A preliminary view at 2020 mortality data from ISTAT

Fabio Ricciato*

16 April 2020

1 Introduction

On 1st April 2020 ISTAT released a data set[1] containing the daily number of deaths by municipality, sex and age in the first four months of each year from 2015 to 2019. Furthermore, for a subset of 1084 municipalities (out of approximately 7900) the data set contains also the deaths for the period 1st January - 21st March 2020. The latter were collected through the Anagrafe Nazionale della Popolazione Residente (ANPR). It is important to note that such subset of 1084 municipalities *is not a random sample* and anyway *can not be taken as a representative sample* of all municipalities. The trends observed in the data set for 2020 are valid only for the municipalities included in the data set, and should not be used to extrapolate figures at other spatial levels.

Such data are aggregate over all causes of deaths, with no explicit distinction for covid19 positivity. However, in several municipalities the magnitude of covid19 impact on mortality is so strong that a simple comparative analysis of the 2020 mortality data against the previous 5 years (2015-2019) provides some coarse insight into the phenomenon, and this is the goal of this short document. What we present hereafter is a simple, descriptive exercise of data exploration and visualisation, and more solid analysis in the direction of modeling and estimation will need to be based on (future) more completed data.

This is a living document subject to corrections and additions. Future version will be uploaded on `https://ec.europa.eu/eurostat/cros/content/preliminary-view-2020-mortality-data-istat_en`.

*The author is an employee of the European Commission working in DG Eurostat. Email: *Fabio.Ricciato@ec.europa.eu*. **The views expressed in this paper are those of the author and do not necessarily represent the official position of his institution.**

[1] The data set was downloaded on 3/4/2020 from `https://www.istat.it/it/archivio/240401`. For a detailed description of the data set see the companion methodological note `https://www.istat.it/it/files//2020/03/Decessi_2020_Nota.pdf`.

2 Presentation of the data set

The data set provides the daily number of deaths for each municipality in each year since 2015. Each death is associated to the municipality where the deceased person was officially registered as resident (which may not correspond to the place of death). We consider here only the 1084 municipalities for which 2020 data are available, which cover roughly 21% of the total population[2]. As explained in the companion methodological note by ISTAT[3], the municipalities included in this data set meet the following three conditions:

- It is included in the ANPR database (currently, ANPR covers around 3/4 of total municipalities in Italy);
- It had no less than 10 deaths in the period 1 January - 21 March 2020;
- It had at least +20% increase in total deaths in the period 1-21 March 2020 over the average number of deaths in the same period during the previous 5 years 2015-2019.

The latter two conditions make evident that we are dealing with a non-random sample, selected on the basis of the data themselves, and therefore affected (by construction) by a selectivity bias towards municipalities with higher covid19 impact. For this reason, as noted already earlier by Rettore and Tonini[4], it is not possible to take these data as representative of what has happened in other municipalities outside the sample (a simple extrapolation would lead to a major over-estimation error).

The map in Fig. 1 shows (marked in green) the territories of the 1084 municipalities included in the data set. The boundaries of regions and provinces are also indicated, respectively, with ticker blue and thinner gray lines. It is evident from Fig. 1 that the geographical coverage of the ANPR data set is not uniform over the country: the regions of Lombardia and Emilia-Romagna, that were most affected by covid19 in the considered period, are over-represented in the data set, while conversely the southern regions, that are less impacted by covid19 deaths, are under-represented. Therefore, the relative mortality figures from this data set cannot be taken as representative of the whole national scenario by simple extrapolation. However, they give a first idea of the order of magnitude of the phenomenon *in those municipalities included in the data set*. The Province of Bergamo (BG), where covid19 had a very strong impact, is highlighted in red and a close-up view is given in the inset image.

We consider only the calendar interval between 1-Jan and 21-Mar. This interval consists of 81 days for the leap years 2016 and 2020, and only 80 days for the other years. In order to align the time-series across different years, we include in the non-leap years a dummy 29-Feb day with all variables padded

[2] According to population figures on 1.1.2019 from `http://dati.istat.it`.

[3] `https://www.istat.it/it/files//2020/03/Decessi_2020_Nota.pdf`.

[4] See `https://www.lavoce.info/archives/65042/decessi-da-covid-facciamo-chiarezza-sui-dati-istat/` and `https://www.lavoce.info/archives/65171/morti-da-coronavirus-calcoli-sul-campione-inadatto/`.

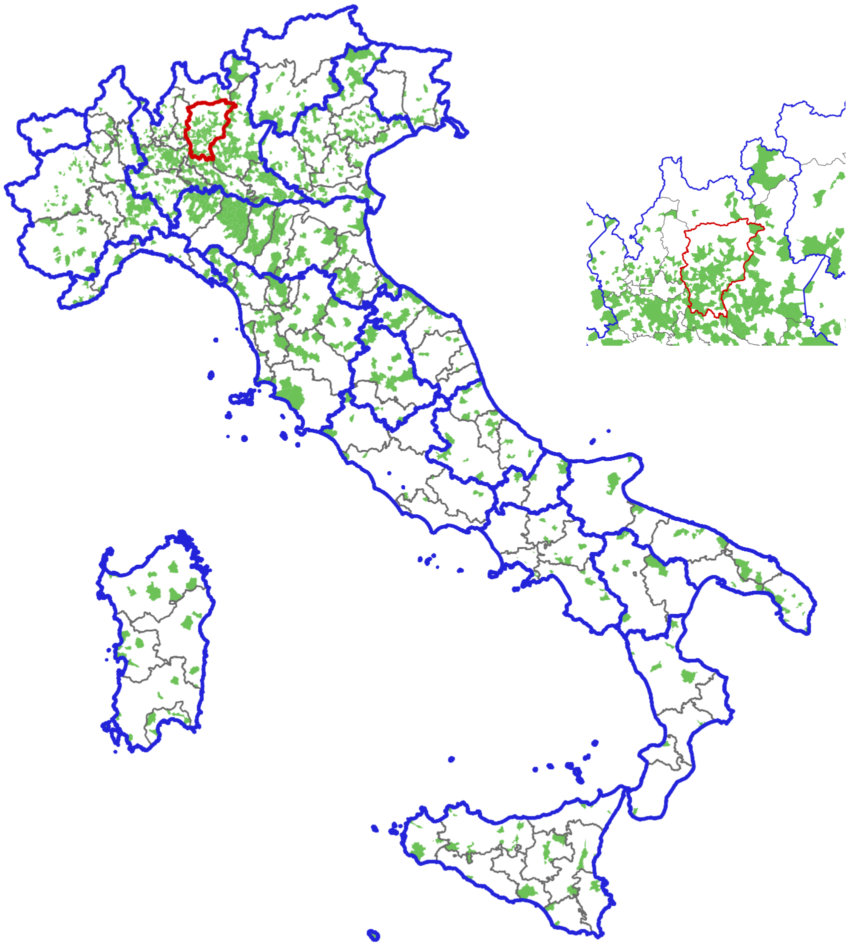


Figure 1: Map of ANPR municipalities included in the data set. Province of Bergamo (BG) highlighted in red (close up in inset image).

to zero. This facilitates the graphical representation of the results without impacting the validity of our findings.

For day $k = 1, \ldots, 81$ (starting from 1-Jan) we shall denote by d_k the number of deaths on that day, and by $D_k \overset{\text{def}}{=} \sum_{i=1}^{k} d_k$ the *cumulated* number of deaths in the period from 1-Jan until day k of each year.

3 Exploration of total deaths across all 1084 municipalities in the data set

In Fig. 2(a) we plot the total daily deaths d_k for all 1084 municipalities in the data set during the first 81 days of each year (recall that $k = 60$ corresponds to 29-Feb and is zero-padded for non-leap years). Different marker colors distinguish the data points for the different years. The temporal profile is pretty stable for the different years, except for a clear diverging trend starting at the very beginning of March 2020. This divergence is due partially, but not entirely to covid19 spreading: as explained by Rettore and Tonini[4] the selective construction of the data set, and specifically the "+20% increment over the previous 5-year baseline" criterion, concur to produce this effect.

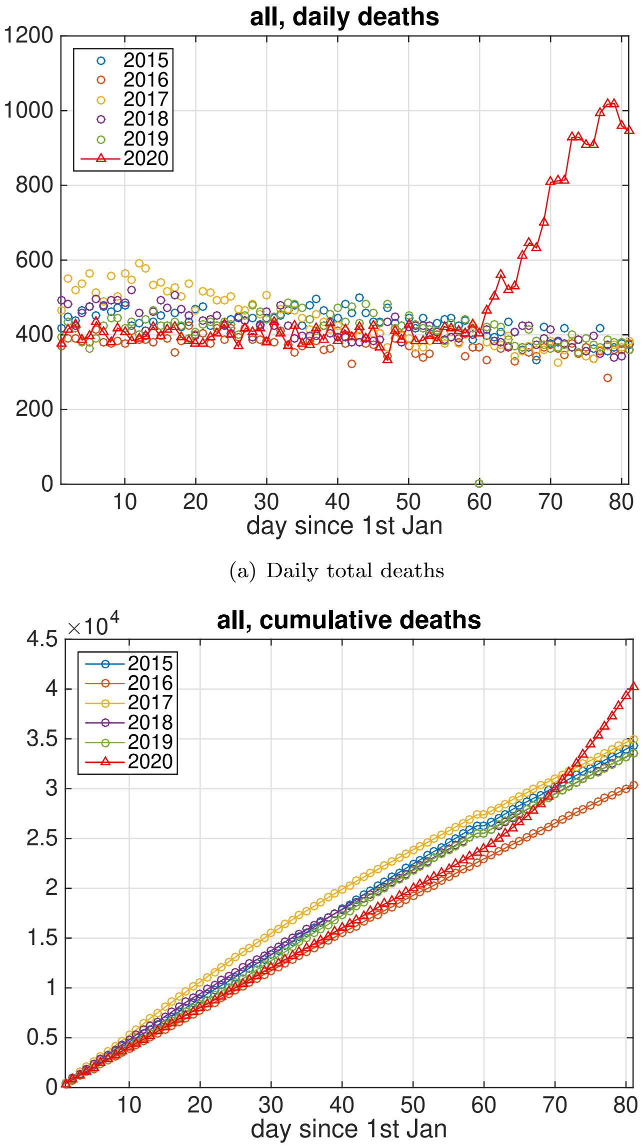
At the end of the observation interval (21-Mar), the number of daily deaths in 2020 is higher by a factor of ×2.5 than the previous years, from around 400 up to c.ca 1000 deaths/day. Without complete data it is impossible to discriminate precisely what share of such growth is explained by data set selection and how much is to be accounted to covid19. However, the preliminary study by Rettore and Tonini[4] (based on the comparison of 2019 figures over previous 4 years baseline) indicates that selectivity may be responsible for roughly +50% increase, leaving a residual +100% increase due to covid19.

In Fig. 2(b) we show the *cumulated* number of deaths D_k. Interestingly, the deaths in 2016 were markedly lower than the other years in the data set during the considered calendar period: this is probably explained by the excess of deaths among elderly people registered throughout the previous year 2015[5]. Taking as baseline the average of the previous 5 years (33373), the number of deaths in 2020 (40244) includes an excess of about 7000 deaths occurring mostly in March.

Next, we look at the incidence of excess deaths by age and sex. To this aim, we focus again on the latter 7 calendar days in the data set (15-21 Mar). For this interval, we plot in Fig. 3 the absolute number of deaths across all municipalities in the data set, in each year, for different age groups (1-4, 5-9, 10-14, 15-19, etc.). The age distribution of deaths is remarkably stable for the previous years 2015-2019, with a peak in the age group 85-89. The close agreement between pre-2020 data justifies taking the average value over the 5-year period 2015-2019 as the baseline reference.

For the upper age group, we plot in Fig. 4 the relative increment of deaths in 2020 over the pre-2020 baseline. For a death count d (value in 2020) and baseline

[5] See `https://www.ncbi.nlm.nih.gov/pubmed/26951698`.



(a) Daily total deaths

(b) Cumulative total deaths

Figure 2: Daily and cumulated deaths for all municipalities in the data set. The x-axis reports the calendar day since 1-Jan ($k = 1$). Day $k = 60$ corresponds to 29-Feb (dummy datapoint for non-leap years). The last day in the dataset is 21-Mar ($k = 81$).

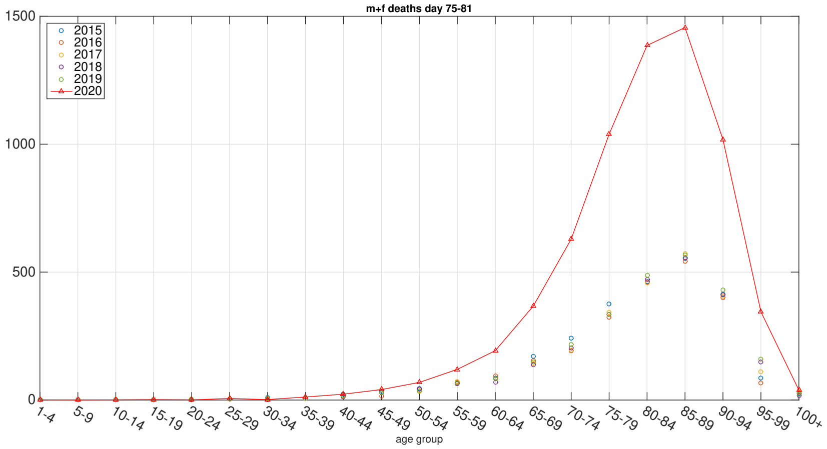


Figure 3: Age distribution of total deaths in the period 15-21 March.

value b (i.e., average of 2015-2019 values) the relative increment is defined as the ratio $\frac{d-b}{b}$. The plot reports separate figures for males (m) and females (f) and the total (m+f). A relative increment of 1 and 2 means, respectively, that the death toll is doubled and tripled with respect to the baseline (100% and 200% increase) in that specific population age group. The component of excess death due to data set selection should not change the distribution of deaths by sex and age, i.e., we would expect that all population groups would experience the same relative increase. Instead, starting from age 60 and above, we observe that (i) males are hit much stronger than females and (ii) the relative increment is considerably higher than the +50% figures that, as said earlier, could be accounted to selectivity. In the considered 14-days interval, the death counts is more than doubled for males aged 60-69 and more than tripled (relative increment above 2) for males aged between 70-89.

The empirical cumulative distribution of excess deaths (defined by the difference between the deaths in 2020 and the baseline average of previous 5-years) per age group is plotted in Fig. 5, showing that 90% of excess deaths are aged 65+.

The distribution of deaths per age and sex has shown that the population group of males aged 65+ yields a stronger mortality to covid19. For this population groups, we plot in Fig. 6(a) the number of daily deaths for all municipalities in the data set. Also this plot, likewise the previous time series in Fig. 2(a), shows a mild indication that the steep raise started at the very beginning of March 2020 has stopped, reaching a peak or plateau around 18 Mar. This finding is in line with the recent analysis by Dalla Zuanna[6] based on the number of intensive care hospitalisations, indicating that the peak of covid19 was reached around the same days under the current containment measures.

[6] G. Dalla Zuanna, *Coronavirus: La luce in fondo al tunnel*, `https://www.neodemos.info/articoli/coronavirus-la-luce-in-fondo-al-tunnel`.

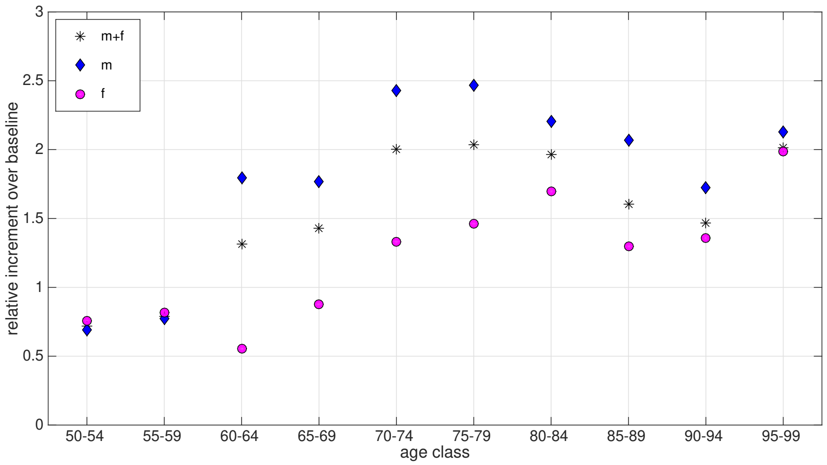


Figure 4: Relative increment of 2020 over baseline in 15-21 March per age group.

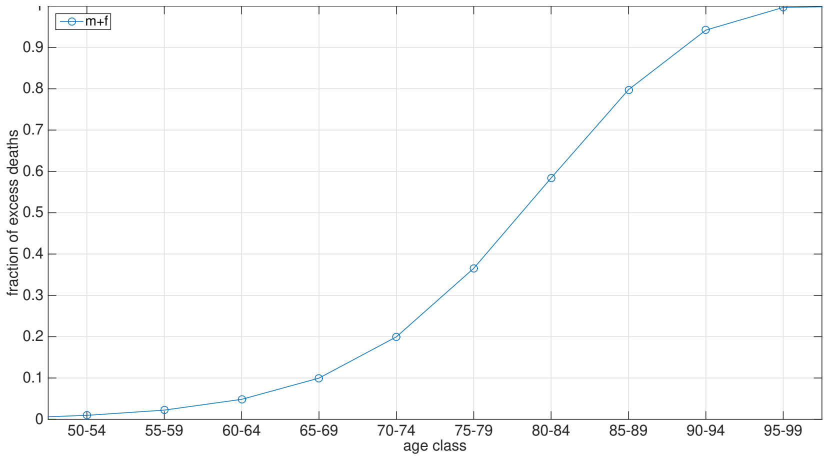
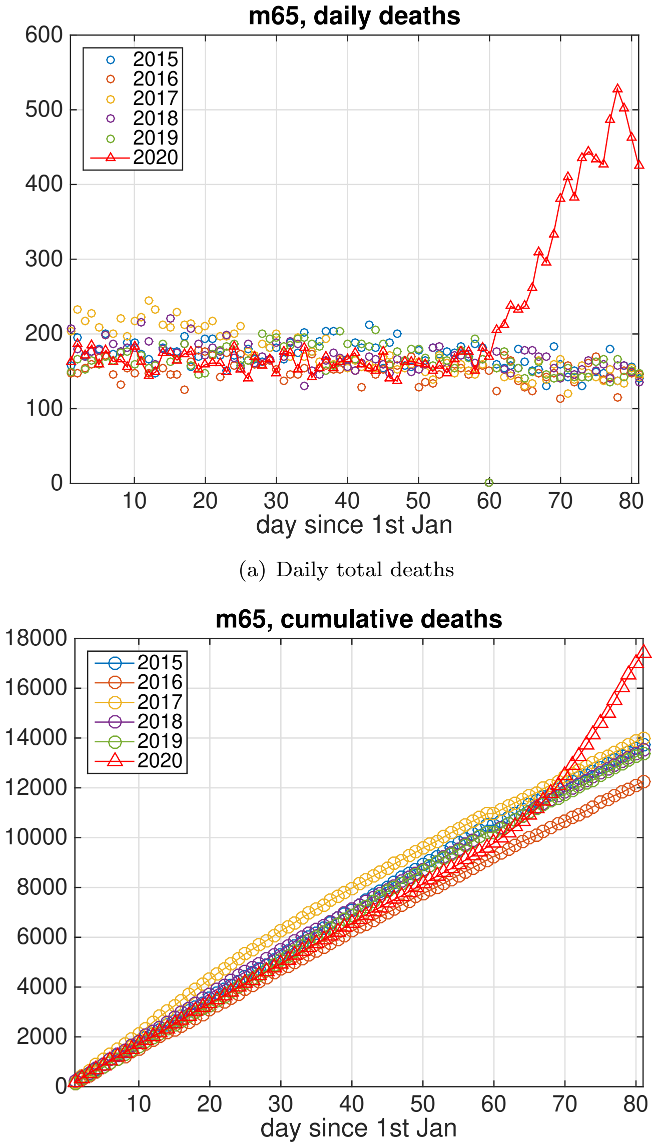


Figure 5: Empirical cumulative distribution of excess deaths in 15-21 March 2020 per age group.



(a) Daily total deaths

(b) Cumulative total deaths

Figure 6: Daily and cumulated deaths of males aged 65+ (all 1084 municipalities in the data set)

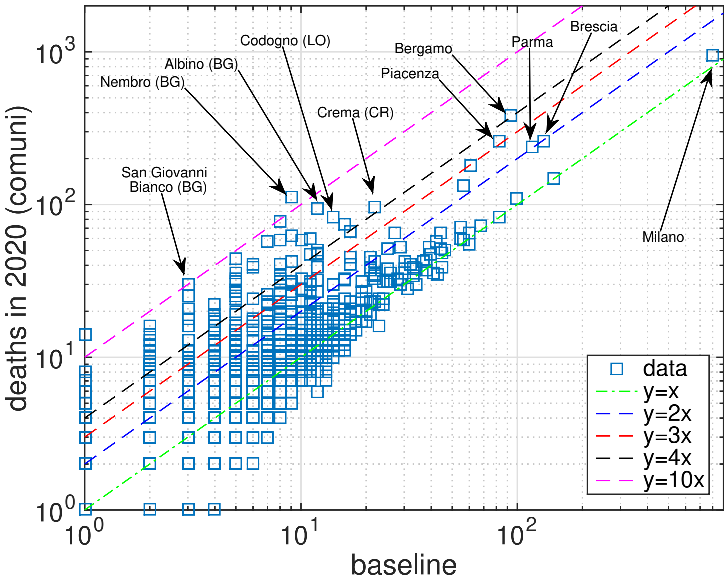


Figure 7: Total deaths in the period 1-21 March per individual municipalities: values of 2020 (vertical axis) against the baseline (horizontal axis). The baseline value is taken as the *maximum* death count during the same period in the previous 5-year. Logarithmic scale. For each province, only a fraction of the municipalities is included in the data set.

4 Preliminary look at individual municipalities

We now focus on the death count for individual municipalities. Focusing on the last three weeks in the data set, i.e. on the calendar period 1-21 March, we compare in Fig. 7 the total deaths in 2020 (y-axis) against the *maximum* value of deaths recorded in the previous 5 years (x-axis).

For some of the municipalities with the highest increase of death count, the Figg. 8-12 show the temporal profile of deaths, daily and cumulative, among the whole population and in the sub-population of males aged 65 and above (65+).

5 Preliminary look at groups of municipalities within the same province

In order to identify other hot-spots areas, we have grouped together municipalities within the same province and summed up their respective death counts. Recall that, for each province, (i) only a subset of the municipalities are included in the data set (not all), and (ii) specifically the municipalities with highest death

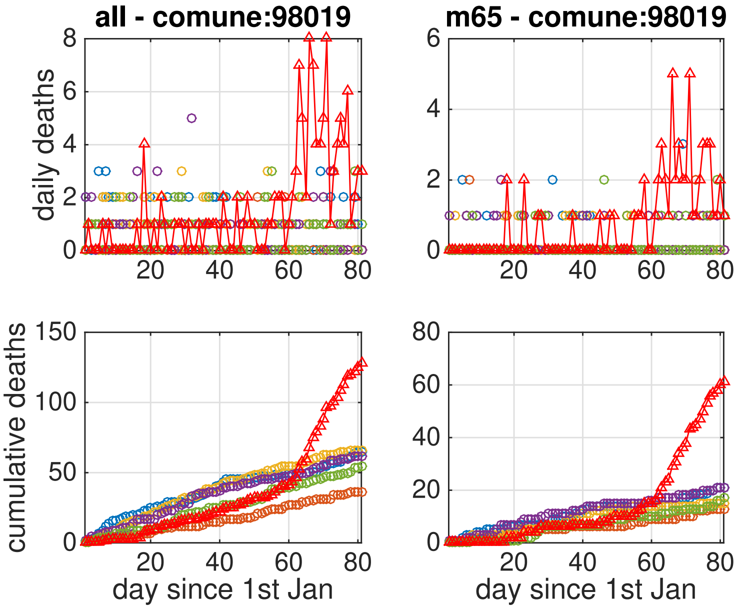


Figure 8: Codogno (LO)

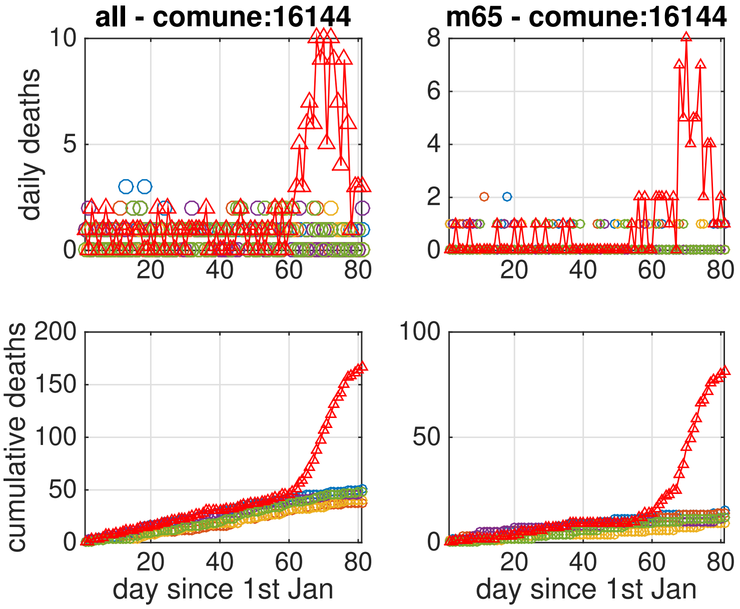


Figure 9: Nembro (BG)

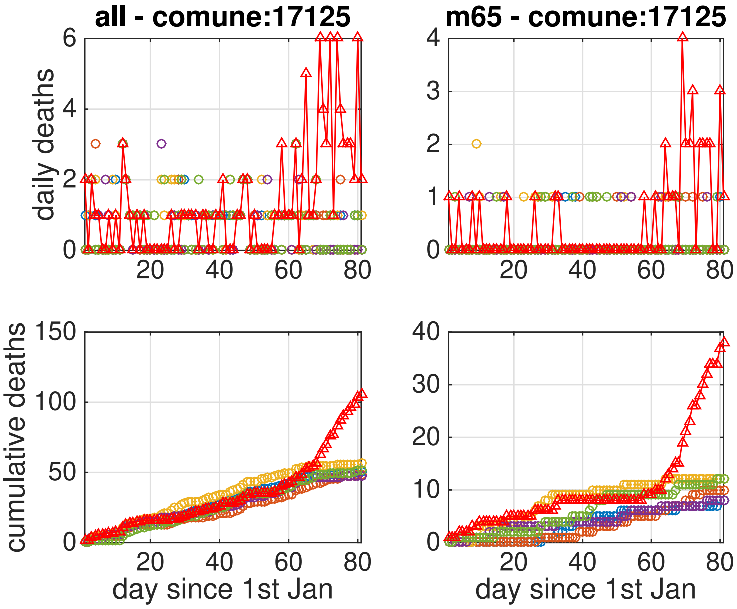


Figure 10: Orzinuovi (LO)

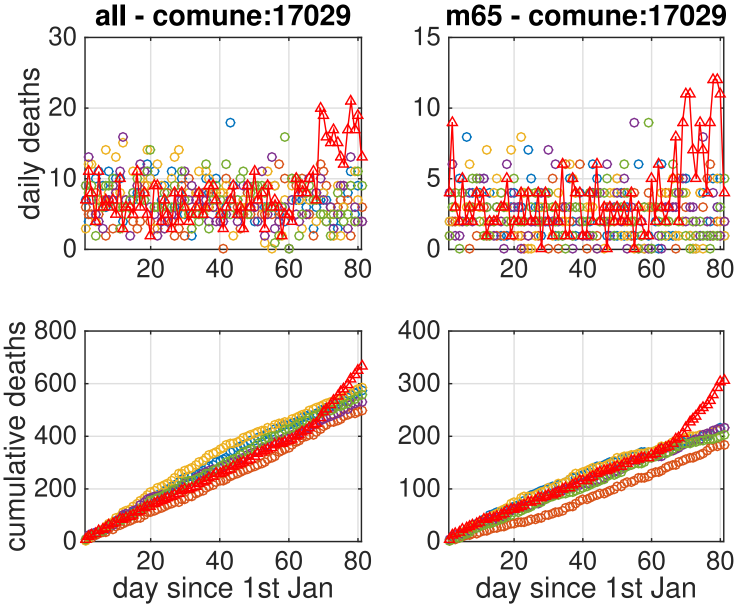


Figure 11: Brescia (BS)

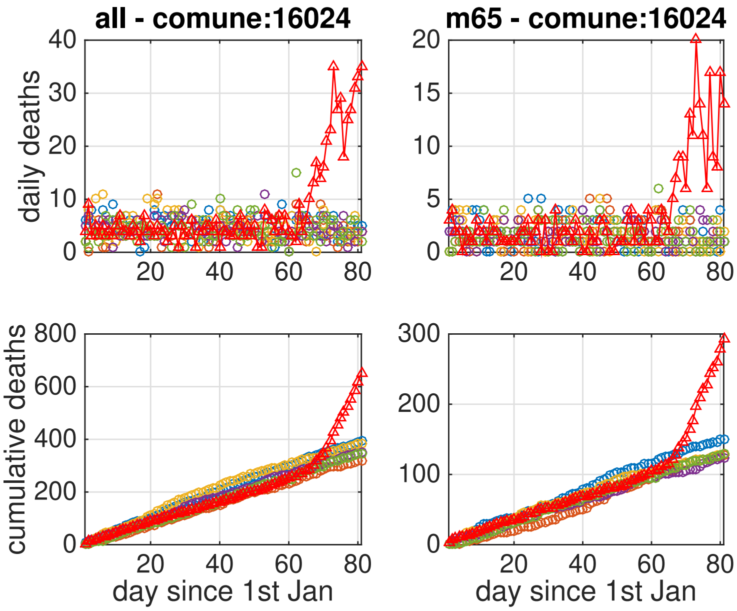


Figure 12: Bergamo (BG)

growth in March 2020 (over the previous 5-year average baseline). Therefore, as far as the *absolute number of deaths* is concerned, the actual value for the whole province is certainly *higher* than the value observed in this data set (due to (i)). On the other hand, as far as the *relative growth* of 2020 death count over the previous years is concerned, the actual value for the whole province is likely *lower* than the value observed in this data set (due to (ii)). In other words, the growth rate shown in the following plots is only valid for the municipalities included in the sample, and should not be taken as representative of the situation across the whole province (for which more complete data will be needed, covering also the remaining ANPR municipalities).

With these caveats in mind, we now present some results for different groups of municipalities within the same province. We report in Fig. 13 the death count for the last 7 days in the observation interval (15-21 March) against the baseline value represented by the *average death count for the previous 5 years* in the same calendar days. The plot is in doubly logarithmic scale and shows only the 55 provinces (out of 102) with more than 10 death events reported in the data set during the considered 7-day interval. Reference lines are added to ease the visual assessment of the growth factor of 2020 death count relative to the baseline values. For 25 provinces the growth factor is above ×2, and for 4 provinces it is above ×4. Collectively, the group of municipalities within the Province of Milano has almost doubled the number of deaths in the last week of the considered period (over the baseline average value of previous 5-years),

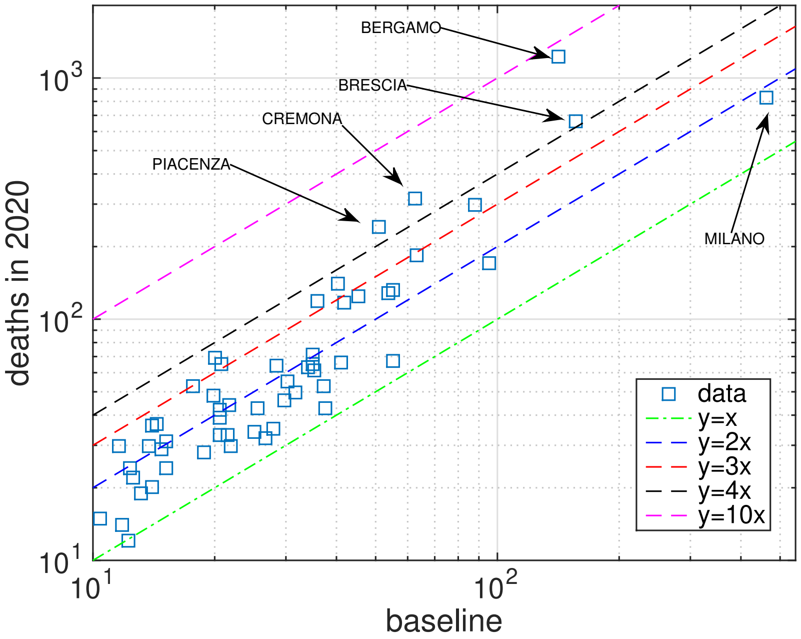


Figure 13: Total deaths in the week 15-21 March by groups of municipalities within the same province: values of 2020 (vertical axis) against the baseline (horizontal axis). The baseline value is taken as the *average* death count during the same period in the previous 5-year. Logarithmic scale.

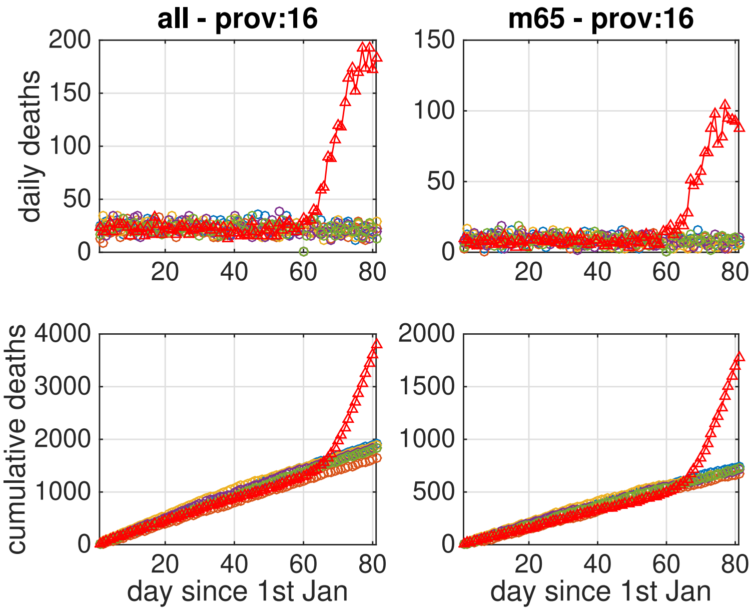


Figure 14: Municipalities in the Province of Bergamo (BG)

while for the group of municipalities within the Province of Bergamo the growth factor is around ×9.

Not surprisingly, the temporal trends that are already evident at the country level are even stronger if we focus on covid19 hot-spot areas. To illustrate, we plot in Figg. 14-16 the temporal profile of daily deaths (top plots) and the accumulated count since 1-Jan (bottom plots) for some of the provinces with highest covid19 impact, selected on the basis of the scatter plot in Fig. 13. The left-hand plots refer to total deaths, while the right-hand plots report only the deaths of males aged 65+.

The Province of Bergamo was hit particularly strong by covid19. From the bottom-right plot in Fig. 14 we can see that the group of municipalities in this province account for around 2000 of the total 7000 excess deaths observed in the data set in the considered period (recall however that these numbers refer to only a subset of the municipalities, therefore the final absolute numbers for the whole province will be even higher). It is striking that the number of daily deaths in this group of municipalities peaks at a value that is roughly ×9 times higher than the normal level observed in the previous years, from around 20 deaths/day up to around 180.

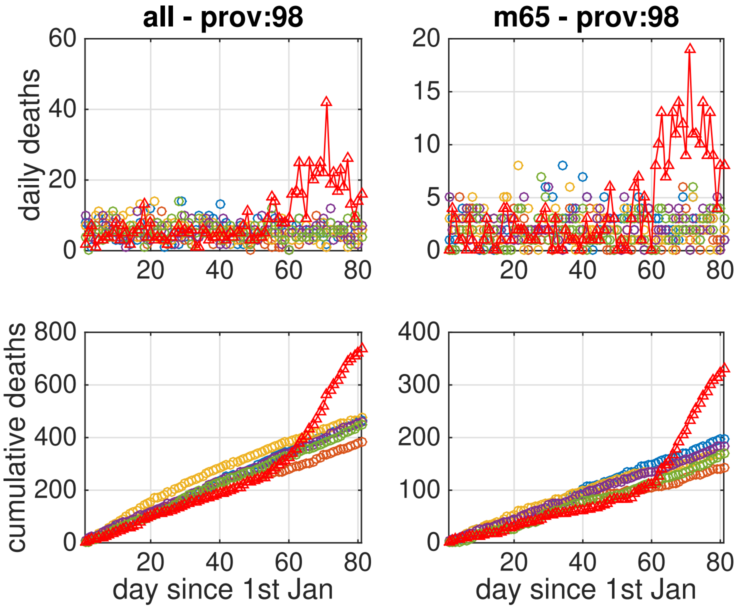


Figure 15: Municipalities in the Province of Lodi (LO)

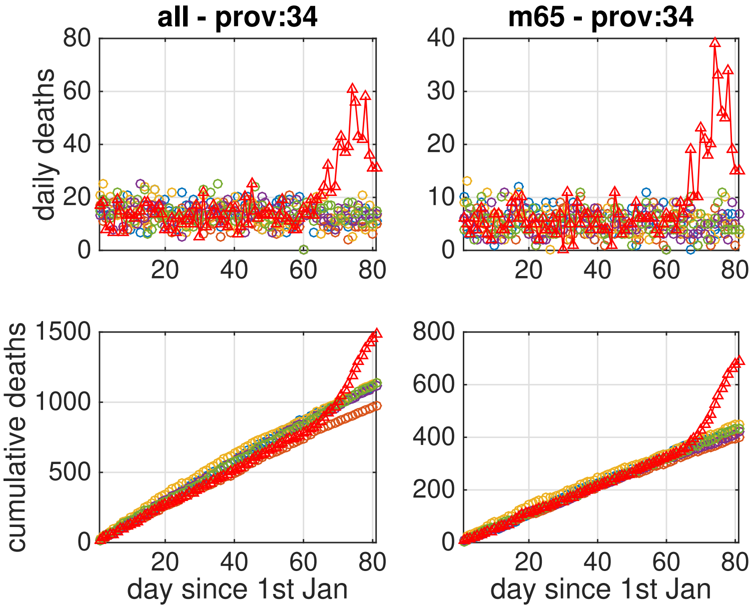


Figure 16: Municipalities in the Province of Parma (PR)

6 Considerations

These quantitative values presented in this simple analysis should be taken with extreme caution, since they refer to a dynamic process that was still in a transitory regime during the observation period. Also, the aggregate counts blend together dynamics that are highly heterogeneous across different spatial regions. Considering (i) the strong bias in the selection of municipalities; (ii) the fact that we have reported absolute death counts, with no re-scaling on the actual population size; (iii) the lack of temporal stationarity; and (iv) the lack of spatial homogeneity, it should be clear that the global values of relative increment cannot be taken as indicative of *frequencies* or *probabilities*, for the estimate of which a more solid analysis based on more complete (future) data will be needed. This goes far beyond the scope of this contribution. However, the fact that clearly distinguishable trend variations emerge even from a basic, merely descriptive analysis of coarsely aggregate data, such as the one presented here, is quite telling.

The number of covid19 deaths recorded officially for the whole Italy up to 22nd March 2020 were 5476[7] (4825 up to 21st March), while we have observed around 7000 additional deaths, in excess of the previous 5-year average baseline, solely for the 1084 municipalities included in the data set. However, as noted by Rettore and Tonini[4], the gap between the two figures may still be explained (at least partially) by the effect of data set selection. In other words, the present data do not allow to conclude that official figures of covid19 deaths based on test positivity are severely under-estimating the phenomenon. Such assessment will need to be based on more complete data referred to the whole set of municipalities (or at least an unbiased sample thereof).

The following considerations can be made, that are indeed rather obvious for experts, but appear even more compelling and evident also to non experts in the light of the presented figures.

Unfortunately, the data used for the present analysis do not include the cause of death, a variable that is not present in the ANPR and is collected by ISTAT through other procedures. More detailed mortality data including cause of death will be available in the future after several months[8]. We expect that a dis-aggregate view of death counts per cause of death (and for all municipalities) could have revealed the anomalous trend before the beginning of March. We anticipate that, when complete and detailed mortality data will be made available to the research community, their retrospective analysis will reveal noticeable signs of epidemic infection occurring in local areas well before March (e.g., anomalous clusters of early deaths due to pneumonia locally concentrated in time and space). The researchers will probably find that such first signs could have been detected with relatively simple algorithms running automatically on the detailed mortality data. An interesting research question is whether an early warning could have been triggered by such algorithms *before* the time of

[7] http://opendatadpc.maps.arcgis.com/apps/opsdashboard/index.html#/b0c68bce2cce478eaac82fe38d4138b1

[8] Anyway within 24 months, see https://www.istat.it/it/archivio/4216.

first official covid19 positive test in Italy, thus enabling a more timely (hence effective) initial response. Such retrospective analysis will not be useless if the results are leveraged to improve the timeliness of data collection and elaboration processes, and in this way get better prepared for the next epidemics.

Besides initial detection, also during the progress of epidemic spread the *timely* availability of such data to experts and researchers would contribute to better picture the evolving scenario and tune the response actions by the competent authorities.

Generally speaking, there is nothing extraordinary in advocating *better data for better response*, but the importance of data *timeliness* is strikingly evident in this specific case. Another lesson to be learned is about unlocking *the value of existing traditional data sources*: the analysis presented here is based on plain traditional administrative records. Driven by the covid19 emergency, we may well discover that gathering traditional data in a truly timely manner — 24 hours instead of 24 months — is not less important and valuable than taking stock of new kinds of non-traditional data. Timeliness of traditional data can and should be improved, and the current situation makes it clear that the cost of doing so (e.g. by stepping up digitalisation of the reporting procedures) is lower than the cost of not doing so.

In the specific case of epidemic emergency, the timely availability of detailed mortality data (and probably also hospitalisation data) from the previous day — rather than from the previous year — combined with an effective process for continuous analysis and automatic warning, could be the key for a more timely detection of (and reaction to) the next epidemic spread.

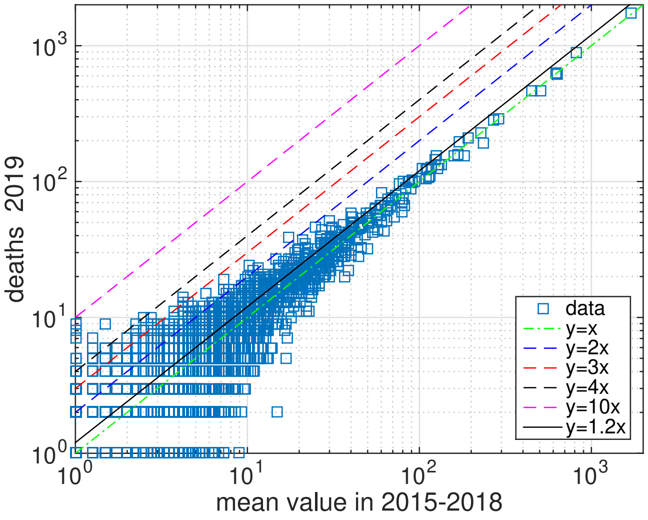
Acknowledgments

Thanks to Albrecht Wirthmann for helping with QGIS and Marija Mamolo for several corrections.

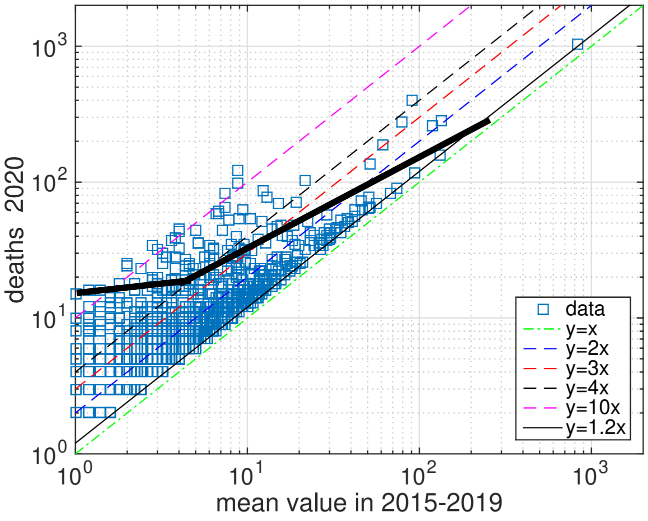
APPENDIX - effect of data set selection

In this section we take a first step to investigate the effect of data set selection, and specifically the "+20% increment criterion". In Fig. 17(a) we plot, *for each municipality in Italy* (approximately 7900 in total), the number of deaths in the period 1-21 March 2019 (y-axis) against the baseline value corresponding to the mean value of the previous 4 years 2015-2018 (x-axis). The black continuous line correspond to $y = 1.2x$: if we had to apply the "+20% increment rule" to such data, all data points below this line would disappear, and the total number of deaths over the residual data points would carry a positive bias (whose value was estimated to be around +50% by Rettore and Tonini[4]). Among the residual data points, small-size municipalities may occasionally yield values that are multiple times larger than the baseline, but as expected the width of the random fluctuations band shrinks rapidly for medium-size data points, and vanishes for large-size municipalities.

In Fig. 17(b) we plot, *for each of the 1084 municipalities for which 2020 data were available*, the number of deaths in the period 1-21 March 2020 (y-axis) against the baseline value corresponding to the mean value of the previous 5 years 2015-2018 (x-axis). As expected, no single data point appears below the $y = 1.2x$ line, due to the selection based on the "+20% increment criterion". However, we observe several data points well above the physiological band of random fluctuations (indicated by a thick line). For these municipalities, the relative increment of 2020 deaths over the previous 5-year baseline can be accounted mostly to the (direct and indirect) effect of covid19.



(a) 2019 values vs. 2015-2018 average, all 7900 municipalities in Italy.



(b) 2020 values vs. 2015-2019 average, 1084 selected municipalities

Figure 17: Scatter plot of deaths in the period 1-21 March for each municipality: 2019 vs 2015-2018 average (top) and 2020 vs 2015-2019 average (bottom). The effect of selection due to "+20% increment" criterion is clearly visible in the bottom plot: missing data points in the lower part, below the $y = 1.2x$ reference line. The effect of covid19 is also clearly visible: data points above the random oscillation band, above the tick line.