Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Information Sciences 574 (2021) 333–348

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Inferring statistical trends of the COVID19 pandemic from current data. Where probability meets fuzziness

Bruno Apolloni *

Department of Computer Science, via Comelico 39/41, 20135 Milano, Italy

ARTICLE INFO

Article history:
Received 12 July 2020
Received in revised form 13 January 2021
Accepted 7 June 2021
Available online 9 June 2021

Keywords:
COVID19 pandemic
Two-phase processes
Shifted-Pareto distribution
Explainable Artificial Intelligence
Statistics from non-iid samples

ABSTRACT

We introduce unprecedented tools to infer approximate evolution features of the COVID19 outbreak when these features are altered by containment measures. In this framework we present: (1) a basic tool to deal with samples that are both truncated and non independently drawn, and (2) a two-phase random variable to capture a game changer along a process evolution. To overcome these challenges we lie in an intermediate domain between probability models and fuzzy sets, still maintaining probabilistic features of the employed statistics as the reference KPI of the tools. This research uses as a benchmark the daily cumulative death numbers of COVID19 in two countries, with no any ancillary data. Numerical results show: (i) the model capability of capturing the inflection point and forecasting the end-of-infection time and related outbreak size, and (ii) the out-performance of the model inference method according to conventional indicators.

1. Introduction

Facing a dramatically imminent phenomenon like the current Covid19 pandemic, we were driven to issue forecasts on its evolution, albeit in the presence of not always clean data [7]. As part of the scientific community's massive effort to put science at an immediate service of society, we focused on epidemic situations where measures to contain the contagion produce tangible effects. This led us to use a two-phase model [5] consisting of a first trait ruled by the physics of Brownian motion followed by a second one in the realm of Lévy flights. The final goal of this modeling is to capture relevant features of the epidemic process that enable a proper protection of public health. As per usual, we identify them with: inflection date, outbreak size and end-of-epidemic date (the *target features* from now on), that we expressly address to the paths of the daily cumulated deaths. The trail to manage this model needs a set of steps as shown in Fig. 1.

In fact, we had to rely on samples of the phenomenon data that are not independent, because they are both truncated to the current date and possibly affected by correlation along the span of time. This accounts for the first two blocks in the above figure, which we get rid of in two ways: by moving from the probability framework to fuzzy set one, to manage non-iid (independently, identically distributed) samples of a pseudo-random variable; and by using the Algorithmic Inference approach [6], to devise a dependency generator under our control. Once these two steps have been accomplished, we are equipped with inference tools for identifying the mentioned two-phase process in terms of a probability distribution of a random variable which is *compatible*, in a proper acceptation, with the observed sample. This closes the trail from death

* Corresponding author.
E-mail address: apolloni@di.unimi.it

https://doi.org/10.1016/j.ins.2021.06.011
0020-0255/

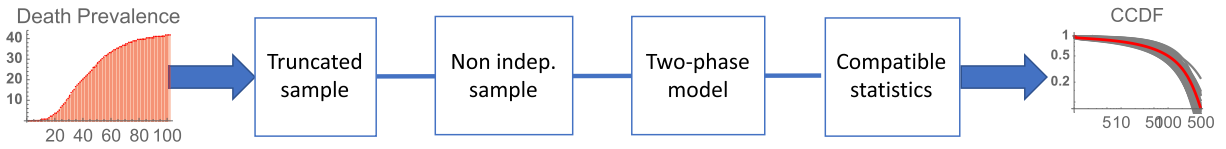

Fig. 1. Conceptual blocks of the trail from data to their model.

prevalences in input to an ensemble of Cumulative Distribution laws, represented through their complements to 1 (CCDF) in LogLog scale on the right end of Fig. 1, which are compatible with the input.

With this output we easily obtain the target features as random variables as well, from which we may compute resuming statistics such as point estimators, spread quantifiers and so on. The deriving procedure is devised to produce relatively long term forecasting, where approximately one fifth of the entire process from start to end of outbreak is used to compute the target features concerning the process as a whole. Approximation of the results will be checked when the data are available of the entire course, akin in the *first wave* of the epidemic. The value of the procedure from data to distribution laws will be compared with specific competitors' in the literature.

The paper is structured as follows. After framing our work in the literature in Section 2, in Section 3 we introduce the theoretical aspects within the statistical framework of Algorithmic Inference and with the focus on our non-iid samples. In Section 4 we recall a phase model introduced elsewhere and adapt it to the current problem. In Section 5 we implement the entire contrivance on actual Covid19 datasets and discuss the value of our results. We devote the last Section (Section 6) to a discussion of the advantages of our approach and the performance of the results in comparison with those of competitors in the literature. We also highlight some advances in the carrying out of the intermediate blocks of the trail in Fig. 1 which are both indispensable for completion of the procedure and unprecedented as for the adopted solutions.

2. Related work

The four steps in Fig. 1 address likewise research tracks calling for advanced solutions to cope with COVID19 data analytics.

Though in some cases used synonymously, what differentiates censored samples from truncated samples is knowledge of the sample size [12]. Thus, of a sample of known size m, with a censored release we may exploit a smaller number of observations because of a threshold on their values or on their number; for instance we may observe only the c smallest values. In a truncated sample we have only r observation available and we do not know how many data we missed because of truncation. While methods of inferring from censored data are well developed, especially in survival analysis [28] and insurance [15], truncated samples are dealt with mainly in specific cases [12]. Actually, the problem has been faced for a long time, with solutions found initially through momentum methods (see for instance [33]) and subsequently through maximum likelihood ones (see for instance [21]). In recent years the problem has taken the features of a machine learning procedure [31], possibly endowed with some unfeasibility lemmas [30]. The benchmark of Covid19 data we use constitutes instances of truncated data with unknown truncation and absence of *oracle* (see [30]).

Samples of non independent observations are the typical subject of longitudinal data analysis in medical data, with wide application precisely in epidemiology [26,24]. *Per se*, sample correlation heavily hampers the suitability of the computed statistics. Hence, methods such as mixed effects [35] and generalized estimating equations [34] partition the samples into subsamples which are internally independent and model the dependence among them via regression models. Despite this, we introduce a rather empirical method to actually process the sample correlation, as we will show in the next section.

Non–homogeneous stochastic processes comprise a wide chapter of mathematical statistics which lists many sophisticated methods in terms either of simulation procedures or of analytical tools, or a combination of them. Basic time descriptors in the case of epidemic processes are the number of people who are: (i) infected, (ii) hospitalized, (iii) dismissed and (iv) deceased. Of these variables we commonly follow the trends in terms of *prevalence*, i.e. the cumulated number of related individuals from the start of the epidemic up to a given day, and the *incidence*, i.e. its daily increment of the number. Rather, we may be interested in some meaningful features such as peak size and date [19] or outbreak size [38], or basic reproduction numbers such as R0 [20].

In studying these kinds of processes, a preliminary choice to make is between regression methods and modeling tools. With the former, we fit in a rather *agnostic* way the available trend to forecast its continuation in the future and possibly derive some analytical features, such as peaks and inflections from the interpolating curve. Many approaches have been specially developed for the questioned epidemic, such as GLM [18], ARIMA[9] and Neural Networks[36]. In this paper we choose the second option in the perspective of Explainable AI [1], hence with the aim of producing a model that may be understood in terms of the actual mechanism behind the pandemic, thus resulting suitable to public health stakeholders in making operational decisions.

Modeling epidemic evolution lies on the two companion threads of ordinary differential equations (ODEs), as for the deterministic approach, and Markov model/stochastic differential equations (SDEs), as for stochastic approach [3]. In a sharp way we may say that the former consider the sole expected values of the random distribution dealt with in the stochastic

approach. The ancestor is the SIR model [39], which models the variables S(t), I(t) and R(t) denoting the fractions of susceptible, infectious and removed individuals at time t, respectively. Later, many variants were developed, possibly specifically for COVID19 [23,29]. At the core we have a birth–death process that we may model on average via ODEs [14]. Their numerical implementation constitutes highly descriptive tools that are susceptible to embedding time dependence and relieving the principled homogeneity of susceptible people, but only when sufficient details of them can be evaluated. The stochastic companions are mainly used to investigate the overall structure of the process, to deduce the presence or absence of major outbreaks and related features from asymptotic state distributions. A bridge between the two approaches is represented by closed analytical forms of the probability distribution of the above variables to be fitted directly into sampled paths. Actually, while homogeneous versions of these processes are at the basis of queuing theory [11], non homogeneous versions prove to be analytically manageable mainly in elementary instances [42], with direct estimation of the involved parameters (see for instance [37]). Moving to more complex instances, we may preserve this benefit through the simple strategy of approximating the analytical solution with functions of the exponential family that are endowed with a sufficient number of parameters to make them fit the experimental companions. For instance, in [27], the case prevalence π is modeled by a Weibull function of *exposure e*where, for short, exposure is the time since the start of the epidemic, while a more complex variant of this distribution is proposed in [40]. The variable π may encompass the non homogeneity of the underlying stochastic process. In a previous work on opportunistic networks [5] we focus only on two phases and marginalize on the phase transition time distribution.

The fourth block in Fig. 1 falls within the sphere of our own approach to statistics. In place of looking for the features of the *true* distribution law (if any) of a random variable, we list features that are *compatible* with the observed sample [6]. We will discuss this at a greater length in the next section.

In this paper we look directly at an approximate distribution law of a spurious random variable. Namely, with reference to the randomness of the prevalence π, it depends on both exposure e and some kind of noise at a given e. Considering a sample of prevalences of the same local infection phenomenon, we have a monotonic relation between observed value and related exposure, so that we may hide this dependence and handle the observation as an ordered sample of the random variable Π. This is a rather spurious variable for the reasons we mentioned in the introduction, which we will deal with in the next sections within the same approximation thread of [27]. By adopting the mentioned two-phase model, we are left with a unique random variable as a reward for the many approximations we carried out in a framework that is partly rooted in probability, partly in fuzziness.

3. Theoretical bases

In this section we introduce a variant of the Algorithmic Inference tools to cope with the non-iid sample of our two-phase model.

3.1. Algorithmic Inference

We adopt a generative approach to random variables, called Algorithmic Inference [6]. In this approach, each such variable [1] X is *explained* through a sampling mechanism $\mathscr{M}_X = (Z, g_\theta)$ where a *random seed* Z translates into a random variable X via an *explaining function* $g_\theta(Z)$, for proper function g and parameter θ. Namely, for a continuous X we adopt the sampling mechanism $\mathscr{M}_X = \left(U, F^{-1}_{X_\theta}\right)$ so that

$$X = g_\theta(U) \tag{1}$$

where $g_\theta = F^{-1}_{X_\theta}$, U is the unitary continuously uniform random variable, F_{X_θ} is the Cumulative Distribution Function (CDF) of X parametrized in θ, and $F^{-1}_{X_\theta}$ is its inverse function, so that $F^{-1}_{X_\Theta}(u) = x | F_{X_\theta}(x) = u$

The main inference tool of the Algorithmic Inference is the *master equation*

$$s = s(x_1, \ldots, x_m) = \ h(\theta, u_1, \ldots, u_m) \tag{2}$$

where s is a statistic properly synthesizing an observed iid sample $\mathbf{x} = \{x_1, \ldots, x_m\}$ and h is its expression that we get by substituting x_i with the right term of (1) evaluated on the corresponding u_i. It is a sort of reverse engineering where we know $\mathbf{x}$ and want to identify θ. Since the seeds $\{u_1, \ldots, u_m\}$ are unknown, apart from their distribution law and their independence, the identification result is a probability distribution on Θ in the role of random parameter.

For instance, for a Negative Exponential distribution law with parameter λ, (1) and (2) read, respectively

$$X = \frac{-\log(1-U)}{\lambda}; \qquad s = \sum_{i=1}^{m} x_i = -\sum_{i=1}^{m} \frac{\log(1-u_i)}{\lambda} \tag{3}$$

[1] By default, capital letters (such as U, X) will denote random variables and small letters (u, x) their corresponding realizations; bold-faced characters will denote vectorial quantities.

Thus, we derive $\lambda = -\sum_{i=1}^{m} \frac{\log(1-u_i)}{s}$ and the corresponding random parameter Λ is a function of the random seeds $\{U_1, \ldots, U_m\}$. By simple analytical considerations we recognize Λ to follow the distribution Gamma with parameters (s, m).

While full details of this inference method may be found in [4], we remark that s must be properly devised. The key feature of such s is to provide the master equation with a solution in θ that is always defined and unique, for whatever seed sample (a condition that denotes it as a *well behaving* statistic).

This feature allows us to establish a bootstrap procedure to derive the parameter distribution law which we will exploit in this paper when analytical tools are not available like for the above Λ. Namely, the following pseudo-code generates a bootstrap population of the random parameter Θ from which to derive its empirical distribution law.

Algorithm 1: Generating Θ parameter population through p-bootstrap.:

1. Identify a statistic $s(\mathbf{x})$ that is well behaving for the parameter θ and its master Eq. (2);
2. repeat for a satisfactory number n_{rep} of iterations:
   (a) draw a sample $\breve{\mathbf{u}}$ of size m from U;
   (b) get $\breve{\theta}$ as a solution in θ of the master equation with seeds $\breve{u}$;
   (c) add $\breve{\theta}$ to Θ population.

3.2. Biasing the seed

The observations of the random process we will handle in the next sections constitute a spurious sample for two reasons:

1. they are truncated
2. they are sequences of data that are correlated.

Since a monotone relation exists between sampled data and their seeds, we may reverse both above defects on the seeds.

Now, truncating a sample to values less than given threshold artificially introduces a negative correlation ρ on the consecutive values of the sample (unitary delay autocorrelation). For instance, the graph in Fig. 2(a) shows the trend of this correlation with the value of the threshold τ mediated on 1000 samples of size 200. Thus point 1 may be seen as a special case of point 2.

A simple rule to produce a general family of correlated samples $\{\mathfrak{u}_1, \ldots, \mathfrak{u}_m\}$ drawn from a sample $\{u_1, \ldots, u_m\}$ of independent observation of U is the following:

$$\mathfrak{u}_i = u_i^{(u_{i-1}/h)^r} \tag{4}$$

with $r \neq 0$ and h as a free parameter. Let us call them *forcedly dependent* samples. Although the induced correlation is not uniform along the sample, for a proper r we may obtain any target value ρ of the unitary-delay-autocorrelation average on the sample, as in Fig. 2(b), with h modulating its course along the sample.

A last step toward a sampling mechanism for non independent samples lies in replacing forcedly dependent samples with independent samples drawn from U^d, let us call this variable *biased U*. On the one hand the empirical CDF of the two samples are similar for proper d and any r (see for instance Fig. 2 (c)); on the other, this allows us to implement a sampling mechanism as by definition and exploit related inference tools.

Moreover, the monotonic feature of the explaining function g in (1) guarantees that the autocorrelation sign is maintained while shifting from non independent seeds to the generated variables. Finally, the sampling mechanism does not distinguish whether the seeds have been drawn from $\mathfrak{U}$ or U^d when the related ECDF almost coincide, as in Fig. 2(c). As a whole, by sampling from U^d we have in our hands a generator of non independent samples working like the speed inverter control lever of a boat, as in Fig. 3.

It is a rather rough device which, with d playing the role of the lever degree, implements the general rule

$$\begin{aligned} &\rho < 0 \leftrightarrow r > 0 \leftrightarrow d < 1 \\ &\rho > 0 \leftrightarrow r < 0 \leftrightarrow d > 1 \end{aligned} \tag{5}$$

and asks for a proper adjustment of its angle during maneuvers, where the main feature is the monotony of d versus r and hence versus ρ. Obviously, the autocorrelation of a sample may be a more articulated function of its items, whereas (4) refers to a special family of autocorrelations. However, it works generally well both with simple random variable models and, on the contrary, with complex models that are endowed with a relatively large number of free parameters to fit the data.

3.2.1. The template algorithm

Since the correlation lever has the effect of transferring the correlation on the seeds of the sampling mechanism, which in turn remain independent, to infer θ from a non independent sample we:

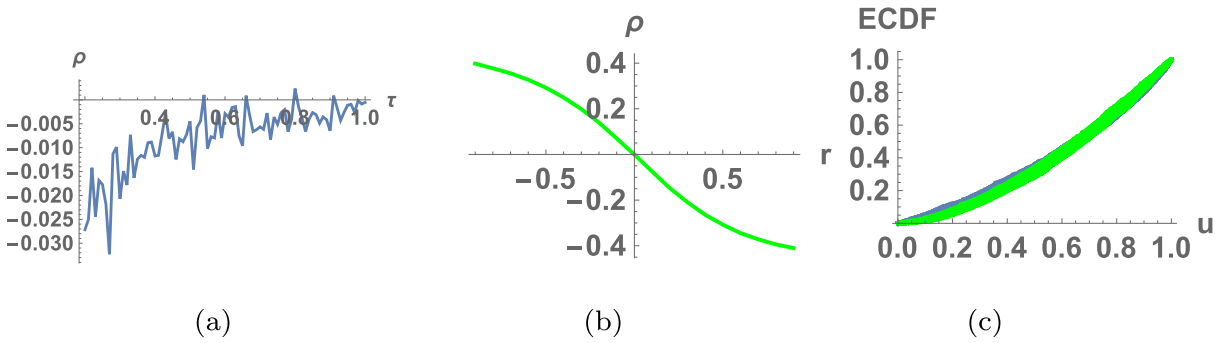

Fig. 2. Three instances of sample correlation. (a): course of the unitary delay autocorrelation ρ of a truncated sample of U with the truncating threshold τ. (b) Course of ρ of very large forcedly dependent samples of U with r as in (4). (c) ECDF of 100 forcedly correlated samples with $r = 0.5$ (blue curves) and 100 biased samples with $d = 0.6$. (green curves).

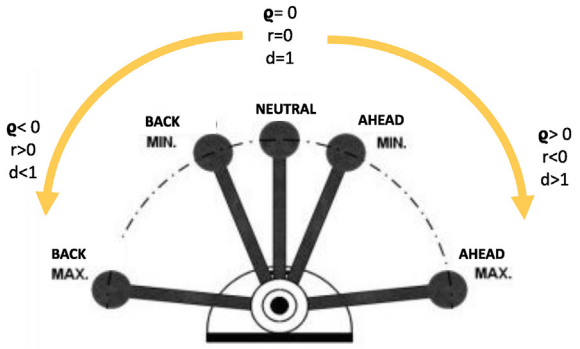

Fig. 3. The correlation lever. We show only the relation between signs of the involved parameters.

1. define a new explaining function g_θ, in place the one in (1):

$$\tilde{g}_{\theta,d}(u) = x \quad \text{s.t.} \quad F_{X_\theta}(x) = u^d \tag{6}$$

2. adopt the general strategy of leaving unchanged the general inference procedures, with the sole full substitution of u_i with u_i^d in the master Eq. (2), where d is an external parameter to be optimized by inspection methods according to a loss function L.

Hence, the new release of the lines (a), (b) of the inner loop of Algorithm 1 read as follows.

Algorithm 2 Generating parameter replicas $\tilde{\theta}$ from non independent samples.

(For the given sample $\{x_1, \ldots, x_m\}$ and well behaving statistic s)
- For d ranging in the inspection set

1.
draw the forced seed $\breve{\mathbf{u}}_d = \{\breve{u}_1^d, \ldots, \breve{u}_m^d\}$
2.
get θ^* by solving in θ the master equation

$$s = s(x_1, \ldots, x_m) = h(\theta, \breve{u}_1^d, \ldots, \breve{u}_1^d) \tag{7}$$

3.
store $par(d) = \theta^*$
4.
compute the loss function $L(d)$
- get $\breve{\theta} = par(d' = argminL(d))$

Refinements and loss function depend on the inference task.

4. A broad class of temporal processes

The third feature we want to manage of the epidemic trends is the possible non homogeneity of the underlying stochastic process. In this paper we come to the more circumscribed task of discovering a change of phase between two dynamics: an initial one that we denote as unintentional and a second that we denote as intentional. Our approach aims to capture a pseudo-equilibrium distribution of the process, i.e. a description of the observed data in terms of a law encompassing their frequencies. Hence our strategy consists of: (i) focusing on two large families of processes, respectively without memory and with memory, (ii) concatenating them in a proper way, and (iii) studying some of the main properties of the resulting process (iv) through a general purpose distribution law (v) endowed with a relatively large number of free parameters, (vi) to be identified through both well done statistics and numerical methods.

4.1. The temporal evolution

In very essential terms, we speak of memory if we have a direction along which to order the events. Now, for any *ordered* variable T, such that events on their sorted values are of interest to us, the following *master equation* holds

$$\mathrm{P}(T > t|T > k) = \mathrm{P}(T > q|T > k)\mathrm{P}(T > t|T > q) \qquad \forall k \leqslant q \leqslant t \tag{8}$$

What is generally the target of the memory divide in stochastic processes is the time $t - k$ elapsing between two events. In this perspective, the template of the memoryless phenomena descriptor is the (homogeneous) Poisson process, whose basic property is $\mathrm{P}(T > t) = \mathrm{P}(T > q)\mathrm{P}(T > t - q)$, if $t > q$. It says that if a random event (for instance a hard disk failure) did not occur before time q and you ask what will happen within time t, you must forget about this former situation (it means that the disk did not become either more robust or weaker), since your true question concerns whether or not the event will occur at a time $t - q$. Hence your local variable is $T - q$, and the above property is satisfied by the (negative) exponential distribution law with

$$\mathrm{P}(T > t) = 1 - F_T(t) = \mathrm{e}^{-\lambda t} \tag{9}$$

for constant $\lambda > 0$, since with this law (8) reads

$$\mathrm{e}^{-\lambda(t-k)} = \mathrm{e}^{-\lambda(q-k)}\mathrm{e}^{-\lambda(t-q)} \tag{10}$$

On the contrary, we introduce a memory of the past (q-long) if you cannot separate $T - q$ from q. In this paper we consider very simple cases where this occurs because the time dependence entails a local variable of the form $(T/q)^{\beta}$. The simplest solution of (8) is represented by

$$\mathrm{P}(T > t|T > k) = (t/k)^{-\alpha} \tag{11}$$

so that the master equation reads

$$(t/k)^{-\alpha} = (t/q)^{-\alpha}(q/k)^{-\alpha} \tag{12}$$

Note that this distribution, commonly called Pareto distribution, is defined only for $t \geqslant k$, with $k > 0$ denoting the true time origin, where α identifies the distribution with the scale of its logarithm. The main difference w.r.t. the exponential distribution is highlighted by the LogLogPlots of their CCDF (complement to 1 of CDF), denoted as $\overline{F}_T$ in Fig. 4: a line segment with a Pareto curve (see picture (a)) in contrast to a more than linearly decreasing curve with the exponential distribution (Fig. 4(b)). A first operational consequence is that, for the same mean value of the variable, we may expect its occurrence in a more delayed time if we maintain memory of it as a target to be achieved (getting a Pareto distribution), rather than relying on chance (getting an exponential distribution).

4.2. The spatial evolution

We relate the above temporal evolution to companion space evolutions represented by Brownian Motion and Lévy flights, respectively. In fact, it is well known that at sufficiently low densities the distribution of times between successive collisions of a molecule in a fluid is approximately exponential, while its trajectory follows a Brownian motion [10]. Analogously, experimental studies show a Pareto distribution reckoning the time intervals between changes in Lévy flight direction that we see in nature. This occurs, for instance, with albatrosses [16] in search of food.

For Covid19 pandemic, we consider the following *wait and chase* model that concatenates the two dynamics. It is iconically described by the course of a dodgem car at a moon park but may be applied to virus activity as well.

Assume you are playing with dodgem cars. You drive around until, from time to time, you decide to bang into a given car which is unaware of your intent. Thus, initial bumps occur by chance, as with molecules in a gas, while the second kind of bumps are intentional, i.e. directed toward the target, albeit disturbed by occasional diversions from the impact trajectory. We may assume the trajectory of each car to be a plane Brownian motion before the triggering of the chase. The change of dynamics during the chase derives from the fact that one dimension of your car motion is now represented exactly by the

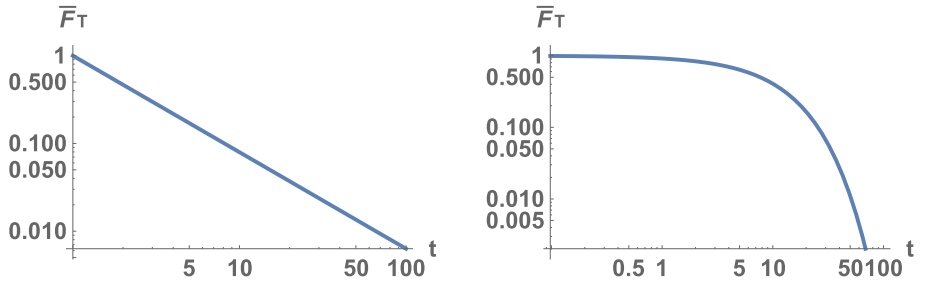

Fig. 4. CCDF LogLogPlot when T follows: Left → a Pareto law with $\alpha = 1.1$ and $k = 1$; Right → an exponential law with $\lambda = 0.09$. Parameters are chosen to have the same mean.

line connecting your car and the target car. In a previous paper [5] we show that the distribution of the random time T elapsed between two subsequent bumps of the dodgem cars (the intercontact time) has the following expression:

$$F_T(t) = 1 - \frac{b+1}{b + (t/c + 1)^a} I_{[0,\infty)}(t) \tag{13}$$

with $a, c > 0$ and $b \geqslant 0$, whose template shape is reported in Fig. 5 in terms of both $F_T(t)$ in normal scale (see picture (a)), and $\overline{F}_T(t) = 1 - F_T(t)$ in LogLog scale (see picture (b)), thus by representing both abscissas and ordinates in logarithmic scale. We call it a *shifted-Pareto* distribution. At a first glance we may identify in Fig. 5(b) an initial nonlinear plateau which we figure to be drawn mainly from the Brownian non intentional motion, followed by a linear part, the tail, referring to the Pareto intentional motion. We assume the vertex of the *knee* of the graph as the indicator of a game changer in the two-phase process. Analytically, its abscissa is close to $b^{1/a}$.

4.3. From inter-contact time to number of inter-contacts

Letting dodgem cars play the role of viruses, we are mainly interested in the evolution of the cumulative number $N(t)$ of contacts (as primers of death) with elapsing time t. For a large number of cars and T following the negative exponential distribution law of the first phase, at each t this variable follows a Poisson distribution law with parameter $\mu(t) = \lambda t$, where the former parameter coincides with expected value $E[N(t)]$ of $N(t)$ and λ is the constant inter-contact rate. Passing to the Pareto distribution of the second phase, the additional $N\prime(t)$ still behaves as a Poisson variable, but the related λ now decreases proportionally to $1/t$, so that $\mu(t) \propto \int_{t_0}^{t} \lambda(\tau) d\tau = cLog[t]$, for a proper c, increases less than linearly with t. Actually, recognizing this function to be the limit of the integration of $\lambda(\tau) = c/(\tau^{\epsilon})$ when ϵ goes to 1, we integrate this function to get μ_t still as a power of t.

Hiding the explicit dependence on t, at the i-th observation, N_i is a function of two random variables: the inter-contact time T and the Poisson spread of N_i around its mean $\mu(T)$. By invoking a locally loose ergodicity of the process in the thread of [32], we approximately equate $\mu(T)$ to the local mean along the temporal path so that $\mu(T)$ constitutes an interpolation $\tilde{N}$ of N.

In this way, in the first phase the CDF of $\tilde{N}$ has the same shape as the CDF of T, while in the second phase this CDF is well approximated by a Pareto distribution as well. Hence we have the same ingredients with which we build up (13) and we are enabled to use this expression as the CDF of $\tilde{N}$ along the entire process.[2] Finally, we replace $\tilde{N}$ with N in our model, so producing the unique effect of filtering minor noise from the observed n_i. This noise would be almost suppressed in the involved statistics and will manifest some ripple when theoretical curves are contrasted with experimental ones.

Fig. 6 shows a pair of simulations (in a reduced scale for obvious reasons) to get evidence of the theory. Namely, starting from T distributed according to either an exponential distribution law or a Pareto one with same mean= 2, on the one hand we drew a sample $\{t_i\}$ of size 60 and a sample of size 100 of a Poisson variable $N(t_i)$ with the mean $\mu(t_i)$ as discussed before. On the other hand we directly drew a sample $\{t_j\}$ of size 6000 and a single value n_j of the related Poisson of mean $\mu(t_j)$. In the picture we compare the ECDF of the local mean of $N(t_i)$, the one of $\{n_j\}$ and the CDF of the starting distribution with parameters estimated on the first kind of simulated data.

[2] Note that, by definition, $\tilde{N}$ is a discrete random variable that we approximate with a continuous one.

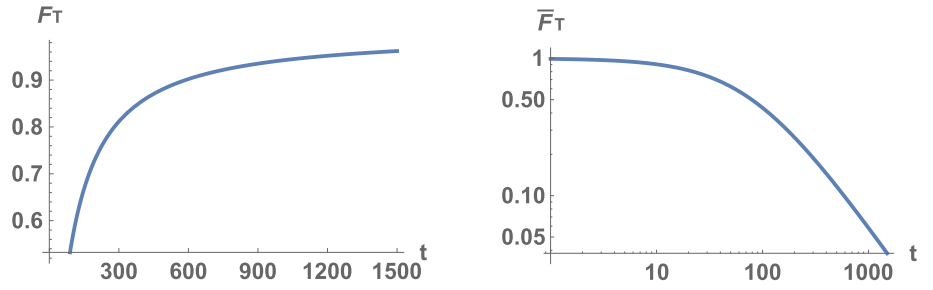

Fig. 5. Left → CDF Plot of a shifted-Pareto distribution. Right → LogLogPlot representation of its complement.

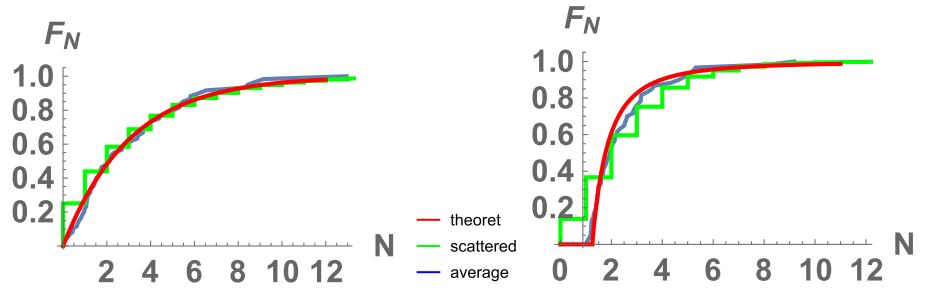

Fig. 6. $N(t)$ distribution approximations starting from: Left → Exponential and Right → Pareto inter-contact distribution laws with same average. Blue line → $ECDF(E[N(t)])$, green line → $ECDF(N)$, red line → $CDF(E[T])$.

4.4. Inferring the parameters

We may consider our estimation problem in terms of drawing a regression curve through the set of pairs $\left(n_i, \hat{\overline{F}}_N(n_i)\right)$, coupling the observed prevalences with the ECCDF computed on it.[3] According to our model, the regression curve depends on three parameters: a, b, c of (13) that we want to estimate with the Algorithmic Inference tools. This requires to establish:

1. *Sampling mechanism.* As stated in the previous section the expression of $F_N(n)$ is the same of (13) with $T$and t replaced by N and n, respectively. Hence, by solving the equation $F_N(n) = u$ in n, the sampling mechanism reads:

$$n = F_N^{-1}(u) = g_{a,b,c}(u) = c\left(\left(\frac{bu+1}{1-u}\right)^{\frac{1}{a}} - 1\right) \tag{14}$$

2. *Relevant statistics.* Denoting by $n_{(i)}$ the i-th element of the sorted sample of N and by med the quantity $\lfloor (m+1)/2 \rfloor$, we adopt the following statistics

$$s_1 = n_{(\text{med})}; \quad s_2 = \frac{1}{m}\sum_{i=1}^{m} n_i - s_1; \quad s_3 = \sum_{i=\text{med}}^{m} \log n_{(i)} \tag{15}$$

They almost completely fulfill the well beahvingness requests. Indeed, thanks to the explaining function in (14), we obtain the master equations

$$s_1 = g_{a,b,c}\left(u_{(\text{med})}\right) \tag{16}$$

$$s_2 = \frac{1}{m}\sum_{i=1}^{m} g_{a,b,c}(u_i) - g_{a,b,c}\left(u_{(\text{med})}\right) \tag{17}$$

$$s_3 = \frac{m}{2}\log c + \frac{1}{a}\sum_{i=\text{med}}^{m} \log\left(\frac{bu_i+1}{1-u_i}\right) \tag{18}$$

[3] ECCDF is the complement to 1 of ECDF.

Since $(a, b, c) \geqslant 0$, from these expressions we see that both s_1 and s_3 are monotonic in the three parameters, thus allowing for unique solutions for whatever seed $\{u_1, \ldots, u_m\}$. s_2 dependences are less univocal, since singularly the two addends denote the same monotonic trends with parameters, but their difference may give rise to non unique solutions.

We solve these master equations in the parameters in correspondence to a large set of randomly drawn seeds $\{u_1, \ldots, u_m\}$. In this way we obtain a sample of fitting curves, as in Fig. 7, which we statistically interpret to be compatible with the observed data. In the figure we also report the 0.90 confidence region for these curves, that we obtain through a standard *peeling* method [4].

4.5. Introducing the correlation lever

To complete the procedure we introduce the *correlation lever* in Section 3.2.

This accounts for fitting the data ECDF with the function:

$$\breve{F}_N(n) = \left(1 - \frac{b+1}{b + (n/c + 1)^a}\right)^{1/d} I_{[0,\infty)}(n) \tag{19}$$

With known d nothing changes on the above statistical procedures, apart from the new seed generation, since the explaining function now reads

$$n = c\left(\left(\frac{bu^d + 1}{1 - u^d}\right)^{\frac{1}{a}} - 1\right) \tag{20}$$

This requires replacing u_i with u_i^d in the formulas (16–18).

5. Numerical results

The latest advances in this research have been fostered by the aim of exploiting data on the current Covid19 pandemic. Statistics on pandemic evolution are relatively abundant, but their quality is not always satisfactory. This stems from the absence of shared protocols with which they are collected and from some *political* biases as well. Therefore we decided to narrow the focus solely on the numbers of deaths, which, while systematically underestimated in the official data, are less affected by variations in data collection procedures. To recall this option, henceforth we will refer to an *infection*, rather than an epidemic, as the phenomenon generating these numbers. No ancillary data have been formally taken into account; rather, we considered two datasets that proved quite familiar to us:

1. deaths from COVID19 in 13 regions of Italy in the period February-December 2020, by obvious reasons given the author's nationality.
2. deaths in Senegal in the same period, due to a special connection with epidemiologists in that country.[4]

5.1. The Italy data

We considered the daily cumulative deaths (prevalence) referring to the 13 most infected Regions in Italy, as they have been reported from the start on GitHub [22], with the miss of Sicily that initially resulted less hit. Table 1 lists the names of these regions, jointly with the related prevalence at the end of the observation period. In Fig. 8-left we report the deriving incidence course in those regions. The gray sections identify the two infection waves that characterized the phenomenon up to the end of year 2020 (but a third wave is expected in 2021).

In greater detail, in Fig. 8-right we report the prevalence ECDF of the first 100 days of infection in each region (first wave). Actually, to gain comparability, we focused on the 100 days since the first one with incidence greater than 0 in the individual regions and rescaled the abscissas to $0 - 100$ so that the graphs have the same starting point and ending point. Moreover, we distinguished two traits, drawn respectively with continuous and dashed lines. The latter begins approximately on the inflection point of the ECDF,[5] thus denoting the start of a further phenomenon not covered by our two-phase model. Hence, we take statistics to estimate the parameters of the shifted-Pareto distribution from the first trait. In addition, we provide an early model for the second trait, still exploiting the tools introduced in Section 3. A similar analysis has been done on the data of the second wave as well. Numerical experiments will focus mainly on the first trait of the first wave, as they are more exploitable. However, numerical considerations will be carried out also on the remaining data for comparison sake.

5.1.1. Compatible statistics

The core of our inference is the estimation of the CDF of the random variable representing the death prevalence along the infection time. From this distribution we will derive the target features: inflection time, end-of-infection time and outbreak

[4] Senegalese Institute of Agricultural Research, Dakar.

[5] nothing to do with the target feature "inflection time".

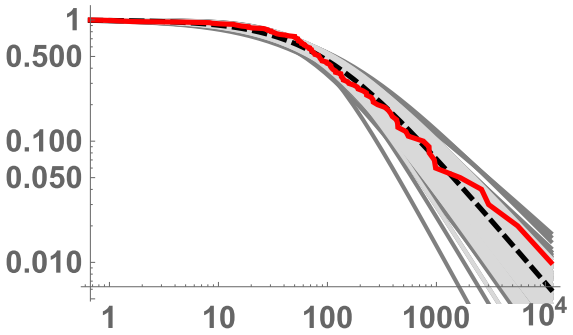

Fig. 7. Curves fitting with parameters compatible with a sample generated by the distribution law (13). Red curve: sample ECCDF; gray curves: 200 population replicas; thick dashed curves: median of the replicas. Light gray region: 0.90 confidence region.

size. Therefore, for each region we draw up to 200 compatible CDFs as a result of likewise estimates of parameters a, b, c, d of our model on the basis of the first traits of the first wave of the above prevalence data. The introduction of d is due to the longitudinal features of the observed data, which induce a positive autocorrelation, and to their truncation, which induces a negative autocorrelation by contrast. Hence the two effects require opposite actions on the correlation lever, i.e. opposite ds, whose prevailing will emerge from the data when we use the explaining function (20). Those values are computed according to Algorithm 2, with a loss function based on a proper distance between experimental CDF and the computed one. As a whole, replicas of the four parameter estimates are computed with the mentioned nesting of Algorithm 2 into Algorithm 1; meanwhile some replicas are discarded because they do not pass a consistency check (due to a bad solution of the Eqs. (16)–(18)).

In the pictures of Fig. 9 we reported these curves in gray for three template regions. In the same pictures we reported in thick-black the curves parametrized with the median of the parameters of the gray curves. Finally, the thick red lines show the related ECCDFs.

From left to right the pictures refer to regions whose data are more to less compliant with our model. In particular Lombardia, jointly with Piemonte and Toscana, are affected by anomalous epidemic trends that are understood only in part by scientists. Though gray curves include the red ones, latter curves show a greater slope toward the end as a consequence of the sample truncation effect.

The pivot of these curves is the mentioned knee vertex of the distribution. We call it the *turning point* and assign it exactly the abscissa $b^{\frac{1}{a}}$; for the median parameters we drew a vertical blue line with this abscissa in the above graphs. This line is variously positioned in those graphs because the data have different scales and are indexed since day one with incidence greater than 0. However the corresponding calendar day for each region is around March 31, 2020. This confirms to us that we may equate this vertex with the phase changer of our process. In fact, if we take into account the contagion latency and the fatal illness elapse time, the changer may represent the outcome of the severe restriction measures issued by the Italian Government on March 8, 2020.

From the turning point we derive the target features as follows:

- Inflection point. We equate it to the turning point.
- Outbreak size. We exploit the linearity of the LogLogPlot of the Pareto CCDF to derive this value at the crossing of the linear interpolation of the CCDF after the turning point in the above representation and a horizontal threshold at the height $(2/outbreaksize)$
- End-of-infection. We compute it as the day on which the outbreak size is reached. It entails that we expect 2 further deaths after this time. It is a rather loose condition for epidemiologists, who generally wait for a certain number of days without new cases to declare the end of the infection, but it is close enough.

Table 2 details the values for all regions and related statistics. In particular, the first column reports the numbers of the data that have been processed to compute these values. They coincide with the lengths of the first traits in Fig. 8 and amount to around one fifth of the end-of infection. This confirms the long term forecasting capability of our procedure. Moreover, we

Table 1
The cumulative Covid19 deaths in 13 Italian regions at day 298 (December 20, 2020).

day	Emilia-R.	Liguria	Lombardia	Marche	Piemonte	Toscana	Veneto
298	6875	2724	24165	1446	7382	3348	5161
day	Abruzzo	Campania	Friuli VG	Lazio	Puglia	PA.Trento	
298	1102	2472	1366	3156	2100	821	

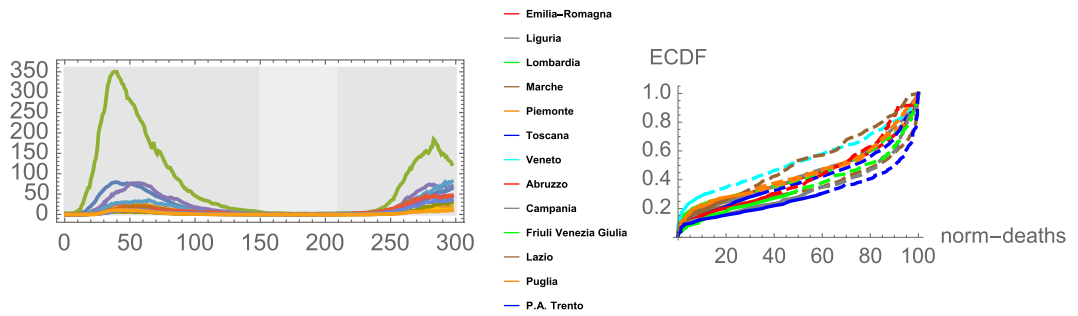

Fig. 8. A sketch of Italian data. Left: two waves shapes of the death incidence by region. Right: ensemble of normalized first wave cumulated deaths by region; dashed line → unprocessed trait.

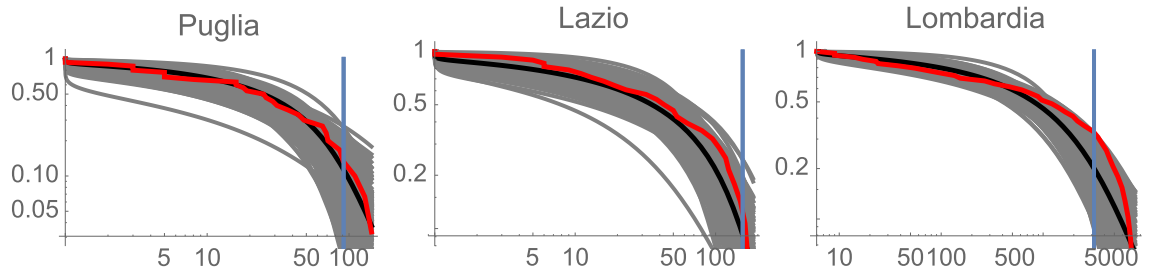

Fig. 9. A synopsis of the estimation results in three template regions. Gray curves → compatible CCDFs, black curves → their medians, red curves → the ECCDFs, blue lines → turning point verticals.

Table 2
Main statistics of death data collected on 13 Italian regions. modeled → days used to compute the statistics; tp → turning point; end-exper → end-of-infection day estimate based on ECCDF; end-med,end-dev → median of end-of-infection day estimates based on compatible CDFs and related mediandeviation; size-med, size-dev → corresponding outbreak size statistics.

Anl day 4/4/2020	modeled	tp	end-exper	end-med	size-med	end-dev	size-dev
Emilia-Romagna	37	31	139.301	158.333	4288.09	4.64919	3.63636
Liguria	31	25	135.92	127.625	1557.27	12.9884	5.54545
Lombardia	39	27	110.067	124.15	16612.6	7.10663	88.3636
Marche	32	24	120.833	125.596	989.818	7.71144	2.72727
Piemonte	29	24	115.943	115.772	4025.68	9.32436	60.2273
Toscana	25	21	121.492	97.852	1037.82	6.52637	28.
Veneto	39	36	*	184.4	2110.27	16.2727	26.3636
Abruzzo	24	21	126.523	124.479	461.364	10.3945	2.63636
Campania	23	21	*	107.536	427.	7.23774	5.09091
Friuli Venezia G.	26	24	*	134.834	345.	6.57265	0.
Lazio	28	26	*	150.787	854.636	12.8452	9.63636
Puglia	30	28	*	112.675	532.909	2.48308	2.77273
P.A. Trento	22	19	109.027	110.789	464.091	4.07883	0.272727

report two estimates of the end-of-infection day; both are routed on the turning point, but the extrapolation is based either on the ECCDF or on the CDFs. The former may prove unfeasible in some cases (denoted by a star in the table) because of too few data after the turning point.

Fig. 10 shows the trend of the two estimates end-of-infection day contrasted with the *true* values for death threshold equal to either 1 or 2. Actually, we get these values at the crossing on the right of the corresponding threshold lines with a smoothed plot of the incidence course. The latter have been derived by an exponential smoothing (with coefficient 0.1) of the discrete values of the daily incidence, with a consequent delay in the interpolated values. Moreover, the outbreak condition we adopted is not related to point values but to the integration of the incidence on the tail (CCDF). We logically compensate for these drifts by considering the two thresholds. The lower pictures detail the positions of the target features in the prevalence and incidence graphs of the three template regions. Actually, the inflection point proves to be a bit questionable, as it depends on the smoothing method we adopt for the prevalence graphs[6] and the policy we choose for its location in the

[6] Recall that we replace $\tilde{N}$ with N in Section 4.3.

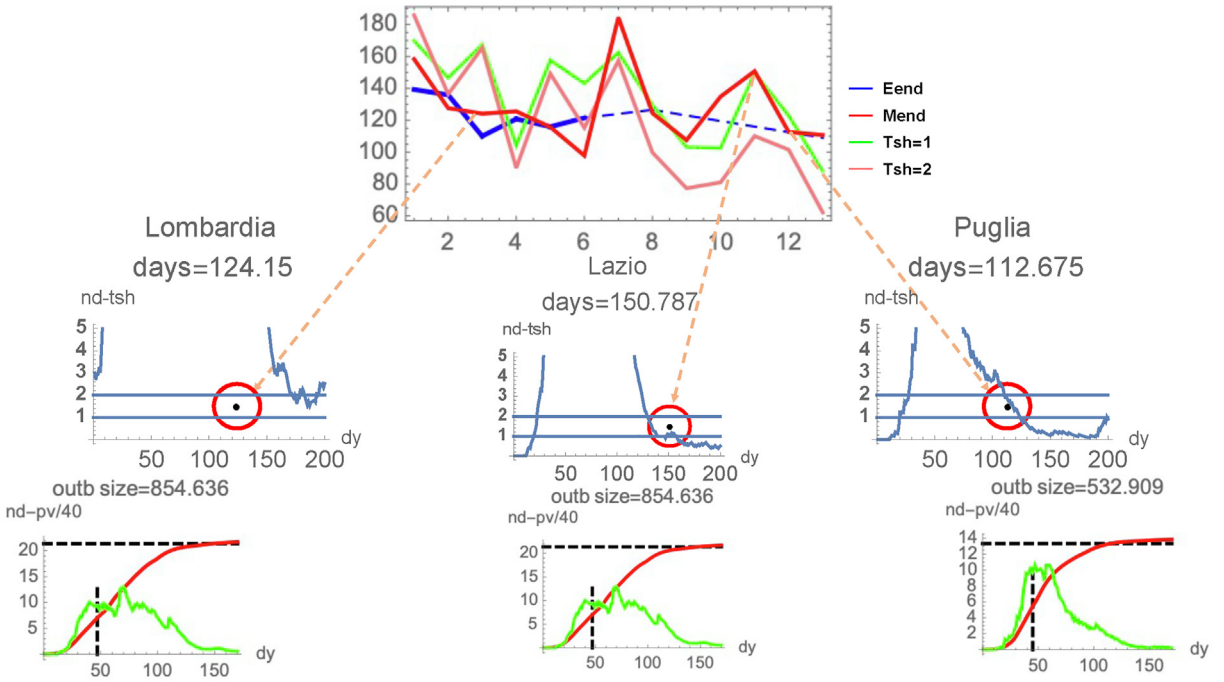

Fig. 10. Contrasting observed with estimated parameters. Upper picture shows the course of end-of-infection time along the regions: Eend→ ECCDF based estimate; Mend → median estimate deriving from compatible CDFs; Tsh = i → crossing of the smoothed incidence with the threshold line with height = i; dashed line corresponds to stars in Table 2. Second lane: blue curve → incidence (nd) graph; point→ (abscissa (dy) = Mend, ordinate=1.5); threshold lines as above. Third lane: red curve → prevalence/40 (pv/40); green curve → smoothed incidence; horizontal dashed line → outbreak size; vertical dashed line → inflection time.

almost rectilinear segment separating the opposite curvatures. According to what discussed, for purposes of representation we locate the end-of-infection point at height= 1.5. Overall, the three regions denote the same model compliance graduation mentioned in regard to Fig. 9.

As a whole, our forecasting achieves root-mean-square-errors on the target features as in Table 3. In spite of the long term forecasting and the absence of any ancillary data, they are on the order of 10%, 3% and 10% of the mean of the predicted value, respectively.

5.1.2. After the first wave

In early September, a second COVID19 wave hit Italy. We repeated the above procedure on the new data from September to December 2020 and found that the two-phase process works relatively well for some regions but with others suffers from some well known anomalies. They are partly due to the fact that the regions managed the second emergence individually, possibly with scarce coordination with the central government. Moreover, in some regions, like Lombardia, new epidemic phenomena arose that are not yet fully explained. Of the remaining template regions, in Fig. 11 we see that the forecasted end-of-infection time makes sense, though far from the observed days, for Puglia and no sense for Lazio. Actually, Puglia seems to follow the trend of the first wave, with a phase changer on October 4th as a result of the reintroduction of the obligatory mask use and other social distancing measures in early September. Lazio took a different path, for instance making face mask use obligatory only on October 2th and achieving a game changer date of October 27th.

5.1.3. The tail data

The second traits of the curves in Fig. 8(b) are not covered by the two-phase model. Rather, they denote a faster decrease of the incidence than found with the Pareto trait. Namely, by analyzing with the same approach the ECCDF of these data we get an incidence R that we approximately handle with a CDF

$$F_R(r) = Exp\left[-\frac{\lambda}{r-a}\right] \tag{21}$$

Albeit it is a very early modeling, with this distribution we recover with an acceptable approximation the ECDF of the death incidence in the various regions, as shown in Fig. 12 with shifts on the right end of the graphs as a consequence of infinite mean of distribution law (21). This trend helps on the one hand to explain the shift between fore-casted and actual

Table 3
The Root mean square error of the target features forecasting computed on: first row → all the regions; second row → all but Lombardia, Piemonte and Toscana.

RMSE	Inflection	Size	End-time
All	2.60177	68.8515	17.7962
Top ten	2.77489	25.9913	10.9317

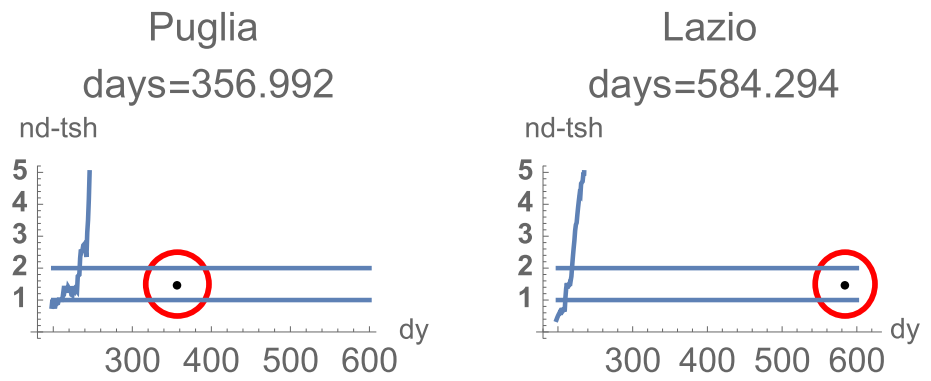

Fig. 11. Daily number of deaths and end-of-infection estimate in two template regions, second wave. Notation as in Fig. 10.

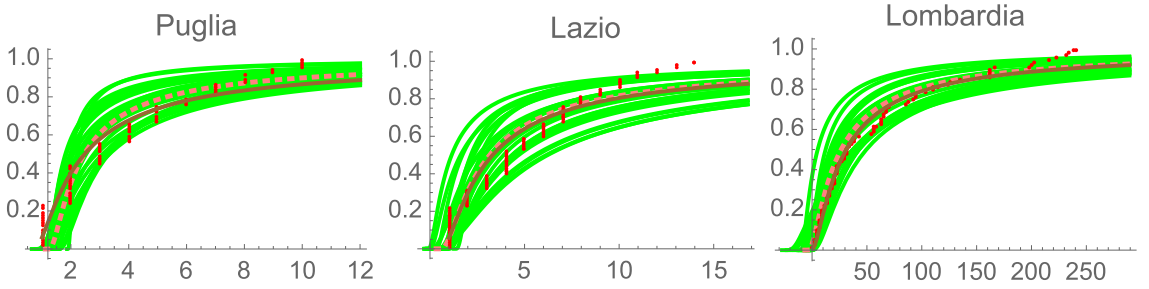

Fig. 12. Daily number of deaths in the tails ending the first wave. Green curves → compatible CDFs; dashed pink curve → their median; brown curve → ML estimator; red points → ECDF.

end-of-infection times; on the other hand, it highlights a rarefaction phenomenon where, besides their reduced number, inter-contacts possibly occur when either the virus has partly exhausted its infection power and/or the disease treatment has improved.

The tail of the second wave is less handy, possibly representing the start of a third infection wave.

5.2. The Senegal data

Table 4 is analogous to Table 2 for Senegal. Actually, the data concern mainly the capital Dakar.

Figs. 13 sums up for this dataset the main features taken into consideration for the case of Italy.

In particular, contrasting them with the Puglia region, where the order of magnitude is double as for both incidence and prevalence, we see that Senegal data denote a slower evolution in terms of traits switch, phase turning point and end-of-infection. In reality, we have no clear control of the Senegal data [8], starting from the percentage of ascertained numbers of covid-death with respect to the actual ones. However, this difference in infection dynamics could derive in part from demographic factors (one third of population density in Senegal w.r.t. Puglia, but an overall population four times larger) and in part from health management reasons (during the infection period the Senegal Government adopted distinctly milder and less controllable restriction measures than those that were put in place in Italy). Apart the vagueness of these considerations, from the picture we may determine that the two-phase model appears to be sufficiently suitable also in this instance.

6. Discussion

Forecasting Covid19 evolution is a formidable challenge as for both a profitably immediate exploitation of the results it produces and the theoretical problems it poses. Within the current trend of Explainable Artificial Intelligence, we decided to

Table 4
Main statistics on the death data collected on 13 Italian Regions.

Anl day 22/07/2020	modeled	tp	end exper	end-med	size-med	end-dev	size-dev
Senegal	105	89	166.026	151.908	282	8.56414	13

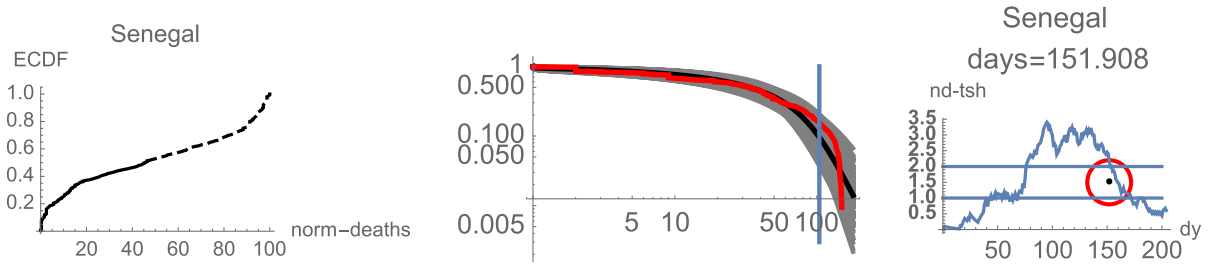

Fig. 13. Senegal data graphical synthesis. Same notation as Figs. 8–10, respectively.

face this challenge with the objective of providing a model that can be understandable from an epidemic perspective, hence exploitable for operational considerations. The two-phase model we propose is not a universal one; rather, it is applicable to epidemic situations where effective measures for the containment of the contagion are introduced. Albeit we implemented it under the constraint of no using ancillary data, we believe we have achieved our objective, getting long term forecasts as for outbreak size, inflection point and end-of-infection time which are accurate for most regions, while identifying some *problematic* ones whose evolution is scarcely covered by our model. We avoid confronting our results with those of the various *agnostic* regression methods, such those mentioned in Section 2, because the underlying models result non interpretable and generally based on a great number of parameters (e.g. the neural network connection weights), with the consequence of producing good short term forecast of the outbreak curves (about 20 days) but no direct evaluation of the above target features. Rather, among the various modeling proposals [17,40], an objective analogous to ours has been pursued in [43], where a non linear regression model with meaningful parameters has been employed. From the inferred parameters the authors compute the inflection point and the outbreak size on Italy data with an accuracy higher than ours; however, they are based on fitted data (with no prediction), exploit data of other countries through mixed effect model and refer to the pandemic in the country as a whole (not to the individual regions). Our approach is an unprecedented one, since we model a univariate pseudo-distribution law of the death prevalence, whose parameters are related to a physical model ruling the contagion process along two specific phases. In this, we too come to a regression problem which concerns, however, the shape of the CDF of this distribution. Identifying it leads, as a fringe benefit, to the estimation of the target features.

Hence we toss the methodological value of our approach exactly on this regression task, having as a competitor two sets of distributions: LogNormal [41], Weibull [27] and Extended Weibull [40], which have been used in the literature for this task, and Pareto, Tapered Pareto [25] and Truncated Pareto [5], which are in the same family as ours. To compare their performances, we use three well-known statistics, namely, the Maximum Likelihood (ML), the Akaike Criterion and the Cramer-von-Mises (CvM) test. In very essential terms, ML expresses the fitness of the distribution in terms of the product of the density functions of the recorded prevalences optimized versus the free parameters of the distribution. The Akaike Criterion is a balanced sum of the above fitness and a second term penalizing the distribution complexity [2]. The CvM test bases the acceptance of a given distribution law as the source of the observed data (the null hypothesis) on the mean square difference between the data ECDF and the hypothesized distribution CDF [13]. We progressively compute these statistics on the first wave of the 13 regions, the same plus the performances in the second wave, and the same with the addition of the Senegal data. Thus in the cells of Table 5 we report the fraction of instances on which the distribution in the row proved to be the winner with respect to the first two statistics in the column and the fraction of instances not rejected by the CvM test, with significance level $= 0.05$, in the third one. Though the additional benefits of the Algorithmic Inference approach [6] are not evidenced by these conventional criteria, the ML column in the first window clearly indicates the outperformance of our method, with Truncated Pareto as the sole competitor. Moreover, the winning fraction of our method becomes 1 if we remove the three anomalous regions (Lombardia, Piemonte and Toscana) from the reckoning. The prevailing score is smoothed in the second column, since the AKAIKE Criterion penalizes the higher number (4 in place of 3) of free parameters of our distribution. Finally, the CvM test highlights the inadequacy of the simplest models (Exponential and Pareto), and NFE-Weibull too, in coping with the complex process under examination. This trend is reversed along the other windows, clearly denoting the selectivity of our model which is not a general purpose one. Namely, while Senegal data are well covered by our method, as mentioned in Section 5.1.2, the second epidemic wave does not fully comply with our two-phase process. This is particularly evident in some regions, while in others it is denoted in any case by some drifts. This turns ML and AKAIKE criteria in favor of the Truncated Pareto in most cases, as a distribution that is more adaptable in the absence of a particular underlying model. The CvM path remains substantially unvaried along the windows.

Table 5
Comparing the performance of our parameter estimates with those of the competitors listed in the first column. Contrasting statistics: ML → Maximum Likelihood, AKAI → Akaike Criterion, CsM → Cramer von Mises test. Window headings → datasets.

	Wave 1			Waves 1 + 2			Waves 1 + 2+Senegal		
	ML	AKAI	CvM	ML	AKAI	CvM	ML	AKAI	CvM
Shift-Pareto	0.769	0.538	0.846	0.423	0.307	0.846	0.444	0.333	0.851
LogNormal	0.	0.	0.769	0.	0.	0.846	0.	0.	0.814
Weibull	0.	0.	1.	0.038	0.038	0.961	0.037	0.037	0.962
NFEWeibull	0.	0.	0.692	0.	0.	0.576	0.	0.	0.555
Exponential	0.	0.	0.692	0.	0.115	0.538	0.	0.111	0.555
Pareto	0.	0.	0.076	0.	0.	0.038	0.	0.	0.037
Tapr Pareto	0.	0.	1.	0.038	0.	0.769	0.037	0.	0.740
Trunct Pareto	0.230	0.461	1.	0.5	0.538	0.923	0.481	0.518	0.925

Besides these operational achievements, this paper introduces an unprecedented way of dealing with non-iid-samples. The iconic image in Fig. 3 recalls this novelty and its limits as well. Moving our imaginary lever ahead we introduce a positive autocorrelation on the data generated through the universal sampling mechanism, and a symmetric effect in the back way. They are specific effects that we analyze without issuing general theorems, albeit framing the analysis in a robust theoretical framework. Rather we propose them as an approximation tool to capture trends of complex processes, as a primer for further investigations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.
[2] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723.
[3] L.J.S. Allen, A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis, Infectious Disease Modelling 2 (2) (May 2017) 128–142.
[4] B. Apolloni, S. Bassis, D. Malchiodi, and P. Witold. The Puzzle of Granular Computing, volume 138 of Studies in Computational Intelligence. Springer Verlag, 2008..
[5] B. Apolloni, S. Bassis, E. Pagani, G.P. Rossi, L. Valerio, Mobility timing for agent communities, a cue for advanced connectionist systems, IEEE Trans. Neural Networks 22 (2011) 2032–2049.
[6] B. Apolloni, D. Malchiodi, S. Gaito, Algorithmic Inference in Machine Learning, 2nd edition., Advanced Knowledge International, Magill, Adelaide, 2006.
[7] A. Ashofteh, J.M. Bravo, A study on the quality of novel coronavirus (covid-19) official datasets, Stat. J. IAOS 36 (2) (2020).
[8] A.M. Babacar, M.N. Mouhamadou, T. Balde, and S. Diaraf. Visualization and machine learning for forecasting of covid-19 in senegal, 2020..
[9] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, M. Ciccozzi, Application of the arima model on the covid-2019 epidemic dataset, Data in brief (2020).
[10] R.J.C. Brown, The collision rate in a dilute classical gas, Can. J. Chem. 44 (1966) 1421–1426.
[11] K. Chen, Introduction to the Queueing Theory, chapter, 7, John Wiley & Sons Ltd, 2015, pp. 129–140.
[12] A.C. Cohen, Truncated and Censored Samples, CRC Press, Boca Raton, 1991.
[13] S. Csorgo, J.J. Faraway, The exact and asymptotic distributions of Cramer-von Mises statistics, Journal of the Royal Statistical Society. Series B (Methodological) 58 (1) (1996) 221–234.
[14] F.A. da Costa Carvalho Chalub and M.O. Souza. From discrete to continuous evolution models: A unifying approach to drift-diffusion and replicator dynamics. Theoretical Population Biology, 76(4):268–277, 12 2009..
[15] M. Denui, O. Purcaru, I.V. Keilegom, Patterns of neuronal migration in the embryonic cortex, J. Actuarial Practice 13 (2006) 5–329.
[16] A.M. Edwards et al, Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer, Nature 449 (2007) 1044–1049.
[17] Al-Ekram E.H., N. Mohammad, N.U. Emon, I.H. Tipo, AS.M. Safayet, A. andAbdullah, and M.S. Islam. Forecasting covid-19 dynamics and endpoint in bangladesh: A data-driven approach. medRxiv, 2020..
[18] C.P. Farrington, M.N. Kanaan, N.J. Gay, Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data, J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 50 (3) (2001) 251–292.
[19] Z. Feng. Final and peak epidemic sizes for seir models with quarantine and isolation, 2007..
[20] L. Ferretti, C. Wymant, and M. Kendall. Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing, 2020..
[21] A. Fisher, Estimation of the parameters in a truncated normal distribution, Math. Tables 1 (1931) 815–852.
[22] Github. https://raw.githubusercontent. com/pcm-dpc/COVID-19/master/dati-regioni/dpc-covid19-ita-regioni.csv..
[23] N. Hens, Modeling infectious disease parameters based on serological and social contact data: A modern statistical perspective, Springer, New York, NY, 2012.
[24] J.A. Hanley, A. Negassa, M.D. Edwardes, J.E. Forrester, Statistical analysis of correlated data using generalized estimating equations: an orientation, Am. J. Epidemiol. 4 (157) (2003) 364–375.
[25] Y.Y. Kagan, F. Schoenberg, Estimation of the upper cutoff parameter for the tapered Pareto distribution, J. Appl. Probability 38 (2001) 158–175.
[26] R.A. Kaslow, The multicenter aids cohort study: rationale, organization, and selected characteristics of the participants, Am. J. Epidemiol. 2 (162) (1987) 161–169.
[27] N. Keiding, Age-specific incidence and prevalence: A statistical perspective, J. R. Stat. Soc. Ser. A (Stat. Soc.) 154 (3) (1991) 371–412.
[28] D.G. Kleinbaum, M. Klein, Introduction to Survival Analysis, Springer, New York, New York, NY, 2012, pp. 1–54.
[29] N.L. Komarova, D. Wodarz. Modeling the dynamics of covid19 spread during and after social distancing: interpreting prolonged infection plateaus. medRxiv, 2020..
[30] V. Kontonis, C. Tzamos, M. Zampetakis, in: Efficient truncated statistics with unknown truncation. In 60th FOCS, IEEE Computer Society, 2019, pp. 1578–1595.

[31] M.H. Laxman, C.D. Ram, Estimation of the parameters in a truncated normal distribution, Published online 2 (162) (2007) 4177–4195.
[32] J.Y. Le Boudec, Understanding the simulation of mobility models with palm calculus, Perform. Evaluation 64 (2) (2007) 126–147.
[33] A. Lee, Table of the gaussian -tail+ functions; when the -tail+ is larger than the body, Biometrika 23 (10) (1914) 208–214.
[34] K.L. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1) (1986) 13–22.
[35] R.A. McLean, W.L. Sanders, W.W. Stroup, A unified approach to mixed linear models, Am. Stat. 45 (1) (1991) 54–64.
[36] L. Mohimont, A. Chemchem, F. Alin, F. Michal, L. Steffenel, Convolutional neural networks and temporal cnns for covid-19 forecasting in france, Appl. Intell. (2020).
[37] H. Mwambi, S. Ramroop, L. White, Age-specific incidence and prevalence: A statistical perspective, Stat Methods Med Res. 20 (5) (2001) 551–570.
[38] D. Qibin, J. Wu, Y. Wu, G. Wang. Predication of inflection point and outbreak size of covid-19 in new epicentres, 2020..
[39] R.M. Anderson, R.M. May, Infectious diseases of humans, Oxford University Press, 1991.
[40] P. Tsui, M. Zuo, S.K. Khosa, Z. Ahmad, Z. Almaspoor, Comparison of covid-19 pandemic dynamics in asian countries with statistical modeling, Comput. Math. Methods Med. (2020).
[41] P.S. Valvo, A bimodal lognormal distribution model for the prediction of covid-19 deaths, Appl. Sci. 10 (2020).
[42] J. van den Broek, H. Nishiura. Using epidemic prevalence data to jointly estimate reproduction and removal. Ann. Appl. Stat., 3(4):1505–1520, 12 2009..
[43] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun. Deep learning methods for forecasting covid-19 time-series data: A comparative study. 140, 2020..