems science and engineering (csse)*. Available at https://github.com/CSSEGISandData/COVID-19, Last seen on 23th November 2022.

[35] W.N. Venables and B.D. Ripley, *Modern Applied Statistics with S*, 4th ed., Springer, New York, NY, 2002.

[36] W. Yang*Early Warning for Infectious Disease Outbreak: Theory and Practice*, Academic Press, Elsevier Inc, London, 2017.

Appendix A. Computed parameters of Poisson distribution for initial conditions

Both type of distributions, HP and NHP, are used in this work to study of initialisation of COVID-19 infection spread. Their rates of arrival are studied by MLE. For the HPP, the daily number of infections during the period of the initial condition gives the rate λ of Poisson distribution as it is shown in Equation (2).

The rates of homogeneous Poisson distribution are computed for two different cases. The first results are obtained for the whole period from the first detected case in every country until the first change point after predefined conditions. Despite this estimate being very biased for the time periods prior last time unit (in our case days $t_{i=n}, i = 1, .., n$), the quantity $n_{t=n}$ is quite a good estimate for the cumulative number of infected at t_n.

The homogeneous Poisson rates are also computed for identified sub-period from the interval until first regime change starting from time zero until the optimal position yielded from change point tool by At Most One Change (AMOC) criteria. This new period is classified as an initial condition and notified as λ_{IC}

All values are computed by the *fitdistr* function from the packet *MASS* and shown in Table A1. The countries are ordered by a combination of geographical and political factors in initial presumption of similarities.

The columns are:

- $\lambda \pm s.d.$ shows the values of λ with their standard deviation errors for first change point data.
- $DaysCP_1$ shows how long the first change point data last.
- $\lambda_{IC} \pm s.d.$ shows the values of λ with their standard deviation errors for the IC period.
- *DaysIC* shows how long initial condition periods last.
- *Days.tot* shows the number of days for the IC period since 15 January 2020.
- Demographics: total population, density
- Location index: categorical (*factor* in *R* language) variable to separate north (label 1) and south (label 2) hemisphere.

B. Computing and statistical inference

The first computed parameter is the growth rate r_i. The used data are downloaded from the dedicated Johns Hopkins University repository [34] and it is computed for daily values x_i by

$$r_i = \frac{x_i - x_{i-1}}{x_{i-1}}. \tag{B1}$$

Then, the computations can be done after change point segments and their local probabilities are determined. The important part in accuracy improvements is the automatic calibration of the branching model. It relies on correction parameter c and the selection of its optimal value relies on the common measure as mean-squared error (MSE), given in [19] by

$$MSE = \frac{1}{n}\sum_{1}^{n}(y_i - \hat{f}(x_i))^2, \tag{B2}$$

where $f(x_i)$ is the prediction that any model f gives for the y_i observation. Usually, it is connected with accuracy parameter *Bias*

$$Bias = \sum_{i=1}^{n}\frac{\hat{f}(x_i)}{n} - \sum_{i=1}^{n}\frac{y_i}{n}$$

by the bias-variance trade-off [19].

The generalised linear model (GLM) regression is used to estimate λ_{IC} dependence on geographical variables. This model is selected from different scenarios by initially defined predictor combinations for Log-Gamma, Gamma and Log-Gaussian distributions. The optimal combination

Table A1. Maximum-likelihood fitting of Poisson distribution for the initial condition.

	Parameter							
Country	$\lambda \pm$s.d.	Days CP1 (number of)	$\lambda_{IC} \pm$s.d.	Days IC (number of)	Days.tot (number of)	Population (in millions)	Density (per m^2)	Locat (index)
North America								
US[a]	59.481±1.05	54	0.447±0.109	38	44	328	87	1
Florida	6.733±0.67	15	6.733±0.67	15	61	21.5	121	1
California	63.59±1.021	61	0.371±0.103	35	45	39.5	97.9	1
New York	994.476±6.882	21	48.692±1.935	13	60	8.2	10716	1
South Dakota	4.737±0.499	19	3.778±0.458	18	73	0.9	4.4	1
Texas	8.538±0.81	13	7.083±0.768	12	61	29	40.6	1
Canada	86.109±1.16	64	0.611±0.13	36	43	38	3.9	1
European Union								
Germany	3.343±0.309	35	0.533±0.133	30	41	83	232	1
France	125.151±1.537	53	0.529±0.125	34	42	67	116	1
Italy	6.458±0.519	24	0.143±0.082	21	36	60	201.3	1
Czechia	7.833±0.808	12	8.273±0.867	11	56	10.7	134	1
Sweden	21.976±0.723	42	0.04±0.04	25	41	10.4	25	1
Netherlands	16±1.414	8	11.714±1.294	7	49	17.4	521	1
Spain	124.571±1.722	42	0.083±0.059	24	40	47.4	94	1
Croatia	7.923±0.552	26	0.933±0.249	15	55	4.1	73	1
Bulgaria	11.643±0.912	14	9.769±0.867	13	65	7	63	1
Greece	17.421±0.958	19	1.125±0.375	8	49	10.7	82	1
Austria	222.871±2.681	31	9.357±0.818	14	54	8.9	106	1
Denmark	20.154±1.245	13	7.5±0.791	12	54	5.8	137.6	1
Rest of Europe								
United Kingdom	31.659±0.879	41	1.37±0.225	27	42	67.9	270.7	1
Russia	1.4±0.176	45	0.114±0.057	35	50	146.7	8.4	1
Turkey	103±2.93	12	60.909±2.353	11	66	83.1	105	1
Norway	58.588±1.856	17	15.769±1.101	13	54	5.4	14	1
Far East and Pacific								
Korea,South	13.531±0.65	32	1.069±0.192	29	35	51.7	507	1
Singapore	90.045±1.006	89	4.5±0.289	54	61	5.7	7804	2
Japan	4.281±0.366	32	1.273±0.241	22	28	126	334	1
Hong Kong	2.87±0.231	54	2.736±0.227	53	60	7.5	6777	2
India	70.265±1.017	68	0.094±0.054	32	46	1352	408.4	2
South America								

(*continued*)

Table A1. Continued.

	Parameter							
Country	$\lambda \pm$s.d.	Days CP1 (number of)	$\lambda_{IC} \pm$s.d.	Days IC (number of)	Days.tot (number of)	Population (in millions)	Density (per m^2)	Locat (index)
Brazil	8.882±0.723	17	3.25±0.451	16	57	210	25	2
Chile	21.143±0.869	28	1.842±0.311	19	57	17.6	24	2
Argentina	36.345±1.119	29	4.938±0.556	16	63	44.9	14.4	2
Australia and New Zealand								
Australia	29±0.707	58	0.441±0.114	34	44	25.7	3.3	2
New Zealand	7.593±0.53	27	0.444±0.157	18	61	5	19	2
Africa								
South Africa	8.286±0.769	14	4.769±0.606	13	62	59.6	42.4	2
Kenya	5.478±0.488	23	0.778±0.294	9	66	54.9	78	2
Middle East								
UAE	10.54±0.409	63	0.763±0.142	38	51	9.9	99	2
Kuwait	4.235±0.499	17	4.312±0.519	16	55	4.4	200.2	2
Israel	85.371±1.562	35	3.333±0.43	18	54	9.5	432	2

[a]Including all states.

of predictors in any model for every scenario is obtained from a step-wise procedure by *AIC* statistics, defined as follows:

$$AIC := nln(RSS/n) + 2k.$$

The preselected explanatory variables are the following columns λ_{IC}, $Data_{IC}$, *Days.tot*, *Population*, *Density*, *Locat* from Table A1.

The selection between different scenarios is based on observation for residuals' heteroscedasticity and finding optimal MSE-Bias proportion by 10-fold crossvalidation. Thus, the GLM final model relies on Log-Gamma identity link $\log(E(Y)) = X\beta$ with weights of squared population density. The final combination of predictors are the location categorical parameter and the fraction $Days.tot/DaysIC$. The are confirmed with quite high Z-value probabilities.