A statistical analysis of death rates in Italy for the years 2015–2020 and a comparison with the casualties reported for 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[7,2]

[1]Università Politecnica delle Marche
[2]INFN Sezione di Bologna
[3]Università degli Studi di Napoli Federico II
[4]INFN Sezione di Napoli
[5]INFN Sezione di Milano Bicocca
[6]INFN Sezione di Padova
[7]Alma Mater Studiorum Università di Bologna
[*]**Corresponding author**, e-mail: `dario.menasce@mib.infn.it`

Abstract

We analyze the data about casualties in Italy in the period 01/01/2015 to 30/09/2020 released by the Italian National Institute of Statistics (ISTAT). Aim of this article is 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 COVID-19, and we point out the differences in the two samples, collected under different assumptions.

1 Introduction

Last year has seen an enhanced global attention concerning the death rates in various countries, a fact due to the outbreak of the COVID-19 pandemics 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 pandemics 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. 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 in 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 the COVID-19 disease.

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 January 1$^{\text{st}}$, 2015 to September 30$^{\text{th}}$, 2020. We have collected these data in histograms where each bin contains the number of deaths for a single day.
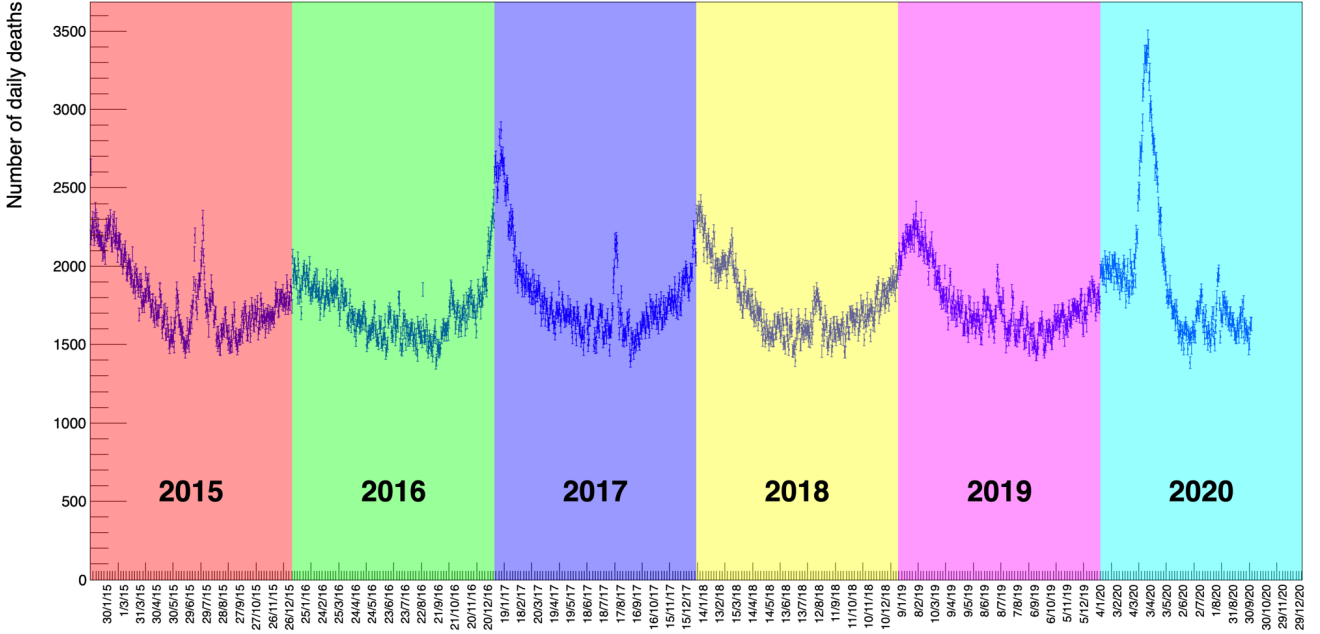

Figure 1: The distribution of deaths collected by ISTAT [1] from 7903 districts in Italy between January 1$^{\text{st}}$ 2015 to September 30$^{\text{th}}$ 2020. These data include both genders and all ages. An annual modulation of the counts is evident with maxima corresponding to winter seasons and minima to summer.

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[1]. This feature remains partly confirmed also by disentangling the data using age as a selection criteria. Fig. 2 shows the distribution of deaths for people in three different age classes: in blue those below 50 years, in green those in range 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 Fig. 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.

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].

---

[1]In the following we will discuss how we established that there is no significant slope of the average value of this wave.
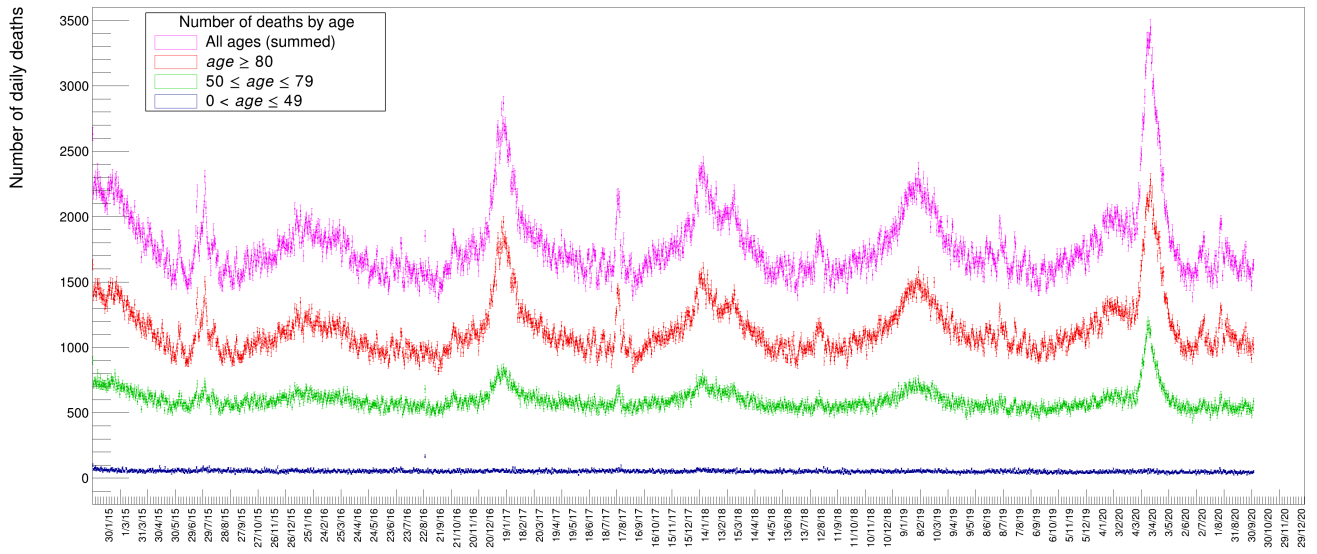
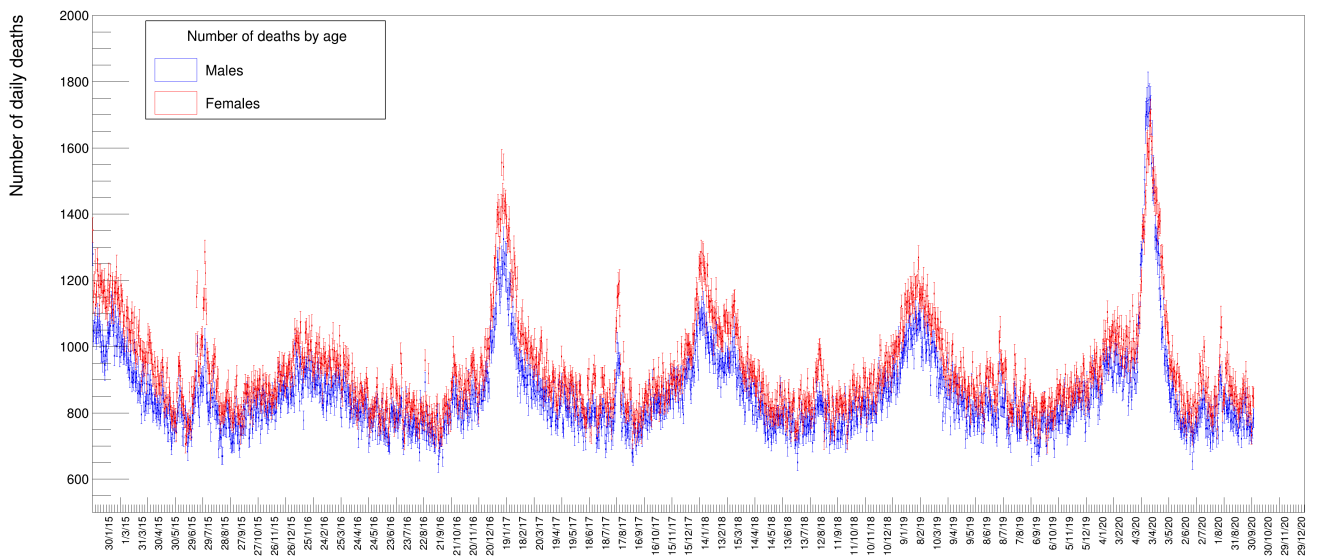Figure 2: Distribution of number of deaths along six years for specific age intervals.

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

In particular we quote analyses [8, 9, 10, 11, 12, 13, 14, 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 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:

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

where $s(t)$, $g_i(t)$ end $\dot{G}(t)$ 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:

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

where t is the day number starting from $t_0 = 1/1/2015$. 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:

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

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 Fig. 4 in the case of the g_6 end 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:

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

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 $g6$ and $g7$ or $g11$ and $\dot{G}$ in Figure 4, something that no other method can accomplish correctly.

4 Results and discussion

In Figure 4 and Tables 1 to 3 we report the results of a fit to the whole data sample, comprising both genders for all
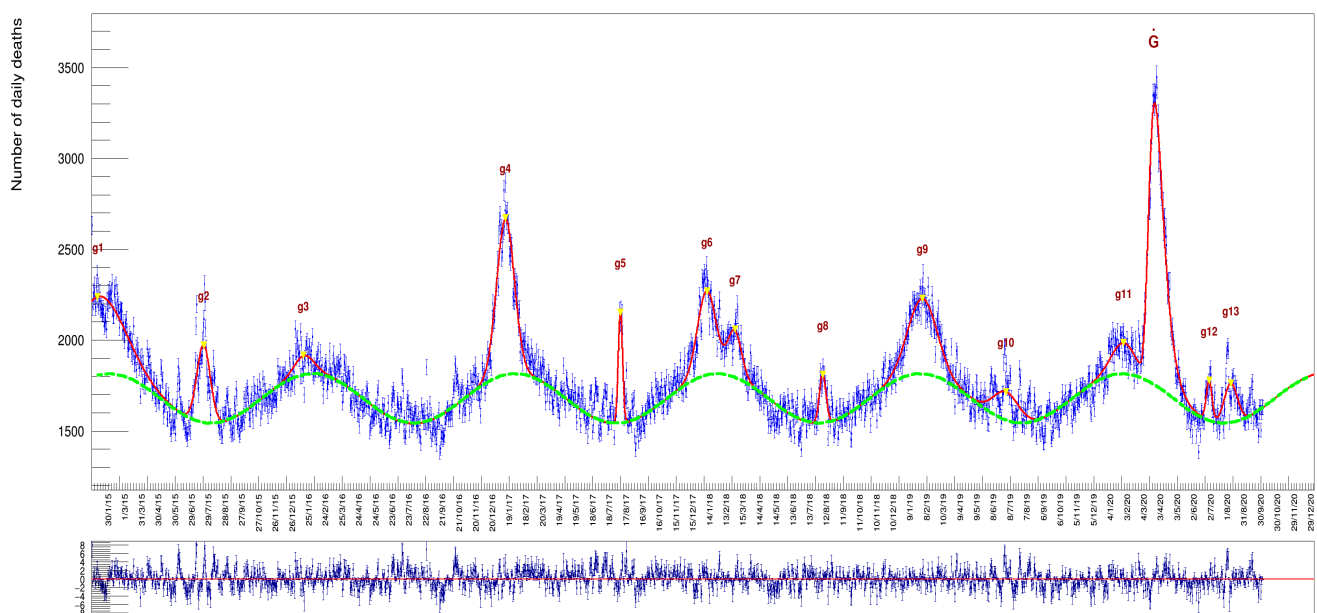

Figure 4: The whole data sample with a superimposed fit function obtained as specified in the text. The plot at the bottom shows the *pulls* (a quantity defined later on in the text) of the fit: their mean value, being compatible with zero and the absence of remarkable localized deviations from this 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 $g1$ to $g13$ labels indicate the 13 Gaussians introduced to describe the data (see eq. 3).

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).

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 outcomes 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.

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	$50\,706 \pm 3\,092$	10.9 ± 3.4	47.1 ± 2.3	9/10/2014	10/1/2015	15/4/2015	189
2	$13\,005 \pm 361$	201.1 ± 0.4	12.0 ± 0.4	25/ 6/2015	20/7/2015	13/8/2015	49
3	$4\,455 \pm 527$	381.6 ± 2.0	17.7 ± 2.1	12/12/2015	16/1/2016	21/2/2016	71
4	$34\,015 \pm 534$	743.5 ± 0.2	15.9 ± 0.3	11/12/2016	12/1/2017	13/2/2017	64
5	$5\,959 \pm 210$	950.1 ± 0.1	3.9 ± 0.2	30/7/2017	7/8/2017	14/8/2017	15
6	$19\,120 \pm 704$	1104.9 ± 0.6	16.5 ± 0.8	6/12/2017	8/1/2018	11/2/2018	67
7	$7\,862 \pm 616$	1155.0 ± 1.0	12.5 ± 1.2	3/2/2018	28/2/2018	24/3/2018	49
8	$4\,084 \pm 256$	1313.3 ± 0.4	6.0 ± 0.5	24/7/2018	5/8/2018	17/8/2018	24
9	$26\,850 \pm 685$	1492.5 ± 0.6	26.1 ± 0.6	10/12/2018	31/1/2019	24/3/2019	104
10	$9\,299 \pm 504$	1642.3 ± 1.4	23.2 ± 1.3	15/5/2019	30/6/2019	15/8/2019	92
11	$9\,020 \pm 613$	1853.3 ± 1.7	21.8 ± 1.5	14/12/2019	27/1/2020	10/3/2020	87
12	$2\,841 \pm 217$	2007.5 ± 0.4	5.1 ± 0.4	19/6/2020	29/6/2020	9/7/2020	20
13	$6\,546 \pm 362$	2046.3 ± 0.7	11.9 ± 0.8	14/7/2020	7/8/2020	31/8/2020	48

Table 1: Results of the fit for individual parameters (and their associated error) for each Gaussian, as modeled by equation number 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 column 7 and 5, also expressed as number of days).

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

Table 2: Results of the fit to the whole data set (no selection applied) for the baseline sinusoidal wave, as modeled by eq. 2. The columns 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.

We report the distribution of the *pulls* in Fig. 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 eq. 3, while the area of the Gompertz derivative is the fit parameter N_G in eq. 4.

Yield	From	Peak	To	Duration (days)
$54\,387 \pm 557$	20/2/2020	24/3/2020	11/5/2020	81

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.

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*.

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 January 31st 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 $4\,455$ excess of casualties). The peak in the spring of 2020, instead, shows characteristics markedly different from the winter excesses of previous year in
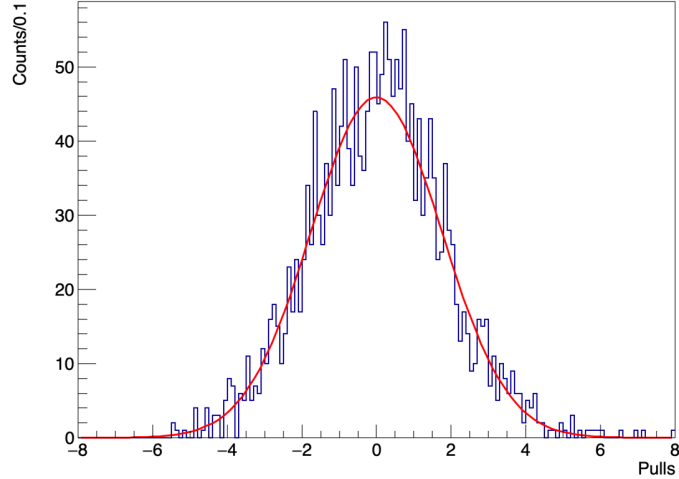
Figure 5: Distribution of the pulls (depicted as a time series in the bottom plot of Fig. 4) fitted with a Gaussian function. The peak of the Gaussian is at $\mu = -0.01 \pm 0.04$ (therefore compatible with zero) while the width id given by $\sigma = 1.75 \pm 0.04$.

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 didn't 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 Fig. 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 Fig. 4, we are forced to adopt a somewhat more stringent fit strategy. 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. Also 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.

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 ~ 32/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 Fig. 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 $60 \leq age < 80$. 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 Fig. 7.

The fit is again pretty similar, in shape but not in amplitude of course, to the full sample shown in Fig. 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 Fig. 6. Values obtained in this case are listed in Tables 6 and 7. The average death rate in this category is $\sim 72/day$. Increasing the age threshold further up, by collecting deaths of people aged ≥ 80, we get a sample with
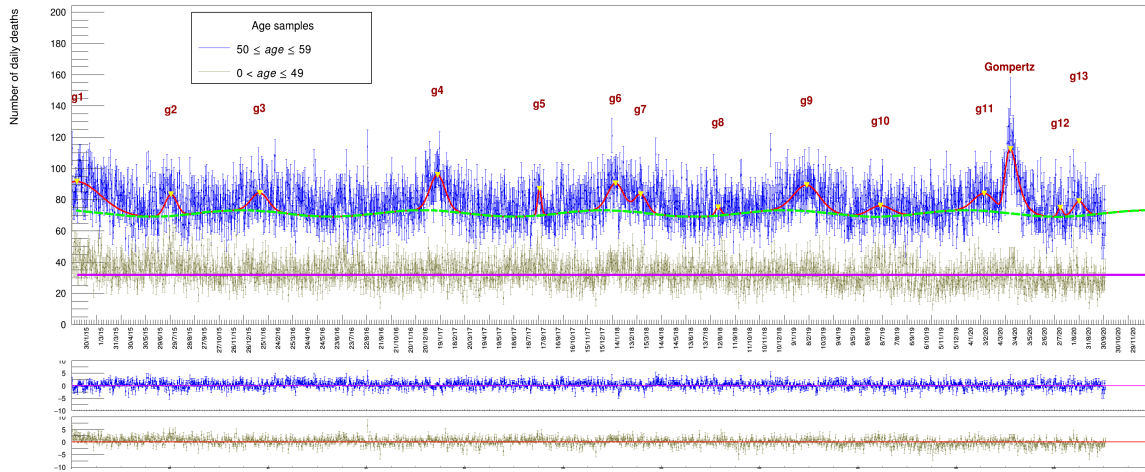
7

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

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 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 Fig. 6

Yield	From	Peak	To	Duration (days)
$1\,373 \pm 87$	28/2/2020	24/3/2020	1/5/2020	63

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 Fig. 6.
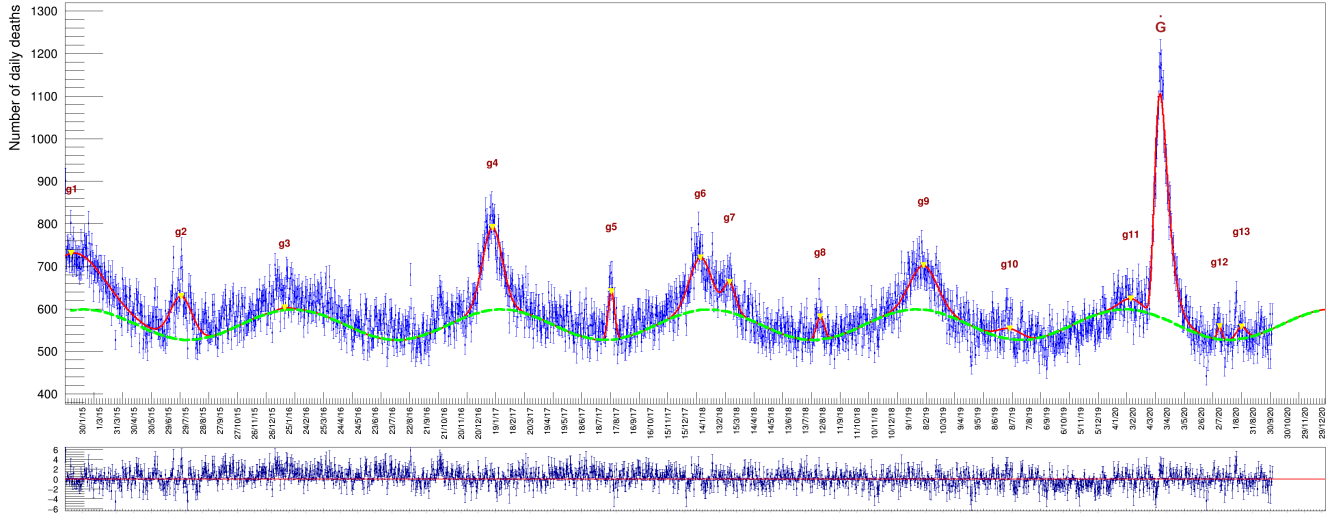
Figure 7: Casualties of people with $60 \le$ age < 80 (with a superimposed fit based on eq.1). Numerical results are listed in Table 6 and 7

g_i	$Yield$
1	$18\,225 \pm 473$
2	$4\,916 \pm 204$
3	100 ± 8
4	$8\,432 \pm 215$
5	$1\,591 \pm 110$
6	$5\,851 \pm 218$
7	$1\,704 \pm 153$
8	921 ± 117
9	$6\,131 \pm 239$
10	$1\,319 \pm 200$
11	$1\,161 \pm 203$
12	320 ± 89
13	657 ± 131

Table 6: Results of the fit for the category of $50 \le$ age < 80 . These values correspond to the fit depicted in Fig. 7

$Yield$	$From$	$Peak$	To	$Duration$ (days)
$15\,951 \pm 242$	25/2/2020	22/3/2020	25/4/2020	60

Table 7: Results of the fit for the category of $60 \le$ age < 80 for the Gompertz derivative function. These values correspond to the fit depicted in Fig. 7
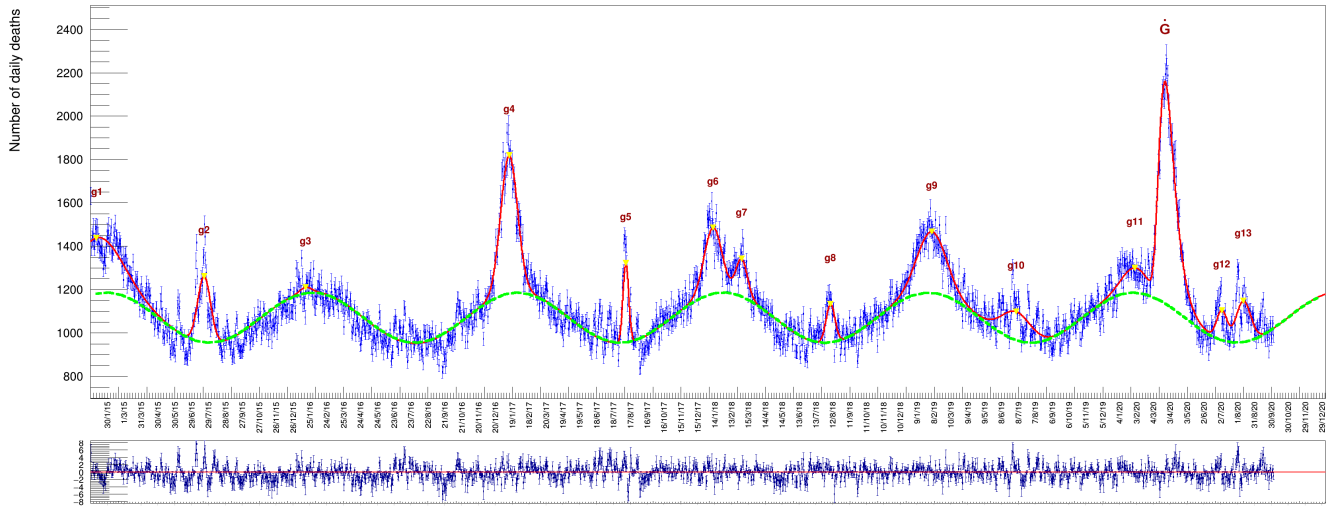
9

Figure 8: Casualties of people with age ≥ 80 (with a superimposed fit based on function 6). Numerical results are listed in Table 8 and 9

g_i	Yield	Peak	Duration (days)
1	$28\,324 \pm 721$	10/1/2015	176
2	$8\,844 \pm 254$	21/7/2015	45
3	630 ± 220	17/1/2016	42
4	$23\,732 \pm 344$	13/1/2017	60
5	$4\,655 \pm 179$	9/8/2017	20
6	$11\,978 \pm 343$	11/1/2018	63
7	$4\,367 \pm 267$	3/3/2018	41
8	$3\,607 \pm 233$	8/8/2018	33
9	$18\,217 \pm 425$	4/2/2019	105
10	$8\,981 \pm 379$	4/7/2019	107
11	$6\,648 \pm 389$	31/1/2020	95
12	$3\,208 \pm 238$	4/7/2020	37
13	$6\,396 \pm 270$	11/8/2020	53

Table 8: Results of the fit to the data set of people aged over 80. These values correspond to the fit depicted in Fig. 8

Yield	From	Peak	To	Duration (days)
$37\,357 \pm 365$	22/2/2020	25/3/2020	5/5/2020	73

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 Fig. 8

very pronounced peaks, see Fig. 8. The average death rate in this last category reaches the high value of $\sim 1070/day$.
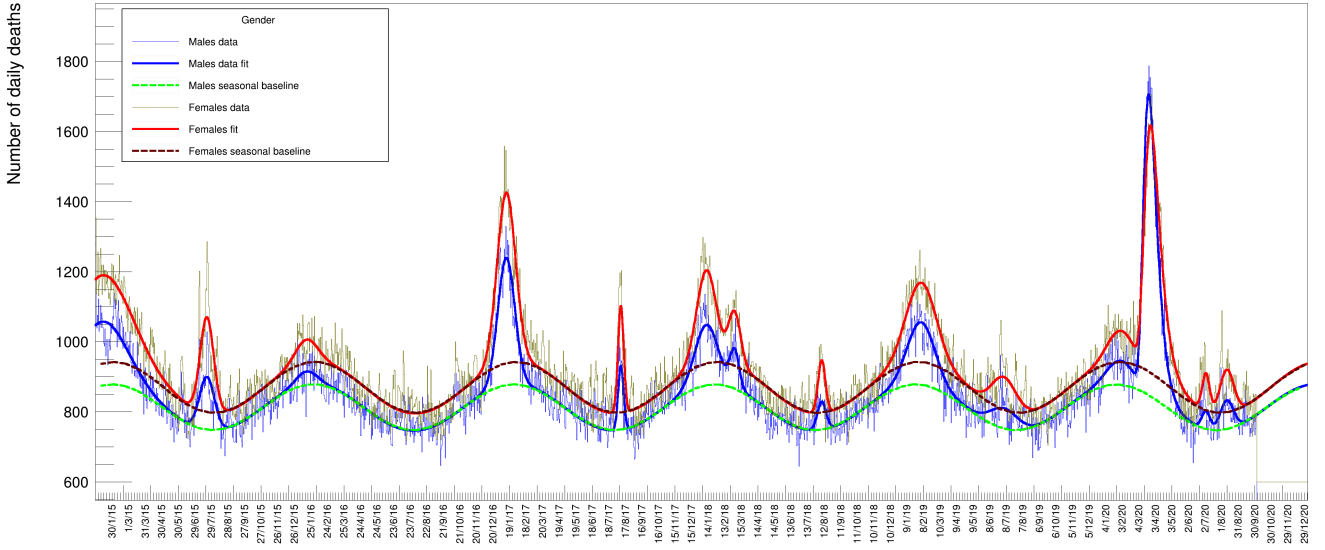

Figure 9: Number of daily casualties for males and females of all ages.

Another information that can be extracted from the data is the relative amount of deaths between genders. Fig 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 January 1st 2020). The resulting values (multiplied by $100\,000$) are listed in Table 10 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.

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.

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 February 24th, 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 Fig. 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 Fig. 4, specifically those between the dates of February 24th and September 30th 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 doesn't requires a subtraction procedure being unaffected by a background.

g_i	Males				Females			
	Yield	Mortality	Peak	Duration	Yield	Mortality	Peak	Duration
1	$20\,279 \pm 788$	69.8	10/1/2015	178	$32\,405 \pm 929$	105.9	10/1/2015	206
2	$5\,132 \pm 258$	17.7	21/7/2015	53	$8\,401 \pm 266$	27.5	21/7/2015	49
3	$1\,734 \pm 394$	6.0	17/1/2016	76	$3\,198 \pm 441$	10.5	17/1/2016	80
4	$13\,907 \pm 357$	47.9	13/1/2017	62	$20\,281 \pm 404$	66.3	13/1/2017	66
5	$2\,252 \pm 156$	7.8	9/8/2017	19	$3\,976 \pm 174$	13.0	9/8/2017	21
6	$8\,875 \pm 468$	30.6	11/1/2018	83	$11\,830 \pm 415$	38.7	11/1/2018	71
7	$2\,192 \pm 268$	7.5	3/3/2018	33	$4\,139 \pm 305$	13.5	3/3/2018	43
8	$1\,612 \pm 194$	5.5	8/8/2018	31	$2\,473 \pm 203$	8.1	8/8/2018	27
9	$11\,066 \pm 462$	38.1	4/2/2019	99	$15\,273 \pm 509$	49.9	4/2/2019	109
10	$3\,728 \pm 341$	12.8	4/7/2019	99	$5\,224 \pm 352$	17.1	4/7/2019	87
11	$4\,112 \pm 443$	14.2	31/1/2020	99	$5\,124 \pm 439$	16.7	31/1/2020	93
12	964 ± 182	3.3	4/7/2020	29	$1\,764 \pm 182$	5.8	4/7/2020	27
13	$2\,489 \pm 236$	8.6	11/8/2020	47	$3\,826 \pm 253$	12.5	11/8/2020	49

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

Males				Females			
Yield	Mortality	Peak	Duration	Yield	Mortality	Peak	Duration
$27\,240 \pm 366$	93.8	22/3/2020	63	$26\,079 \pm 395$	85.2	26/3/2020	72

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 Fig. 9. The meaning of the *Mortality* column is described in the text.

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 August 15th, 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 August 16th, 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 Fig. 10 as a vertical green arrow.

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

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.
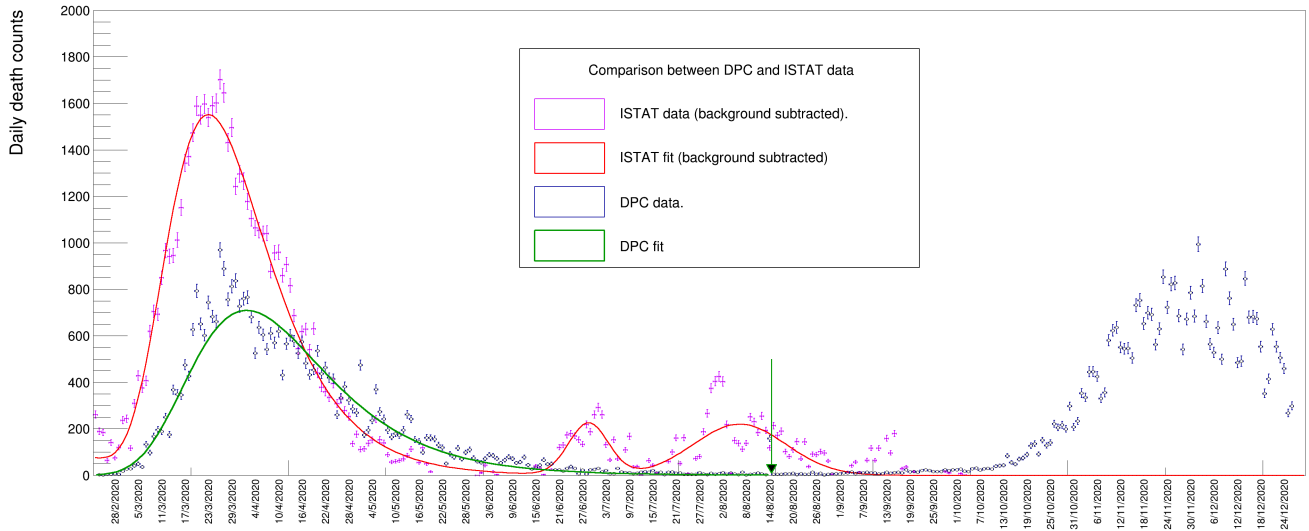
Figure 10: Comparison between ISTAT and DPC data samples

7 Additional considerations

The rich data sample provided by ISTAT allows for various additional visualizations. In Fig. 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 Fig. 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 Fig.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.
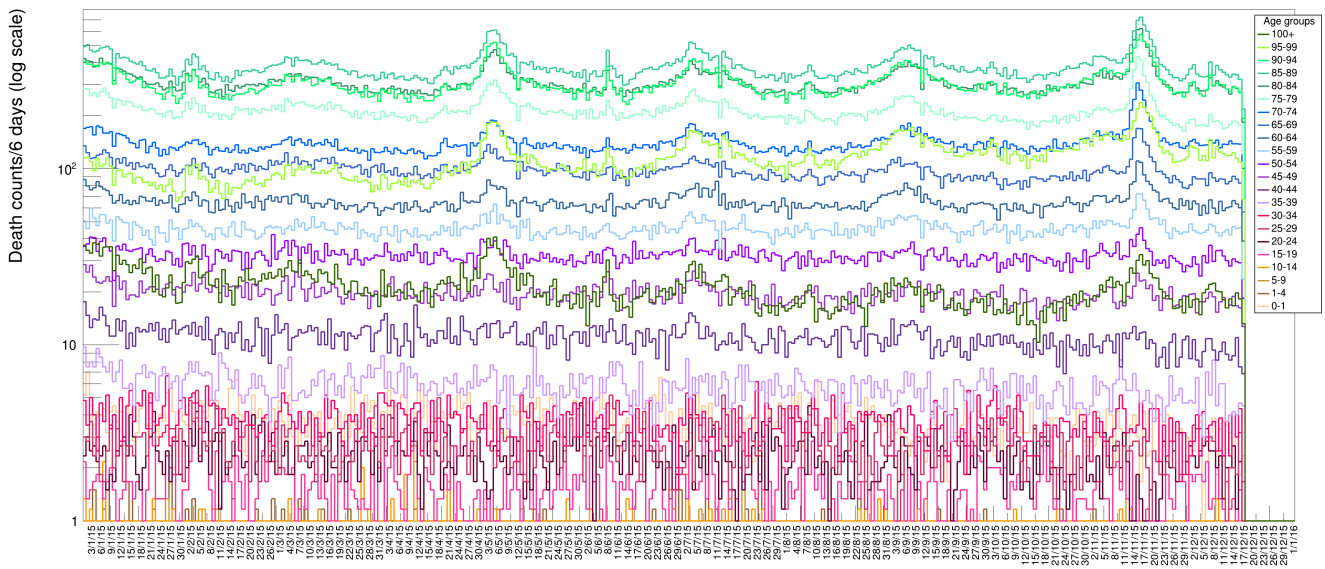

Figure 11: ISTAT data set with disentangled age categories. The data are binned in groups of six days each for an enhanced visualization clarity.

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.

Fig. 13 shows the fits for each of the four zones in Italy, namely *North*, *Center*, *South* and *Islands*[2]. The fits in
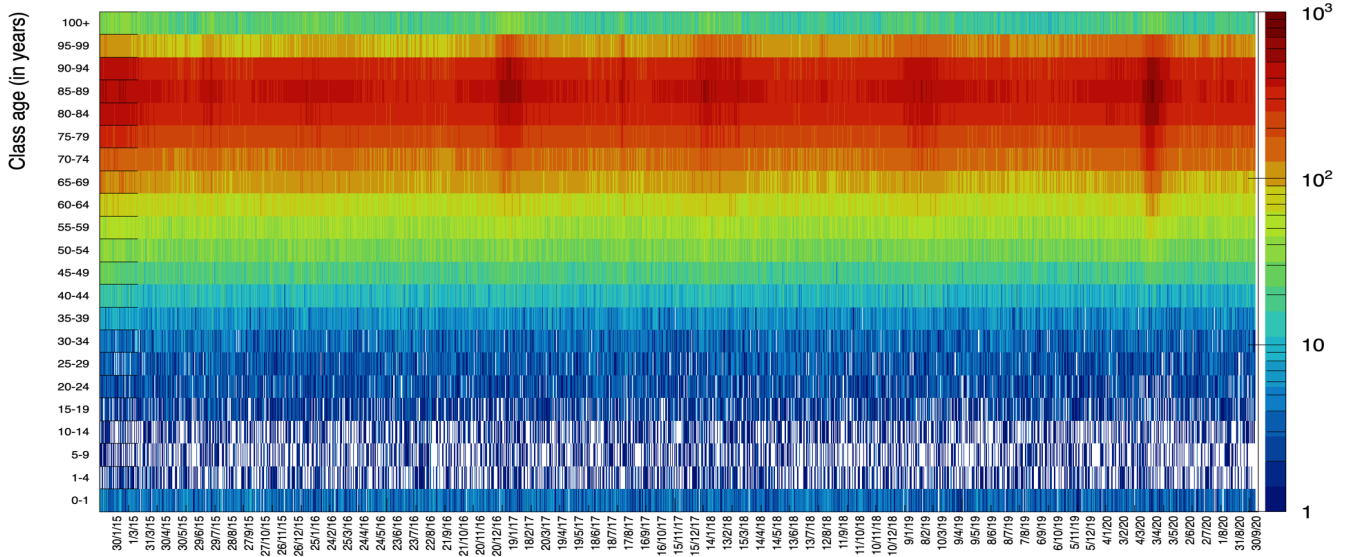

Figure 12: Scatter plot of the class age versus day of death. The dark red spots show the age and the date corresponding to highest values of deaths incidence.

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 Fig. 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).

---

[2]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*
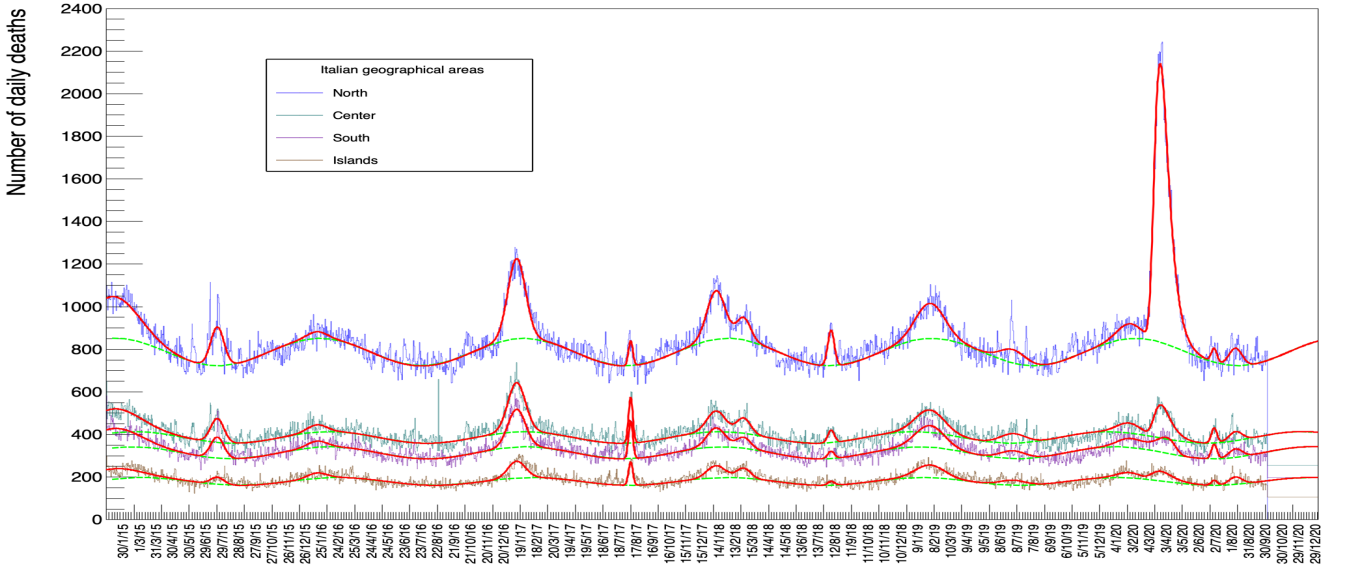
Figure 13: ISTAT data set with the Italian geographical areas disentangled (their definition is detailed in the text).

g_i	North	Central	South	Islands
1	$23\,046 \pm 518$	$13\,293 \pm 366$	$10\,670 \pm 333$	$5\,730 \pm 249$
2	$5\,520 \pm 193$	$3\,409 \pm 139$	$2\,995 \pm 126$	999 ± 91
3	$1\,333 \pm 235$	$1\,501 \pm 166$	$1\,250 \pm 151$	$1\,120 \pm 115$
4	$14\,924 \pm 254$	$9\,184 \pm 183$	$7\,058 \pm 165$	$3\,135 \pm 121$
5	$1\,143 \pm 106$	$2\,078 \pm 85$	$1\,716 \pm 77$	$1\,066 \pm 58$
6	$9\,393 \pm 246$	$4\,027 \pm 170$	$3\,676 \pm 156$	$2\,344 \pm 119$
7	$3\,213 \pm 203$	$2\,233 \pm 144$	$1\,576 \pm 129$	$1\,451 \pm 102$
8	$2\,511 \pm 135$	876 ± 93	511 ± 82	274 ± 61
9	$10\,692 \pm 302$	$6\,865 \pm 213$	$6\,620 \pm 197$	$3\,884 \pm 150$
10	$3\,711 \pm 255$	$2\,631 \pm 180$	$2\,168 \pm 161$	$1\,374 \pm 122$
11	$3\,737 \pm 264$	$2\,505 \pm 186$	$2\,033 \pm 171$	$1\,361 \pm 130$
12	836 ± 119	898 ± 87	871 ± 79	300 ± 57
13	$2\,477 \pm 182$	$1\,286 \pm 131$	$1\,188 \pm 117$	$1\,026 \pm 90$
Gompertz	$45\,350 \pm 33$	$6\,063 \pm 201$	$3\,559 \pm 205$	$2\,158 \pm 150$

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

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

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 December 31st, 2020.

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 18919 ± 557 deaths over a total of 54387 ± 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 pandemics 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.

9   Acknowledgements

References

[1] Public data on deaths in the Italian municipalities for the years 2015-2020, provided by ISTAT, *https://www.istat.it/it/files/2020/03/Dataset-decessi-comunali-giornalieri-e-tracciato-record_dati-al-30-settembre.zip*

[2] ISTAT, Istituto Nazionale di Statistica, *https://www.istat.it/en/*

[3] Hajat, S., & Gasparrini, A. (2016). The excess winter deaths measure: why its use is misleading for public health understanding of cold-related health impacts. Epidemiology (Cambridge, Mass.), 27(4), 486.

[4] Weinberger, K. R., Harris, D., Spangler, K. R., Zanobetti, A., & Wellenius, G. A. (2020). Estimating the number of excess deaths attributable to heat in 297 United States counties. Environmental Epidemiology (Philadelphia, Pa.), 4(3).

[5] Dati COVID-19 Italia, Dipartimento della Protezione Civile, *https://github.com/pcm-dpc/COVID-19*

[6] Dipartimento di Protezione Civile (DPC), *http://www.protezionecivile.gov.it/*

[7] A pandemic primer on excess mortality statistics and their comparability across countries (and further references therein), *https://ourworldindata.org/covid-excess-mortality*

[8] The Human Mortality Database, *https://www.mortality.org/*

[9] Weekly Death Statistics, *https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Weekly_death_statistics*

[10] Estimating the burden of SARS-CoV-2 in France, Salje H.,Kiem C.T.Lefrancq N.Courtejoie N.Bosetti P.Paireau J.Andronico A.Hozé N.Richet J.Dubost C.-L.Strat Y.L.Lessler J.Levy-Bruhl D. Fontanet A. Opatowski L. Boelle P.-Y.Cauchemez S., Science 2020

[11] Temporal dynamics in total excess mortality and COVID-19 deaths in Italian cities. Michelozzi P. de'Donato F. Scortichini M. Pezzotti P. Stafoggia M.Sario M.D. Costa G. Noccioli F. Riccardo F. Bella A. Demaria M. Rossi P. Brusaferro S. Rezza G. Davoli M., Bmc Public Health. 2020

[12] Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the United States, March to May 2020. Weinberger D.M. Chen J. Cohen T. Crawford F.W. Mostashari F. Olson D. Pitzer V.E. Reich N.G. Russi M. Simonsen L. Watkins A. Viboud C., Jama Intern Med. 2020

[13] A Cross-Regional Analysis of the COVID-19 Spread during the 2020 Italian Vacation Period: Results from Three Computational Models Are Compared Luca Casini Marco Roccetti, Sensors 2020

[14] Empirical model for short-time prediction of COVID-19 spreading, Martí Català,Sergio Alonso ,Enrique Alvarez-Lacalle,Daniel López,Pere-Joan Cardona,Clara Prats, Plos Computational Biology 2020

[15] Monitoring for outbreak-associated excess mortality in an African city: Detection limits in Antananarivo, Madagascar, Fidisoa Rasambainarivo, Anjarasoa Rasoanomenjanahary, Joelinotahiana Hasina Rabarison, Jean Michel Heraud, C. Jessica E. Metcalf, Benjamin L. Rice, International Journal of Infectious Diseases, 2020

[16] EuroMOMO, European MOrtality MOnitoring, *https://www.euromomo.eu/*

[17] James, F., & Roos, M. (1975). MINUIT: a system for function minimization and analysis of the parameter errors and corrections. Comput. Phys. Commun., 10(CERN-DD-75-20), 343-367.

[18] Jacoboni, C. W. T. Eadie, D. Dryard, F. E. James, M. Roos e B. Sadoulet — Statistical methods in experimental physics. Nuov Cim A 40, 235 (1977). *https://doi.org/10.1007/BF02776791*

[19] Gompertz, B. (1825). 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. Philosophical transactions of the Royal Society of London, (115), 513-583.

[20] Ohnishi, A., Namekawa, Y., & Fukui, T. (2020). Universality in COVID-19 spread in view of the Gompertz function. Progress of Theoretical and Experimental Physics, 2020(12), 123J01. doi:10.1093/ptep/ptaa148

[21] *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/*

[22] *https://www.tuttitalia.it/statistiche/popolazione-eta-sesso-stato-civile-2018/*

[23] Resident Italian population, ISTAT, *http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1*

[24] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). *https://doi.org/10.1038/sdata.2016.18*

[25] INFN Open Access Repository *https://www.openaccessrepository.it/*