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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A predictive model and country risk assessment for COVID-19: An application of the Limited Failure Population concept

Themistoklis Koutsellis, Alexandros Nikas*

School of Electrical and Computer Engineering, National Technical University of Athens, Greece

ABSTRACT

This article provides predictions for the spread of the SARS-CoV-2 virus for a number of European countries and the United States of America, drawing from their different profiles, both socioeconomically and in terms of outbreak and response to the 2019–2020 coronavirus pandemic, from an engineering and data science perspective. Each country is separately analysed, due to their differences in populations density, cultural habits, health care systems, protective measures, etc. The probabilistic analysis is based on actual data, as provided by the World Health Organization (WHO), as of May 1, 2020. The deployed predictive model provides analytical expressions for the cumulative density function of COVID-19 curve and estimations of the proportion of infected subpopulation for each country. The latter is used to define a Risk Index, towards assessing the level of risk for a country to exhibit high rates of COVID-19 cases after a given interval of observation and given the plans of lifting lockdown measures.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Time series forecasting is of paramount importance for many real-life scenarios [1]. Often, it is the base ground for many decision-making procedures [2], including health-related issues [3], be that human resource requirements [4], expenditures calculation [5,6], or pandemic preparedness [7]. However, a variation of forecasting methods, based on different assumptions and data utilisation techniques, yield different predictions and statistical inferences per case scenario, significantly affecting the precision of forecasting and the decisions to be taken [8]. In this study, drawing from the associated urgency, we deploy a predictive model customised for the case of the COVID-19 pandemic.

Sometimes, forecasting models cannot represent real-world processes [9], with some state-of-the-art methods, e.g. ARIMA models [10], failing to capture the underlying trend of a sequence of events; for example, an analytical expression of trend may not exist or either differentiation or log-transformation may not yield trend removal. In that case, such analysis would not suffice to abstract the trend and make the process stationary, which would enable the analyst to proceed thereafter to forecasting. Moreover, such models, which are typically trained by previous time series curve shapes, do not analyse the procedure of the underlying apparatus that creates the sequence of events.

Besides the drawbacks of state-of-the-art techniques on how they handle and extract information from observed data, there are limitations derived from unrealistic assumptions. In the case of the 2019–2020 novel coronavirus pandemic, drastic changes in patient testing and case recording approaches [11], focus on specific symptoms [12], assumptions driven by previous experiences [13], or technical limitations to daily testing capacity [14] may lead to optimistic forecasts. In contrast, assumptions that the entire population will eventually display COVID-19 symptoms, or unfounded expectations of when this will happen and how it can be handled [15,16], may result in overestimation of near-future COVID-19 cases and therefore in varying, often unrealistic levels of alert for governments [17]. Not to mention that assumptions are made based on international knowledge, despite the employment of significantly different testing approaches across the globe [18]. As with other health emergencies and pandemics, enhancing accuracy of forecasting will facilitate governments in preparing for the COVID-19 pandemic [19].

Motivated by these challenges, this research proposes a new approach to forecasting the COVID-19 pandemic spread and putting together country risk profiles, based on the principles of Limited Failure Population (LFP) [20–34] and Truncated Data (TD) [30–33]. In particular, the proposed approach aims to tackle the challenge associated with the unrealistic assumption that the entire population will eventually display COVID-19 symptoms. In this respect, this research classifies the population into infected and healthy sub-

---

* Corresponding author: Decision Support Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, Iiron Politechnio 9, 157 80, Athens, Greece.

E-mail address: anikas@epu.ntua.gr (A. Nikas).

https://doi.org/10.1016/j.chaos.2020.110240
0960-0779/© 2020 Elsevier Ltd. All rights reserved.
populations; it also provides a risk index to assess the levels of risk for a country to exhibit high rates of COVID-19 cases after a given interval of observation. To our knowledge, such an index has not yet been introduced or used in the literature.

1.1. Definition of Limited Failure Population (LFP)

The principle of LFP is applied when the population is not homogeneous, i.e. when its statistics are not described by a unimodal distribution. In case of a pandemic catastrophe, it is unrealistic to assume that the entire population will exhibit symptoms or be carrier of the SARS-CoV-2 virus. Only a proportion of the population has the propensity to get sick; the rest will be uninfected (“healthy”) subpopulation. In that case, the underlying distribution of getting infected is bimodal. The second mode — corresponding to “healthy” subpopulation — lies on a region close to infinity (see Fig. 1), which practically means that they will never show COVID-19 symptoms. On the other hand, the first mode is for those who are, at some point, going to get infected.

Each mode is separated from and, practically, do not overlap with one another. Also, the interval of observation is considered short and all observed COVID-19 cases are only related to the first subpopulation (“infected”) subpopulation. These assumptions define the LFP model and ensure that any recorded COVID-19 case is only related to the “infected” subpopulation.

In Fig. 1, the support, $T_s$, of the first mode (“infected” subpopulation) is defined as the time interval where most – approximately 99% – of COVID-19 cases will occur. That is, the interval where approximately 99% of the area of the probability density function (PDF) of the “infected” subpopulation lies on. The censoring time $T_c$ is the time interval where all observed data lies on; by definition, this data is related to the first mode. The observed data is used to train the deployed predictive model. In most practical cases, the following inequality holds between censoring time and support:

$$T_c \ll T_s,$$

which means that the observing data is limited.

If $p$ indicates the portion of COVID-19 cases up to the end of the pandemic, then the area of the first mode PDF is equal to $p$ and the area of the second mode (“healthy” subpopulation) PDF equals to $1 - p$, so that the underlying area of the entire PDF equals to $1$. If $f_t$ the first mode PDF, $f_h$ the second mode PDF and $f$ the entire PDF, then

$$f(t) = p \cdot f_t(t; \theta) + (1 - p) \cdot f_h(t).$$

where $\int_0^\infty f(t) dt = 1$ and $\theta$ the parameters of the underlying $f_t$ distribution.

1.2. Definition of truncated data (TD)

In real-life applications, there is a bound beyond which we intentionally do not, or due to limitations cannot, collect data in the experimental procedure. In such cases, there are two possible sets of data: censored and truncated.

The difference between these two data types lies in the level of knowledge one may have regarding the number and time of events to occur after said bound. If we know the exact number of events that will occur right after a certain bound (the censoring time $T_c$; see Fig. 1), but we do not know when the corresponding events will occur, the collected data in $[0, T_c]$ is called Censored Data (CD). In contrast, if we do not know neither the exact number of events at time $t > T_c$ nor when the corresponding events will occur, the collected data is called Truncated Data (TD). By definition, the CD is a special case of TD if the number of events at time $t > T_c$ is known.

In the case of SARS-CoV-2, the proportion of “infected” population (first mode; Fig. 1) is unknown; equivalently, the length of the support of the first mode, interval $[0, T_c]$, is unknown. Therefore, the collected data in $[0, T_c]$ is truncated – we do not know neither how many nor when COVID-19 cases beyond bound $T_c$ will occur.

1.3. Parameter estimation of LFP model

By LFP definition, all observed COVID-19 cases are only related to the “infected” subpopulation (see Section 1.2). Therefore, the second term of the right-hand side of Eq. (2) is almost zero in the time domain where the first mode is dominant and therefore Eq. (1) becomes

$$f(t; p, \theta) \approx p \cdot f_t(t; \theta).$$

In terms of the Cumulative Distribution Function (CDF), Eq. (3) becomes

$$F(t; p, \theta) \approx p \cdot F_t(t; \theta).$$

If $r_0$ is the observed COVID-19 cases in $[0, T_c]$, then the value of $F(T_c; p, \theta)$ at $T_c$ is

$$F(T_c; p, \theta) \approx \frac{r_0}{n},$$

where $n$ is the total number of people. Combination of Eqs. (4) and (5) yields

$$p \approx \frac{r_0}{n \cdot F_t(T_c; \theta)}.$$  

To estimate the parameters of the LFP model, we only need to estimate the parameters $\theta$ of $F_t$. Then, the proportion $p$ can be estimated by using Eq. (6).

We cannot apply the widely used Maximum Likelihood Estimation (MLE) approach to estimate parameters $\theta$ due to several limitations. The MLE likelihood function is often incomplete [25]. Moreover, the correct likelihood function formula (see [25,26] and [30–32]) yields erroneous results in the case of TD [33]. In this article, the deployed method uses an approach similar to the one described in [33]. Various types of underlying distributions (Weibull, Log-normal, Gamma, Dagum, Chi and Rayleigh) are tested against observed data to find optimal solution. The optimal solution is observed when the Dagum underlying distribution is assumed. Following the estimation of parameters $\theta$ and $p$, the provided predictive model gives a) a predictive curve for future COVID-19 cases and b) an estimation of the proportion $p$ of the infected population. The latter, in comparison with the observed data in $[0, T_c]$, will give a Risk Index (RI) per country. RI indicates how many COVID-19 cases would be recorded after time instance $[0, T_c]$. It is, say, an infection potential related to the number of unknown but
already existing COVID-19 cases plus new COVID-19 cases, all of which will emerge after $[0, T_c]$. Section 2 defines the LFP model with the Dagum underlying distribution and lists all assumptions employed in the research. It also describes the proposed approach and gives estimates of $\theta$, $p$ and $R_l$. In Section 3 the proposed method is applied to a diverse pool of countries, including European countries of various profiles and the United States of America (USA), using data from January 20, 2020 to May 1, 2020; the latter is the day of gradually lifting lockdown for most countries. Finally, Section 4 summarises the research findings and recommends future work.

2. Truncated data and truncated CDF

In most practical applications, the interval of observation $[0, T_c]$ is short (see Fig. 1) and the observed data limited for statistical inferences. In the case of LFP, the truncated data includes realisations from the left tail of the underlying distribution $F_t$ (see Eq. (4)). At the same time, the truncated data includes realisations from a conditional CDF, $F_T$, which we call “truncated CDF.” $F_T$ is the distribution of realisations of $F_t$ in the $[0, T_c]$ interval and provides the probability of COVID-19 cases under the condition that these cases occur in $[0, T_c]$. Evidently, the conditional distribution $F_T$ is only defined in $[0, T_c]$. The $F_T$ distribution depends on the $F_t$ distribution. Their relation is used in this article to estimate the parameters of the latter if the empirical estimation of the former is known. An empirical estimator, $\hat{F}_t$, of $F_t$ is not available because we do not have sample points after time $T_s$, i.e., the data is truncated. However, we do have sample points in $[0, T_c]$, where all realisations of $F_t$ lie (see Fig. 2). The symbol “x” in Fig. 2 indicates a COVID-19 event in $[0, T_c]$. Note that $F_T(T_c) = 1$ and $F_T(T_c) < 1$, therefore $F_T \leq F_t$ in $[0, T_c]$.

If $r_0$ is the number of observed failures in $[0, T_c]$, the empirical estimator of $F_T$ is

$$\hat{F}_{T,i} = \frac{i}{r_0},$$

where $i$ is the $i$th time instance of observed COVID-19 cases in $[0, T_c]$, after all COVID-19 cases are sorted in increasing order in terms of time.

2.1. Relationship between CDF of infected population $F_t$ and truncated CDF $F_T$

The relationship between $F_t$ and $F_T$ is derived as follows. The probability a realization of $F_t$ is in $[0, T_c]$ is

$$Pr[T < t \cap T < T_c] = \frac{Pr[T < t]}{Pr[T < T_c]}, t < T_c.$$  \hspace{1cm} (8)

Because $t < T_c$, the numerator of Eq. (8) is

$$Pr[T < t \cap T < T_c] = Pr[T < t].$$ \hspace{1cm} (9)

yielding

$$Pr[T < T_c \mid T < t < T_c] = \frac{Pr[T < t]}{Pr[T < T_c]}, t < T_c.$$ \hspace{1cm} (10)

All collected data in $[0, T_c]$ follows this conditional CDF which is the truncated CDF $F_T$. The denominator of Eq. (10) is equal to $F_T(T_c)$ and the numerator is equal to $F_T(t)$. Thus,

$$F_T(t; \theta) = \frac{F_t(t)}{F_t(T_c)}, 0 \leq t \leq T_c.$$ \hspace{1cm} (11)

From Eq. (11), it can be derived that, if $T_c \to \infty$, then $F_T(t; \theta) \to F_t(t)$. 2.2. Estimating LFP parameters from observed data

The observed COVID-19 cases can be used to provide estimates of LFP parameters. By using the relationship between the CDF of the COVID-19 population $F_t$ and truncated CDF $F_T$, the problem of finding the LFP parameters ends up to an optimisation problem of finding a best fitting curve.

Combining Eqs. (6) and (11) yields

$$F_T(t; \theta, p) = \left( \frac{n - p}{T_0} \right) \cdot F_t(t; \theta).$$ \hspace{1cm} (12)

Given the values of $\hat{F}_{T,i} -$ derived from data of observation — and after assuming a type of $F_t(t; \theta)$ distribution, the LFP parameters, $\theta$ and $p$, can be found by solving the following optimisation problem

minimize : $0.5 \sum_{i=0}^n R\left( F_T(t; \theta, p) - \hat{F}_{T,i} \right)^2.$ \hspace{1cm} (13)

subject to:

$$0.2 - 0.05 \leq F_T(t_s; \theta, p) \leq 0.2 + 0.05$$

$$0.6 - 0.05 \leq F_T(t_s; \theta, p) \leq 0.6 + 0.05$$

$$0.8 - 0.05 \leq F_T(t_s; \theta, p) \leq 0.8 + 0.05$$

$$F_T(T_c; \theta, p) = 1$$

$$\frac{2}{3} \leq p \leq 1$$

where $F_T(t; \theta, p) = \left( \frac{n - p}{T_0} \right) \cdot \hat{F}_t(t; \theta)$ (see Eq. (12)) and $\hat{F}_{T,i}$ are derived from Eq. (7). Function $R(\cdot)$ is used to reduce the influence of outliers on the solution. In this article we use the smooth approximation of the $R(\cdot)$ loss function

$$R(z) = 2 \cdot \left( \sqrt{1 + z^2} - 1 \right).$$ \hspace{1cm} (15)

Other types of $R(\cdot)$ functions were also tested but their performance proved insufficient in the case of our SARS-CoV-2 spread application. We also used the Trust Region Reflective algorithm to solve the optimisation problem. Values $t_a$, $t_b$ and $t_f$ are the time instances, where $\hat{F}_t$ equals to 0.2, 0.6 and 0.8, respectively.

The objective function of the optimisation problem is based on the least squares error method, which is a linear regression technique. Therefore, when applied to non-linear models ($F_T(t)$ is not linear), it may suffer from non-linearity issues; the fitting curve will not be able to capture the relationship between $\hat{F}_T$ and $t$. To avoid such cases, we constrain the solution of the optimisation of problem in a region close to the observed $\hat{F}_T$ (see three first constraints). We constrain the optimal solution in $a \pm 0.05$ region from $\hat{F}_T$. The fourth constraint is required, so that the estimated $F_T(t; \theta, p)$ curve satisfies the fundamental property of COVID-19 at the end of its right tail. Finally, the fifth constraint bounds the $p$ parameter. The percentage of infected people cannot be less than the sample percentage, $p_0 = \frac{r_0}{n}$, in $[0, T_c]$ nor greater than one, by definition ($p_{\text{max}} = \frac{n}{n} = 1$).
Fig. 3. USA - Predictive curve against training and validity data.

Fig. 4. Sweden - Predictive curve against training and validity data.

Fig. 5. Belgium - Predictive curve against training and validity data.
2.3. The Dagum distribution

In Section 2.2, the deployed optimisation problem requires to pre-assume a certain type of $F_i(t)$ CDF. To do so, we set up the following procedure. We test various continuous distributions with support equal to $[0, \infty]$; Weibull, Log-normal, Gamma, Dagum, Chi and Rayleigh. A subset of the historical (observed) values must be used for training the model; although there is no specific rule (in the literature the training data can be anything, from 80% [35] to 95% [36]), our aim is to adequately feed the model with input data as well as be able to validate the predictions with a sufficient dataset. We therefore used 85% of the observed data to train our predictive model (input of the optimisation problem) for each country and the remaining 15% of data to validate the predictions (validity data). To access the model and the corresponding assumption of the underlying distribution, we used the Root Mean Square Error (RMSE) between predicted and validity data. For all countries and various assumed intervals of observations, the Dagum distribution yielded results closer to actual COVID-19 data (smallest RMSE values). Therefore, we concluded that the underlying distribution of the pandemic resembles Dagum, which has the following formula:

$$F_i(t) = \left(1 + \left(\frac{x}{\beta}\right)^{\frac{1}{\alpha}}\right)^{-\gamma}.$$  

(16)

The parameters $\alpha$, $\beta$ and $\gamma$ of the Dagum distribution constitute the $\theta$ vector of the optimisation problem of Section 2.2 (see Eqs. (13) and (14)) and are to be found based on the observed data.

2.4. Definition of risk index ($RI$)

Amongst the estimated LFP parameters of the coronavirus pandemic, $p$ provides the total infected subpopulation. This proportion includes three factors: the first factor is the observed proportion, $p_0 = \frac{T_C}{T}$, in $[0, T_C]$; the second factor, $p_u$, is the number amongst the proportion of infected population during $[0, T_C]$ representing individuals that have not yet exhibited any symptoms — they will probably show symptoms at time instance $t < T_C$; and the third factor, $p_f$, refers to all those that will be infected after the censoring time $[0, T_C]$.

$$p = p_0 + p_u + p_f.$$  

(17)

The difference $p - p_0 = p_u + p_f$ indicates how many more COVID-19 cases will happen after the interval of observation $[0, T_C]$, called potential of infection hereafter. After solving the optimisation problem of Section 2.2 (see Eqs. (13) and (14)), the estimated $\hat{p}$ provides an estimate of the infection potential. The higher the difference $p - p_0$, the higher the risk to exhibit a higher rate of COVID-19 cases in the near future. We define Risk Index (RI) as the
absolute error between \( p_o \) and \( \hat{p} \)

\[
RI = \left( \frac{\hat{p} - p_o}{p_o} \right) 
\]· 100%. \hspace{1cm} (18)

Note that \( RI \) indicates the risk about the future of a country and indicates whether we are safe to assume that the cumulative pandemic curve reached a point of saturation. It does not indicate, though, whether a country already has high levels of COVID-19 cases. This is described by the observed data only. Therefore, there might be countries with already high existing levels of COVID-19 cases but low estimated \( RI \). This means that despite the high levels of COVID-19 cases there are few unrecorded infected people circulating around amongst the population and, therefore, few extra people will get sick in the near future. The pandemic cumulative curve reached an end.

3. Application of proposed predictive model

We applied the method of Section 2 for twelve countries: Austria, Belgium, France, Germany, Italy, the Netherlands, Portugal, Spain, Sweden, Switzerland, the United Kingdom (UK), and USA. These countries constitute an interesting pool, as they had all been infected by the novel coronavirus pandemic during the examined period, to different scales, while featuring significantly diverse conditions and profiles, in terms of state responses to the pandemic.

The data used comes from the officially released reports of the World Health Organization (WHO) (Reports 1–102), as of May 1, 2020 [37]. The censoring time \( T_C \) is May the 1st, 2020; the day of lifting lockdown measures for most of the examined countries. We assumed the Dagum underlying distribution for all countries with \( \alpha, \beta \) and \( \gamma \) parameters (see Section 2.3). We split data into training and validity, with an 85:15 ratio. After solving the optimisation problem (see Section 2.2), we estimated all LFP parameters (\( \hat{\alpha}, \hat{\beta}, \hat{\gamma} \) and \( \hat{p} \)) per country. All LFP parameters are then used to evaluate the predictive curve. The underlying CDF provides an estimation of the sequence of occurrences of COVID-19 cases. Therefore, the underlying distribution multiplied by the number of the estimated infected population gives the predictive curve \( W(t) \).

\[
W(t) = \hat{F}_s(t; \alpha, \beta, \gamma) \cdot \hat{p} \cdot n, \hspace{1cm} (19)
\]

where \( n \) is the total volume of population of a country. Finally, the estimated \( \hat{p} \) was also used to evaluate \( RI \) per country. It should be noted that there is a previous attempt (see [38]) to identify the parameter \( \hat{p} \) for some European countries by using the terminology of “attack rate.” We believe that the proposed method and the concepts of LFP and TD completes this attempt from both a practical and a theoretical perspective.

Finally, Section 3.1 illustrates the predicted curves against the training and validity data and provides the estimates of LFP pa-
rameters. Section 3.2 summarises the risk index per country and provides a physical interpretation of the results.

3.1. Predictions per country

In Figs. 3–14, the ratio between training (red dots) and validity (green squares) data remains the same per country (85:15). As there is a different starting point (the time instance when the first COVID-19 case occurred) and number of observed data for each country, the number of validity date in not the same in every Figure.

Figs. 3–14 provide evidence of the validity of the proposed predictive model. In general, the proposed method provides forecasting values close to validity data for most of the countries, i.e. USA, Sweden, Portugal, Italy, Spain, Switzerland, the UK, and Austria. For Belgium and Germany, the predicted values are close to validity data up to a certain point, after which there is a slight overestimation of future COVID-19 cases. Only France and Netherlands, i.e. two out of twelve examined countries, seem to give overestimated results for all validity data. However, these represent a small share of the country pool and display an overestimation performance of the proposed model, thereby posing insignificant doubts over the accuracy of the proposed method in general.

Below, Table 1 summarises the parameters of the underlying Dagum distribution (see Eq. (16)) per country.

| Country   | Dagum parameters α | β     | γ     |
|-----------|--------------------|-------|-------|
| USA       | 3.22               | 32.65 | 29.24 |
| Sweden    | 1.44               | 18.13 | 21.58 |
| Belgium   | 3.07               | 19.02 | 53.64 |
| Portugal  | 2.20               | 19.00 | 4.90  |
| Italy     | 2.67               | 14.61 | 49.97 |
| Spain     | 3.88               | 27.60 | 25.61 |
| Germany   | 4.01               | 29.29 | 26.21 |
| France    | 4.21               | 33.35 | 25.49 |
| Switzerland | 2.65          | 11.34 | 14.95 |
| GLOBAL    | 1.54               | 21.00 | 25.50 |
| UK        | 3.09               | 28.03 | 27.81 |
| Austria   | 3.30               | 13.23 | 11.98 |
| Netherlands| 0.93              | 9.65  | 10.85 |

3.2. Risk index (RI) per country

Table 2 summarises the RI results for each country. Here, the interval of observation is from January 20, 2020 to May 1, 2020, representing the day of lifting lockdown in most of the considered
these countries (see Table 2). The observed COVID-19 cases, $N_o = p_o \cdot n$, where $n$ is the population of USA, amount to $N_o = 0.00316 \cdot 327,200,000$ or $N_o = 1,033,952$. The estimated infected population, $\hat{N}_{\text{infected}} = \hat{p} \cdot N_o$, is equal to $\hat{N}_{\text{infected}} = 0.00586 \cdot 327,200,000$, or $\hat{N}_{\text{infected}} = 1,917,392$, almost twice the size of $N_o$. Besides USA, Sweden appears to also exhibit high levels of both COVID-19 rates and $RI$. Sweden’s coronavirus response was based to the “herd immunity” principle and thus the model expectedly yields high levels of $RI$. Following these two, other countries exhibiting rela-

Table 2

| Country | Observed proportion of infected population, $p_o$ | Estimated proportion of infected population, $p_e$ | Risk Index ($RI$) |
|---------|-----------------------------------------------|-----------------------------------------------|------------------|
| USA     | 0.00316                                      | 0.00586                                      | 85.44            |
| Sweden  | 0.00208                                      | 0.00330                                      | 58.65            |
| Belgium | 0.00424                                      | 0.00647                                      | 52.60            |
| Portugal| 0.00243                                      | 0.00347                                      | 42.80            |
| Italy   | 0.00340                                      | 0.00481                                      | 41.47            |
| Spain   | 0.00521                                      | 0.00703                                      | 34.93            |
| Germany | 0.00192                                      | 0.00234                                      | 21.88            |
| France  | 0.00191                                      | 0.00225                                      | 17.80            |
| Switzerland | 0.00344                                  | 0.00397                                      | 15.41            |
| GLOBAL  | **0.00041**                                  | **0.00047**                                  | **14.63**        |
| UK      | 0.00258                                      | 0.00279                                      | 08.14            |
| Austria | 0.00175                                      | 0.00187                                      | 06.86            |
| Netherlands | 0.00229                              | 0.00243                                      | 06.11            |

countries. Hence, $RI$ provides a risk assessment of exiting quarantine.

Globally, the $RI$ is relatively low. This means that few unknown COVID-19 cases are circulating around the globe. However, this $RI$ level is not uniformly distributed. There are countries exhibiting high $RI$, which means that the termination of the lockdown period may yield high COVID-19 rebounds, if no additional preventive measures are taken to outbalance the effect of lifting lockdown measures. USA also appears to feature the highest risk amongst...
tively high risk levels include Belgium, Portugal, Italy and Spain, i.e. countries with large international transit areas, multiple and large clusters, and/or aged population that is concentrated in nursing home facilities instead of family-orientated treatment.

In conclusion, it seems that the physical interpretation of the risk index is consistent with the actual background, underlying conditions and pandemic profile of the examined countries, providing strong evidence of the usefulness of the proposed index.

4. Summary and future work

In this research, we introduced a new predictive model, aimed at helping forecast COVID-19 cases using the principles of Limited Failure Population and Truncated Data within an interval of observation. The proposed framework was applied to twelve countries of diverse profile, in socioeconomic and geographic terms but also in terms of infection and response to the COVID-19 pandemic. These included Austria, Belgium, France, Germany, Italy, the Netherlands, Portugal, Spain, Sweden, Switzerland, the UK, and the USA. The model provides acceptable accuracy, when compared against real data (validity data). A risk index is also introduced to assess the level of risk for a country to exhibit high rates of COVID-19 cases in the near future, based on the cut-off date of validity data representing the actual or approximate date of lifting the strictest lockdown measures across the country pool.

It should be noted that, although the risk index results seem to replicate or be consistent with the underlying conditions and COVID-19 spread profiles of the examined countries, attributes additionally to the data inputs described were not explicitly modelled. For example, circumstances related to testing and hospitalisation capacity or health system resilience are hardly represented implicitly in the actual data. The same can be said for the uncertainty of future advancements, which is completely overlooked in the proposed model, meaning that our forecasting exercise from a strictly engineering perspective assumes that certain conditions will remain the same after a given interval of observation: the SARS-CoV-2 transmission rate does not change with changes in weather conditions, the population has not become immune to the novel coronavirus (i.e. no “herd immunity”), there emerge no additional, more dangerous SARS-CoV-2 mutations with different spread capacity or severity of symptoms, etc.

In the future, we aim to draw from state-of-the-art forecasting models in the literature to carry out a comparative analysis, apply the model in ex-post analysis based on different benchmarks, and provide confidence intervals for each parameter of the Limited Failure Population model as well as prediction interval for the provided predictive curve. It should also be noted that the predictive model does not necessarily apply in consideration of coronavirus-related deaths and recovered cases, and the underlying Dagum distribution is found suitable only for the case of the infected population. With regard to the latter, the overall population of a country is a priori known; this is not the case for the deaths and recovered cases, which essentially are a proportion of the infected population rather than the overall population. Applying the proposed model, in consideration of the deaths and recoveries as a function of the infected population may yield small values for the parameter $p$ (see Table 2), which could in turn yield inconclusive results. As such, the deployment of a predictive model for the recovery and death cases is another subject of future research to ensure the accuracy and efficiency of our predictions. Finally, another prospect lies in a sensitivity analysis of the Risk Index in respect to the Dagum parameters.

Like other global emergencies and sustainability challenges [39,40], understanding and effectively tackling the 2019–2020 novel coronavirus pandemic is a long-term process, which requires that a diversity of tools be employed, from various scientific areas and interdisciplinary perspectives [41–46]. The model proposed in this research simply aims to contribute by providing some mathematical tools from an engineering and data science perspective. Hopefully, this model, in combination with theory and tools from the areas of epidemiology and bio-engineering, can pave the ground in understanding this pandemic.

Credit author statement

Themistoklis Koutsellis: Conceptualization, Methodology, Software, Writing, Original draft preparation, Visualization.
Alexandros Nikas: Data curation, Writing, Supervision, Final draft preparation, Reviewing and Editing.

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.

CRediT authorship contribution statement

Themistoklis Koutsellis: Conceptualization, Formal analysis, Methodology, Supervision, Validation, Visualization, Writing - origi-
inal draft. **Alexandros Nikas**: Data curation, Supervision, Validation, Writing - original draft, Writing - review & editing.

**Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**References**

[1] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. OTexts; 2018.
[2] Shim JK. Strategic business forecasting: the complete guide to forecasting real world company performance. CRC Press; 2000.
[3] Soysin IN, Reidpath DD. An overview of health forecasting. Environ Health Prev Med 2013;18(1):11.
[4] O’Brien-Pallas L, Baumann A, Donner G, Murphy GT, Lochhaas-Gerlach J, Luba M. Forecasting models for human resources in health care. J Adv Nurs 2001;33(1):120–9.
[5] Lee R, Miller T. An approach to forecasting health expenditures, with application to the US Medicare system. Health Serv Res 2002;37(5):1365–86.
[6] Weiner JP. Forecasting the effects of health reform on US physician workforce requirement: evidence from HMO staffing patterns. JAMA 1994;272(3):222–30.
[7] Myers MF, Rogers DJ, Cox J, Pahalaut A, Hay SI. Forecasting disease risk for increased epidemic preparedness in public health. Adv Parasitol 2000;47:309.
[8] Sims CA. Are forecasting models usable for policy analysis? Q Rev, (Win) 1986:2–16.
[9] Dagum P, Galper A, Horvitz E, Seiver A. Uncertain reasoning and forecasting. Int J Forecast 1995;11(1):73–87.
[10] Khashie M, Bijiari M, Ardali GAR. Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (ANNs). Neurocomputing 2009;72(4–6):956–67.
[11] Lipstich M, Sverdlov D, Finelli L. Defining the epidemiology of Covid-19—Studies needed. New Engl J Med 2020;382(13):1194–6.
[12] Klompas, M. (2020). Coronavirus Disease 2019 (COVID-19): protecting hospitals from the invisible. Ann Intern Med.
[13] Nishiura, H., Linton, NM., & Akhmetzhanov, A.R. (2020). Serial interval of novel coronavirus (COVID-19) infections. Int J Infect Dis.
[14] Rodriguez, P.F. (2020). Predicting whom to test is more important than more tests-modeling the impact of testing on the spread of COVID-19 virus by true positive rate estimation. medRxiv.
[15] Kwok, K.D., Lai, F., Wei, W.J., Wong, S.Y.S., & Tang, J.W. (2020). Herd immunity—estimating the level required to halt the COVID-19 epidemics in affected countries. J Infect.
[16] de Vlas, S.J., & Coffeng, L.E. (2020). A phased lift of control: a practical strategy to achieve herd immunity against COVID-19 at the country level. medRxiv.
[17] Hunter DJ. Covid-19 and the stiff upper lip—The pandemic response in the United Kingdom. N Engl J Med 2020;382(16):e31.
[18] Cohen, J., & Kupferschmidt, K. (2020). Countries test tactics in ‘war’ against COVID-19. Science.
[19] Chretien JP, George D, Shama J, Chitale RA, McKenzie FE. Influenza forecasting in human populations: a scoping review. PLoS ONE 2014;9(4).
[20] Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J Roy Statist Soc Ser A 1948;11(8):15–53.
[21] Fourn LA, Woodlock JW. Early prediction of market success for new grocery products. J Mark 1960;25(2):31–8.
[22] Anscombe FJ. Estimating a mixed exponential response law. J Am Stat Assoc 1961;56:493–502.
[23] Maltz MD, McCleary R. The mathematics of behavioral change: recidivism and construct validity. Eval Rev 1977;1:421–38.
[24] Lloyd MR, Joe GW. Recidivism comparisons across groups, methods of estimation and tests of significance for recidivism rates and asymptotes. Eval Q 1979;3:105–17.
[25] Stollmack S. Comments on the mathematics of behavioral change. Eval Q 1979;3:118–23.
[26] Blumenthal S, Marcus R. Estimating population size with exponential failure. J Am Stat Assoc 1975;70:913–22.
[27] Farewell VT. A model for a binary variable with time-censored observations. Biometrika 1977;64(1):43–6.
[28] Steinhurst WR. Hypothesis tests for limited failure survival distributions. Eval Rev 1981;5:699–711.
[29] Meeker W. Limited failure population life tests: application to integrated circuit reliability. Technometrics 1987;29(1):51–65.
[30] Johnson NL. Estimation of sample size. Technometrics 1962;4(1):59–67.
[31] Sanathanan L. Estimating the size of a multinomial population. Ann. Math Stat 1972;43(1):142–52.
[32] Sanathanan L. Estimating the Size of a Truncated Sample. J Am Stat Assoc 1977;72(359):669–72.
[33] Koutsellis T, Mourelatos ZP. Parameter estimation of limited failure population model with a weibull underlying distribution. ASME. ASME J Risk Uncertain Part B 2020;6(2):021007. https://doi.org/10.1115/1.4044715.
[34] Koutsellis T, Mourelatos Z, Hijawi M, Guo H, Castanier M. Warranty forecasting of repairable systems for different production patterns. SAE Int J Mater Manuf. 2017;10(3):264–73.
[35] Santhosh TV, Gopika V, Ghosh AK, Fernandes BG. An approach for reliability prediction of instrumentation & control cables by artificial neural networks and Weibull theory for probabilistic safety assessment of NPPIs. Reliab Eng Syst Saf 2018;170:31–44.
[36] Maaher MI, Kaja S, Kaebnerich H. Remaining life estimation of used components in consumer products: life cycle data analysis by Weibull and artificial neural networks. J Oper Manage 2007;25(6):1184–93.
[37] World Health Organization (2020). Coronavirus disease (COVID-19) situation reports. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.
[38] Zhang X, Ma R, Wang L. Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. Chaos Solitons Fractals 2020;135:109829.
[39] Doukas H, Nikas A. Decision support models in climate policy. Eur J Oper Res 2020;280(1):1–24.
[40] Doukas H, Nikas A, González-Eguino M, Arto I, Anger-Kraavi A. From integrated to integrative: delivering on the Paris Agreement. Sustainability 2018;10(7):2299.
[41] Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. Chaos Solitons Fractals 2020:109850.
[42] Ndaro F, Area I, Nieto JJ, Torres DF. Mathematical modeling of COVID-19 transmission dynamics with a case study of wuhan. Chaos Solitons Fractals 2020:109846.
[43] Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos Solitons Fractals 2020:134:109761.
[44] Barmparis GD, Tsironis GP. Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach. Chaos Solitons Fractals 2020:109842.
[45] Boccialetti S, Ditto W, Mindlin G, Atangana A. Modeling and forecasting of epidemic spreading: the case of Covid-19 and beyond. Chaos Solitons Fractals 2020.
[46] Martelloni G, Martelloni G. Modelling the downhill of the Sars-CoV-2 in Italy and a universal forecast of the epidemic in the world. Chaos Solitons Fractals 2020:139:110064.