**ACTA IMEKO**
**ISSN: 2221-870X**
*September 2023, Volume 12, Number 3, 1 - 15*

# Incidence rate estimation of SARS-COVID-19 via a Polya process scheme: a comparative analysis in Italy and European countries

**Filippo Carone Fabiani[1], Rosario Schiano Lo Moriello[2], Davide Ruggiero[3]**

[1] *University Milano-Bicocca, Department of Economics, Management and Statistics, Milano, Italy*
[2] *University Federico II Napoli, Department of Industrial Engineering, 80125, Napoli, Italy*
[3] *STMicroelectronics, Analog. MEMS and Sensor Group R&D, Arzano, 80022, Napoli, Italy*

ABSTRACT
During an ongoing epidemic, especially in the case of a new agent, data are partial and sparse, also affected by external factors, as for example climatic effects or preparedness and response capability of healthcare structures. Despite that, we showed how, under some universality assumptions, it is possible to extract strategic insights by modelling the pandemic through a probabilistic Polya urn scheme. Adopting a Polya framework, we provided both the distribution of infected cases and the asymptotic estimation of the incidence rate, showing that data are consistent with a general underlying process at different scales. Using European confirmed cases and diagnostic test data on COVID-19, we also provided an extensive comparison among European countries and between Europe and Italy at regional scale, for both the two big waves of infection. We globally estimated an incidence rate in accordance with previous studies.

**Section:** RESEARCH PAPER

**Keywords:** COVID-19; Polya urn; negative binomial; incidence rate

**Citation:** Filippo Carone Fabiani, Rosario Schiano Lo Moriello, Davide Ruggiero, Incidence rate estimation of SARS-COVID-19 via a Polya process scheme: a comparative analysis in Italy and European countries, Acta IMEKO, vol. 12, no. 3, article 54, September 2023, identifier: IMEKO-ACTA-12 (2023)-03-54

**Section Editor:** Francesco Lamonaca, University of Calabria, Italy

**Received** April 6, 2023; **In final form** June 14, 2023; **Published** September 2023



**Corresponding author:** Filippo Carone Fabiani, e-mail: filippo.caronefabiani@unimib.it

## 1. INTRODUCTION

On December 2019, a novel coronavirus (SARS-CoV-2)-infected pneumonia (COVID-19) was first identified in Wuhan, Hubei (China). Due to the extensive spreading of the infection, on March 11, the World Health Organization (WHO) declared it a pandemic [1]. Since then, more than 80 million of confirmed cases and about 2 million of deaths have been reported worldwide. Just in Europe, confirmed and deaths have been respectively about 25 and 0.5 million of cases, mostly concentrated in Russia, France, Italy, U.K. and Spain. Since the outbreak of the pandemic, many national and international health organizations have collected daily data about the COVID-19 pandemic, although following different policymaking, in terms of specific information provided, temporal and geographical aggregation of data and efficiency of tests. For example, Spain, Germany, Netherlands, and Sweden have provided only weekly data and few of them have provided aggregated data on a small regional scale. There are also countries that started later to provide tests, or, in some cases, they didn't report any data about performed tests (Albania, Moldova, Montenegro). Moreover, due to the emergency caused by the ongoing epidemic, helplessness in detecting concurrent causes of death and, in some cases, delays or corruptions in data reporting, have affected the database on regional and national level. Last but not least, due to the excess load of the healthcare systems and to the unknown nature of the pathogen, direct counting of the total number of infected patients (confirmed, asymptomatic and pauci-symptomatic cases) is impeded [2], [3]. This leads to a biased estimation of key indexes, as the case fatality rate (*CFR*) and the infection fatality rate (*IFR*) [4]-[6], usually used during a pandemic to measure the lethality of the incipient infection. Due to the above limitations, performing a rigorous analysis to assess the patterns of the epidemic, is difficult, mostly in the case of a new pathogen. Nevertheless, despite the scarcity and uncertainty of the available data, a simple analysis of empirical curves of European confirmed cases suggests some universality in the epidemic spreading. Multiple waves of infection and closed

values of key indexes [7]-[9] observed in countries under a wide social and geographical conditions, suggest the presence of distinctive features of the infection, according to the same underlying process at different scales.

Here we present an extensive analysis, of the COVID-19 pandemic spread, at national and regional scales, comparing results from 37 European countries and 21 Italian regions, for both the two observed big waves of infection, from 24 February 2020 to 30 January 2021. We showed that the dynamics of the COVID-19 susceptible-infection system can be appropriately modelled by a generalized Polya urn model. Different Polya schemes were previously used in epidemiology to model disease transmission for infectious diseases, like SARS or smallpox [10], and to implement more general transmission models [11]. The probabilistic Polya urn scheme [12], [13] describes the stochastic process of drawing from an urn, with reinforcement, a number of balls with two different colours (here labelling healthy and infected cases), thus allowing to explain important features of the COVID-19 pandemic's scenario. In fact, so far as is known, in the long-time limit, this model directly provides, both the probability density function (PDF) of infected cases and an estimation of the asymptotic composition $\rho(\infty)$ of the urn.

In order to implement the Polya scheme, we applied a simplified multiple waves approach, in which each wave of infection is considered an independent process. For each selected wave of infection and for each selected country (or region), confirmed cases observations were used to fit the Negative Binomial PDF that governs the Polya drawings. We statistically tested PDFs resulting from our fitting, obtaining about 75% of successful PDFs for European countries and about 65% for Italian regions, over both the waves. A comparison between estimated parameters, for both the waves of infection, showed that they are consistent with some general underlying behaviour, for a wide range of external conditions. Actually we proved that all the infection processes are governed by distributions characterized by parameters with the same initial conditions. As a consequence, both national and regional data can be considered as independent sequences of trials, related to the same process at different scales. This enabled us to estimate the asymptotic composition $\rho(\infty)$ of the urn, by means the expectation of the Beta PDF which governs the limiting distribution of the sample mean of a Polya process. In order to estimate the Beta PDF we used the time series of the empirical ratio between the cumulative number of confirmed cases and the cumulative number of performed diagnostic tests. The entire procedure has been repeated for both the two big infection waves of the pandemic, for each group of the European countries and for each group of the Italian regions. The comparative results between European and Italian spread are consistent with the assumption of universality stated for our procedure, enabling us to provide an asymptotic estimation of the incidence rate (*IR*), by adopting $\rho(\infty)$ as a global variable. The computed $\rho(\infty)$ of confirmed cases represents a very important feature in the pandemic dynamics, in that it could play a crucial role as a proxy variable to estimate the *IR* of the total infected cases (symptomatic and asymptomatic), which in turns represents a key value to get strategic information required for the public health policy-making processes [14].

## 2. METHODOLOGY

Hereinafter, we described how the spreading of the COVID-19 infection can be ascribed to a contagion process, based on the Polya urn scheme. In its basic version, a single urn is considered, initially containing a number *N* of balls: *w* white balls plus *b* black balls. A ball is drawn at random and then replaced together with a number *d* of balls of the same colour of the drawn one. The parameter *d* simulates the reproducibility number, it means the average number of people which are infected when in touch with a sick person. The procedure is repeated *n* times, with *n* unlimited. In such a scheme it is known that, as *n* goes to infinite and for large *N*, the probability mass function of drawing *m* white balls after the *n*-th draw can be approximated by a Negative Binomial PDF $NB(r,p)$, with parameters $r = \frac{w}{d}$, and $p = \frac{N}{N+nd}$ [15]. Furthermore, once denoting the process by the indicator function *I*(*n*) (equal to 1 or 0 if the drawn ball is respectively white or not), the fraction of the total number of white balls, inside the urn after *n*-th draw, can be written as:

$$\rho(n) = \frac{\rho_0 + n\delta Z(n)}{1 + n\delta} \tag{1}$$

where the parameters

$$\begin{aligned} \delta &= \frac{d}{N} \\ \rho_0 &= \frac{w}{N} \end{aligned} \tag{2}$$

are, respectively, the normalized reproduction number and the initial fraction of white balls, while

$$Z(n) = \frac{1}{n}\sum_{t=1}^{n} I\,(t) \tag{3}$$

is the sample average of the process. It can be proved that, setting $\theta = \frac{N}{d} - r$, the processes $\rho(n)$ and $Z(n)$ are martingale, both converging almost surely, as *n* goes to infinite, to a random variable distributed according to a Beta $B(r,\theta)$, with a mean value $\rho(\infty)$ coinciding with the initial fraction $\rho_0$ of white balls ($\rho(\infty) = \rho_0$) [16]-[18]. Note that the above parametrization explicitly links the distribution of the random process with the distribution of its sample mean, in that they have the same *r* and $\theta$ linked by definition. The above properties can be extended to more general schemes: balls labeled with an arbitrary number of colors, different replacement rules (*d*) [18], or different number *m* of drawn balls at once (multiple drawings) [19], [20]. In the latter case it can be also proved that the total fraction of white balls $\rho(n)$ is still converging to a random variable with a Beta-like distribution. This supports the idea to describe confirmed cases data by a Polya process in whatever time or geographical format they have been aggregated.

### 2.1. Assumptions and statements

In order to assure suitable conditions for implementing the entire scheme, we need to set the following assumptions.

1. Regular and persistent features, common to many countries, with different levels of disease severity, highlight a general underlying dynamic of the disease spread. We suppose it would exist some universality property of the process, connected with the specific nature of the virus. As observed in many countries, although with different social and geographical conditions, the presence of different peaks in the curve of confirmed cases suggests that the disease spread can be modelled by multiple waves of infection, considering each of them as a single independent infection process. We concentrated our analysis on the two biggest waves

observed in almost all considered areas (countries or regions).

2. We applied this multiple-waves scheme describing each of the two waves by a single uni-modal distribution. In order to separate data set of each country, we selected time intervals around the two highest peaks, with endpoints coinciding with the lowest values before and after the peak. For practical calculation, we assume these values represent respectively the onset and the close of a single wave of infection, although in many cases there is an overlap between consecutive waves.
3. We restricted our analysis excluding abrupt change in dynamical conditions and applying our procedures to consecutive waves of infection. In other words, we assume the pandemic, after a finite time interval, has exhausted its actual pathogenic load, due to natural causes or to containment rules adopted. This allows us to deal with asymptotic quantities in the limit of long-time *t*.
4. We adopted the apex *g* (with *g*=*E*,*I*) to label the group of the European countries (*E*) or the Italian regions (*I* ); the apex $\nu$ (with $\nu$=*f*, *s*) to label the first (*f* ) or the second (*s*) waves of infection and the apex *i* to label the *i*-th European country (with i=1,…,37 if *g*=*E*) or the *i*-th Italian region (with *i*=1,…,21 if *g*=*I*). Using this notation, we define $c_i^{g,\nu}(n)$ the daily number of confirmed cases, at each step *n*, for each *i*-th country, *g*-th group and $\nu$-th wave of infection. Identifying the white balls with the variable $c_i^{g,\nu}(n)$ and the number of drawn balls, at each step *n*, with the daily tests, we described the infection spread by a Polya process, assuming that each time series $c_i^{g,\nu}(n)$ follows a Negative Binomial distribution $NB(r_i^{g,\nu}, p_i^{g,\nu})$, with parameters $r_i^{g,\nu}$ and $p_i^{g,\nu}$, defined as in [15]. We assume an ideal efficiency of the diagnostic tests; it means tests are 100% sensitive for detecting COVID-19 infections.
5. We applied all our procedures separately to the set of European countries and Italian regions, in order to analyse and compare the spreading of COVID-19 at different scales, in terms of number of population and geographical size. We also repeated the procedures for the first and the second wave of infection.
6. According to the scenario presented in assumption 1, for each *g* and $\nu$, we assume it exists a general process underlying the disease spreading process observed in all the *i* European countries (Italian regions). This corresponds to assume the *i* European countries (Italian regions) as independent urns, that generate different processes characterized by different initial conditions (*N*, $\omega$ and *d*) but governed by a single compound distribution. In order to implement such a scheme, we chose to model the main process by a Negative Binomial compound distribution $NB(r^{g,\nu}, p^{g,\nu})$, in which the characteristic parameters $r^{g,\nu}$ and $p^{g,\nu}$ are also random variables, normally distributed. In such a scheme, at fixed *g* and $\nu$, each *i*-th parameter $r_i^{g,\nu}$ and $p_i^{g,\nu}$, characterizing the *i*-th distribution $NB(r_i^{g,\nu}, p_i^{g,\nu})$, can be seen as a samples of the normal distributions $r^{g,\nu}$ and $p^{g,\nu}$. These normal distributions are characterized by the sample means $\mu_r^{g,\nu}$ and $\mu_p^{g,\nu}$ and the sample variances $\sigma_r^{g,\nu}$ and $\sigma_p^{g,\nu}$.
7. Let define:

$$R_i^{g,\nu}(n) \equiv \frac{\sum_{t=1}^{n} c_i^{g,\nu}(t)}{\sum_{t=1}^{n} t} \tag{4}$$

the ratio between cumulative cases and cumulative diagnostic tests. In our scheme each $R_i^{g,\nu}(n)$ replaces the sample mean $Z_i^{g,\nu}(n)$ of Equation 2 for each process and, according to general results [19], [20], we can assume that each $R_i^{g,\nu}(n)$ follows a certain Beta limiting distribution $B(r_i^{g,\nu}, \theta_i^{g,\nu})$. Following the statements in assumption 6, we suppose to model such distributions by a Beta compound $B(r^{g,\nu}, \theta^{g,\nu})$, whose parameters are linked to the parameters of the Negative Binomial compound, as specified in section 2. Actually, the two above compounds have the same $r^{g,\nu}$ and $\theta^{g,\nu} = \frac{N}{d} - r^{g,\nu}$, assuming also $\theta^{g,\nu}$ normally distributed.

The above assumptions define a compound model for the disease spread that, under specific initial conditions, can be used to simply evaluate relevant parameters of the epidemic. According to the proposed scheme, for each *g* and $\nu$, we propose to estimate the IR of each *i*-th country (region) by using the limiting mean value $\rho_i^{g,\nu}(\infty)$ as the fraction of total number of infected cases over the total population (Equation 1). In order to estimate $\rho_i^{g,\nu}(\infty)$ we should have to estimate each distribution $B(r_i^{g,\nu}, \theta_i^{g,\nu})$ and then to calculate its limiting mean value. Unfortunately, this is not empirically possible, since different trials of the same *i*-th country (region) are in principle inaccessible. In a recent paper [21] the authors proposed a rescaled version of Polya model, that, taking into account the correlation among data, due to the reinforcement mechanism, shows almost sure convergence of the empirical mean, under suitable conditions. However, since we are interested in deriving the shape of the distribution of the confirmed cases and of the fraction $\rho(n)$, we adopt a different strategy. We first observe that, in the special case in which the variances $\sigma_r^{g,\nu}$ and $\sigma_p^{g,\nu}$ can be considered small enough (it means $r_i^{g,\nu}$ are very close each other, the same holds for $p_i^{g,\nu}$), the distributions $r^{g,\nu}$ and $p^{g,\nu}$ can be replaced respectively with their mean values $\mu_r^{g,\nu}$ and $\mu_p^{g,\nu}$:

$$r^{g,\nu} \simeq \mu_r^{g,\nu} \tag{5}$$

$$p^{g,\nu} \simeq \mu_p^{g,\nu}. \tag{6}$$

According to the definitions of the distributions $r^{g,\nu}$ and $p^{g,\nu}$, in the limit of large *n*, the above conditions correspond to consider $\rho_0^{g,\nu}$ and $\delta^{g,\nu}$ as normal distributions, concentrated around their mean values: $\rho_0^{g,\nu} \simeq \mu_{\rho_0}^{g,\nu}$ and $\delta^{g,\nu} \simeq \mu_\delta^{g,\nu}$. This occurs if all countries (regions), are settled in the same local conditions, namely, the initial ratio between infected and susceptible cases and the normalized reproduction number are invariant with respect to *i*. We call these conditions, or equivalently the Equation 5 and Equation 6, *universality* properties.

Under the above conditions, for each $g$ and $\nu$, the Negative Binomial compounds collapse into a single $NB(\mu_r^{g,\nu}, \mu_p^{g,\nu})$ and, as a consequence, the same happens for Beta compound, that reduces to a single $B(\mu^{g,\nu}, \mu_\theta^{g,\nu})$, where $\mu^{g,\nu} = \mu_r^{g,\nu}$ and $\mu_\theta^{g,\nu} = \frac{N}{d} - \mu^{g,\nu}$. This corresponds to consider the *i* European countries (Italian regions) as independent identical urns and each *i*-th contagion sequence $c_i^{g,\nu}(n)$ as random trials of the same underlying process, with the same Beta limiting distribution. In

this scenario, the $R_i^{g,\nu}(n)$ of Equation 4 can be considered as different trials of the same distribution, so enabling us to estimate the distribution $B(\mu^{g,\nu}, \mu_\theta^{g,\nu})$ and its limiting mean value $\rho^{g,\nu}(\infty)$.

In the next sections, we'll provide a procedure to show how the proposed Polya scheme, applied on European COVID-19 data, is consistent with the above conditions, so enabling us to obtain an estimation of the ratio $\rho^{g,\nu}(\infty)$ by accessible data.

### 2.2. Infection rate estimation based on universality properties

For our calculations, confirmed cases and diagnostic tests data of the 37 accessible European countries were collected, from 24 February 2020 to 30 January 2021. Data have reported by Humanitarian Data Exchange (HDX) database [22], that provides national data worldwide. For the Italian case, we consider all the 21 regions using data provided by the Italian Civil Protection Agency (CPA) database [23], on daily time scales. We first pre-processed raw data, avoiding data inconsistency and smoothing time series, if necessary. Then, we fitted the $NB(r_i^{g,\nu}, p_i^{g,\nu})$ PDFs for each of the 37 European selected countries and for each of the 21 Italian regions. For properly selecting observations, for each wave of infection and for each country, we proceeded as described in assumption 2. In most of the countries (regions), a multiple waves path is recognizable from daily curves of the confirmed cases. Although shifted by a time delay, all those curves show a first isolated wave followed by a second or a third big wave, overlapping each other, and a number of smaller peaks that seem to be not ascribable to random fluctuations. We consider only the two biggest waves, locating the onset and close points simply considering the minima around each peak. The maximum likelihood estimation algorithm was used to compute the fitting parameters $r_i^{g,\nu}$ and $p_i^{g,\nu}$. For practical calculation, we adopted the `fitdist` function implemented in Matlab, training the models with the confirmed cases data $c_i^{g,\nu}(n)$. In order to select PDFs that successfully fit confirmed cases data, we performed both the Kolmogorov-Smirnov (KS-test), using the `kstest` Matlab function, and the $\chi^2$-test using the `chi2gof` Matlab function, to confirm KS-tests results. For the sake of simplicity, in Table 1 and Table 2, as well as in Table 5 and Table 6, we only report p-values of KS-tests confirmed by $\chi^2$-tests. As a result, for both the two waves, we were able to select a subset of European countries (about 74% of successful PDFs) and a subset of Italian regions (about 64% of successful PDFs) that passed both the tests. Once such successful PDFs are selected, we have to assess the validity of universality conditions (Equation 5 and Equation 6) on the real Covid data. Relation 5 can be implemented requiring a sharpness hypothesis, it means the variances $\sigma_r^{g,\nu}$ are negligible with respect to the sample error of $r_i^{g,\nu}$. According to such criterion, we first computed sample means and variances ($\mu_r^{g,\nu}$ and $\sigma_r^{g,\nu}$), from the estimated parameters $r_i^{g,\nu}$ and then, under normality assumption, we tested whether the sample variances $\sigma_r^{g,\nu}$ were significantly smaller than the maximum confidence interval (c.i.) among all c.i. of $r_i^{g,\nu}$. To this aim, for each wave $\nu$, we

Table 1. Fitting parameters $r_i^{E,f}$ and $p_i^{E,f}$ $(i = 1\ldots27)$ of the Negative Binomial PDFs for the selected European countries, in the first wave of infection. Best $p$-values resulting from both $KS$ and $\chi^2$ test are also reported.

| Countries | $r_i^{E,f} \pm 95\%$ c.i. | $p_i^{E,f} \pm 95\%$ c.i. in $10^{-2}$ | $p-value$ |
|---|---|---|---|
| Austria | 1.70 ± 0.57 | 2.33 ± 0.92 | 0.15 |
| Belgium | 1.11 ± 0.28 | 0.19 ± 0.06 | 0.54 |
| Bos.-Herzeg. | 3.29 ± 1.31 | 9.33 ± 3.62 | 0.72 |
| Croatia | 1.12 ± 0.36 | 3.44 ± 1.30 | 0.24 |
| Denmark | 0.70 ± 0.17 | 0.67 ± 0.22 | 0.02 |
| Estonia | 0.59 ± 0.14 | 3.59 ± 1.17 | 0.04 |
| Finland | 0.83 ± 0.19 | 1.48 ± 0.43 | 0.15 |
| France | 1.73 ± 0.70 | 0.25 ± 0.12 | 0.06 |
| Germany | 1.35 ± 0.40 | 0.06 ± 0.03 | 0.22 |
| Greece | 1.06 ± 0.35 | 2.57 ± 1.02 | 0.49 |
| Hungary | 1.30 ± 0.35 | 3.12 ± 0.97 | 0.70 |
| Iceland | 0.69 ± 0.24 | 2.34 ± 1.04 | 0.66 |
| Ireland | 0.62 ± 0.14 | 0.26 ± 0.09 | 0.38 |
| Italy | 1.04 ± 0.23 | 0.06 ± 0.02 | 0.12 |
| Latvia | 1.08 ± 0.32 | 8.98 ± 2.94 | 0.44 |
| Lithuania | 1.27 ± 0.36 | 6.55 ± 2.11 | 0.08 |
| Luxembourg | 0.50 ± 0.12 | 1.24 ± 0.46 | 0.04 |
| Norway | 0.77 ± 0.16 | 1.11 ± 0.33 | 0.03 |
| Poland | 17.57 ± 6.53 | 4.74 ± 1.69 | 0.66 |
| Portugal | 1.02 ± 0.32 | 0.25 ± 0.10 | 0.06 |
| Romania | 4.20 ± 1.38 | 1.59 ± 0.55 | 0.71 |
| Serbia | 0.62 ± 0.17 | 0.49 ± 0.19 | 0.29 |
| Slovakia | 1.10 ± 0.39 | 4.41 ± 1.82 | 0.87 |
| Spain | 0.94 ± 0.34 | 0.09 ± 0.05 | 0.32 |
| Sweden | 3.32 ± 1.27 | 1.46 ± 0.59 | 0.56 |
| United Kingdom | 2.57 ± 0.75 | 0.15 ± 0.04 | 0.23 |
| Ukraine | 22.04 ± 9.01 | 4.73 ± 1.87 | 0.73 |

Table 2. Fitting parameters $r_i^{E,s}$ and $p_i^{E,s}$ $(i = 1\ldots28)$ of the Negative Binomial PDFs for the selected European countries, in the second wave of infection. $p$-values resulting from fitness $KS$-tests are also reported.

| Countries | $r_i^{E,s} \pm 95\%$ c.i. | $p_i^{E,s} \pm 95\%$ c.i. in $10^{-2}$ | $p-value$ |
|---|---|---|---|
| Austria | 0.67 ± 0.10 | 0.04 ± 0.01 | 0.04 |
| Belgium | 1.13 ± 0.28 | 0.03 ± 0.01 | 0.14 |
| Bulgaria | 0.97 ± 0.20 | 0.07 ± 0.02 | 0.30 |
| Croatia | 1.48 ± 0.34 | 0.09 ± 0.03 | 0.12 |
| Cyprus | 0.59 ± 0.12 | 0.31 ± 0.09 | 0.03 |
| Czechia | 1.79 ± 0.67 | 0.03 ± 0.01 | 0.51 |
| Denmark | 0.51 ± 0.08 | 0.07 ± 0.02 | 0.12 |
| Estonia | 2.03 ± 0.52 | 0.53 ± 0.16 | 0.13 |
| France | 0.58 ± 0.10 | 0.01 ± 0.01 | 0.77 |
| Germany | 1.54 ± 0.33 | 0.02 ± 0.01 | 0.14 |
| Greece | 2.57 ± 0.68 | 0.20 ± 0.06 | 0.34 |
| Hungary | 0.88 ± 0.16 | 0.05 ± 0.02 | 0.18 |
| Iceland | 0.77 ± 0.18 | 2.77 ± 0.82 | 0.15 |
| Ireland | 1.06 ± 0.34 | 0.06 ± 0.03 | 0.70 |
| Italy | 1.47 ± 0.37 | 0.01 ± 0.01 | 0.05 |
| Latvia | 0.46 ± 0.08 | 0.14 ± 0.04 | 0.04 |
| Lithuania | 0.90 ± 0.18 | 0.08 ± 0.02 | 0.10 |
| Netherlands | 0.72 ± 0.14 | 0.03 ± 0.01 | 0.04 |
| Nor. Macedonia | 1.53 ± 0.32 | 0.29 ± 0.08 | 0.11 |
| Norway | 0.87 ± 0.15 | 0.34 ± 0.08 | 0.04 |
| Poland | 1.17 ± 0.27 | 0.02 ± 0.01 | 0.18 |
| Portugal | 0.87 ± 0.15 | 0.03 ± 0.01 | 0.06 |
| Romania | 3.02 ± 0.67 | 0.07 ± 0.02 | 0.90 |
| Serbia | 0.87 ± 0.19 | 0.04 ± 0.01 | 0.05 |
| Slovakia | 3.04 ± 0.92 | 0.15 ± 0.05 | 0.33 |
| Sweden | 1.01 ± 0.20 | 0.04 ± 0.01 | 0.05 |
| United Kingdom | 4.13 ± 1.35 | 0.02 ± 0.01 | 0.27 |
| Ukraine | 1.25 ± 0.20 | 0.03 ± 0.01 | 0.03 |

performed a one-tailed $\chi^2$-test for variance, obtaining successful results from all the performed statistical tests. The `vartest` Matlab function was used in this case. Moreover, coherently with our assumptions we expect sample means $\mu_r^{g,\nu}$ to be also invariant under change of wave $\nu$. To verify such property, we statistically tested the null hypothesis: $\mu_r^{g,f} = \mu_r^{g,s}$, by performing a two-sample two-tailed t-test implemented in Matlab by the `ttest2` function. Also, in this case successful results confirmed the validity of Equation 5. A two-sample two-tailed t-test were also performed with success, to simultaneously compare couples of $\mu_r^{g,\nu}$ picking all the combinations of *g* and $\nu$. On the other hand, in order to complete the proof, condition in Equation 6 must be tested by the same procedure. Unfortunately, parameters $p_i^{g,\nu}$ depend on *n* which depends on *i*, *g* and $\nu$, so they can't be directly compared, within and between the waves, or estimated from $B(\mu^{g,\nu}, \mu_\theta^{g,\nu})$. In order to bypass such limitation, and without demanding completeness, we showed that, for each *g* and $\nu$, $p_i^{g,\nu}$ slightly differ within groups of countries (regions), having close numbers of performed diagnostic tests *n*, so that we can restrict the comparison procedure for $p_i^{g,\nu}$ within such groups. To this aim, we first identified such homogeneous groups by applying different clustering algorithms: K-Means, Density-Based Spatial Clustering and Gaussian Mixture Model (GMM) Clustering algorithms were compared, identifying countries (regions) by variables *n* and $r_i^{g,\nu}$. The best partitioning was obtained using GMM algorithm implemented in Matlab by two main functions: `gmdistribution` and `cluster`. Afterwards, we tested the sharpness hypothesis by performing a one-tailed $\chi^2$-test on the sample variances $\sigma_p^{g,\nu}$ calculated for each cluster. Results show that, in each cluster, the distribution of $p^{g,\nu}$ can reduce to a single value, hence suggesting also a weak dependency of $\delta_i^{g,\nu}$ from *i*.

Once the distributions of confirmed cases have been verified, the problem of estimating the asymptotic value $\rho^{g,\nu}(\infty)$ reduces to compute the expectation value of the asymptotic distributions $B(\mu^{g,\nu}, \mu_\theta^{g,\nu})$. According to Equation 5 and Equation 6, in fact, we were able to fit the Beta PDFs and their parameters $\mu^{g,\nu}$ and $\rho^{g,\nu}(\infty)$, by using the set of the *i* limiting values $R_i^{g,\nu}(\infty) = \lim_{n\to\infty} R_i^{g,\nu}(n)$ (see (b) and (c)). A cut-off value has been also introduced to establish a finite size convergence criterion. However, in most of the cases $R_i^{g,\nu}(\infty)$ were obtained directly by the ratio computed on the end points of waves of infection, since they can be considered far enough to represent the limit for large times, where the stability of $R_i^{g,\nu}(n)$ is reached. Also in this case, maximum likelihood estimation algorithm and `fitdist` function was adopted to estimate the parameters $\mu^{g,\nu}$ and $\theta^{g,\nu}$, for each Beta PDF. KS-test and $\chi^2$-test were also successfully conducted to assess the goodness of the estimated Beta. In order to assure the consistency of the whole estimation procedure we finally tested whether $\mu_r^{g,\nu}$, estimated by $NB(r^{g,\nu}, p^{g,\nu})$, were significantly different from the respective $\mu^{g,\nu}$ estimated by $B(\mu^{g,\nu}, \mu_\theta^{g,\nu})$. For each group *g* and each wave of infection $\nu$ we implement a one-sample two-tailed t-test, under the null hypothesis $\mu^{g,\nu} = \mu_r^{g,\nu}$, by using the `ttest` Matlab function.

The above comparative results account for the universality properties, and they support the resulting property of ~~scale~~ invariance of the asymptotic fraction $\rho^{g,\nu}(\infty)$, under change of $\nu$ or *g*. In the next sections we provide detailed numerical results of all the above procedures.

## 3. RESULTS

### 3.1. European countries

Before engaging in parameters estimation, a pre-processing phase was performed. We restricted to the countries with total population greater than one million, selecting 40 of the 48 countries from European continent reported in the HDX database. We also discarded countries that did not provide diagnostic test data at all (Albania, Montenegro, Moldova). We processed remaining 37 European countries, first removing inconsistent data (outliers and negative data) than selecting only the observations provided with their diagnostic tests. Wherever possible, data were collected on daily time scale, otherwise we smoothed weekly data, $c_i^{E,\nu}(n)$ and $n_i^{E,\nu}$, as in the case of Germany, Spain, Netherlands, France and Ukraine. Due to different starting day of the pandemic, delays or corruptions in data reporting, four countries (Bulgaria, Cyprus, Czechia, North Macedonia ) did not provided sufficient data for the first wave of infection. As a result, we obtain two sets composed by 33 countries for the first wave and by 37 for the second wave. Fitting procedures were applied on the above two groups and both KS-test and $\chi^2$-test were performed, with a confidence level $\alpha$=0.01, to select successful PDFs. KS-tests p-values were reported in Table 1 and Table 2. Resulting *i*=27 successful $NB(r_i^{E,f}, p_i^{E,f})$ PDFs for the first wave are showed in Figure 1, with their relative CDFs (Figure 3 and Figure 4). The same selecting procedure were performed for the second wave obtaining *i*=28 successful $NB(r_i^{E,s}, p_i^{E,s})$, reported in Figure 2 (CDFs in Figure 5 and Figure 6). Fitting parameters $r_i^{E,\nu}$ and $p_i^{E,\nu}$ of all successful PDFs were reported in Table 1 (first wave) and Table 2 (second wave). In Figure 3 and Figure 4, as well as in Figure 5 and Figure 6, it is also reported the graphical comparison between fitted CDFs and empirical distribution functions, computed with a 95% confidence interval. By using the successful subset of data, we computed sample means and variances ($\mu_r^{E,\nu}$, $\sigma_r^{E,\nu}$) from estimated $r_i^{E,\nu}$. Excluding the two outlier values coming from Ukraine and Poland, we obtained $\mu_r^{E,f}$=1.38 (95% c.i. : 0.98-1.78) , for the first wave. Similar results were obtained for the second wave: $\mu_r^{E,s}$ =1.35 (95% c.i. : 1.05-1.65) (see Table 4). For each wave of infection we assessed the sharpness of the normal distributions $r^{E,\nu}$ by using two distinct one-tailed $\chi^2$-tests, with a significance level $\alpha$=0.02, under the null hypothesis that $\sigma_r^{E,\nu}$ were lower than the larger confidence interval among all $r_i^{E,\nu}$ in each wave $\nu$ ( actually: $\sigma_r^{E,f} < 0.92$ and $\sigma_r^{E,s} < 0.75$, see Table 1 and Table 2). Successful results were obtained with p-values respectively 0.11 and 0.09. Focusing on parameters $r_i^{E,\nu}$ we also note that $\mu_r^{E,f}$ is very close to $\mu_r^{E,s}$ within the c.i., suggesting it is invariant also under change of the wave $\nu$ considered. In order to verify the latter hypothesis ($\mu_r^{E,f} = \mu_r^{E,s}$) a two-sample two-tailed t-test, with a significance level $\alpha$=0.05, was also performed, obtaining a p-value=0.55. Concerning the parameters $p_i^{E,\nu}$, they depend on *n*, hence they must be evaluated within homogeneous groups obtained by a GMM clustering algorithm, as discussed above. The optimization option (Calinski-Harabasz) was also adopted to optimally choose the number of clusters. As a result (see Table 3), we obtained four clusters (A, B, C, D) for the first wave and three cluster for the second wave (A, B, C). As we can see, $p_i^{E,f}$ show a very small variance inside each cluster. We performed a one-tailed $\chi^2$-tests,

with a significance level $\alpha$=0.05, on variances $\sigma_p^{E,v}$, to test the sharpness hypothesis in each cluster. All clusters successfully passed the test and p-values were reported in the last columns of Table 3. Some membership changes are present, but they are not in disagreement with our analysis, in that new external conditions could always arise, changing the spreading dynamic of a specific country. Obviously further comparisons between $\mu_p^{E,f}$ and $\mu_p^{E,s}$ are prevent since $p_i^{E,v}$ depend on *n*.

Once the convergence of $R_i^{E,v}$ was reached, asymptotic values of the ratio $R_i^{E,v}(\infty)$ were used to fit the two Beta PDFs $B(\mu^{E,v}, \mu_\theta^{E,v})$, (see Figure 7). KS-test and $\chi^2$-test were successfully performed on both the Beta PDF, with a significance $\alpha$=0.02 obtaining p-values of 0.45 and 0.12, respectively for the first and second wave. Fitting parameters $\mu^{E,v}$ and $\mu_\theta^{E,v}$ were estimated and reported in Table 4. We found a value of 1.80 (95% c.i.: 1.07-3.02) for $\mu^{E,f}$ and 1.95 (95% c.i.: 1.10-3.46) for $\mu^{E,s}$, showing that, within the c.i., $\mu^{E,v}$ are close to the sample means obtained from estimated Negative Binomials. In order to show the latter statement, a two one-sample two-tailed t-tests were also conducted, testing the null hypothesis

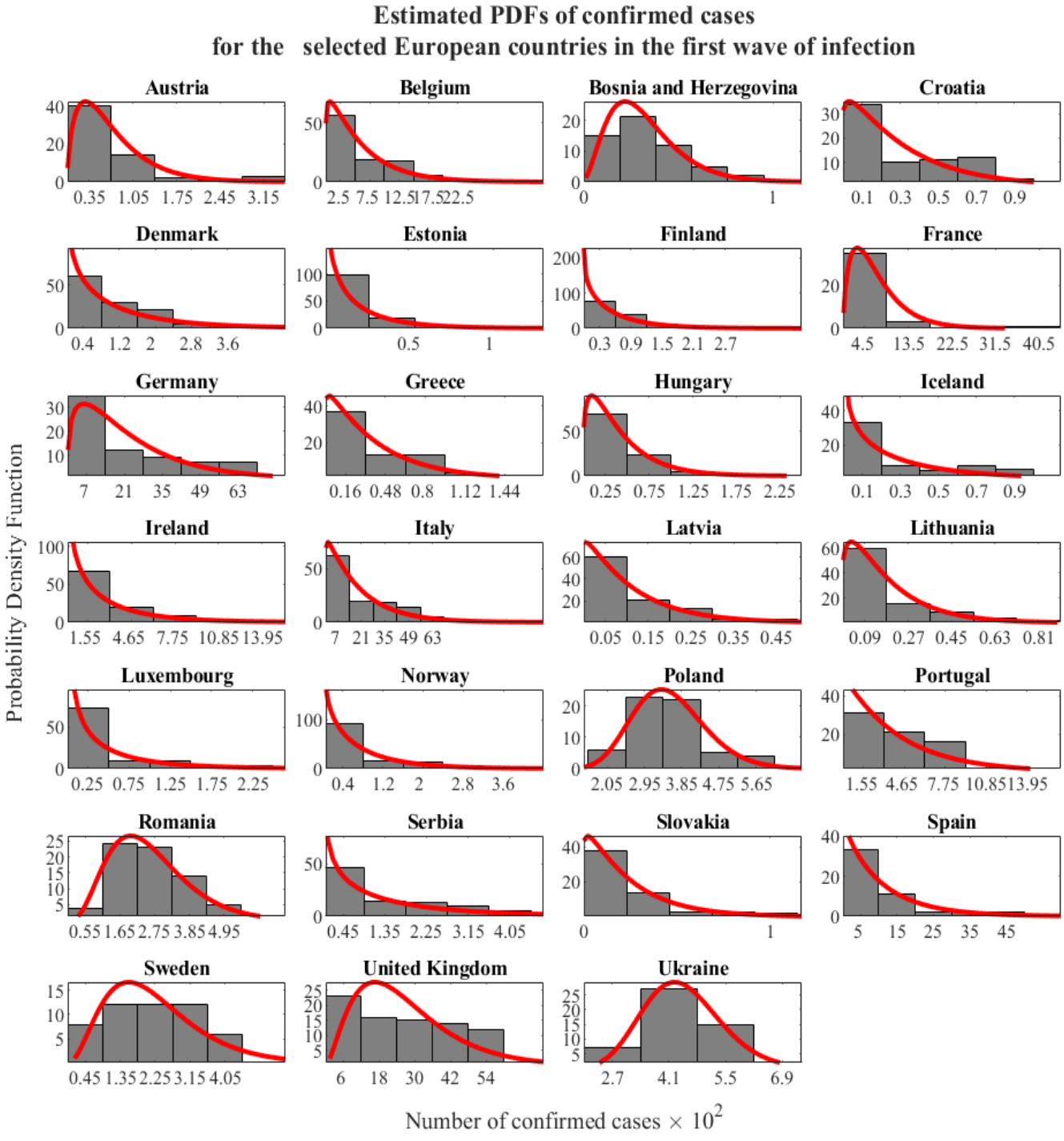


Figure 1. Negative Binomial PDFs (red line) of the confirmed cases for the group of 27 successful European countries, during the first waves of infection, was considered. Histograms of observed confirmed cases were also reported (bar plot). Black dashed lines represent upper and lower bounds, with 95% c.i..

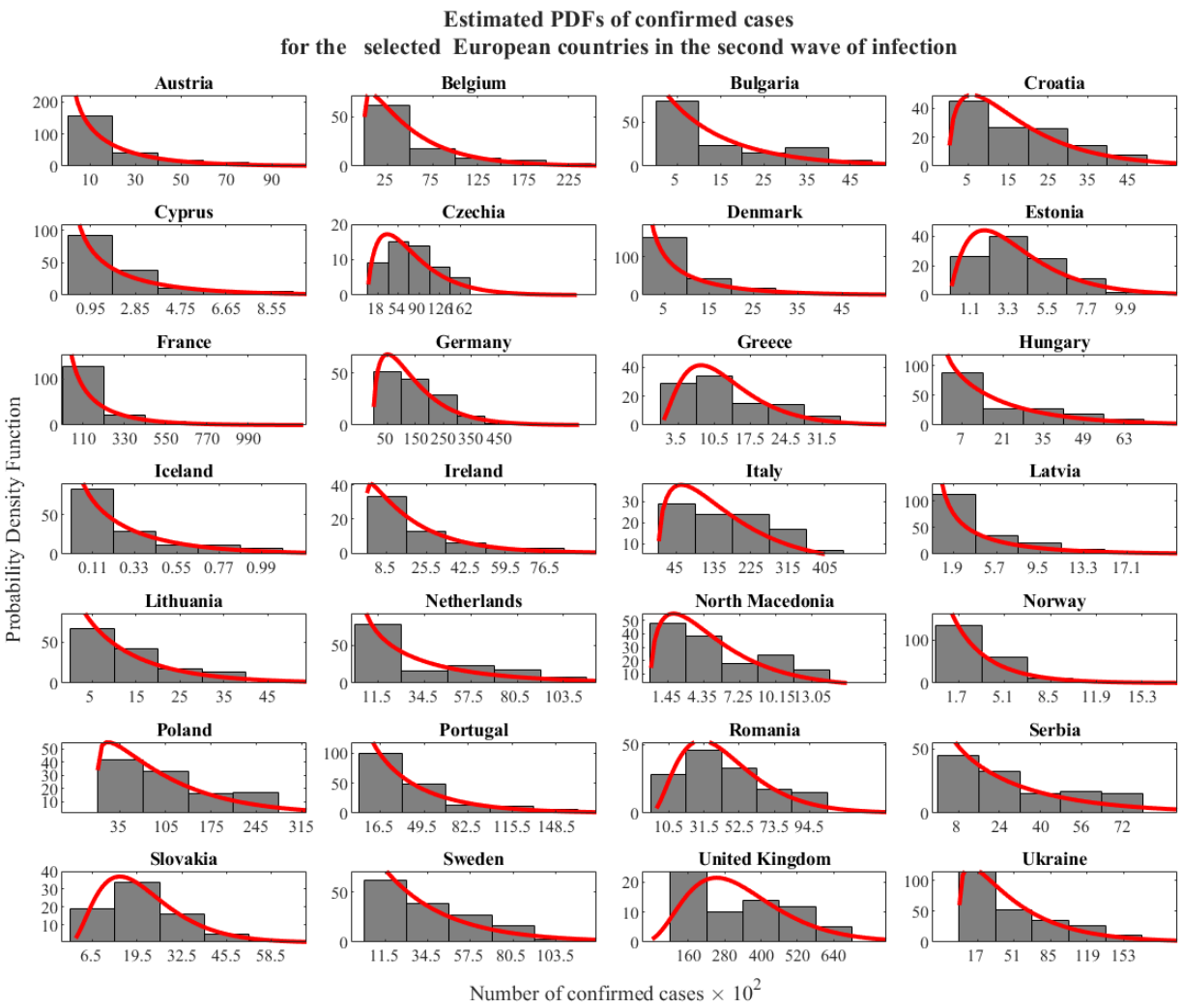


Figure 2. Negative Binomial PDFs (red line) of the confirmed cases for the group of 28 successful European countries, during the second waves of infection, was considered. Histograms of observed confirmed cases were also reported (bar plot). Black dashed lines represent upper and lower bounds, with 95% c.i..

Table 3. Clusters of European countries, in both waves of infection, based on number of executed tests and values of $r_i^{E,v}$ of Table 1 and Table 2.

| First wave of infection | | | | Second wave of infection | | | |
|---|---|---|---|---|---|---|---|
| **Cluster** | **Countries** | $\sigma_p^{E,f}$ | $p-value$ | **Cluster** | **Countries** | $\sigma_p^{E,s}$ | $p-value$ |
| A | Austria, Croatia, Estonia, Finland, Greece, Hungary, Iceland, Ireland, Latvia, Lithuania, Luxembourg, Norway, Poland, Serbia, Slovakia | $0.57 \cdot 10^{-3}$ | 0.24 | A | Austria, Belgium, Bulgaria, Cyprus, Hungary, Iceland, Ireland, Latvia, Lithuania, Netherlands, Norway, Poland, Portugal, Serbia, Sweden, Ukraine | $0.46 \cdot 10^{-4}$ | 0.14 |
| B | Belgium, Denmark, France, Portugal, Spain | $0.49 \cdot 10^{-5}$ | 0.45 | B | Croatia, Czechia, Estonia, Greece, North Macedonia, Romania, Slovakia | $0.30 \cdot 10^{-5}$ | 0.31 |
| C | Germany, Italy, United Kingdom | $0.54 \cdot 10^{-7}$ | 0.15 | C | Denmark, France, Germany, Italy, United Kingdom | $0.06 \cdot 10^{-6}$ | 0.43 |
| D | Bosnia and Herzegovina, Romania, Sweden, Ukraine | $0.19 \cdot 10^{-2}$ | 0.34 | | | | |

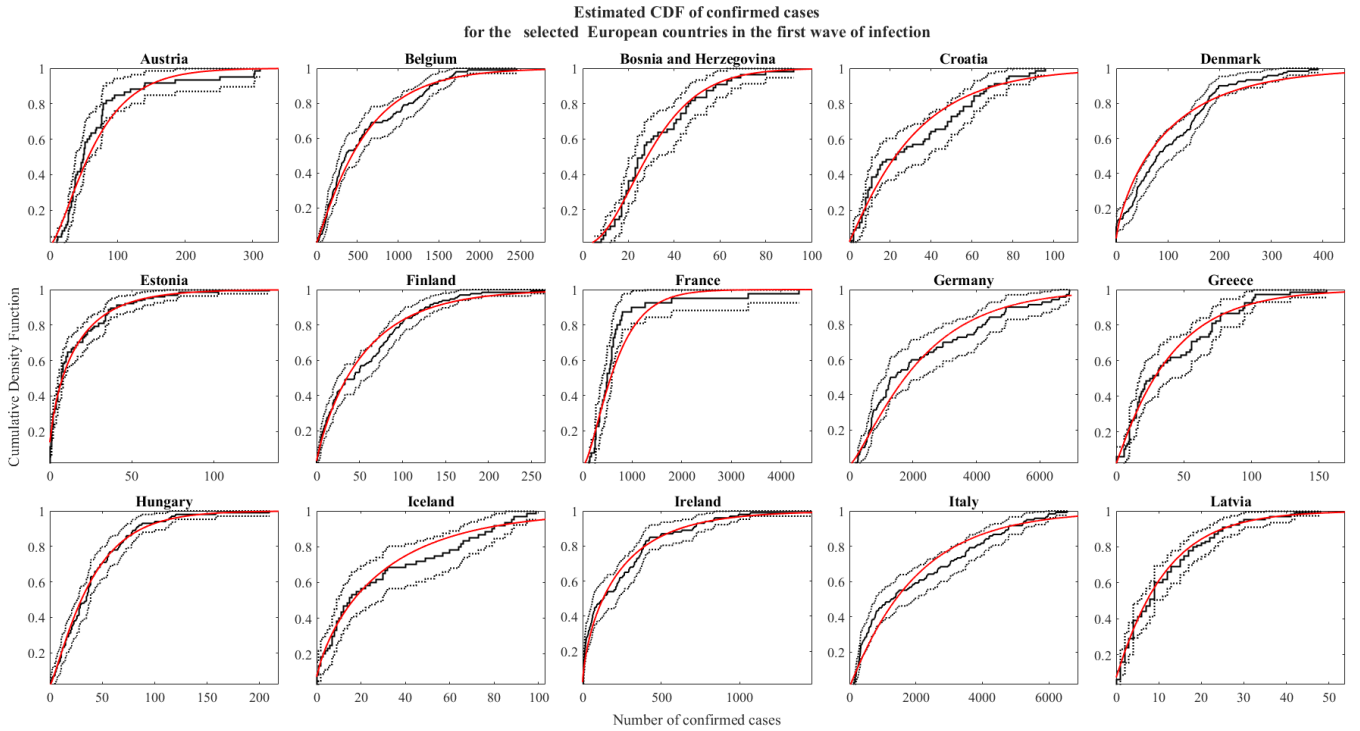


Figure 3. Negative Binomial CDFs (red line) of the confirmed cases for the group of 27 successful European countries, during the first waves of infection, was considered. Empirical CDFs of confirmed cases was reported (black line); black dashed lines represent upper and lower bounds, with 95% c.i..

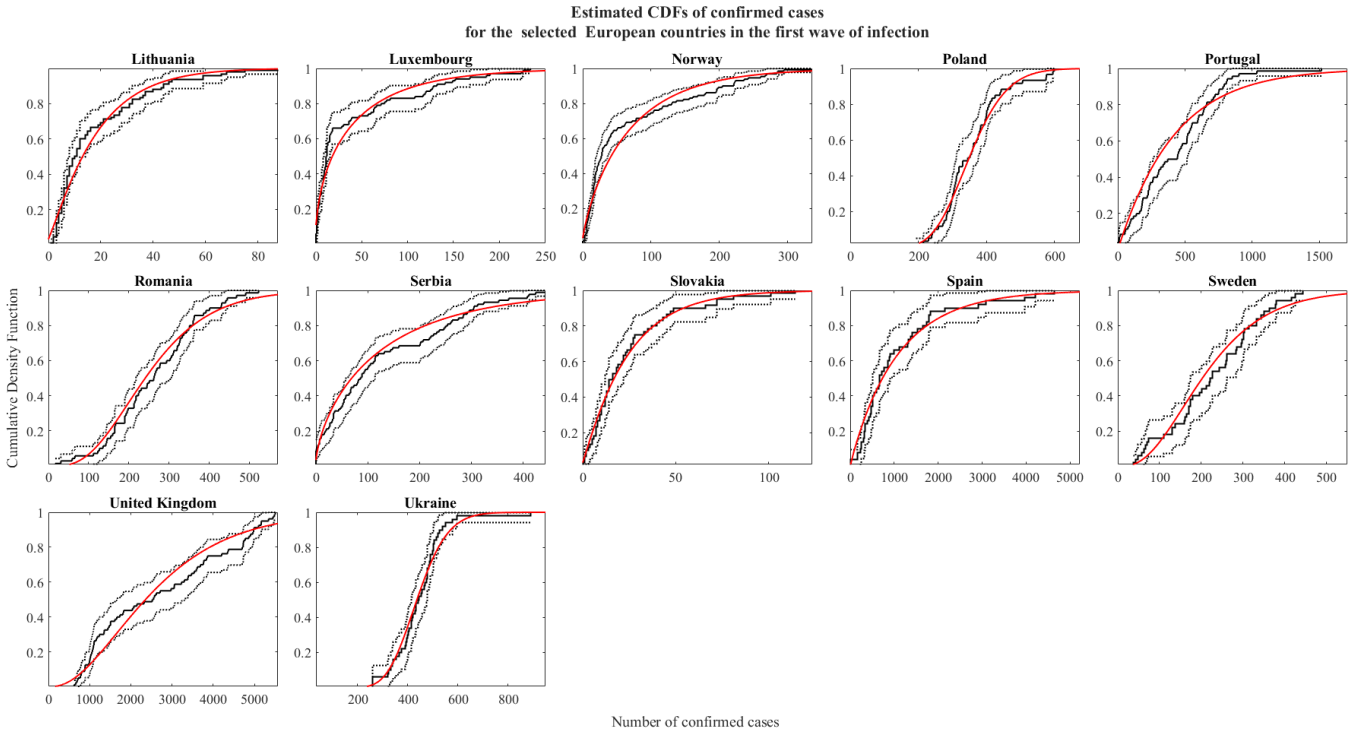


Figure 4. Negative Binomial CDFs (red line) of the confirmed cases for the group of 27 successful European countries, during the first waves of infection, was considered. Empirical CDFs of confirmed cases was reported (black line); black dashed lines represent upper and lower bounds, with 95% c.i..

Table 4. Fitting parameters $\mu_r^{E,\nu}$ and $\mu_\theta^{E,\nu}$ of the Beta PDFs for the first (f.w.) and second (s.w.) waves of infection for the European countries. The asymptotic means $\rho^{E,\nu}(\infty)$ are also reported. In the last two columns we also report the sample means $\mu_r^{E,\nu}$ and variances $\sigma_r^{E,\nu}$ obtained by Table 1 and Table 2.

| **Waves of infection First (f.w.), Second(s.w.)** | $\boldsymbol{\mu^{E,\nu}}$ $(95\%\boldsymbol{c.i.})$ | $\boldsymbol{\mu_\theta^{E,\nu}}$ $(95\%\boldsymbol{c.i.})$ | $\boldsymbol{\rho^{E,\nu}}(\infty)$ $(95\%\boldsymbol{c.i.})$ | $\boldsymbol{\mu_r^{E,\nu}}$ $(95\%\boldsymbol{c.i.})$ | $\boldsymbol{\sigma_r^{E,\nu}}$ $(95\%\boldsymbol{c.i.})$ |
|---|---|---|---|---|---|
| f.w. | 1.80 (1.07 - 3.02) | 34.92 (21.08 - 57.46) | 0.05 (0.03 - 0.07) | 1.48 ( 1.34 - 1.62 ) | 0.93 (0.88 - 0.98 ) |
| s.w. | 1.95 (1.10 - 3.46) | 12.86 (6.07 - 27.02) | 0.11 (0.08 - 0.14) | 1.68 ( 1.54 - 1.82 ) | 0.74 (0.66 - 0.82 ) |

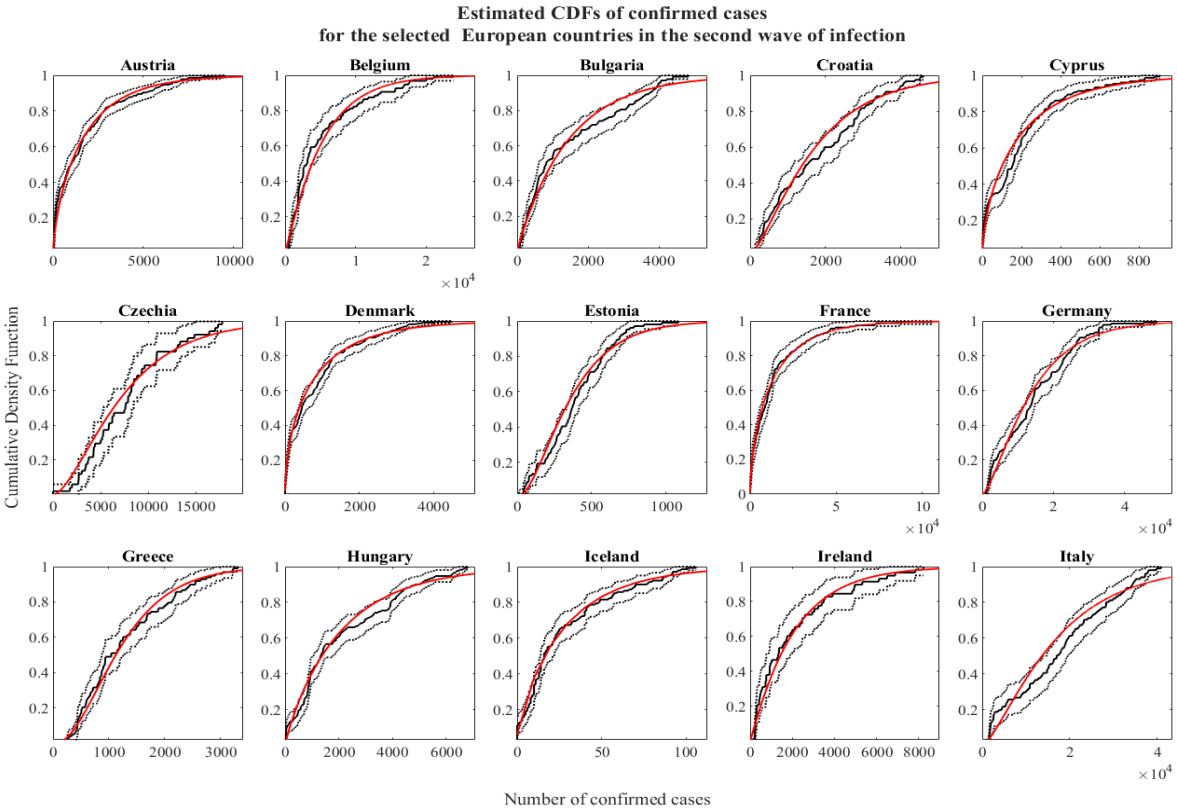


Figure 5. Negative Binomial CDFs (red line) of the confirmed cases for the group of 28 successful European countries, during the second waves of infection, was considered. Empirical CDFs of confirmed cases was reported (black line); black dashed lines represent upper and lower bounds, with 95% c.i..

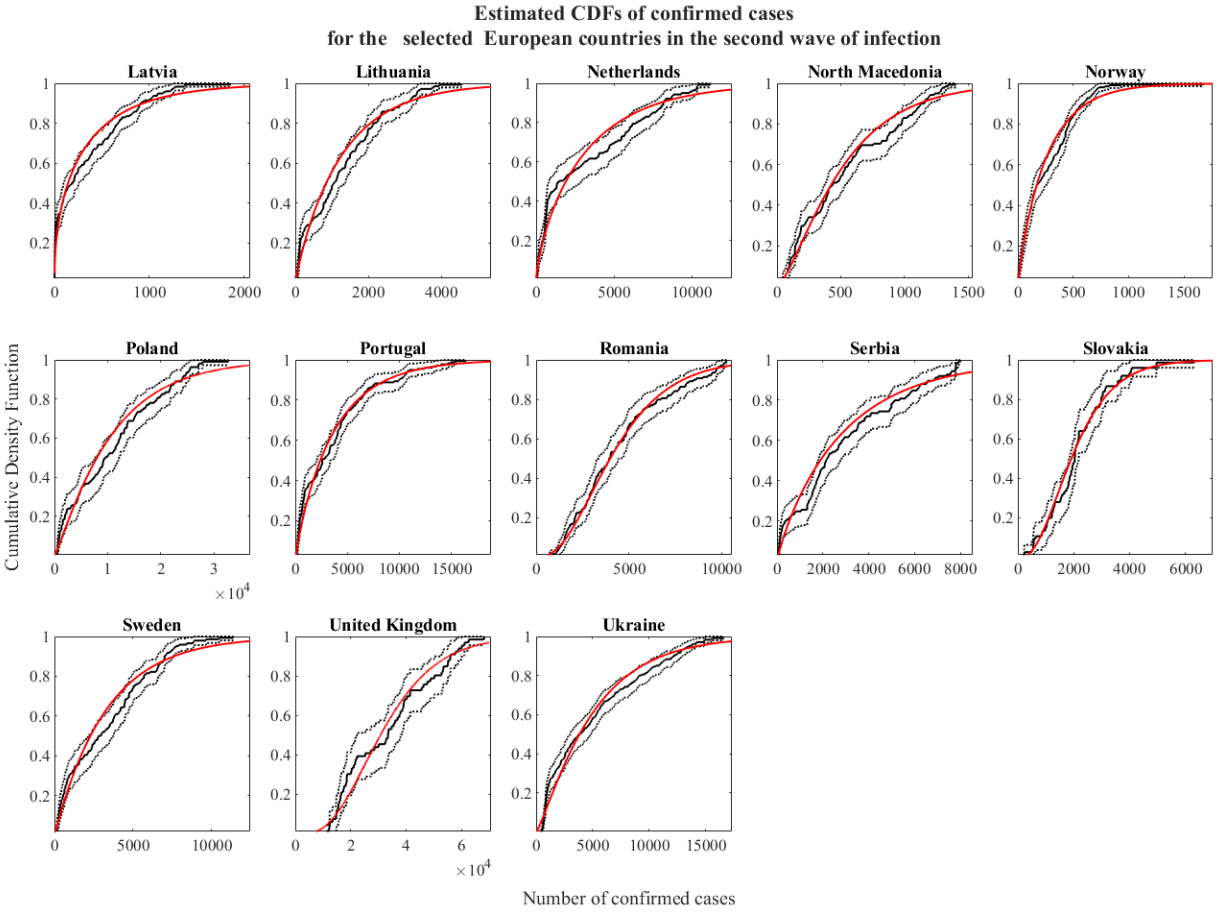


Figure 6. Negative Binomial CDFs (red line) of the confirmed cases for the group of 28 successful European countries, during the second waves of infection, was considered. Empirical CDFs of confirmed cases was reported (black line); black dashed lines represent upper and lower bounds, with 95% c.i.

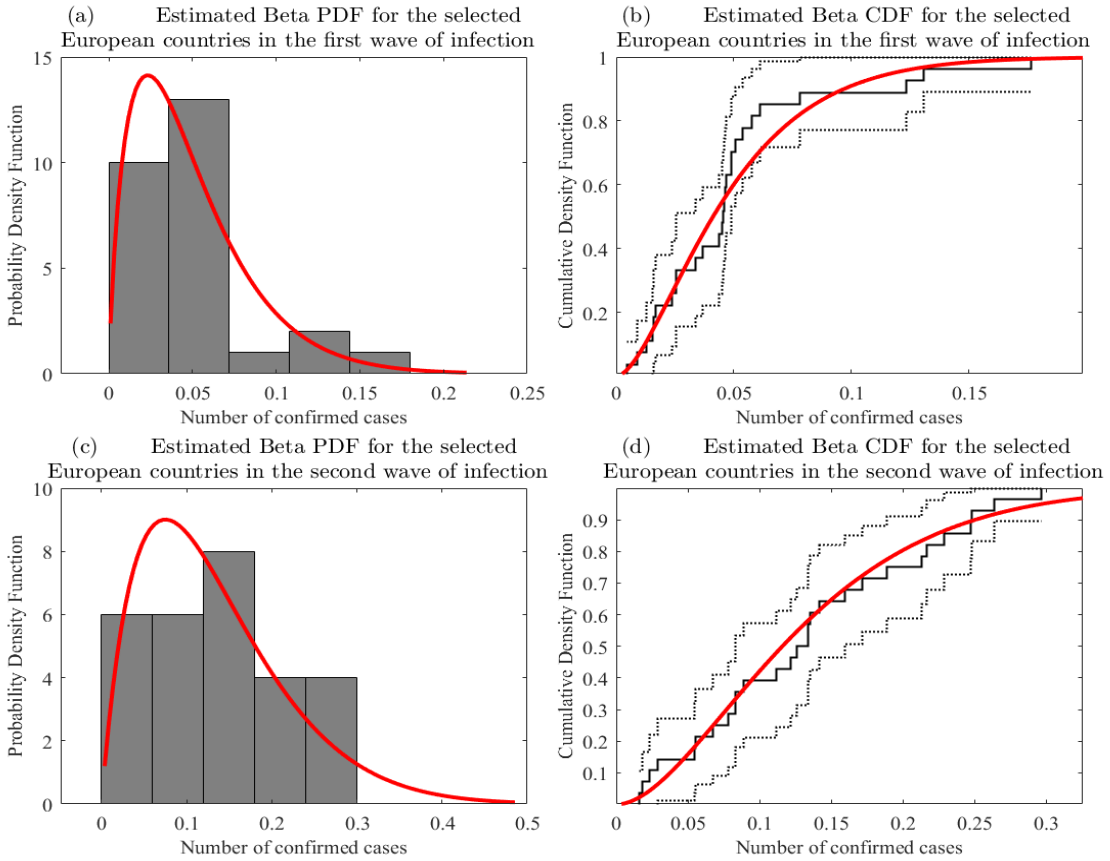


Figure 7. Beta PDFs of the ratio$R^{E,\nu}(\infty)$ (red line) with relative histogram (bar plot) during the first (a) and second (c) waves of infection for the European countries. Beta CDFs , respectively (b)and (d), corresponding to the above PDFs (red line); black dashed lines represent upper and lower bound, with 95% c.i.

that respectively $\mu^{E,f} = \mu_r^{E,f}$ and $\mu^{E,s} = \mu_r^{E,s}$. Both tests passed successfully with a significance $\alpha$=0.05 and p-values equal to 0.62 and 0.22, respectively for the first and second wave. Note that the latter equalities represent further evidence that the Covid-19 spreading is consistent with the universality conditions of Equation 5 and Equation 6. Finally, we computed expected values $\rho^{E,\nu}(\infty) = \rho_0^{E,\nu}$ (Table 4). We obtained a value $\rho^{E,f}(\infty)$ = 0.05 (95% c.i.: 0.03-0.07) for the first wave and $\rho^{E,s}(\infty)$ = 0.11 (95% c.i.: 0.08-0.14) for the second one. Both results seem in good agreement with the observations provided worldwide for the incidence rate on monthly scale [24] and the deviation between the two values $\rho^{E,\nu}(\infty)$ accounts for the observed growing of the incidence rate passing from the first to the second wave of infection. Moreover, since parameters $\mu^{E,f}$ and $\mu^{E,s}$ are very close, the difference between values of estimated $\rho^{E,\nu}(\infty)$ is mainly connected with the variation of the parameter $\mu_\theta^{E,\nu}$, which passes from 34.92 (95% c.i.: 21.08-57.46) for the first wave to 12.86 (95% c.i.: 6.07-27.02) for the second one. A decreasing $\mu_\theta^{E,\nu}$ means an increasing $\delta_i^{E,\nu}$, which corresponds to an increasing *d*, the population number being constant. According to increasing value of $\rho^{E,\nu}(\infty)$, the latter assertion could be explained by an increasing viral load, conceivably due to the inflow of a more aggressive variant of the virus, associated with an increasing initial sick population *w*, since the parameter $\mu^{E,\nu}$ is proved to remain constant varying $\nu$. All the above results support our assumptions although, as discussed above, a complete proof is prevented since $\delta_i^{g,\nu}$ cannot be directly estimated from $p_i^{g,\nu}$ or $\theta^{g,\nu}$. Finally, within the specificity described above, all the $\rho^{E,\nu}(\infty)$, resulting from the first and second wave, for all considered countries, are in accordance with population-based studies, as in [25], [26], that reports an overall antibody prevalence respectively of 6% (95% c.i.: 5.8-6.1) and 4.6% (95% c.i.: 4.3-5.0).

### 3.2. Italian regions

The entire procedure, described above for the European case, was repeated for the 21 Italian regions. First, we performed the sequence of statistical fitness KS-tests and $\chi^2$-tests, with a significance level $\alpha$=0.02, in order to select the successful PDF. Figure 8 to Figure 11 show, respectively, 14 and 13 successful Negative Binomials PDFs and CDFs, corresponding to the selected regions during respectively the first and the second wave of infection. Table 5 and Table 6 report the estimated fitting parameters and p-values of the KS-tests, for both the two waves. Sample means and variance of $r_i^{I,\nu}$ were estimated obtaining the following results: $\mu_r^{I,f}$=0.62 (95% c.i.: 0.39-0.65), $\mu_r^{I,s}$=0.45 (95% c.i.: 037-0.53) . As in the European case, a one-sample two-tailed $\chi^2$-tests on both the variances, with a $\alpha$=0.05 were successfully performed ($\sigma_r^{I,f}$ <0.20 and $\sigma_r^{E,s}$<0.11, see Table 5 and Table 6) with p-values respectively 0.21 and 0.73. Afterwards, a two-sample two-tailed t-tests for comparing $\mu_r^{I,f} = \mu_r^{I,s}$ were also successfully performed with a significance $\alpha$=0.05 and a *p*-value equal to 0.34. Also in this case, in order to estimate the distribution $p^{I,\nu}$, the GMM clustering algorithm was applied to individuate homogeneous groups: cluster A and B for the first wave and clusters A, B and C for the second wave (see Table .7).

Table 5. Fitting parameters $r_i^{I,f}$ and $p_i^{I,f}$ $(i = 1 \ldots 14)$ of the Negative Binomial PDFs for the selected Italian regions, in the first wave of infection. Best $p$-values resulting from both $KS$ and $\chi^2$ test are also reported.

| Regions | $r_i^{I,f} \pm 95\%\ c.i.$ | $p_i^{I,f} \pm 95\%\ c.i.$ in $10^{-2}$ | $p-value$ |
|---|---|---|---|
| Abruzzo | 0.48 ± 0.11 | 1.78 ± 0.61 | 0.03 |
| Campania | 0.52 ± 0.12 | 1.32 ± 0.44 | 0.12 |
| Emilia-Romagna | 0.82 ± 0.18 | 0.37 ± 0.11 | 0.02 |
| Friuli | 0.51 ± 0.12 | 1.89 ± 0.64 | 0.03 |
| Lazio | 0.84 ± 0.19 | 1.29 ± 0.39 | 0.29 |
| Liguria | 0.81 ± 0.18 | 1.01 ± 0.30 | 0.11 |
| Lombardia | 0.99 ± 0.20 | 0.16 ± 0.04 | 0.19 |
| Marche | 0.50 ± 0.11 | 0.91 ± 0.31 | 0.14 |
| Piemonte | 0.65 ± 0.14 | 0.26 ± 0.08 | 0.03 |
| Puglia | 0.44 ± 0.10 | 1.21 ± 0.43 | 0.02 |
| Sicilia | 0.47 ± 0.11 | 1.68 ± 0.58 | 0.03 |
| Toscana | 0.54 ± 0.12 | 0.66 ± 0.21 | 0.16 |
| Trento | 0.41 ± 0.10 | 1.14 ± 0.41 | 0.03 |
| Veneto | 0.54 ± 0.11 | 0.35 ± 0.12 | 0.13 |

Table 6. Fitting parameters $r_i^{I,s}$ and $p_i^{I,s}$ $(i = 1 \ldots 13)$ of the Negative Binomial PDFs for the selected Italian regions, in the second wave of infection. Best $p$-values resulting from both $KS$ and $\chi^2$ test are also reported.

| Regions | $r_i^{I,s} \pm 95\%\ c.i.$ | $p_i^{I,s} \pm 95\%\ c.i.$ in $10^{-2}$ | $p-value$ |
|---|---|---|---|
| Abruzzo | 0.39 ± 0.07 | 0.23 ± 0.07 | 0.02 |
| Bolzano | 0.40 ± 0.07 | 0.28 ± 0.08 | 0.02 |
| Calabria | 0.38 ± 0.07 | 0.34 ± 0.10 | 0.02 |
| Campania | 0.40 ± 0.07 | 0.04 ± 0.02 | 0.19 |
| Lazio | 0.50 ± 0.09 | 0.07 ± 0.02 | 0.02 |
| Liguria | 0.51 ± 0.09 | 0.19 ± 0.06 | 0.36 |
| Lombardia | 0.66 ± 0.11 | 0.03 ± 0.01 | 0.03 |
| Piemonte | 0.49 ± 0.07 | 0.06 ± 0.02 | 0.02 |
| Sardegna | 0.57 ± 0.10 | 0.34 ± 0.09 | 0.02 |
| Sicilia | 0.41 ± 0.07 | 0.09 ± 0.03 | 0.03 |
| Toscana | 0.42 ± 0.07 | 0.08 ± 0.03 | 0.02 |
| Umbria | 0.35 ± 0.07 | 0.24 ± 0.08 | 0.03 |
| Veneto | 0.53 ± 0.08 | 0.05 ± 0.01 | 0.03 |

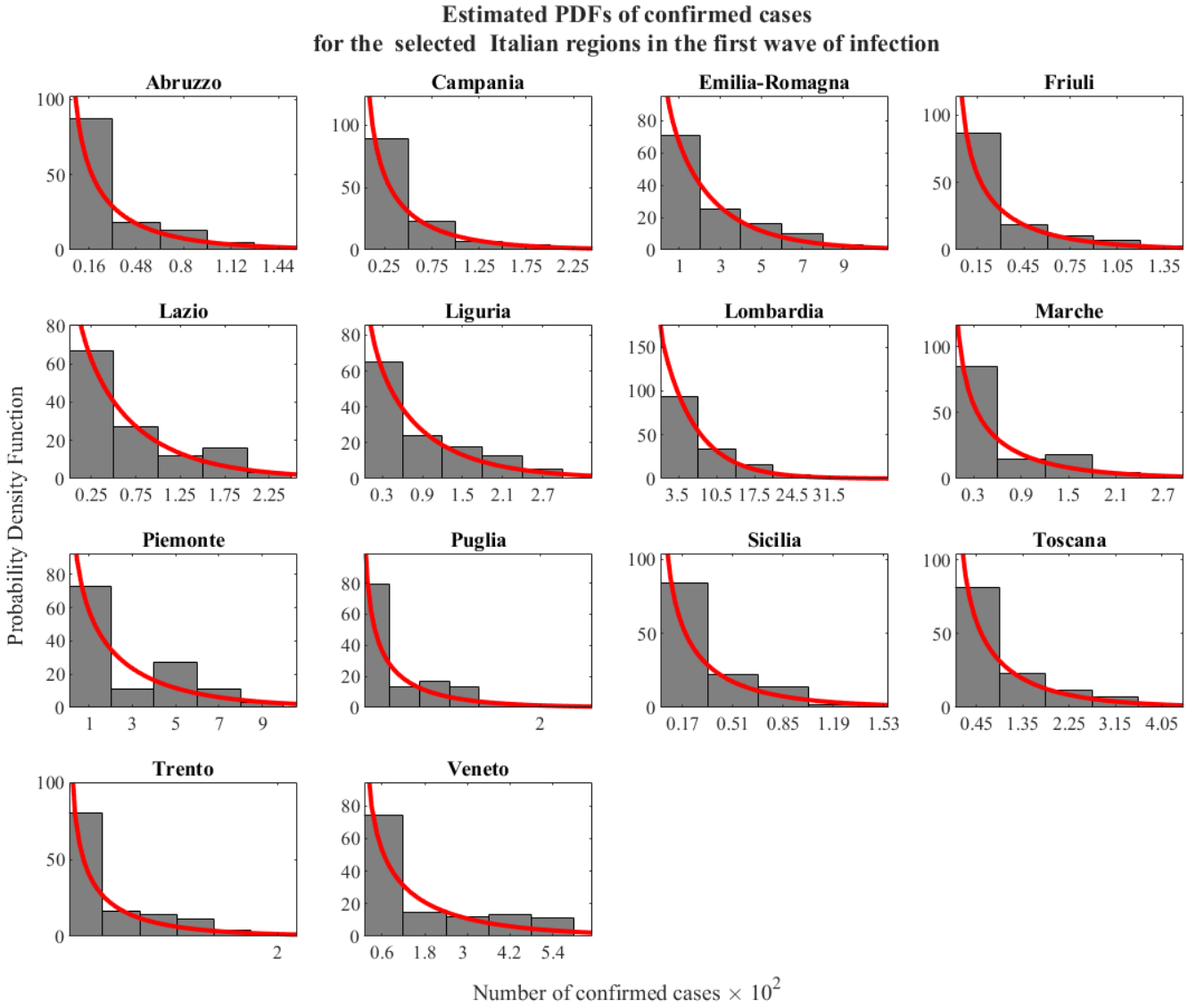


Figure 8. Negative Binomial PDFs (red line) of the confirmed cases for the group of 14 successful Italian regions, during the first waves of infection, was considered. Histograms of observed confirmed cases were also reported (bar plot). Black dashed lines represent upper and lower bounds, with 95%c.i..

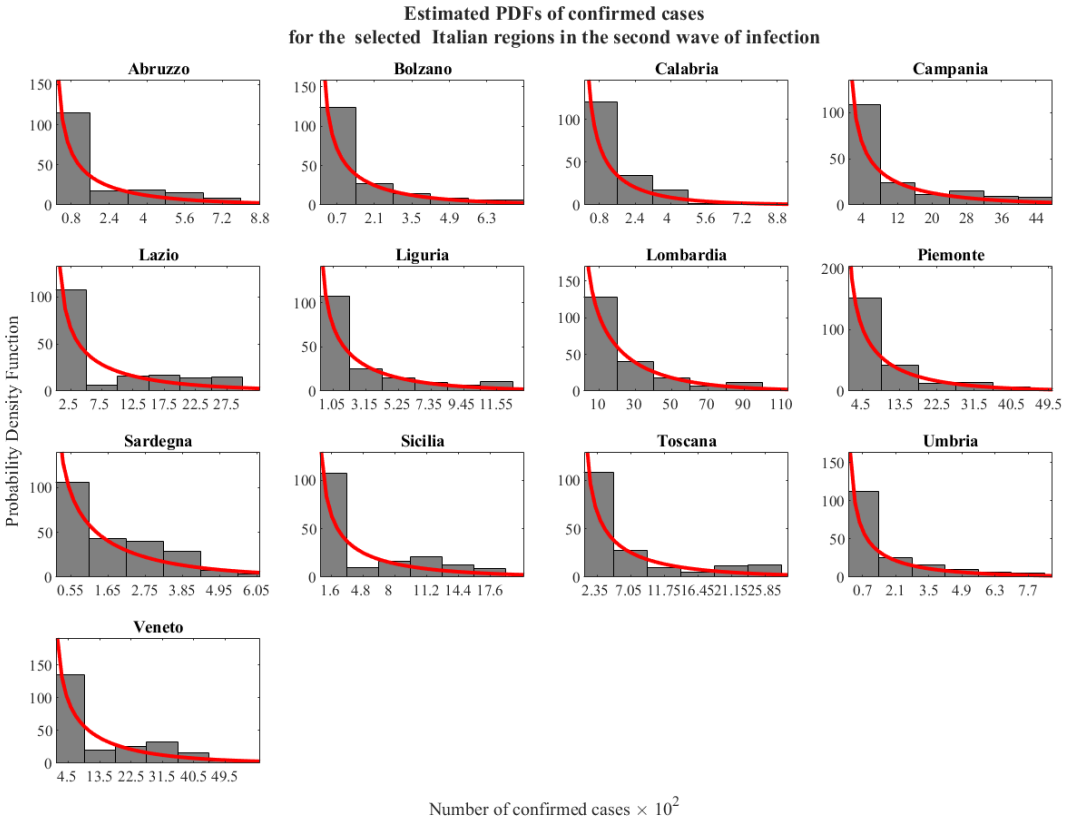


Figure 9. Negative Binomial PDFs (red line) of the confirmed cases for the group of 13 successful Italian regions, during the second waves of infection, was considered. Histograms of observed confirmed cases were also reported (bar plot). Black dashed lines represent upper and lower bounds, with 95%c.i..

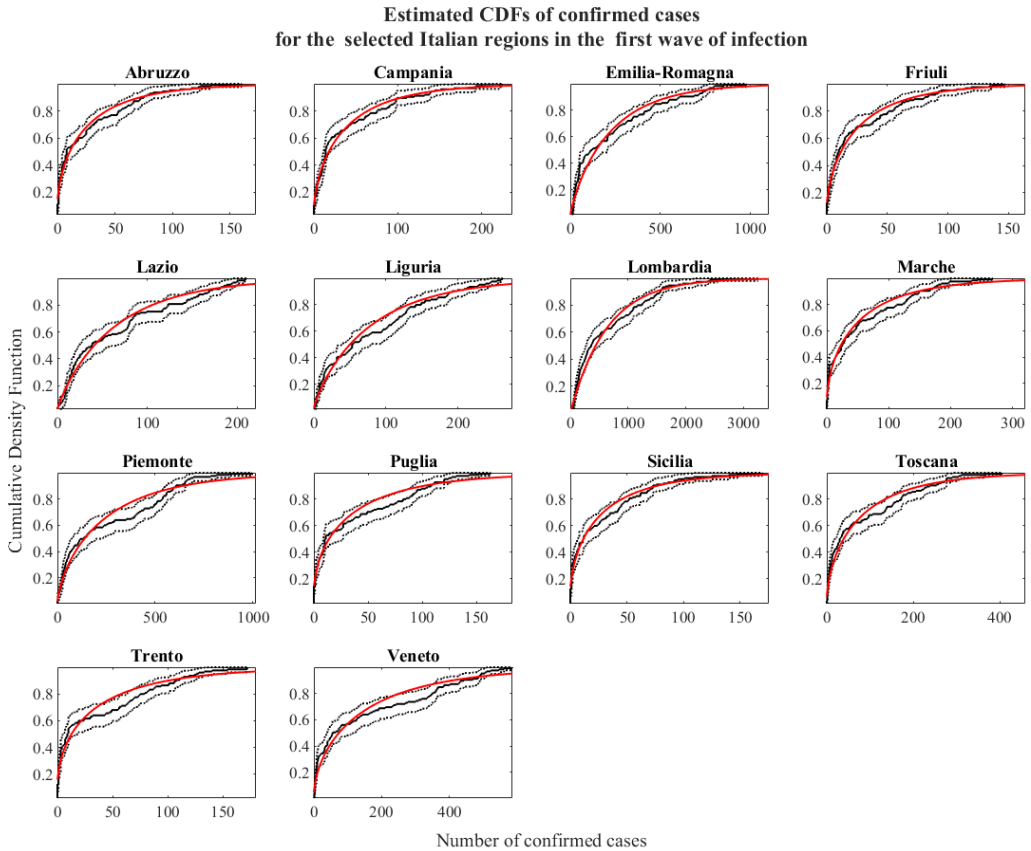


Figure 10. Negative Binomial PDFs (red line) of the confirmed cases for the group of 14 successful Italian regions, during the first waves of infection, was considered. Histograms of observed confirmed cases were also reported (bar plot). Black dashed lines represent upper and lower bounds, with 95%c.i..

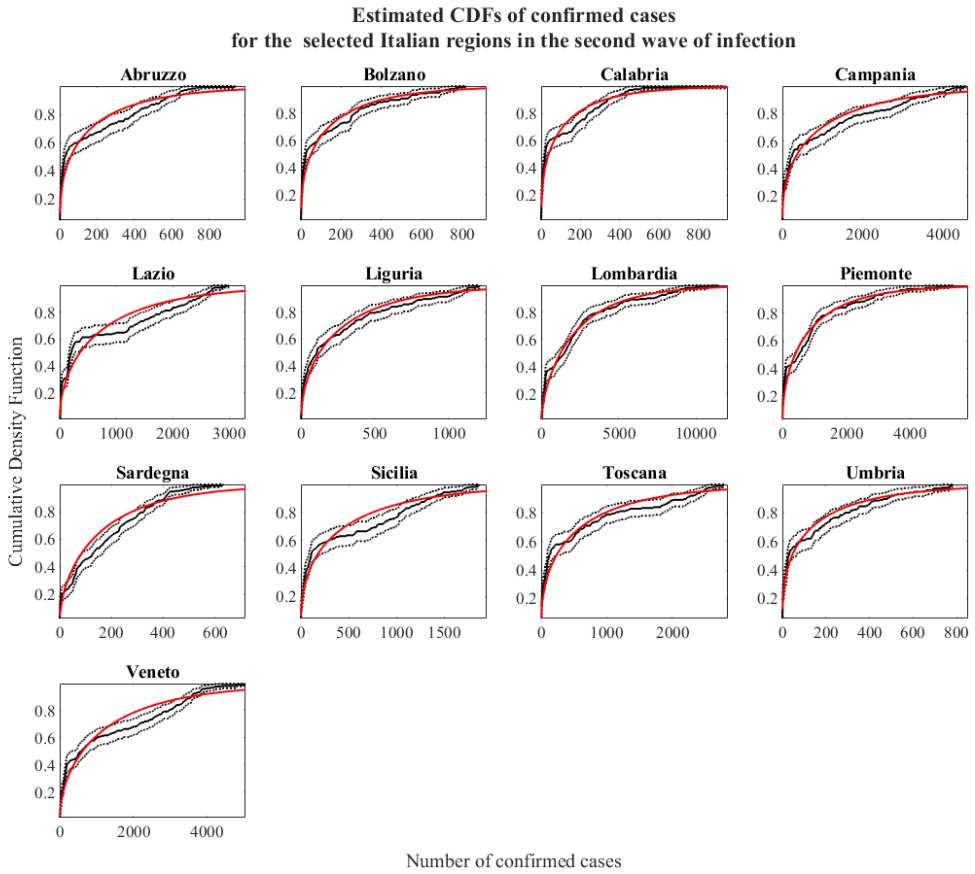


Figure 11. Negative Binomial CDFs (red line) of the confirmed cases for the group of 13 successful Italian regions, during the second waves of infection, was considered. Empirical CDFs of confirmed cases was reported (black line); black dashed lines represent upper and lower bounds, with 95%c.i..

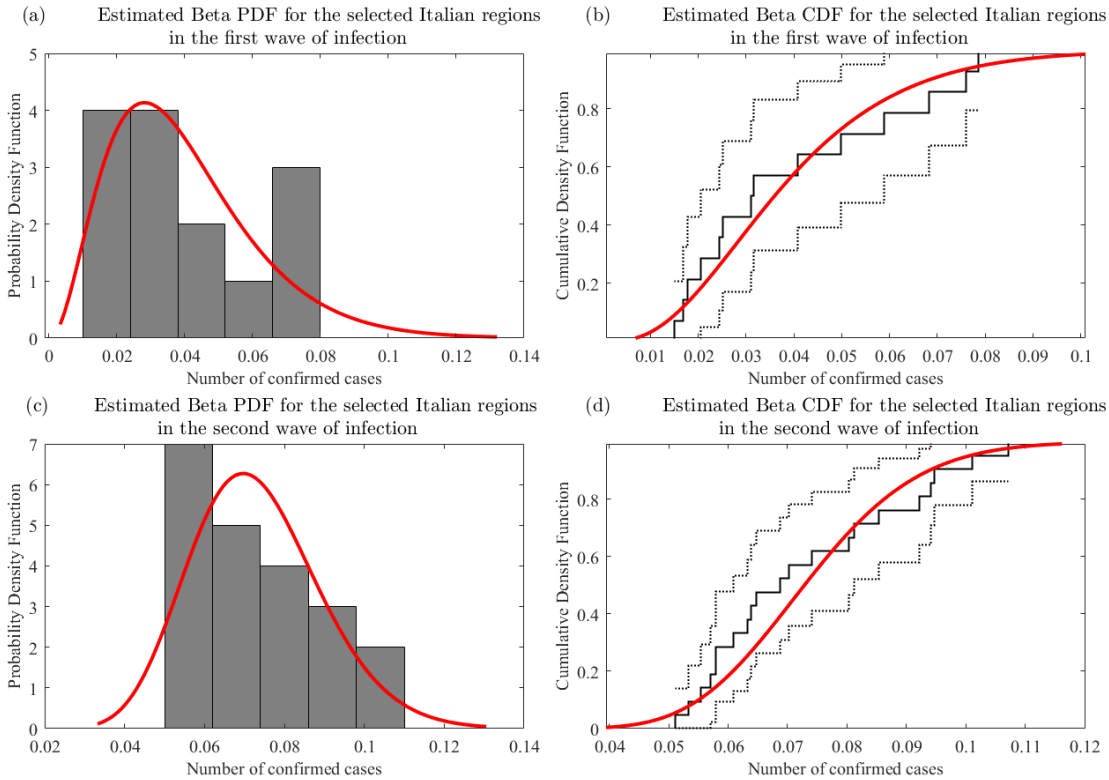


Figure 12. Beta PDFs of the ratio $R^{I,v}(\infty)$ (red line) with relative histogram (bar plot) during the first (a) and second (c) waves of infection for the Italian regions. Beta CDFs , respectively (b) and (d), corresponding to the above PDFs (red line); black dashed lines represent upper and lower bound, with 95%c.i.

Table 7. Clusters of Italian regions, in both waves of infection, based on number of executed tests and values of $r_i^{I,\nu}$ of Table 5 and Table 6.

| Cluster | Regions | $\sigma_p^{I,f}$ | $p-value$ | Cluster | Regions | $\sigma_p^{I,s}$ | $p-value$ |
|---|---|---|---|---|---|---|---|
| A | Abruzzo<br>Campania<br>Friuli<br>Marche<br>Puglia<br>Sicilia<br>Toscana<br>Trento | $1.87 \cdot 10^{-5}$ | 0.25 | A | Abruzzo<br>Campania<br>Emilia-Romagna<br>Friuli<br>Puglia<br>Sicilia | $1.29 \cdot 10^{-6}$ | 0.11 |
| B | Emilia-Romagna<br>Lazio<br>Liguria<br>Lombardia<br>Piemonte<br>Veneto | $2.12 \cdot 10^{-5}$ | 0.41 | B | Lazio<br>Liguria<br>Marche<br>Piemonte | $1.76 \cdot 10^{-6}$ | 0.34 |
| | | | | C | Lombardia<br>Trento | $6.67 \cdot 10^{-9}$ | 0.56 |

Table 8. Fitting parameters $\mu_r^{I,\nu}$ and $\mu_\theta^{I,\nu}$ of the Beta PDFs for the first (f.w.) and second (s.w.) waves of infection for the Italian regions. The asymptotic means $\rho^{I,\nu}(\infty)$ are also reported. In the last two columns we also report the sample means $\mu_r^{I,\nu}$ and variances $\sigma_r^{I,\nu}$ obtained by Table 5 and Table 6.

| Waves of infection<br>First (f.w.) \| Second(s.w.) | $\mu^{I,\nu}$ $(95\%c.i.)$ | $\mu_\theta^{I,\nu}$ $(95\%c.i.)$ | $\rho^{I,\nu}(\infty)$ $(95\%c.i.)$ | $\mu_r^{I,\nu}$ $(95\%c.i.)$ | $\sigma_r^{I,\nu}$ $(95\%c.i.)$ |
|---|---|---|---|---|---|
| f.w. | 2.85 (1.01 - 6.10) | 68.13 (30.31 - 112.02) | 0.04 (0.02 - 0.06) | 0.62 (0.39 - 0.65) | 0.03 (0.02 - 0.04) |
| s.w. | 21.81 (11.32 - 39.12) | 273.82 (115.04 - 401.10) | 0.07 (0.05 - 0.09) | 0.45 (0.37 - 0.53) | 0.02 (0.01 - 0.03) |

Successively we performed a one-tailed $\chi^2$-tests, with a significance level $\alpha$=0.05, on variances $\sigma_p^{I,\nu}$. In Table 7 we report clusters and p-values resulting from relative $\chi^2$-test. Moreover, these clusters seem to persist, with some exception, passing from the first to the second wave. Results are internally consistent and comparable with the European scenario. However, parameters $\mu^{I,\nu}$ and $\mu_\theta^{I,\nu}$, estimated by the $B(\mu^{I,\nu}, \mu_\theta^{I,\nu})$ of Figure 12, show some relevant deviation from European case that need to be discussed (see Table 8). In fact, while $\mu^{I,f}$ value can be still considered in agreement with previous results, $\mu^{I,s}$ is clearly far from all respective value changing $g$ or $\nu$. Actually, the reason why the strong differences between the two values $\mu^{I,\nu}$, estimated from Beta distributions, and the mean values $\mu_r^{I,\nu}$, obtained from the single regions, must be sought in that a small amount of data are available. Moreover, similarly to the European case, in the second wave the ratio $R_i^{I,s}$ have not reached yet stable values. Crossed t-tests comparing corresponding Italian parameters, as described in the section 3.1, were prevented due to the above discrepancies. Nevertheless, a look to the shows that values of $\rho^{I,f}(\infty)$=0.04 (95% c.i.: 0.02-0.06) and $\rho^{I,s}(\infty)$=0.07 (95% c.i.: 0.05-0.09) seem in agreement respectively with the first and second waves related to the European case , confirming the invariance property of the quantity $\rho^{g,\nu}(\infty)$ with respect to $g$ and its slight dependence from wave $\nu$.

## 4. CONCLUSION

Guided by heuristic evidence of some universality property, here we propose to describe the spread of (SARS-CoV-2)-infected pneumonia (COVID-19) within a probabilistic Polya urn scheme. Under universality conditions on initial value of the model parameters and applying a multiple waves approach, we analysed European data reported on confirmed cases and diagnostic test performed. A comparative analysis at regional and national scales was performed showing the presence of distinctive features according to the same underlying process at different scales. general characteristics can be extracted. Specific patterns and key indicators slightly depend on social or geographical conditions. On the other hand, some parameters seem to hold universality properties, properly identifying COVID-19 infection. A sequence of statistical tests was performed to prove our hypothesis. Based on test results, for each wave of infection, we were able to consider data from each European country, and Italian region, as *i* different sequences of trials of a process with different population but characterized by the same sample mean distribution. This allows us to estimate the incidence rate by using the asymptotic mean $\rho^{g,\nu}(\infty)$ of the sample average of the process. Resulting estimation of $\rho^{g,\nu}(\infty)$, related to the first and second wave of infection for the Italian case are broadly in line with European one and in agreement with real observations. Future directions of our research include the estimation of *CFR* and *IFR* indexes, since we believe that the quantity $\rho^{g,\nu}(\infty)$ could play a crucial role as a proxy variable for an unbiased estimation of the real incidence rate, including symptomatic and asymptomatic cases.


## ACKNOWLEDGEMENT

The authors would like to thank Dr. Maurizio Crippa for helpful discussion about insight of the present paper, and Riccardo Gioia, C.E.O of H-DATA S.r.l.s. for technical support for the graphical elaborations.


## REFERENCES


[1] World Health Organization, Report of the WHO-China joint mission on coronavirus disease 2019, 2019.

[2] J. Papenburg, M. Baz, M. E. Hamelin, Ch. Rhéaume, J. Carbonneau (+ 8 more authors), Household transmission of the 2009 pandemic A/H1N1 influenza virus: Elevated laboratory-confirmed secondary attack rates and evidence of asymptomatic infections, Clinical Infectious Diseases (2010), 51, pp. 1033–1041. DOI: 10.1086/656582

[3] B. J. Cowling, S. Ng, E. S. K. Ma, C. K. Y. Cheng, W. Wai (+ 6 more authors) Protective efficacy of seasonal influenza vaccination against seasonal and pandemic influenza virus infection during 2009 in Hong Kong, Clinical Infectious Diseases (2010), 51, pp. 1370–1379.
DOI: 10.1086/657311

[4] J. Ma, P. Van Den Driessche, Case fatality proportion, Bulletin of Mathematical Biology (2008), 70, pp. 118–133.
DOI: 10.1007/s11538-007-9243-8

[5] H. Nishiura, Case fatality ratio of pandemic influenza, Lancet. Infect. Dis. (2010), 10, pp. 443–444.
DOI : 10.1016/S1473-3099(10)70120-1

[6] S. N. Wood, E. C. Wit, M. Fasiolo, P. J. Green, Covid-19 and the difficulty of inferring epidemiological parameters from clinical data, Lancet. Infect. Dis. (2020), 10, pp. 443–444.
DOI: 10.1016/S1473-3099(20)30437-0

[7] T. W. Russell, J. Hellewell, C. I. Jarvis, K. van Zandvoort, S. Abbott, R. Ratnayake, CMMID COVID-19 working group (+ 4 more authors), Estimating the infection and case fatality ratio for coronavirus disease (covid-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020, Euro Surveill. (2020), 25.
DOI: 10.2807/1560-7917.ES.2020.25.12.2000256

[8] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, Ch. Whittaker (+ 28 more authors), Estimates of the severity of Coronavirus disease 2019: a model-based analysis, Lancet Infect. Dis. (2020), 20, pp. 669–677.
DOI: 10.1016/S1473-3099(20)30243-7

[9] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie (+ 26 more authors), Impact of non-pharmaceutical interventions (NPIS) to reduce Covid-19 mortality and healthcare demand, Imperial College London COVID-19 (2020).
DOI: 10.25561/77482

[10] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence, Nature (2005), 438, pp. 355–359.
DOI: 10.1038/nature04153

[11] M. Hayhoe, F. Alajaji, B. Gharesifard, A polya contagion model for networks, IEEE Transactions on Control of Network Systems (2018), 5, pp. 1998–2010.
DOI: 10.1038/nature04153

[12] G. Polya, Sur quelques points de la théorie des probabilités, Annales de l'institut Henri Poincare (1930), 1, pp. 117–161. Online [Accessed 21 September 2023] [In French]
http://eudml.org/doc/78952

[13] F. Eggenberger, G. Polya, Uber die Statistik Verketteter Vorgänge, Z. Angew. Math. Mech. (1923), 3, pp. 279–289. [In German]
DOI: 10.1002/zamm.19230030407

[14] M. Lipsitch, S. Riley, S. Cauchemez, A. C. Ghani, N. M. Ferguson, Managing and reducing uncertainty in an emerging influenza pandemic, New England Journal of Medicine (2009), 361, pp. 112–115.
DOI: 10.1056/NEJMp0904380

[15] K. Teerapabolarn, An improved bound for negative binomial approximation with z-functions, AKCE International Journal of Graphs and Combinatorics (2017), 14, pp. 287–294.
DOI: 10.1016/j.akcej.2017.04.005

[16] W. Feller, An introduction to probability theory and its applications, vol. ii, 2nd edn ,John Wiley, New York, 1971.
DOI: 10.1063/1.3062516

[17] N. L. Johnson, S. Kotz, Urn models and their application, John Wiley, New York, 1977.

[18] S. Kotz, N. Balakrishnan, Advances in combinatorial methods and applications to probability and statistics, Birkhauser, Boston, 1997.

[19] D. Aoudia, F. Perron, A new randomized Pólya urn model, Applied Mathematics (2012), 3, pp. 2118–2122.
DOI: 10.4236/am.2012.312A292

[20] M. Chen, C. Z. Wei, A new urn model, Journal of Applied Probability at JSTOR (2005), 42, pp. 964–976.
DOI: 10.1038/nature04153

[21] G. Aletti, I. Crimaldi, Generalized Rescaled Pólya urn and its statistical application, Electron. J. Statist. (2022), 16(1), pp. 1635-1680.
DOI: 10.1214/22-EJS1993

[22] Humanitarian Data Exchange, United Nations Office for the Coordination of Humanitarian Affairs. Online [Accessed 21 September 2023]
https://data.humdata.org/event/covid-19

[23] Italian civil protection agency, Italian COVID-19 data. Online [Accessed 21 September 2023]
https://github.com/pcm-dpc/COVID-19

[24] European centre for disease prevention and control, COVID-19. Online [Accessed 21 September 2023]
https://www.ecdc.europa.eu/en/covid-19

[25] H. Ward, Chr. Atchison, M. Whitaker, K. E. C. Ainslie, J. Elliott (+ 9 more authors), Sars-Cov-2 antibody prevalence in England following the first peak of the pandemic, Nature Communications (2021), 12, 905.
DOI: 10.1038/s41467-021-21237-w

[26] M. Pollán, B. Pérez-Gómez, R. Pastor-Barriuso, J. Oteo, M. A. Hernán (+ 104 more authors), Prevalence of Sars-Cov-2 in Spain (ene-covid): a nationwide, population-based seroepidemiological study, Lancet (2021), 396(10250), pp. 535–544.
DOI: 10.1016/S0140-6736(20)31483-5