

Article

A Statistical Analysis of Death Rates in Italy for the Years 2015–2020 and a Comparison with the Casualties Reported from the COVID-19 Pandemic

Gianluca Bonifazi ^{1,2}, Luca Lista ^{3,4}, Dario Menasce ^{5,*}, Mauro Mezzetto ⁶, Alberto Oliva ², Daniele Pedrini ⁵, Roberto Spighi ² and Antonio Zoccoli ^{2,7}

- ¹ Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Via Brecce Bianche, 12, 60131 Ancona, Italy; g.bonifazi@staff.univpm.it
 - ² INFN Sezione di Bologna, Viale C. Berti Pichat, 6/2, 40127 Bologna, Italy; alberto.oliva@bo.infn.it (A.O.); roberto.spighi@bo.infn.it (R.S.); antonio.zoccoli@bo.infn.it (A.Z.)
 - ³ Dipartimento di Fisica E. Pancini, Università degli Studi di Napoli Federico II, Complesso di Monte Sant'Angelo, ed. 6, Via Cintia, 80126 Napoli, Italy; luca.lista@na.infn.it
 - ⁴ INFN Sezione di Napoli, Complesso di Monte Sant'Angelo, ed. 6, Via Cintia, 80126 Napoli, Italy
 - ⁵ INFN Sezione di Milano Bicocca, Piazza della Scienza 3, 20126 Milano, Italy; daniele.pedrini@mib.infn.it
 - ⁶ INFN Sezione di Padova, Via Francesco Marzolo, 8, 35131 Padova, Italy; mauro.mezzetto@pd.infn.it
 - ⁷ Alma Mater Studiorum, Università di Bologna, Via Irnerio, 46, 40126 Bologna, Italy
- * Correspondence: dario.menasce@mib.infn.it



Citation: Bonifazi, G.; Lista, L.; Menasce, D.; Mezzetto, M.; Oliva, A.; Pedrini, D.; Spighi, R.; Zoccoli, A. A Statistical Analysis of Death Rates in Italy for the Years 2015–2020 and a Comparison with the Casualties Reported from the COVID-19 Pandemic. *Infect. Dis. Rep.* **2021**, *13*, 285–301. <https://doi.org/10.3390/idr13020030>

Academic Editor: Nicola Petrosillo

Received: 7 February 2021

Accepted: 23 March 2021

Published: 1 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: We analyze the data about casualties in Italy in the period 1 January 2015 to 30 September 2020 released by the Italian National Institute of Statistics (ISTAT). The aim of this article was the description of a statistically robust methodology to extract quantitative values for the seasonal excesses of deaths featured by the data, accompanying them with correct estimates of the relative uncertainties. We will describe the advantages of the method adopted with respect to others listed in literature. The data exhibit a clear sinusoidal behavior, whose fit allows for a robust subtraction of the baseline trend of casualties in Italy, with a surplus of mortality in correspondence to the flu epidemics in winter and to the hottest periods in summer. The overall quality of the fit to the data turns out to be very good, an indication of the validity of the chosen model. We discuss the trend of casualties in Italy by different classes of ages and for the different genders. We finally compare the data-subtracted casualties, as reported by ISTAT, with those reported by the Italian Department for Civil Protection (DPC) relative to the deaths directly attributed to the Coronavirus Disease 2019 caused by the SARS-CoV-2 virus (COVID-19), and we point out the differences in the two samples, collected under different assumptions.

Keywords: statistics; Coronavirus; death excesses

1. Introduction

An enhanced global attention concerning the death rates in various countries has been seen, a fact due to the outbreak of the COVID-19 pandemic at the beginning of 2020 and the ensuing alarm it generated worldwide. A first visual inspection of the historical data archives [1], publicly made available by the Italian Istituto Nazionale di Statistica (ISTAT) [2], shows a periodic variation of the death rate depending upon seasons, well represented by a stable and regular sinusoid. Superimposed to the sinusoid trend, there may be additional death excesses, most likely due to seasonal diseases like influenza in winter or to very intense heat waves in the summer [3,4].

The approach adopted here for an estimate of the seasonal excess of deaths is an interpolation of the data with a fit function exhibiting an ad hoc modeling of the main features of the curve. While the excess peaks are symmetric in shape, the peak in coincidence with the COVID-19 pandemic is asymmetric and more pronounced. We fit the former with a

Gaussian function and the latter with a Gompertz function, in order to quantify number of casualties, the duration, and the position of all causes of excess deaths. The focus of this paper is the method to compute the number of deaths from the data rather than a discussion about the particular functional model chosen for this task or an interpretation of the outcomes of such an evaluation. A least-squares fit of a single function, encompassing both the background (the periodical seasonal variation of deaths) and the specific additional excesses above this background, allows for a very robust evaluation of the latter, both for the numerical values and for the relative uncertainties. We present results of this method applied to the data provided by ISTAT during the period 2015–2020. A comparison is then carried out with a different data sample [5], provided by the Dipartimento della Protezione Civile (DPC) [6], which provides counts for deaths directly attributed to COVID-19.

2. The Data Sample

This study is based on publicly available data provided by ISTAT [1] as time series of recorded deaths by the National Registry Office. The data, collected from all the 7903 districts located in the 20 Italian regions, covers the period from 1 January 2015 to 30 September 2020. We have collected these data in histograms where each bin contains the number of deaths for a single day.

The data are collected by gender, age, and location for each individual death. Figure 1 shows the number of deaths in all the categories in the considered period. What is already striking by a simple visual inspection of the distribution is a periodic seasonal variation that behaves approximately like a sinusoidal wave of constant amplitude on top of an equally constant offset value (in the following, we will discuss how we established that there is no significant slope of the average value of this wave.). This feature remains partly confirmed also by disentangling the data using age as a selection criteria.

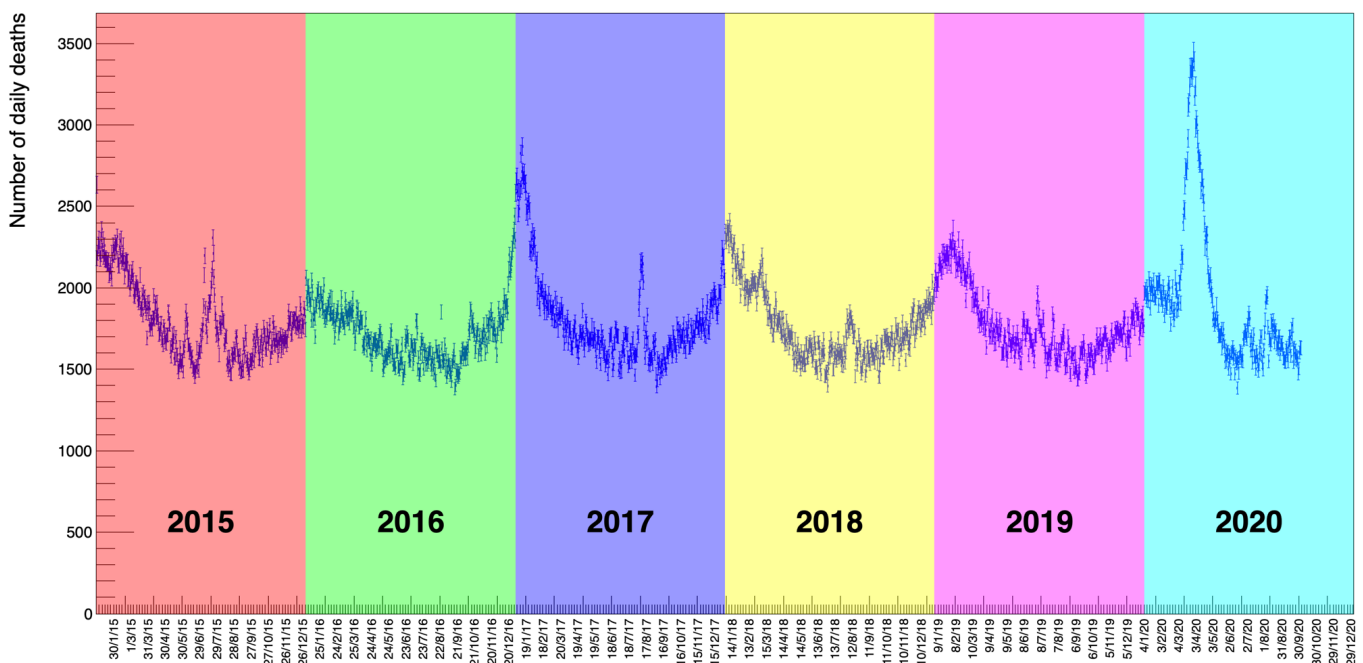


Figure 1. The distribution of deaths collected by the Italian National Institute of Statistics (ISTAT) [1] from 7903 districts in Italy between 1 January 2015 to 30 September 2020. These data include both genders and all ages. An annual modulation of the counts is evident with maxima corresponding to winter seasons and minima to summer.

Figure 2 shows the distribution of deaths for people in three different age classes: in blue those below 50 years, in green those in the range of 50 to 79; in red those above 80; and in magenta the sum of all these three classes. It is evident from these distributions that people with an age below 50 die, to a good approximation, with a constant average probability in any given day of the year while those above that age tend to have a varying,

periodic probability of death with a maximum in winter and a minimum in summer. The older the age, the larger the excess of death in particular periods of the year, appearing in the form of Gaussian-like excesses over the sinusoidal wave. Disentangling the data by gender, see Figure 3, there seems to be a slight prevalence of female deaths with respect to males, except for the COVID-19 peak, where the situation happens to be reversed. These are just raw values, though, not corrected to take into account the ratio between males and females in the Italian population. Later on, in this paper, we will quantify and appropriately weight these data.

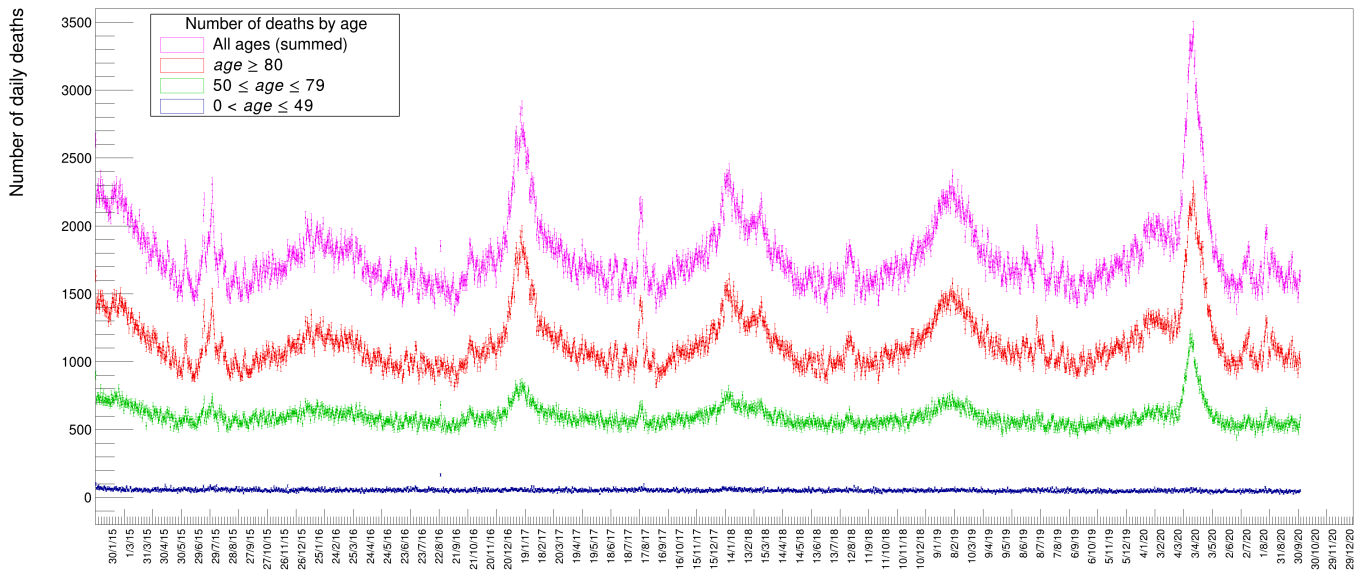


Figure 2. Distribution of number of deaths along six years for specific age intervals.

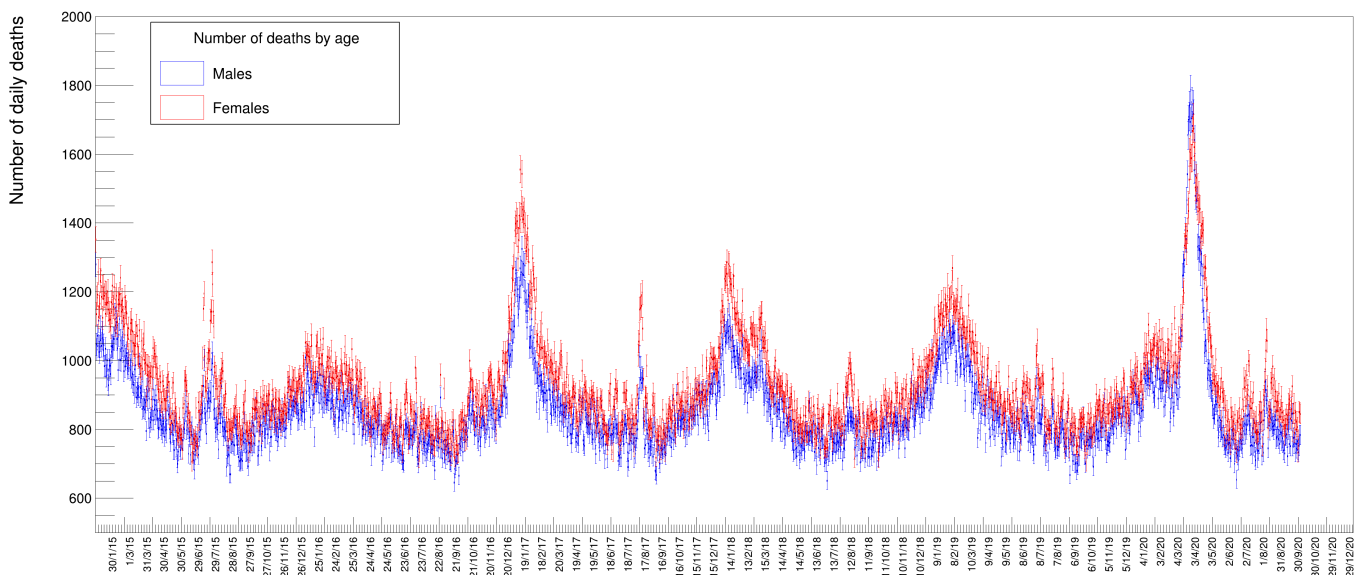


Figure 3. Distribution of number of deaths for males and females in the same time period as Figure 2.

3. Methodology of the Data Analysis

We perform a global fit of the data, where we simultaneously estimate the sinusoidal baseline of the distribution, the seasonal death excesses and the 2020 peak in correspondence of the COVID-19 pandemic. This method significantly differs from other methods often reported in literature [7].

In particular, we quote analyses [8–15] in which the background is subtracted by computing the average number of counts in the same period of the past 5 years. In this way, the excesses of seasonal pandemics, like the flu, are expressed against the average counts of the same pandemics in the previous years and not in absolute terms.

In [16], the sinusoidal baseline is computed by fitting the data during periods of time where the excesses are not evident, like in spring or autumn. While this approach can be easily automated, it is subject to a certain degree of arbitrariness due to the specific choice of the periods included in the fit.

A global fit to the time series has, instead, the merit of simultaneously using all the available data to shape both the excesses and the baseline, without any degree of arbitrariness. Furthermore, the least-squares method provides a complete and fully correct covariance matrix that allows to compute the uncertainties involved in the final result. Eventually the goodness of the fit and the absence of biases can be quantified by the final χ^2 of the interpolation and by the pulls distribution, respectively.

We therefore used a χ^2 fit to interpolate the data with an appropriate function meant to model the data in order to determine the value of the unknown parameters of the model along with their uncertainties. The actual minimization is carried out by the MINUIT [17] package, while the adopted statistical methodology is described in [18].

The overall fit function has been defined as the sum of individual components in the following form:

$$F(t) = s(t) + \sum_{i=1}^k g_i(t) + \dot{G}(t) \quad (1)$$

where $s(t)$, $g_i(t)$, and $\dot{G}(t)$ are defined and described below.

The $s(t)$ function is meant to model the wave-like variation of deaths with seasons, the $g_i(t)$ function describes the excess peaks visible above the wave and $\dot{G}(t)$ represents the rightmost excess peak (spring 2020), which, unlike the others, is asymmetrical. The index i runs from 1 up to k , the number of excess peaks featured by the data distribution except the last one on the far right ($k = 13$ peaks in this particular case).

The general wave-like behavior of the data is modeled by a sinusoidal function of the form:

$$s(t) = c(t) + a \sin\left(\frac{2\pi t}{T} + \varphi\right) \quad (2)$$

where t is the day number starting from $t_0 = 1$ January 2015. The parameter $c(t)$ represents the slowly-varying offset from zero deaths, a the amplitude of the oscillation, T is the period of variation (the time delay between consecutive maxima) and finally φ the phase. We tried to model $c(t)$ allowing for a linear dependence on t , as $c(t) = c_0 + c_1 t$, but the fit determines a slope c_1 compatible to zero within uncertainty. We therefore decided to maintain the c term constant and independent of time in the final fit.

Each individual excess above this $s(t)$ wave can be modeled by a Gaussian distribution of the canonical form:

$$g_i(t) = \frac{N_i}{\sqrt{2\pi\sigma_i}} e^{-(t-\mu_i)^2/2\sigma_i^2} \quad (3)$$

The choice of a Gaussian function here is only justified by being the simplest symmetrical function to describe these excesses representing, at the same time the distribution of a random variable. Modeling the excess peaks in the described way has the advantage that the individual g_i fit parameters correspond to a Gaussian with the background contribution already taken into account in the overall fit model. The N_i parameter of each Gaussian corresponds to the number of excess deaths with respect to the wave-like background, whose values are also determined optimally by the fit itself. An advantage of this approach is that in the case of adjacent, overlapping Gaussians (as can be seen in Figure 4 in the case of the g_6 and g_7 peaks but also the g_{11} and the big peak on the far right of the distribution), each individual area is computed correctly by taking into account the nearby contributing ones.

While the excess peaks look highly symmetrical around their maximum and can thus be reasonably well modeled with Gaussians, as described before, the peak of the spring 2020, associated with the COVID-19 pandemic, is clearly asymmetric. We have tried several possible parametrizations for that distribution, such as bifurcated Gaussians with a common peak, generalized logistics, or else, to reflect the asymmetry, but in the end we resolved to adopt the derivative of a Gompertz function [19,20] simply because it is customarily adopted by epidemiologists to describe epidemic evolution's over time and we therefore considered it more suitable to our purpose.

A Gompertz function is parametrized in the following way:

$$G(t) = N_G e^{-b} e^{-ht} \tag{4}$$

Equation (4) represents a cumulative distribution. Since our data represent instead daily counts, we used its derivative, given by:

$$\dot{G}(t) = \frac{dG(t)}{dt} = N_G bh e^{-b} e^{-ht} e^{-ht} \tag{5}$$

where the parameter N_G is the value of the integral of this function. It is worthwhile to note that a global fit can correctly take into account contributions from partially overlapping peaks, like g_6 and g_7 or g_{11} and \dot{G} in Figure 4, something that no other method can accomplish correctly.

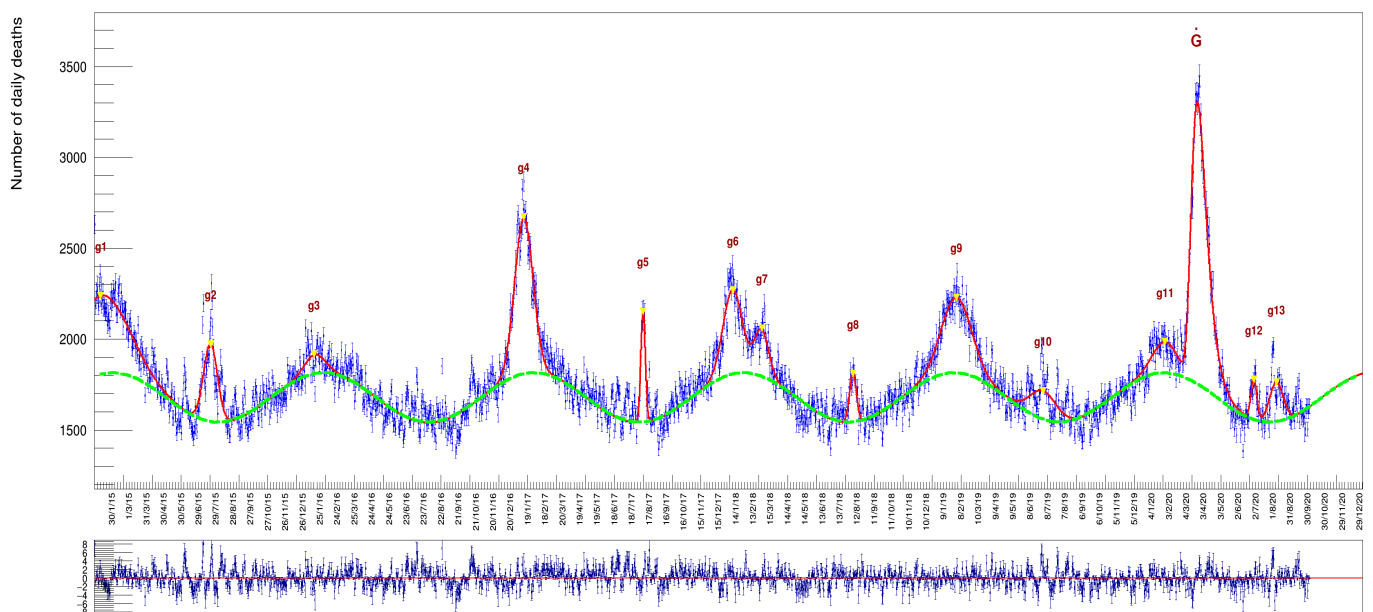


Figure 4. The whole data sample with a superimposed fit function obtained as specified in the text. The plot at the bottom shows the *pulls* (a quantity defined later on in the text) of the fit: their mean value, being compatible with zero and the absence of remarkable localized deviations from this value along the whole time series, is a testimony of the appropriate choice of the particular model adopted. Numerical values for the integral of each individual Gaussian are provided in Table 1. The g_1 to g_{13} labels indicate the 13 Gaussians introduced to describe the data (see Equation (3)).

4. Results and Discussion

In Figure 4 and Tables 1–3, we report the results of a fit to the whole data sample, comprising both genders for all ages in the six years from 2015 to 2020. The column labeled ' μ_i ', in Table 1, indicates the day when the maximum of an excess has been reached while those labeled ' $\mu_i \pm 2\sigma_i$ ' indicate, respectively, the day of onset and demise from the average background, a time interval in which occur 95% of the death cases (expressed with calendar

dates). The column labeled ‘Duration’ is the time difference between onset and demise (namely 4σ , expressed as number of days).

Table 1. Results of the fit for individual parameters (and their associated error) for each Gaussian, as modeled by Equation (3). The columns header indicates the Gaussian number (g_i), the yield (its area, N_i), its peak position (μ_i), the width (the one standard-deviation duration expressed in number of days, σ_i), and the duration within 4σ (the difference between the values of columns 7 and 5, also expressed as number of days). Date: dd/mm/yyyy.

g_i	N_i	μ_i (Days)	σ_i (Days)	$\mu_i - 2\sigma_i$ (Date)	μ_i (Date)	$\mu_i + 2\sigma_i$ (Date)	Duration (Days)
1	50,706 ± 3092	10.9 ± 3.4	47.1 ± 2.3	09/10/2014	10/01/2015	15/04/2015	189
2	13,005 ± 361	201.1 ± 0.4	12.0 ± 0.4	25/06/2015	20/07/2015	13/08/2015	49
3	4455 ± 527	381.6 ± 2.0	17.7 ± 2.1	12/12/2015	16/01/2016	21/02/2016	71
4	34,015 ± 534	743.5 ± 0.2	15.9 ± 0.3	11/12/2016	12/01/2017	13/02/2017	64
5	5959 ± 210	950.1 ± 0.1	3.9 ± 0.2	30/07/2017	07/08/2017	14/08/2017	15
6	19,120 ± 704	1104.9 ± 0.6	16.5 ± 0.8	06/12/2017	08/01/2018	11/02/2018	67
7	7862 ± 616	1155.0 ± 1.0	12.5 ± 1.2	03/02/2018	28/02/2018	24/03/2018	49
8	4084 ± 256	1313.3 ± 0.4	6.0 ± 0.5	24/07/2018	05/08/2018	17/08/2018	24
9	26,850 ± 685	1492.5 ± 0.6	26.1 ± 0.6	10/12/2018	31/01/2019	24/03/2019	104
10	9299 ± 504	1642.3 ± 1.4	23.2 ± 1.3	15/05/2019	30/06/2019	15/08/2019	92
11	9020 ± 613	1853.3 ± 1.7	21.8 ± 1.5	14/12/2019	27/01/2020	10/03/2020	87
12	2841 ± 217	2007.5 ± 0.4	5.1 ± 0.4	19/06/2020	29/06/2020	09/07/2020	20
13	6546 ± 362	2046.3 ± 0.7	11.9 ± 0.8	14/07/2020	07/08/2020	31/08/2020	48

Table 2. Results of the fit to the whole data set (no selection applied) for the baseline sinusoidal wave, as modeled by Equation (2). The column header indicates the average value of the sinusoid C , the amplitude, a , the period, T , and the phase, ϕ as further explained in the text.

c	a	T (Days)	ϕ (rad)
1678 ± 1.5	139.4 ± 2.593	364 ± 0.4	-5.27 ± 0.02

Table 3. Results of the fit to the whole data set (no filters applied) for the Gompertz derivative function. The meaning of the columns labeled From, Peak and To is explained in the text. Date: dd/mm/yyyy.

Yield	From	Peak	To	Duration (Days)
54,387 ± 557	20/02/2020	24/03/2020	11/05/2020	81

The pulls, p_i , are defined as:

$$p_i = \frac{d_i - F(t_i)}{\epsilon_i} \tag{6}$$

where d_i is the number of death counts in a given day i and ϵ_i the corresponding amount of statistical fluctuation. The data, being the outcome of counting values, are assumed to follow a Poisson distribution, hence $\epsilon_i = \sqrt{d_i}$.

The χ^2/n_{DOF} of the fit turns out to be 3.271.

We report the distribution of the pulls in Figure 5 fitted with a Gaussian function. The mean value of the fit is -0.01 ± 0.04 , compatible with zero, while the standard deviation of the Gaussian fit turns out to be 1.75 ± 0.04 , confirming the significant underestimate of the uncertainties. This deviation from unity, of about 75%, gives an approximate amount of the increase that could be applied to the data errors to make them compatible with Poissonian values.

The area of each Gaussian function i is given by the fit parameter N_i defined in Equation (3), while the area of the Gompertz derivative is the fit parameter N_G in Equation (4).

The width of the Gompertz is computed from the first day in which the integral of the function exceeds 2.5% of the total to the day in which the integral reaches 97.5% of the total. These two days are reported in Table 3 under the columns labeled From and To.

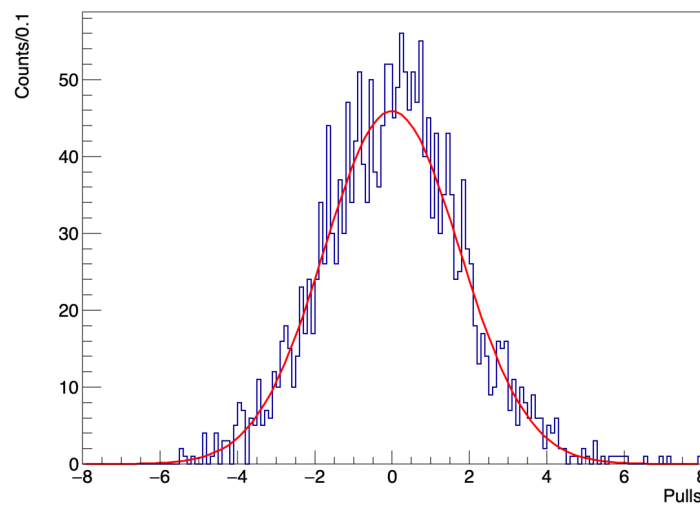


Figure 5. Distribution of the pulls (depicted as a time series in the bottom plot of Figure 4) fitted with a Gaussian function. The peak of the Gaussian is at $\mu = -0.01 \pm 0.04$ (therefore compatible with zero) while the width is given by $\sigma = 1.75 \pm 0.04$.

The value of the period $T = 364.0 \pm 0.4$ days of the sinusoidal wave is compatible with a full year cycle within about three standard deviations. The offset value $c = 1678 \pm 1.5$ can be assumed to represent the average number of deaths per day (the overall vertical offset of the sinusoid with respect to the zero value). Finally, from the results of Table 2, it turns out that the peak of the sinusoid (the maximum number of deaths) falls on 31 January of every year.

These results highlight an interesting feature of the COVID-19 deaths excess. As already noted, almost every winter there is a surplus of deaths with respect to the baseline, with the notable exception of the years 2015–2016 (a period with a particularly balm winter [21], with a relatively small value of 4455 excess of casualties). The peak in the spring of 2020, instead, shows characteristics markedly different from the winter excesses of previous year in terms of amplitude, width, and day of the year when the maximum is reached. In the following, we will mention the possible implications of these differences.

As far as we could investigate in literature, we did not find any mention of usage of the interpolation methodology we indicate in this paper, whereas the most common approach adopted is a subtraction of the baseline from previous years.

5. Age and Gender Mortality

We have also disentangled the data by age and gender and fit the distributions in these different categories to obtain accurate numerical values.

We start with a cumulative plot for all people aged between 50 (included) and 60 years (excluded) who died between 2015 and 2020, shown in Figure 6. This plot shows that the average value of daily deaths for people in this age range is about 70 casualties/day. In order to get a fit comparable with the one in Figure 4, we are forced to adopt a somewhat more stringent fit strategy.

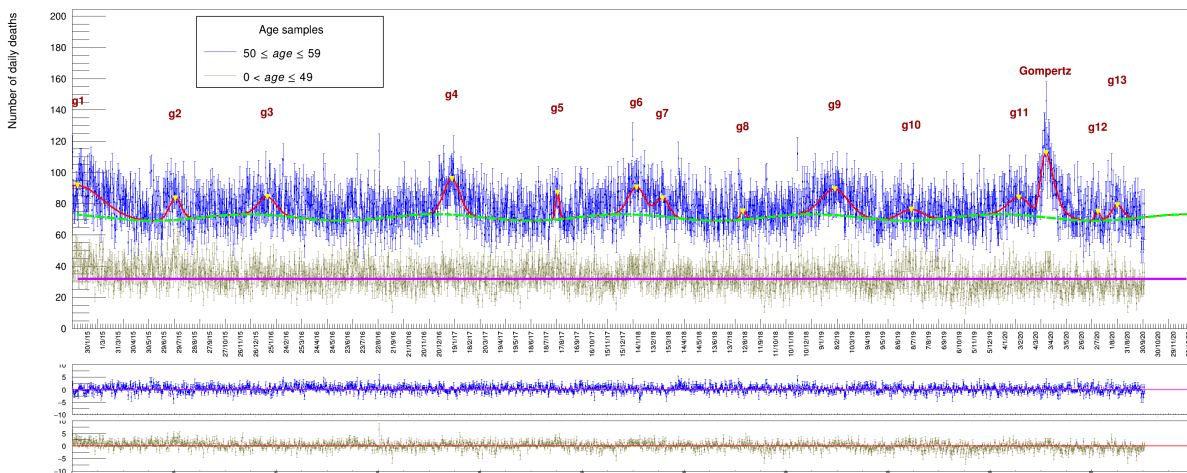


Figure 6. Casualties of people with age in the range $50 \leq \text{age} \leq 59$ with a superimposed fit based on function 6 (blue points) while gray points are the category $0 < \text{age} \leq 49$. Numerical results are listed in Tables 4 and 5. The plots at the bottom show the *pulls* of the two samples.

The wave parameter corresponding to the phase has been fixed to the value established for the full data sample (the other three are left free to float in the $s(t)$ function). This guarantees that *maxima* are reached in the winter and *minima* in the summer and no spurious time translation is introduced by the fit procedure when a local minima can eventually be found. In addition, the peak position and the width of the 13 Gaussians have been fixed to the values established by a fit to the whole data sample while the Gompertz parameters are all left free. The corresponding fit results are listed in Tables 4 and 5.

Table 4. Results of the fit to the data set of people aged between 0 and 60 (excluded). These values correspond to the fit depicted in Figure 6.

g_i	Yield
1	2252 ± 154
2	417 ± 58
3	508 ± 71
4	910 ± 71
5	166 ± 34
6	724 ± 71
7	373 ± 60
8	80 ± 39
9	1105 ± 90
10	413 ± 78
11	603 ± 79
12	71 ± 37
13	276 ± 57

Table 5. Results of the fit to the whole data set (no filters applied) for the Gompertz derivative function. These values correspond to the fit depicted in Figure 6. Date: dd/mm/yyyy.

Yield	From	Peak	To	Duration (Days)
1373 ± 87	28/2/2020	24/3/2020	1/5/2020	63

The picture shows two categories of age at the same time: those in the range 0–49 (in gray) do not show any sign of seasonal variation around the mean value of $\sim 32/\text{day}$ (they were fit with a simple constant term). A sinusoidal variation begins to be noticeable only in the range 50–59 (blue dots), along with the presence of the corresponding death excesses indicating a continuous increase in magnitude with age starting around 50. The results are affected by larger uncertainties with respect to the full sample of Figure 4, reflecting the smaller size of population in this range.

The excess peaks and the sinusoid amplitude become more evident in a sample of even higher ages, namely $60 \leq \text{age} < 80$. The average number of deaths in this category is much larger, due to an enhanced health fragility for people of progressively higher age, as seen in Figure 7.

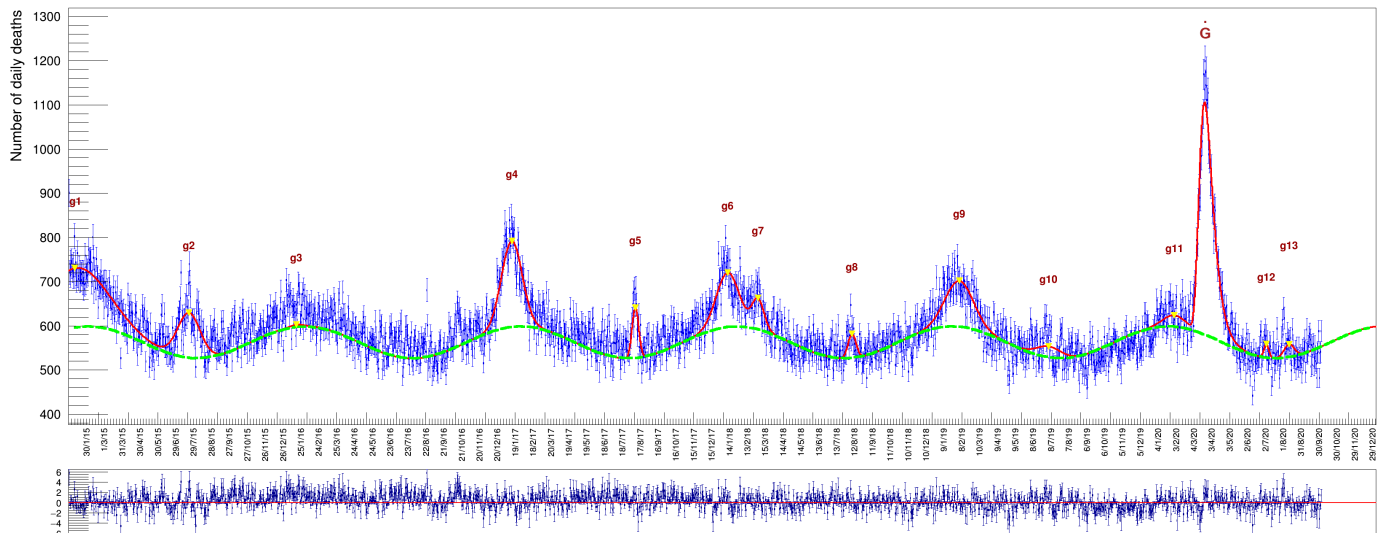


Figure 7. Casualties of people with $60 \leq \text{age} < 80$ (with a superimposed fit based on Equation (1)). Numerical results are listed in Tables 6 and 7.

The fit is again pretty similar, in shape but not in amplitude of course, to the full sample shown in Figure 4. The *pulls* feature a mean value compatible with zero also in this case. The fit strategy is the same as the one described before for Figure 6. Values obtained in this case are listed in Tables 6 and 7. The average death rate in this category is $\sim 72/\text{day}$.

Table 6. Results of the fit for the category of $50 \leq \text{age} < 80$. These values correspond to the fit depicted in Figure 7.

g_i	Yield
1	18,225 \pm 473
2	4916 \pm 204
3	100 \pm 8
4	8432 \pm 215
5	1591 \pm 110
6	5851 \pm 218
7	1704 \pm 153
8	921 \pm 117
9	6131 \pm 239
10	1319 \pm 200
11	1161 \pm 203
12	320 \pm 89
13	657 \pm 131

Table 7. Results of the fit for the category of $60 \leq \text{age} < 80$ for the Gompertz derivative function. These values correspond to the fit depicted in Figure 7. Date: dd/mm/yyyy.

Yield	From	Peak	To	Duration (Days)
15,951 \pm 242	25/02/2020	22/03/2020	25/04/2020	60

Increasing the age threshold further up, by collecting deaths of people aged ≥ 80 , we get a sample with very pronounced peaks, see Figure 8, Tables 8 and 9. The average death rate in this last category reaches the high value of $\sim 1070/\text{day}$.

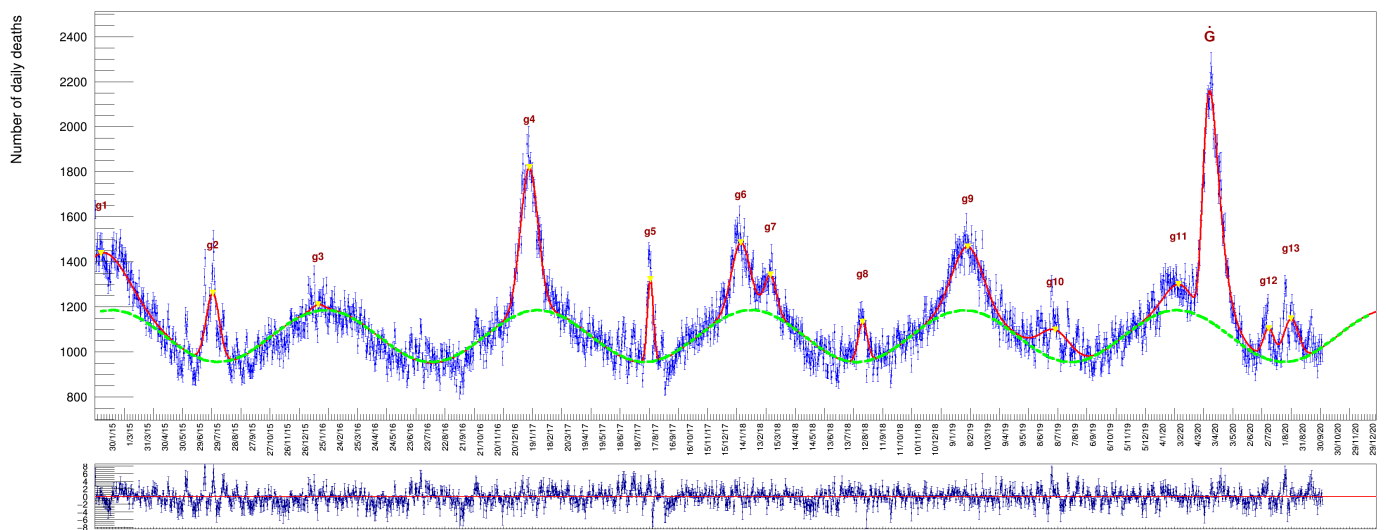


Figure 8. Casualties of people with $age \geq 80$ (with a superimposed fit based on function 6). Numerical results are listed in Tables 8 and 9.

Table 8. Results of the fit to the data set of people aged over 80. These values correspond to the fit depicted in Figure 8. Date: dd/mm/yyyy.

g_i	Yield	Peak	Duration (Days)
1	28,324 ± 721	10/01/2015	176
2	8844 ± 254	21/07/2015	45
3	630 ± 220	17/01/2016	42
4	23,732 ± 344	13/01/2017	60
5	4655 ± 179	09/08/2017	20
6	11,978 ± 343	11/01/2018	63
7	4367 ± 267	03/03/2018	41
8	3607 ± 233	08/08/2018	33
9	18,217 ± 425	04/02/2019	105
10	8981 ± 379	04/07/2019	107
11	6648 ± 389	31/01/2020	95
12	3208 ± 238	04/07/2020	37
13	6396 ± 270	11/08/2020	53

Table 9. Results of the fit to the data set of people aged over 80 for the Gompertz derivative function. These values correspond to the fit depicted in Figure 8. Date: dd/mm/yyyy.

Yield	From	Peak	To	Duration (Days)
37,357 ± 365	22/02/2020	25/03/2020	05/05/2020	73

Other information that can be extracted from the data is the relative amount of deaths between genders. Figure 9 shows the distribution of males and females (summed over all ages) superimposed with the relative fits. In this case, since the two samples have a rather large statistical amount, both fits have been performed with all parameters free to vary. These numbers need to be corrected by the relative number of males and females in the Italian population. The fraction of males in 2020 was 48.7% while females were 51.3% [22]: we compute a mortality factor (for each gender) by normalizing the yields to 29,050,096 and 30,591,392 (the respective number of males and females of the total Italian population by 1 January 2020). The resulting values (multiplied by 100,000) are listed in Tables 10 and 11 under the columns *Mortality*. While the absolute number of female deaths is higher than the males one in every year of the time series, the opposite seems true for the 2020 peak. After re-weighting this small discrepancy between genders, this assertion remains basically true for all peaks except the 2020 one, where the mortality turns out to be larger for males than for females.

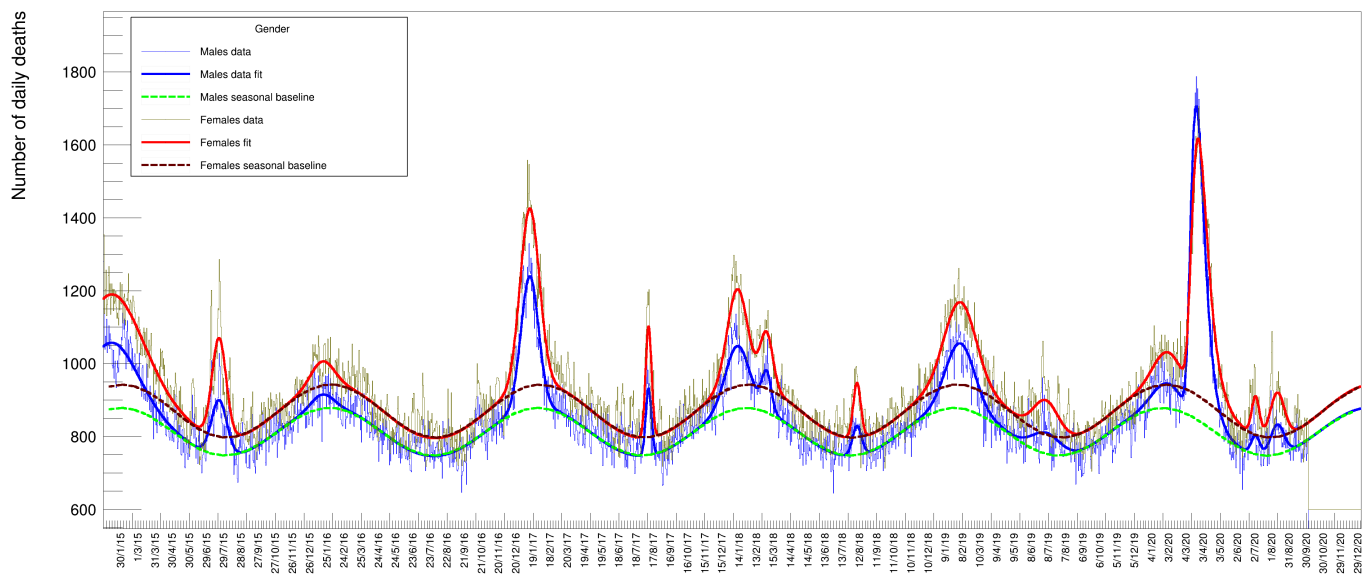


Figure 9. Number of daily casualties for males and females of all ages.

The fraction of casualties for the two genders turns out to be about the same, at the level of one standard deviation in all the years, till 2019 included.

Table 10. Results of the fit to the data set divided in a sample of males and another of females (of all ages) in Figure 9. The meaning of the *Mortality* column is described in the text. Date: dd/mm/yyyy.

g_i	Males				Females			
	Yield	Mortality	Peak	Duration	Yield	Mortality	Peak	Duration
1	20,279 ± 788	69.8	10/01/2015	178	32,405 ± 929	105.9	10/01/2015	206
2	5132 ± 258	17.7	21/07/2015	53	8401 ± 266	27.5	21/07/2015	49
3	1734 ± 394	6.0	17/01/2016	76	3198 ± 441	10.5	17/01/2016	80
4	13,907 ± 357	47.9	13/01/2017	62	20,281 ± 404	66.3	13/01/2017	66
5	2252 ± 156	7.8	09/08/2017	19	3976 ± 174	13.0	09/08/2017	21
6	8875 ± 468	30.6	11/01/2018	83	11,830 ± 415	38.7	11/01/2018	71
7	2192 ± 268	7.5	03/03/2018	33	4139 ± 305	13.5	03/03/2018	43
8	1612 ± 194	5.5	08/08/2018	31	2473 ± 203	8.1	08/08/2018	27
9	11,066 ± 462	38.1	04/02/2019	99	15,273 ± 509	49.9	04/02/2019	109
10	3728 ± 341	12.8	04/07/2019	99	5224 ± 352	17.1	04/07/2019	87
11	4112 ± 443	14.2	31/01/2020	99	5124 ± 439	16.7	31/01/2020	93
12	964 ± 182	3.3	04/07/2020	29	1764 ± 182	5.8	04/07/2020	27
13	2489 ± 236	8.6	11/08/2020	47	3826 ± 253	12.5	11/08/2020	49

Table 11. Results of the fit to the data set (for the Gompertz peak only) divided in two samples of males and females (of all ages) in Figure 9. The meaning of the *Mortality* column is described in the text. Date: dd/mm/yyyy.

Males				Females			
Yield	Mortality	Peak	Duration	Yield	Mortality	Peak	Duration
27,240 ± 366	93.8	22/03/2020	63	26,079 ± 395	85.2	26/03/2020	72

6. Comparison between Different Data Sets

The data set provided by ISTAT [1] and used for the present analysis is not the only one publicly available: the Dipartimento della Protezione Civile (DPC) data set [5] provides a somewhat different kind of information regarding the number of deaths in the context of the COVID-19 pandemic. In particular, the data record, which begins 24 February 2020, contains the number of daily deaths directly attributed to the current pandemic, whereas the ISTAT one only refers to recorded deaths regardless of their cause.

A plot of the data from these two disparate sources is shown in Figure 10. The magenta points (and the accompanying fit result of a Gompertz derivative function in red) correspond to the ISTAT data sample: these data are a subset of those displayed in Figure 4, specifically those between the dates of 24 February and 30 September 2020, with the entries

in each bin replaced with the difference between the actual counts and the contribution due to the underlying wave. This subtraction of the background of the data allows for a direct comparison between the ISTAT and DPC data, the latter does not requires a subtraction procedure being unaffected by a background.

The DPC sample is shown as blue dots (with the corresponding Gompertz fit superimposed in green). A clear peak is visible around spring 2020 together with a second one during fall 2020, corresponding, respectively, to the first and the second wave of the 2020 pandemic. It is worthwhile to note that the DPC data reports the day when the death was finally registered, unlike the case of the ISTAT data, which records the actual day of death, thus introducing a potential delay of a few days between the two samples, visible as a translation of the green line with respect to the red one.

The DPC data shows a spike corresponding to 15 August, due to the fact that a certain number of deaths were not correctly reported in the preceding weeks and were recovered assigning that day as the actual death date. In order to compare the yield returned by the fit to the value provided by the ISTAT data, we had to exclude the contributions from the second pandemic peak: we decided to introduce a cutoff value while computing the sum of entries of the DPC sample in correspondence to 16 August, a day when the minimum number of casualties was reached between the two pandemic waves, therefore including also the spike. The cutoff date is shown in Figure 10 as a vertical green arrow.

The sum in that period (the blue dots) results to be 35,468.

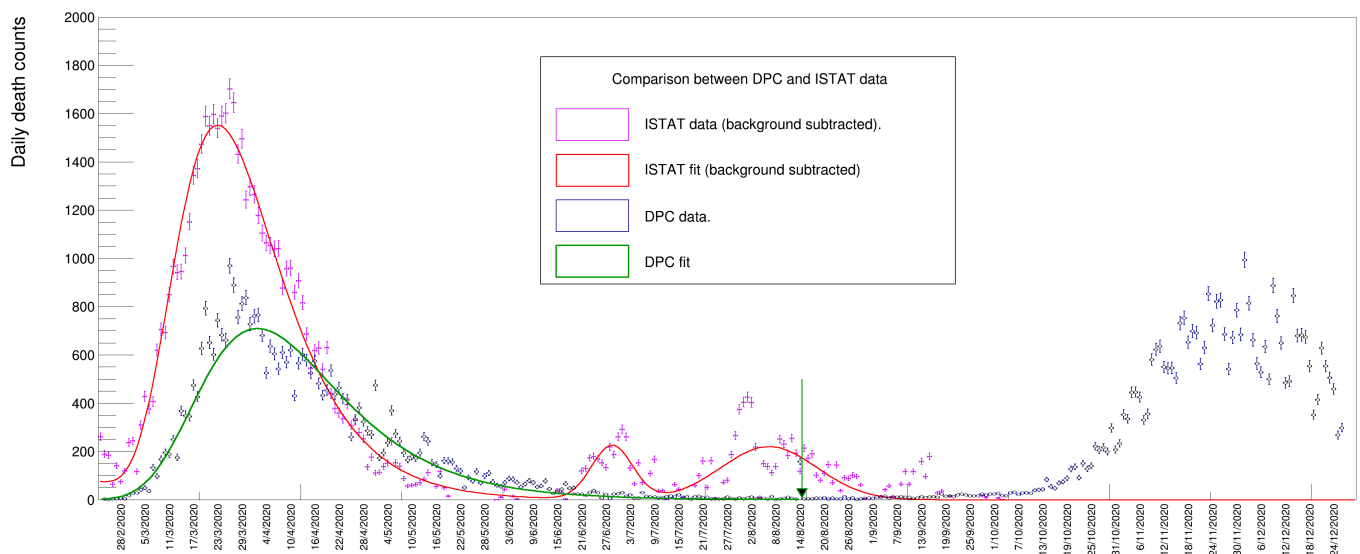


Figure 10. Comparison between ISTAT and Dipartimento della Protezione Civile (DPC) data samples.

On the other hand, the yield obtained for the ISTAT sample is the one reported in Table 3, namely $54,387 \pm 557$, resulting from the integral of the Gompertz peak (the yields of two peaks at around July and August are therefore not included). The difference in the number of deaths from these two samples amounts to $18,919 \pm 557$. This strikingly large difference could be due to several different reasons, such as an excessive pressure on the Italian health system in the early stages of the pandemic which prevented a certain number of patients with diseases other than COVID-19 to be safely treated in hospitals and emergency rooms. We have no elements in the data that can allow us to discern the different contributions to this discrepancy and an exhaustive discussion about this outcome is beyond the scope of this article.

7. Additional Considerations

The rich data sample provided by ISTAT allows for various additional visualizations. In Figure 11 we display data for ages in groups of 4 years to visualize the increase of

death probability with age: it becomes more evident what was already shown in Figure 2, namely the fact the young people tend to die with a rather flat probability along each year, while progressively higher age tend to suffer more from illnesses in specific seasonal periods. Each bin in this plot contains the number of deaths lumped together from six contiguous days. In Figure 12, we present a scatter plot of death rates as a function of the day of the year (for the six years from 2015 to 2020) versus the age category. This graphical representation clearly illustrates the higher probability of death for the age category 70–95 with respect to the others.

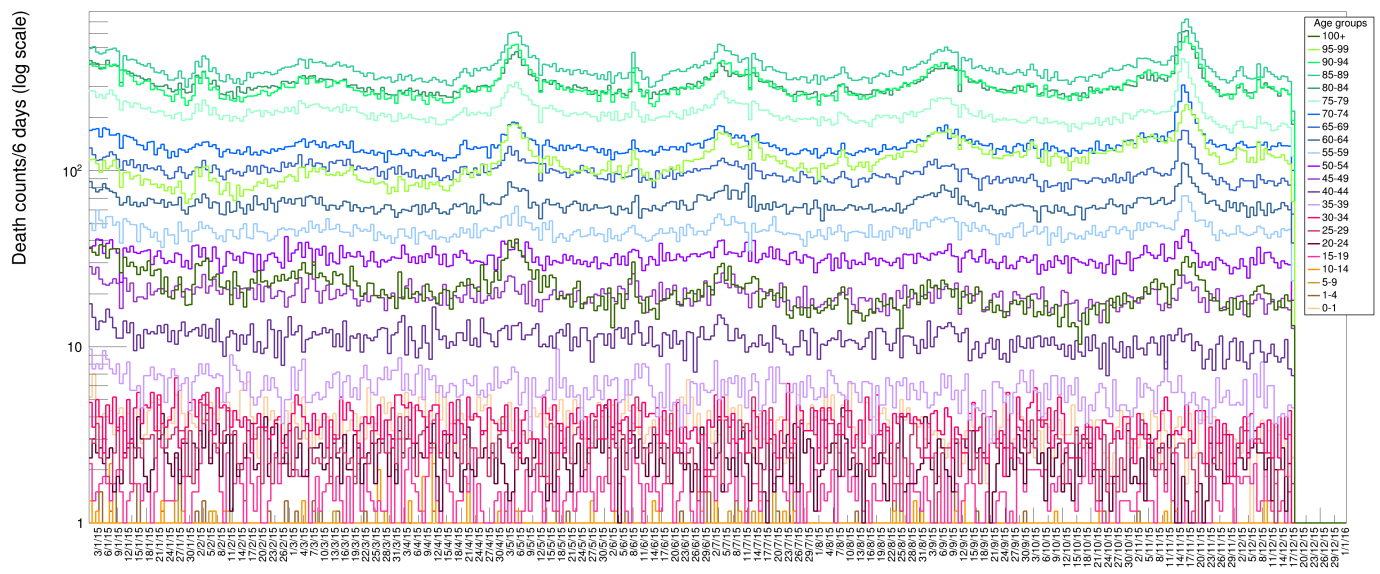


Figure 11. ISTAT data set with disentangled age categories. The data are binned in groups of six days each for an enhanced visualization clarity.

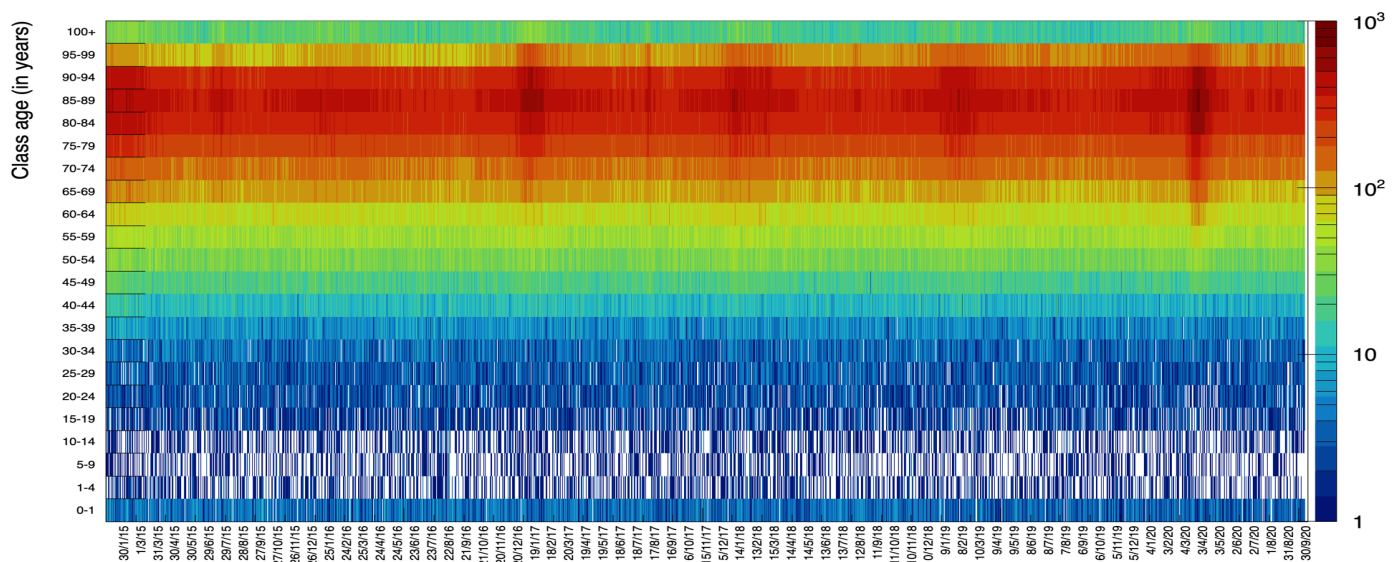


Figure 12. Scatter plot of the class age versus day of death. The dark red spots show the age and the date corresponding to highest values of deaths incidence.

Each value in the ISTAT data sample comes with a geographical tagging marker, allowing for a categorization of the number of deaths in different parts of Italy.

Figure 13 shows the fits for each of the four zones in Italy, namely *North*, *Center*, *South*, and *Islands* (The subdivision is arbitrary and we have defined *North* as the sum of values for the following regions: *Piemonte*, *Valle d’Aosta*, *Liguria*, *Lombardia*, *Trentino-Alto Adige*, *Veneto*, *Friuli-Venezia Giulia*, *Emilia-Romagna*. *Center* comprises *Toscana*, *Umbria*, *Marche*,

Lazio, Abruzzo and Molise, **South** includes Campania, Puglia, Basilicata and Calabria. Finally Islands corresponds to Sicilia and Sardegna).

The fits in these plots correspond to minimization procedures with all parameters free.

Table 12 reports the value of the Gaussian and Gompertz integral for these different regions. In order to compare these values between different zones, in Table 13 we report the same values but normalized to the relative amount of registered inhabitants [23].

A visual inspection of Figure 13 shows the magnitude of the peak in the winter/spring of 2020 for the North of Italy which is not matched by a comparably populated peak for the Center, South, and Islands. Table 13 confirms this impression numerically: while values of each column, for a given row (normalized by population), are comparable between zones, the value of the Gompertz peak in the North remains much bigger (actually by a factor from 10 to over 20).

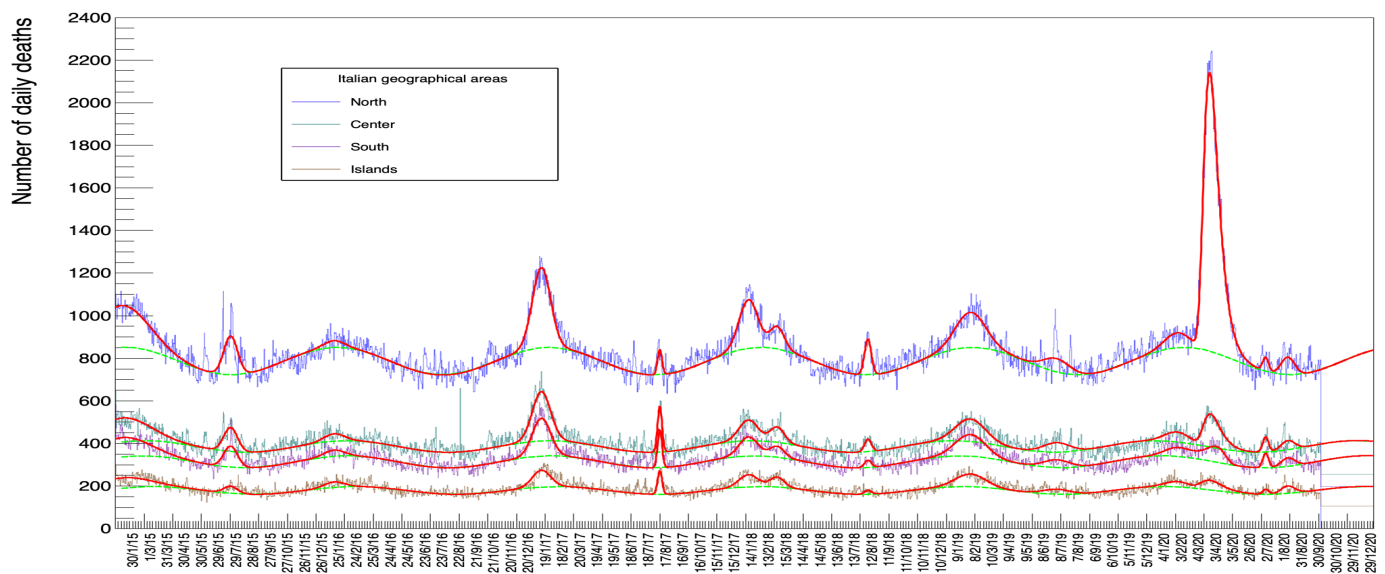


Figure 13. ISTAT data set with the Italian geographical areas disentangled (their definition is detailed in the text).

Table 12. Integral of the various peaks of Figure 13 detailed for the 4 Italian geographical areas defined in the text.

g_i	North	Central	South	Islands
1	23,046 ± 518	13,293 ± 366	10,670 ± 333	5730 ± 249
2	5520 ± 193	3409 ± 139	2995 ± 126	999 ± 91
3	1333 ± 235	1501 ± 166	1250 ± 151	1120 ± 115
4	14,924 ± 254	9184 ± 183	7058 ± 165	3135 ± 121
5	1143 ± 106	2078 ± 85	1716 ± 77	1066 ± 58
6	9393 ± 246	4027 ± 170	3676 ± 156	2344 ± 119
7	3213 ± 203	2233 ± 144	1576 ± 129	1451 ± 102
8	2511 ± 135	876 ± 93	5 11 ± 82	274 ± 61
9	10,692 ± 302	6865 ± 213	6620 ± 197	3884 ± 150
10	3711 ± 255	2631 ± 180	2168 ± 161	1374 ± 122
11	3737 ± 264	2505 ± 186	2033 ± 171	1361 ± 130
12	836 ± 119	898 ± 87	871 ± 79	300 ± 57
13	2477 ± 182	1286 ± 131	1188 ± 117	1026 ± 90
Gompertz	45,350 ± 33	6063 ± 201	3559 ± 205	2158 ± 150

Table 13. Mortality in the four Italian zones: the quoted values are obtained by normalizing the values of Table 12 to the number of inhabitants in those same regions taken from [23], corresponding to the population registered at 31 December 2020.

g_i	North (%)	Central (%)	South (%)	Islands (%)
1	0.0835 ± 0.0019	0.0990 ± 0.0027	0.0881 ± 0.0027	0.0883 ± 0.0038
2	0.0200 ± 0.0007	0.0254 ± 0.0010	0.0247 ± 0.0010	0.0154 ± 0.0014
3	0.0048 ± 0.0009	0.0112 ± 0.0012	0.0103 ± 0.0012	0.0173 ± 0.0018
4	0.0540 ± 0.0009	0.0684 ± 0.0014	0.0583 ± 0.0014	0.0483 ± 0.0019
5	0.0041 ± 0.0004	0.0155 ± 0.0006	0.0142 ± 0.0006	0.0164 ± 0.0009
6	0.0340 ± 0.0009	0.0300 ± 0.0013	0.0303 ± 0.0013	0.0361 ± 0.0018
7	0.0116 ± 0.0007	0.0166 ± 0.0011	0.0130 ± 0.0011	0.0224 ± 0.0016
8	0.0091 ± 0.0005	0.0065 ± 0.0007	0.0042 ± 0.0007	0.0042 ± 0.0009
9	0.0387 ± 0.0011	0.0511 ± 0.0016	0.0547 ± 0.0016	0.0599 ± 0.0023
10	0.0134 ± 0.0009	0.0196 ± 0.0013	0.0179 ± 0.0013	0.0212 ± 0.0019
11	0.0135 ± 0.0010	0.0187 ± 0.0014	0.0168 ± 0.0014	0.0210 ± 0.0020
12	0.0030 ± 0.0004	0.0067 ± 0.0006	0.0072 ± 0.0007	0.0046 ± 0.0009
13	0.0090 ± 0.0007	0.0096 ± 0.0010	0.0098 ± 0.0010	0.0158 ± 0.0014
Gompertz	0.164 ± 0.001	0.045 ± 0.001	0.029 ± 0.002	0.033 ± 0.002

8. Conclusions

The data provided by ISTAT allow for a detailed quantitative estimate of the number of deaths excesses with respect to a baseline. This baseline is represented by a sinusoidal variation of the number of deaths which turns out to be almost perfectly in phase with the yearly seasonal cycle. We presented a study of these excesses evaluated by a statistical interpolation of the data based on a χ^2 minimization method using a function which is the sum of a sinusoidal wave, a number of Gaussian distributions to represent the excesses above the sinusoid and, finally, a Gompertz derivative to model the asymmetric peak of spring 2020. The overall fit resulted satisfactory in terms of the final χ^2 and pull distributions, describing the 2014 data points with just 46 parameters. This allows for a quantitative definition of the properties of all the peaks, along with a precise determination of the errors. In this study, we discussed the methodology adopted for the interpolations and analyze different samples by disentangling genders, ages and locations. A comparison has also been carried out between the number of deaths provided by ISTAT in the period corresponding to the first wave of the pandemic and the numbers provided by DPC in the same period for the deaths directly attributed to COVID-19. We found a rather large discrepancy, amounting to $18,919 \pm 557$ deaths over a total of $54,387 \pm 557$. We have no elements in the data that can allow us to discern the different contributions to this discrepancy and an exhaustive discussion about it is beyond the scope of this article.

As a final remark, we think this study once more underlines the importance of a unified protocol of data collection and the online availability of these same data under a sheared Open Data international agreement. Open Data repositories with useful data already exist (ISTAT and DPC are good examples) but they are not exhaustive in the number of information provided. Other repositories, containing valuable data for improved analyses are usually restricted or not compliant with the FAIR [24] approach, one of the prerequisites of the Open Data paradigm. These shortcomings hamper the possibility of further in-depth studies of the pandemic effects and its evolution by a large number of scholars. INFN is very active in this field and has recently implemented an Open Access/Open Data repository [25], containing also, among many other documents and data sets, the whole ensemble of results produced by our group.

Author Contributions: Conceptualization, formal analysis and software, D.M. All other authors contributed equally to methodology, validation, investigation and curation. Draft-preparation, D.M., all other authors contributed to editing and finalization of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study, since this work does not include studies on individual human beings, it only relies on recorded timelines data.

Informed Consent Statement: Patient consent was waived due to the fact that no data concerning specific patients was used in the preparation of this study.

Data Availability Statement: Sources of the data employed in this study are thoroughly reported in the bibliography.

Acknowledgments: The present work has been done in the context of the INFN CovidStat project that produces an analysis of the public Italian COVID-19 data. The results of the analysis are published and updated on the website <https://covid19.infn.it>, accessed on 12 February 2021. We are grateful to Mauro Albani, Marco Battaglini, Gianni Corsetti and Sabina Prati of ISTAT for useful insights and discussions. We wish also to thank Daniele Del Re and Paolo Meridiani for useful discussions. The project has been supported in various ways by a number of people from different INFN Units. In particular, we wish to thank, in alphabetic order: Stefano Antonelli (CNAF), Fabio Bredo (Padova Unit), Luca Carbone (Milano-Bicocca Unit), Francesca Cuicchio (Communication Office), Mauro Dinardo (Milano-Bicocca Unit), Paolo Dini (Milano-Bicocca Unit), Rosario Esposito (Naples Unit), Stefano Longo (CNAF), and Stefano Zani (CNAF). We also wish to thank Domenico Ursino (Università Politecnica delle Marche) for his supportive contribution.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Public Data on Deaths in the Italian Municipalities for the Years 2015–2020. Provided by ISTAT. Available online: https://www.istat.it/it/files/2020/03/Dataset-decessi-comunali-giornalieri-e-tracciato-record_dati-al-30-settembre.zip (accessed on 12 February 2021).
- ISTAT, Istituto Nazionale di Statistica. Available online: <https://www.istat.it/en/> (accessed on 12 February 2021).
- Hajat, S.; Gasparrini, A. The excess winter deaths measure: Why its use is misleading for public health understanding of cold-related health impacts. *Epidemiology* **2016**, *27*, 486. [[CrossRef](#)] [[PubMed](#)]
- Weinberger, K.R.; Harris, D.; Spangler, K.R.; Zanobetti, A.; Wellenius, G.A. Estimating the number of excess deaths attributable to heat in 297 United States counties. *Environ. Epidemiol.* **2020**, *4*. [[CrossRef](#)] [[PubMed](#)]
- Dati COVID-19 Italia, Dipartimento della Protezione Civile. Available online: <https://github.com/pcm-dpc/COVID-19> (accessed on 12 February 2021).
- Dipartimento di Protezione Civile (DPC). Available online: <http://www.protezionecivile.gov.it/> (accessed on 12 February 2021).
- A Pandemic Primer on Excess Mortality Statistics and Their Comparability across Countries (and Further References Therein). Available online: <https://ourworldindata.org/covid-excess-mortality> (accessed on 12 February 2021).
- The Human Mortality Database. Available online: <https://www.mortality.org/> (accessed on 12 February 2021).
- Weekly Death Statistics. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Weekly_death_statistics (accessed on 12 February 2021).
- Salje, H.; Kiem, C.T.; Lefrancq, N.; Courtejoie, N.; Bosetti, P.; Paireau, J.; Andronico, A.; Hozé, N.; Richet, J.; Dubost, C.-L.; et al. Estimating the burden of SARS-CoV-2 in France. *Science* **2020**. [[CrossRef](#)] [[PubMed](#)]
- Michelozzi, P.; de' Donato, F.; Scortichini, M.; Pezzotti, P.; Stafoggia, M.; Sario, M.D.; Costa, G.; Noccioli, F.; Riccardo, F.; Bella, A.; et al. Temporal dynamics in total excess mortality and COVID-19 deaths in Italian cities. *BMC Public Health* **2020**, *20*, 1238. [[CrossRef](#)] [[PubMed](#)]
- Weinberger, D.M.; Chen, J.; Cohen, T.; Crawford, F.W.; Mostashari, F.; Olson, D.; Pitzer, V.E.; Reich, N.G.; Russi, M.; Simonsen, L.; et al. Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the United States, March to May 2020. *JAMA Intern. Med.* **2020**. [[CrossRef](#)] [[PubMed](#)]
- Casini, L.; Rocchetti, M. A Cross-Regional Analysis of the COVID-19 Spread during the 2020 Italian Vacation Period: Results from Three Computational Models Are Compared. *Sensors* **2020**, *20*, 7319. [[CrossRef](#)] [[PubMed](#)]
- Català, M.; Alonso, S.; Alvarez-Lacalle, E.; López, D.; Cardona, P.-J.; Prats, C. Empirical model for short-time prediction of COVID-19 spreading. *PLoS Comput. Biol.* **2020**, *16*, e1008431. [[CrossRef](#)] [[PubMed](#)]
- Rasambainarivo, F.; Rasoanomenjanahary, A.; Rabarison, J.H.; Ramiadantsoa, T.; Ratovoson, R.; Randremanana, R.; Randrianarisoa, S.; Rajeev, M.; Masquelier, B.; Heraud, J.M.; et al. Monitoring for outbreak-associated excess mortality in an African city: Detection limits in Antananarivo, Madagascar. *Int. J. Infect. Dis.* **2020**, *103*, 338–342. [[CrossRef](#)] [[PubMed](#)]
- EuroMOMO, European Mortality Monitoring. Available online: <https://www.euromomo.eu/> (accessed on 12 February 2021).
- James, F.; Roos, M. MINUIT: A system for function minimization and analysis of the parameter errors and corrections. *Comput. Phys. Commun.* **1975**, *10*, 343–367. [[CrossRef](#)]
- Jacoboni, C.W.T.; Eadie, D.; Dryard, F.E.; James, M.R.; Sadoulet, B. Sadoulet—Statistical Methods in Experimental Physics. *Nuov. Cim. A* **1977**, *40*, 235. [[CrossRef](#)]
- Gompertz, B. XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c. *Philos. Trans. R. Soc. Lond.* **1825**, *115*, 513–583.

20. Ohnishi, A.; Namekawa, Y.; Fukui, T. Universality in COVID-19 spread in view of the Gompertz function. *Prog. Theor. Exp. Phys.* **2020**, 123J01. [[CrossRef](#)]
21. Available online: <https://www.centrometeoitaliano.it/notizie-meteo/clima-inverno-2015-temperature-ancora-oltre-la-media-in-italia-per-tutta-la-stagione-invernale-15-03-2015-25727/> (accessed on 12 February 2021).
22. Available online: <https://www.tuttitalia.it/statistiche/popolazione-eta-sesso-stato-civile-2018/> (accessed on 12 February 2021).
23. Resident Italian Population, ISTAT. Available online: http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1 (accessed on 12 February 2021).
24. Wilkinson, M.; Dumontier, M.; Aalbersberg, I.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [[CrossRef](#)] [[PubMed](#)]
25. INFN Open Access Repository. Available online: <https://www.openaccessrepository.it/> (accessed on 12 February 2021).