UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE,
ECONOMICHE E SOCIALI

CORSO DI LAUREA IN SCIENZE POLITICHE

CURRICULUM: POLITICS AND ECONOMICS

"How to critically read COVID-19 data: Some coordinates not to get lost"

Elaborato finale di: Andrea Pio Cutrera

Relatrice: Prof.ssa Silvia Salini

Anno Accademico: 2019/2020

Index:

“Cominciate col fare ciò che è necessario, poi ciò che è possibile,

e all’improvviso vi sorprenderete a fare l’impossibile”

San Francesco

1 Premise

As a premise, I want to be clear that this work does not presuppose to be an epidemiological study about the *Covid-19* disease. It only wants to be a guide in the reading of some data and a vademecum to be more aware in front of some models that statisticians and epidemiologists present to the civil society. Statistics, even in this field, is crucial and it will be our tool to see in a clearer way what is happening around us.

In these days we are living something that anybody has never seen. A virus, that is only a single strand of *RNA* (ribonucleic acid) linked to a protein, is spreading death across the world more than how any other one can be remembered by human beings.

The World Organisation of Health (WHO), in the January 9th of 2020 declared that a new virus – denominated "*Corona*" for his crown in the more external part of the capsid – has been identified, and its name was "*2019-nCoV*". Then "*SARS-CoV-2*" was its official name, and the associated disease "*Covid-19*". This name came from the pneumonia outbreak in December 31st of 2019 localised in Hubei (China) to which the *SARS-CoV-2* was linked; in Italy only one month later 2 cases of disease were assessed, but the autochthonous case came up in 21st of February[1].

From that moment on the entire country (Italy) was called to arms for dealing with this. Chronologically, Italy is one of the first countries that had to switch all the priorities to the health care system support, but no other countries can be excluded from it. It's only a matter of time whether the virus has appeared or not.

[1] Istituto Superiore di Sanità. (April 2020). website: https://www.iss.it/coronavirus

I posed the question of whether it was possible to study the epidemiological diffusion between Provinces or Regions in Italy in a quantitative way and how these studies are reported and read by citizens. Many statisticians were called for help in the creation of models jointly with epidemiologists; researchers, experts and all the people of science need do put all the efforts in order to well advise decision makers. Time for discussion is very little, cost-benefit analysis are even difficult to do. But who are the scientists entitled to explain, rationalise and forecast what is happening around? Only physicians and epidemiologists? In these times a new kind of science is needed, so the 'post-normal' paradigm, presented in the 90's of the last century by Funtowicz and Ravetz, seems to be an optimal solution for this period: an "*extended peer-community*", instead of a *technocratic* one[2] (Waltner-Toews et al, 2020), is the requirement to answer very important questions when "*facts are uncertain, stakes high, values in dispute and decisions urgent"*[3] (Funtowicz and Ravetz, 1993).

Citizens, in the meanwhile are restricted at home; who is not afflicted from the contagion has nothing to do or worse nothing to eat. They only have the opportunity to see mere aggregated data, but are they able to read them? Data are playing an essential role more than ever in these times, but people do not have a sufficiently good data literacy for understanding and discerning what is true from what is fake in an *infodemic* flood of news.

With this paper I want to make a critical discussion of some descriptive analysis of the phenomenon under study – nominally *Covid-19* – in particular how the trends of the cases are used to make comparisons and explanations, but also to report some more elaborate

[2] Waltner-Toews, Biggeri, De Marchi, Funtowicz et al. (preprint - April 2020). *Post-normal Pandemics: Why Covid-19 requires a New Approach to Science*. ESRC STEPS Centre blog.

[3] Funtowicz and Ravetz. (September 1993). *Science for the post normal age*. Futures, 25 (7), retrieved from: https://search-proquest-com.pros.lib.unimi.it:2050/docview/1292249765?accountid=12459

models that can suggest us insights in what cannot be seen in the raw data that we hear from the daily report of Civil Protection Department of Italy or journals. It wants also to be a warning for a cautious reading of any explanatory or forecasting model we will be in front of.

2 Introduction

As we have seen, the autochthonous first case in Italy was assessed in the 21st of February 2020 in a little city of Lombardy[1] and it spreads as fast as nobody can imagine. Once appeared the only thing we can do is to stay at home and not to have social relations, since no cure or prevention exists, this is what experts of the National Superior Health Institute (ISS) suggest to the Prime Minister in Italy: all strategies that are called "*Non-Pharmaceutical Interventions – NPIs*" (Agosto et al, 2020)[4].

He has to learn from researchers how much we do need to postpone our return to our lives in safe balancing all the economic losses that each economic sector is making by lockdown. Citizens read and hear every day the increasing trends in the aggregated numbers of national cases, deaths associated with the epidemy and the recovered from the disease. Numbers that are given daily, with a press conference by the Civil Protection Department; they increase every day: they are unbelievable. Journals make a run for being the first to report data and get visualizations; but they are reported almost always as they are given, with no filters of understanding. So, what are we learning from these data? Something crucial or something irrelevant? This is what we are going to discuss about in this paper.

Other data have been released by Istat (Italian National Institute of Statistics); they are a complete and exhaustive source of information about the annual deaths of individuals, but they require a lot of time for being collected in their entirety[5], and still they cover a small share of Municipalities.

[4] Agosto et al. (March 2020). *Monitoring Covid-19 contagion growth in Europe*. CEPS Working Document, 1-2.
[5] Istat. (April 2020). website: https://www.istat.it/it/archivio/4216

In our discussion we are going to see first which kind of data are reported every day by Civil Protection and why they are biased and need some corrections with the help of a paper written by researchers of University of Florence[6]. The same will be done for Istat data on deaths in 2020.

We will go deeper in the analysis of some studies that further explain how aggregate data of Istat can be a good benchmark for understanding the underestimation of people died, being always aware of the limits of inferences.

In addition another more comprehensive epidemiological study, called *SIR*, will be presented. It is very useful to better understand the life of a virus through the rate at which other people get infected, rate at which people recover and also the rate at which people die. And at the end Basic reproduction number ("*R-with-zero*") parameter will be explained as a crucial signalling factor of the spreading of the virus for political choices.

[6] Menchetti et al. (2020). *Guida alla lettura e all'interpretazione dei dati Covid-19*. Università degli Studi di Firenze, Italy.

3.1 An understanding of the data about the spread of Covid-19 in Italy: Department of Civil Protection data

From the 31st of January 2020 the Italian Council of Ministries declared the Emergency status for Health risk for a time span of six months. The Civil Protection Department was called for coordinating all the intervention needed. One of their tasks was to inform citizens about the numbers of individuals infected, died and recovered; they make this by a *system of surveillance*, managed by Superior Institute of Health (ISS) from 28 of February[7], that aggregates epidemiological data collected by Regions and Autonomous Provinces. They have the duty to put these data (that represent a report of the day before) on the platform every day with the deadline of 11:00 (a.m.). Microbiologic surveillance is instead directly done by Superior Institute of Health which collect samples sent by hospitals from all the regions, then analyses and assesses the presence of the *SARS-CoV-2* and at least report it in the platform. Clinical studies are done jointly with the "Spallanzani National Institute of Infective Diseases" from the large database in which all the clinical characteristics are reported. Then in order to make data available, Department of Civil Protection has been given the "*CC-BY-4.0*" licence. But what do they exactly report daily?

[7] Dipartimento della Protezione Civile. (27th February 2020). *Ulteriori interventi urgenti di protezione civile in relazione all'emergenza relativa al rischio sanitario connesso all'insorgenza di patologie derivanti da agenti virali trasmissibili*. Ocdpc n. 640. http://www.protezionecivile.gov.it/documents/20182/823803/OCDPC+N.+640+del+27+febbraio+2020.pdf/a329a393-fcaa-4982-9f29-9efae47e1342

The numbers they report in the press conference, that are a national aggregate with each record representing a day from 24/02/2020, are all updated every day at 6.30 p.m., and they are like a photo on the day before. Specifically, they are[8]:

- **Hs** = Number of people with symptoms at hospital
- **Hic** = Number of people in Intensive Care Units
- **H** = Total number of people Hospitalised

$$\mathbf{H = Hs + Hic}$$

- **I** = Number of people in home isolation
- **P** = Number of people currently positive (at home and hospital)
- **dP** = absolute variation from the day before in currently positive individuals

$$\mathbf{P = I + H}$$

$$\mathbf{dP = P_t - P_{t-1}}$$

Then:

- **R** = Number of people recovered (cumulative)
- **D** = Number of people died (cumulative)
- **N** = Total confirmed cases (cumulative)
- **dN** = absolute variation from the day before in people that got the Virus

$$\mathbf{R + D + P = N}$$

$$\mathbf{dN = N_t - N_{t-1}}$$

Then:

- **S** = Total number of nasopharyngeal swabs done until that day (cumulative)

[8] Variables' names are arbitrary and have been abbreviated for the sake of simplicity in the most representative way.

- **dS** = daily swabs in that day (not in the original dataset)

$$\mathbf{dS = S_t - S_{t-1}}$$

Orally they also report:

- **dR** = new recovered (daily increment)

$$\mathbf{dR = R_t - R_{t-1}}$$

- **dD =** new deaths (daily increment)

$$\mathbf{dD = D_t - D_{t-1}}$$

The last variable:

- **tc** = number of people that have been tested (it has been introduced from 19th of April) (daily increment)

Here the complete spreadsheet with the last 5 days of April reported:

Date	Hs	Hic	H	I	P	dP	R	D	N	dN	S	dS	dR	dD	tc
2020-04-26	21372	2009	23381	82722	106103	256	64928	26644	197675	2324	1757659	49916	1808	260	1210639.0
2020-04-27	20353	1956	22309	83504	105813	-290	66624	26977	199414	1739	1789662	32003	1696	333	1237317.0
2020-04-28	19723	1863	21586	83619	105205	-608	68941	27359	201505	2091	1846934	57272	2317	382	1274871.0
2020-04-29	19210	1795	21005	83652	104657	-548	71252	27682	203591	2086	1910761	63827	2311	323	1313460.0
2020-04-30	18149	1694	19843	81708	101551	-3106	75945	27967	205463	1872	1979217	68456	4693	285	1354901.0

Figure 1 - complete spreadsheet with Civil Protection data (modified)[8] in the last 5 days of April

They made a repository in the GitHub website, in order to make possible some research[9] where they report also data aggregated at regional and provincial level. For the Regions we have the same quantitative variables as for the national aggregate, but for Provinces they

[9] GitHub. (April 2020). website: https://github.com/pcm-dpc/COVID-19

give only numbers of total cases. The reason why they do not report them (or maybe do not even have them) is that it is difficult for example to assign properly a death to a province for many reasons (e.g. off-site residency, too large dataset for being made available). Then the more we want to have a very detailed data the greater is the difficulty to collect them.

These data nonetheless contain some errors; we cannot think to have been given the truth from the heaven, so we must be careful in reading them.

The variables 'N' (total cases), 'R' (recovered), 'D' (deaths) and 'S' (swabs) are cumulative series that must be growing or at least, in the best situation, equal to the day before; if, for example, counts of today are arrived to 25000 deaths, we cannot have a number that is lower tomorrow (at least it should be equal to 25000) or if hospitals have submitted 1.000.000 nasopharyngeal swabs until time *t*, that count cannot be lower at time *t+1*. Despite this fact, for several reasons like compiling errors, we can find fallacies in our spreadsheet under study. Some statisticians (Menchetti et al, 2020) have found them in the regional aggregate one. They reported the Regions and the linked incongruences found [Figure 2] with the specification of the date and the variable contested[6], but this analysis can be easily replicated looking for some negative or null changes in the cumulative series above mentioned.

The dispersed responsibility across regions is maybe the primary source of this kind of mistakes. All the regions have been given the duty to make the surveillance since the statutory regulation of Italy pose the Health care system to the regional control (*Title 5th* of *Constitution of Italian Republic*). So, every region, in a monopsonist perspective, managed the data collection on their own with the sources they had.

Region	Incongruences
Calabria	Recovered up to 12/03; deaths up to 23/03
Campania	Recovered up to 21/03
Emilia-Romagna	Swabs up to 30/03
Friuli Venezia Giulia	Recovered up to 17/03, up to 18/03, up to 24/03; swabs up to 19/03
Liguria	Recovered up to 15/03; total cases up to 01/03 and 02/03
Lombardia	Swabs up to 26/02
Piemonte	Total cases up to 27/02 and 09/03
Puglia	Recovered up to 14/03
Sicilia	Total cases up to 02/03
Valle d'Aosta	Swabs up to 15/03

Figure 2 - Incongruences in the "dati-regioni" cumulative counts done by Menchetti et al (2020), 'Guida alla lettura e all'interpretazione dei dati Covid-19', Università degli Studi di Firenze, Italy

Another kind of error, more evident, is linked to the late communications. They are more evident for the reason that there is advertisement (in GitHub repository itself) to the inefficiently latest daily report done with the specification of what are the data mistaken or unreported and the region to which they are referred.

For what concern the count of nasopharyngeal swabs there is the problem of duplication, because positive individuals are done the test many times, until they are given a negative result. The variable 'S' that counts the swabs, does not take into account whether there are people that are doing the test for the second or third time. For this reason, from 19th of April, it has been added the variable that counts not the swabs, but the people that have been submitted to the test.

3.2 An understanding of the data about the spread of Covid-19 in Italy: Istat subsidiary data

National Institute of Statistics in Italy cannot be subtracted from its support in this period. Istat workers needed to make a great effort for completing, in a couple of months, a job that would have taken 2 years. Among the great amount of data they collect there are also the daily deaths at Municipality level, for which they require a very long procedure in order to be in their database[10]. Regimented by *DPR n. 285* of 10th September 1990, a physician has the duty to compile a death report of the patient who is assisting (or of a person for which he is called to certificate the death) no later than 24 hours after the death. In this report, "[…] *established by the Ministry of Health jointly with the National Institute of Statistics*", there is the specification of the causation (and other eventual co-causations) of the death of the individual who is the unit of analysis. This module is then sent to the Municipality (by a member of the dead individual's family or by Health Management in the case the death happens in Hospital) in order to be completed with the socio-demographic data by the official designated registrar. At least it arrives to the Istat institute database.

Following the usual time required we would have the complete and detailed dataset in the end of 2022. Aware of the necessity, and of how these data can be useful in a period like this, they tried to anticipate the times, and they made a first release of some data in the last 1st of April 2020. These data are part of a mortality report until 21st of March 2020. Some of them are very detailed but they have a problem that we must keep in mind: the sample

[10] Istat. (April 2020). website: https://www.istat.it/it/archivio/4216

under study is not very representative (Menchetti, et al, 2020). They are a very biased sample because they are chosen between:

1. Municipalities that are part of the ANPR (National Registry of Resident Population); and the municipalities that adhere to ANPR are 5866 out of the 7904 municipalities in Italy (74% of them).
2. Municipalities that communicated on time data related to 2020.
3. Municipalities that had a number of people died superior to 10 units in the period 1st January – 21st March 2020 and that in the month of March 2020 had an increase in the mortality by a minimum of 20% in respect to the mean mortality calculated in the 5 years 2015-2019.

At least we have a selected sample of 1084 municipalities out of 7904 (14% of them).

In a second release Istat gave the data referring to the period 1st March – 4th April again with similar rules of selection:

1. Municipalities of ANPR that were efficient in communications of data of 2020
2. at least 10 deaths in the period 4th January – 4th April 2020 (the longer time span increased the number of municipalities in our sample);
3. 20% increase of deaths in the period 1st March – 4th April 2020 with respect to mean of the five years 2015-2019;

In this second release, as expected, we have a greater sample (still not very representative for the reasons above mentioned) of 1689 municipalities out of 7904 (21% of them). Finally another limit of these data is the missing value in some weekly report: it can be found as a '999' but it means that this information is not already available for that municipality in that date (Istat website, 2020).

4 A simple analysis of Civil Protection dataset (National aggregate)

If we take data as they are given, (Civil Protection Department ones) they do not say anything, if not something wrong, especially if we want to make comparisons between Regions or Provinces. Numbers should be normalised to a common ground in order to be compared, but firstly let's have a look into the national numbers.

As all epidemies, the spread of *Covid-19* in Italy follows a "*S*" shaped curve that starts from zero and grows with an increasing rate (positive second derivative). From a moment on, that we can individuate as the "turning point" of the epidemy (where its second derivative is equal to zero) [Figure 3.1], its daily increase stop growing and start declining [Figure 3.2] where the convex curve becomes concave [Figure 3.1]. What we expect is an end with a point of tangency of the curve when there are no more additional cases in the cumulative count (daily new cases equal to zero).

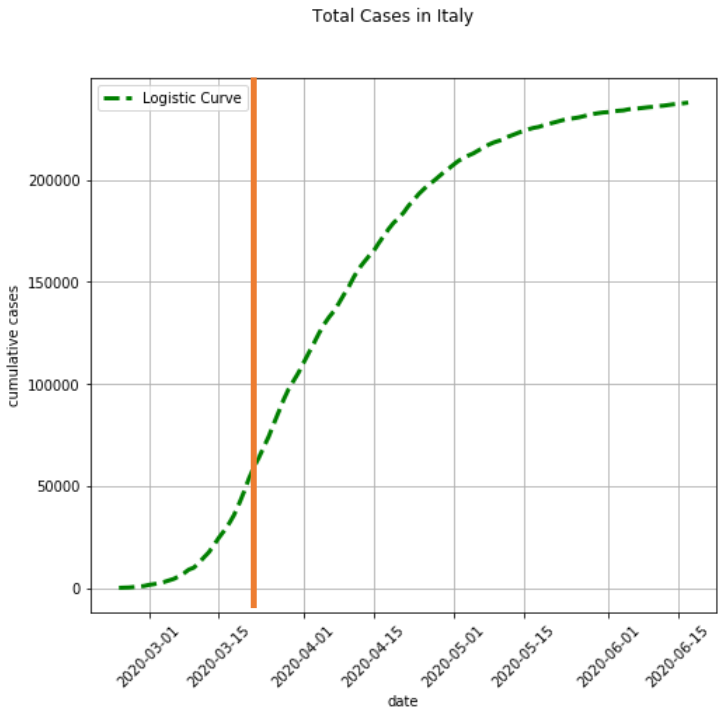


Figure 3.1 – Total cases in Italy - cumulative counts

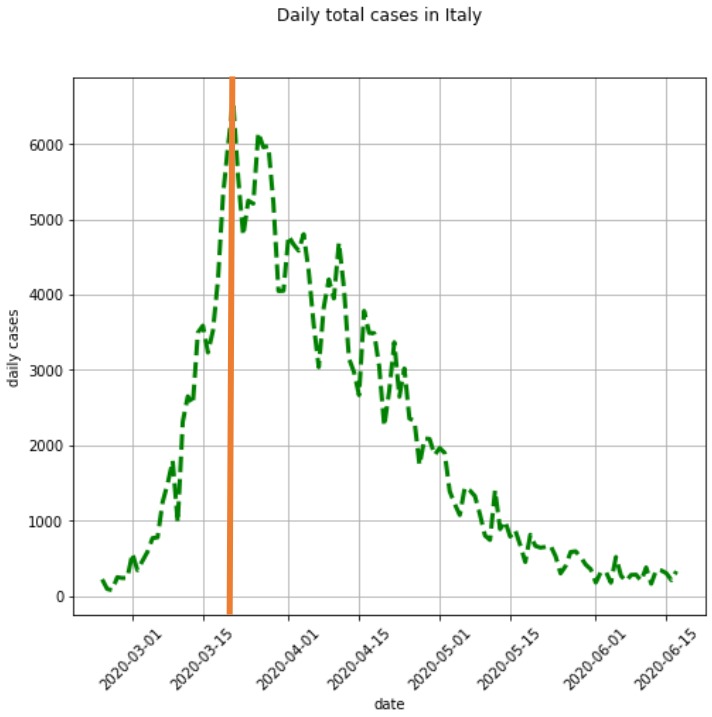


Figure 3.2 – Total cases in Italy - daily counts

In the period represented in that graphs, among total cases there have been many people recovered or died. Both have exited the disease, so scatter-plotting the variable that counts the daily change in currently positive individuals we can see also a bell shape curve that has reached a point in which the sign is negative, meaning a reduction in that count. In green I highlighted the points that represents the positive (but arithmetically negative) numbers which bring a decline in currently positive from the last 10 days of April, and what we expect to see in that graph is a totally (axial) symmetric bell shaped curve scatter-plotted in green when the reduction in positive individuals fully covers all the cases observed in the red points [Figure 4].

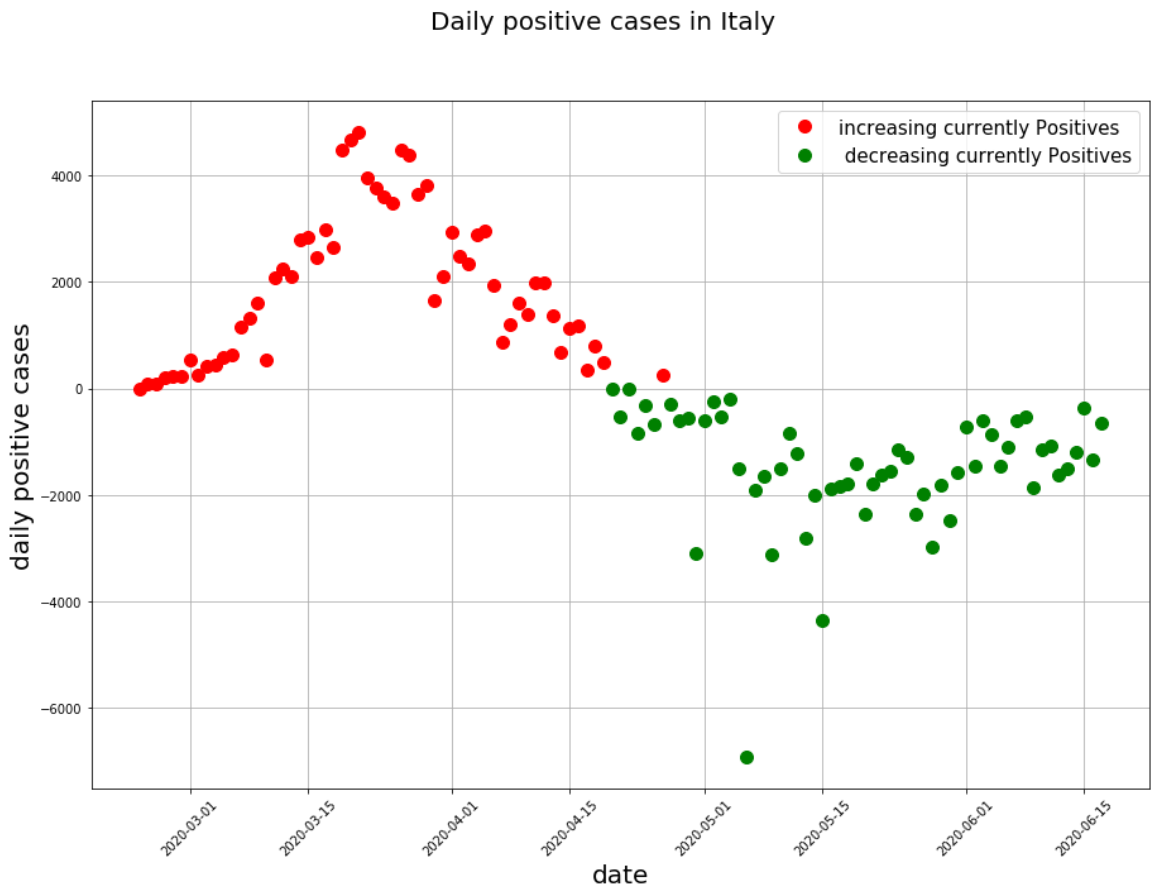


Figure 4 - Daily change in positive people (in red an increase, in green a decrease)

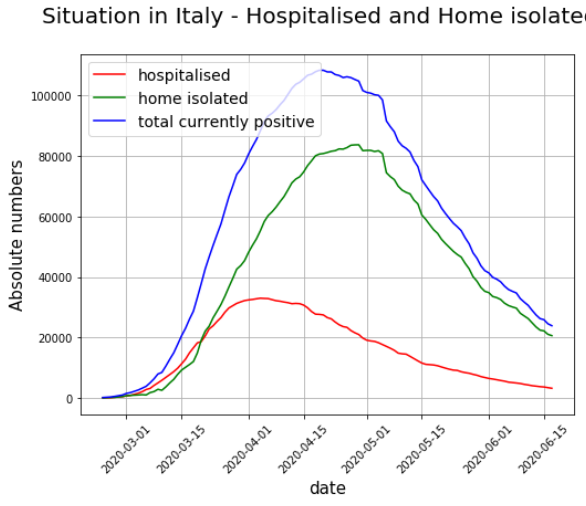


Figure 5 - trends of hospitalised and home isolated in comparison

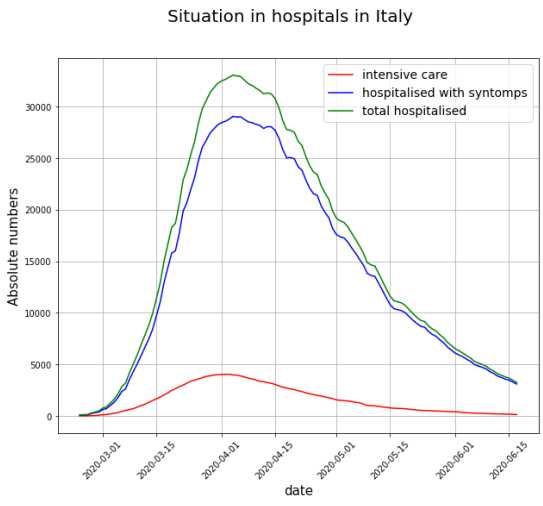
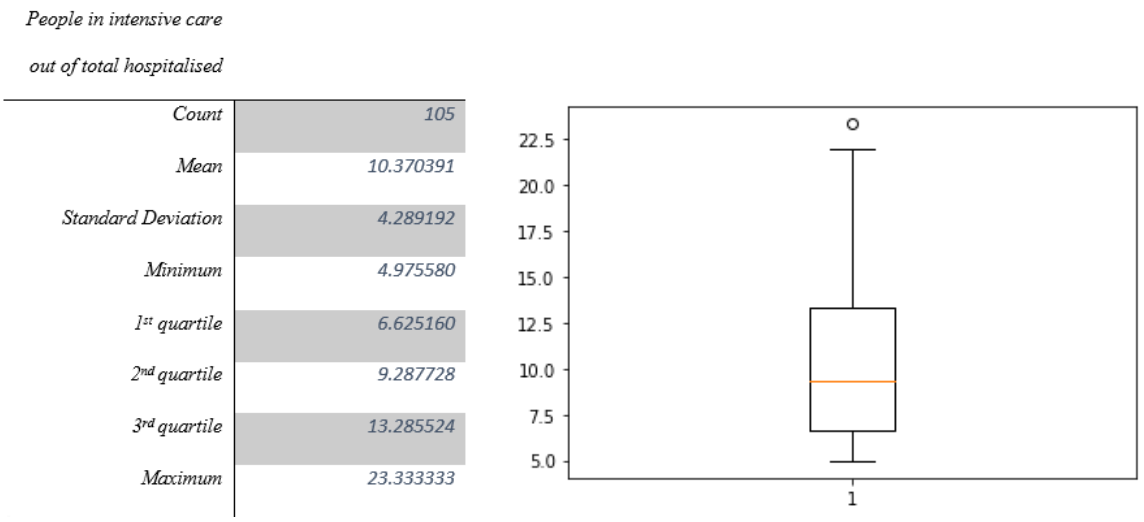


Figure 6 - trends between the total number of hospitalised people, intensive care or hospitalised just with symptoms

Between that currently positives individuals most of them are home isolated. This can be said to be true only from the end of March [Figure 5] when the green line overcome, in height, the red one; instead we can also compare how many people among the hospitalised are in intensive care units and how many are just hospitalised with symptoms [Figure 6]; here we are tempted to say that the people that are taken in intensive care units are relatively less and less, but we should take in consideration the variable that makes a percentage of people in intensive care out of all the people who are cared by hospitals in order to have a correct idea on how much people are actually risking life in hospitals (maybe it is more meaningful rather than considering the percentage out of the total cases). It seems that this indicator is keeping a decreasing trend from the mid of April being below 10%, and at the end of May it has gone below a minimum of 5% [Figure 8]. Still, even though it seems that it has been very high at the beginning, the distribution has a mean value of 10.37% with a standard deviation of 4.29 points. The only point above the 22% is an outlier as explained by the boxplot [Figure 7].



People in intensive care out of total hospitalised	
Count	*105*
Mean	*10.370391*
Standard Deviation	*4.289192*
Minimum	*4.975580*
1st quartile	*6.625160*
2nd quartile	*9.287728*
3rd quartile	*13.285524*
Maximum	*23.333333*

Figure 7 – Description of the distribution of the values of indicator (intensive care units out of hospitalised); a boxplot.

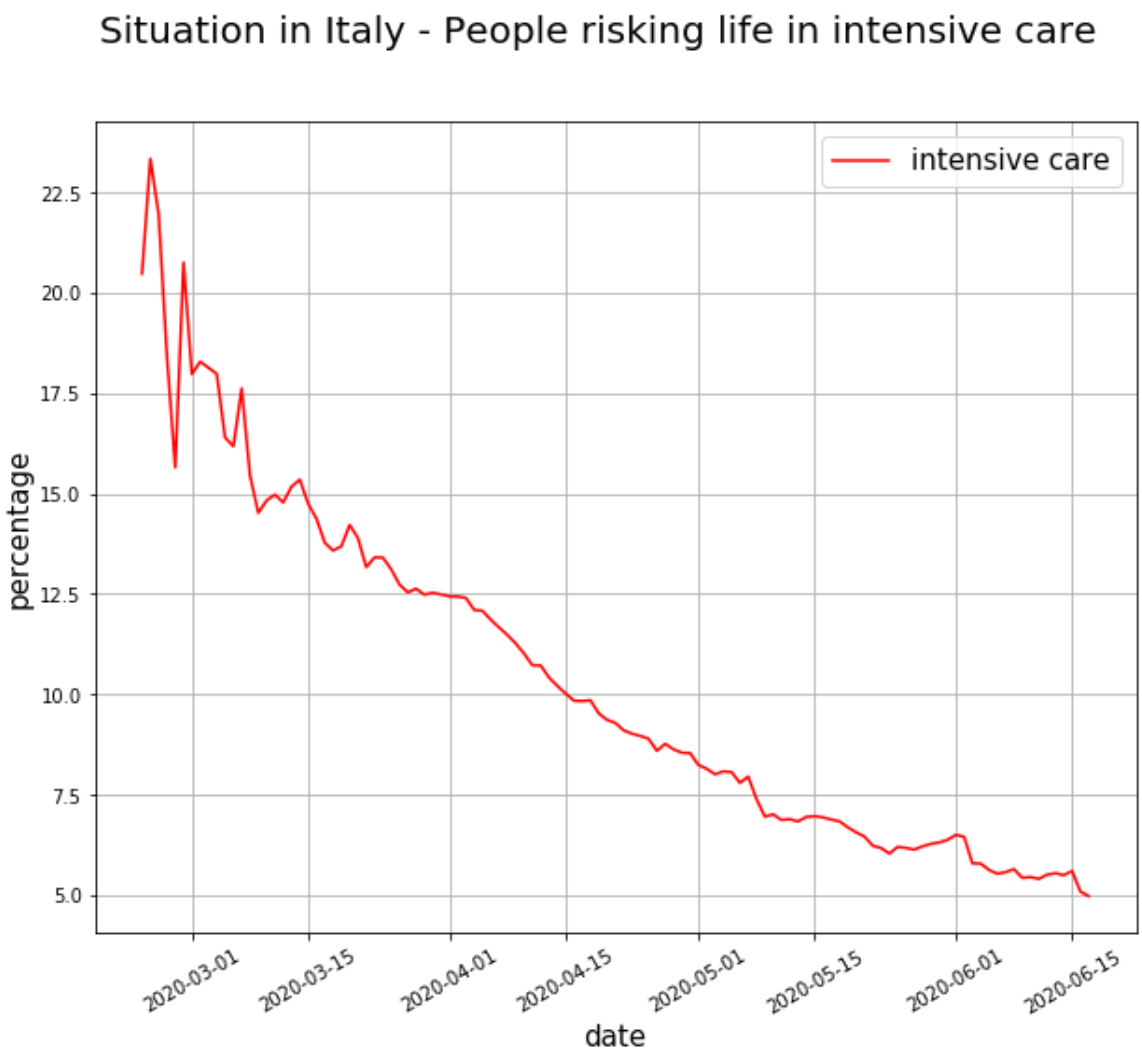


Figure 8 – trend in intensive cared individuals out of total hospitalised

For what concerns the number of people recognized as positive we should always take in mind that in this moment the only instrument to detect the presence of the *SARS-CoV-2* is the nasopharyngeal swab. The more swabs are done, the more infectious people are found (Menchetti et al, 2020), and this argument seems pretty obvious. Here I report what Menchetti and Norjean did in their research, calculating the daily percentage changes in how many people are found positives and the same for the number of swabs (but in the national aggregate); we can easily see that in the 4 regions selected (Lombardia, Veneto, Emilia-Romagna and Toscana) the two indicators seem to go hand in hand. We can maybe

improve this indicator by comparing the percentage changes of positive individuals found in a day with the number of daily swabs done in the day before rather than in the same day; this choice could improve indicator because we are sure that from the moment in which a swab is done passes a minimum of a day to the result of positiveness being reported.

Another indicator proposed is the possibility to find a positive individual among a certain number of tests done (Menchetti et al, 2020); it is nothing more than the ratio between the indicators of positive and total swabs but in a time span predefined (for example a week): positive cases found over the tests done; here again I propose to consider the same amount of time, but with a delay of a day in the responsiveness of the test (for the denominator). My proposal in the formula for the indicator is a new parameter of delay d = 1; even if the test require 4-5 hours for a response (as 'Gruppo San Donato' of the private hospital 'San Raffaele' reports in his website[11]) we can imagine a delay of 1 day that elapse from the test to the report of the result.

.

$$\rho_k(n) = \frac{C(n) - C(n-k)}{T(n) - T(n-k)}$$

$$p_k(n) = \frac{C(n) - C(n-k)}{T(n-d) - T(n-k-d)}$$

[11] Gruppo San Donato. (April 2020). *Tampone faringeo: come funziona il test per la diagnosi del coronavirus*, website: https://www.grupposandonato.it/news/2020/marzo/tampone-faringeo-coronavirus

Furthermore what we should take care is how swabs are done; we know that when you're tested positive in order to be considered officially recovered you must have 2 negative tests: as written in the *European Centre for Disease Prevention and Control* (ECDC) technical report (*Novel coronavirus – Sars-Cov-2*, 2020)[12], used by Italian Health system as guideline, "*A COVID-19 patient can be considered cured after the resolution of symptoms and 2 negative tests for SARS-CoV-2 at 24-hour intervals*"; so in the cumulative numbers of swabs reported by Civil Protection we have also the replication of the tests to the same individual; for this purpose it has been introduced the variable *tested cases* which can be maybe more useful.

In Italy these tests are employed in a very heterogeneous way and this is one of the reasons why we must always be cautious in the reading of trends. A region can have an increase or a drop in the cases, due to the number of tests done. This high sensitivity to the number of tests done that we can see [Figure 9] by the alignment of the 2 lines (percentages changes in swabs and positives) can be an empirical evidence of the importance of doing many tests in order to discover new positives cases (Menchetti et al 2020). Given this relation, it is very important not to read them as they are reported, but a normalization as the one proposed above is crucial for the understanding of how many are the contagions in terms of the quantity of tests done.

[12] European Centre for Disease Prevention and Control. (2020). *Novel coronavirus (SARS-CoV-2) - Discharge criteria for confirmed COVID-19 cases*. technical report.

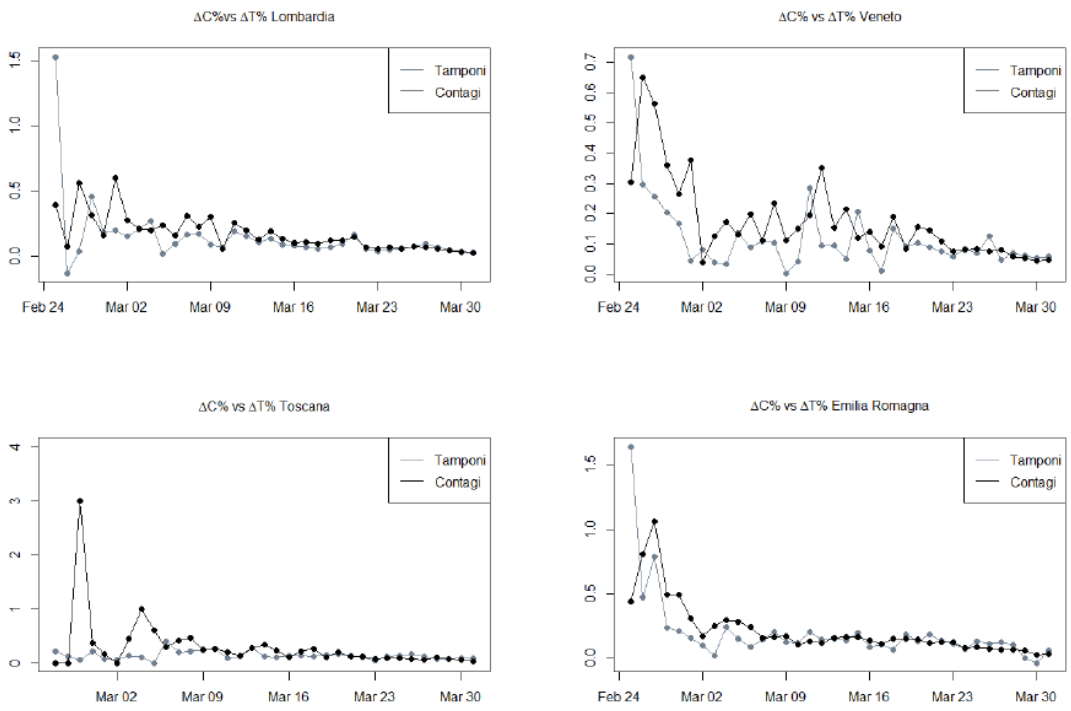


Figure 9 – daily change in percentages for tested cases as positives and tests done (Menchetti et al, 2020)

Moving to the disaggregated data among Regions and Provinces can help us for making comparisons and understand which are really the most afflicted zones in geographical terms. Following the advice of a dashboard developed by the CEEDS (*Centre of Excellence in Economics and Data Science*) it is easy, but not immediate, to see the differences between Provinces in the *cumulative cases* and the *cumulative rates*. What distinguishes the latter from the former is a normalization for the population of the Province under study. This normalization is the ratio between the cumulative cases and the population multiplied by 100.000 (Ferrari et al, 2020)[13]. In Figure 10 we can easily see that the very high numbers

[13] Luisa Ferrari, Giuseppe Gerardi, Giancarlo Manzi, Alessandra Micheletti, Federica Nicolussi, Silvia Salini. (2020). *Covid-Pro in Italy: A dashboard for a province-based analysis*. Department of Economics, Management and Quantitative Methods – Department of Environmental Science and Policy, University of Milan

of Milan, for example, are not so high as they seem; they are the same of Turin, given that they follow the same trends. Instead Cremona and Piacenza that have the lowest *cumulative cases* in absolute terms, are in reality the more afflicted as the graph of cumulative rates suggests. At least what I want to stress here is that data cannot be taken as they arrive to us; we have always to put a filter of understanding in order not to read them wrongly or make erroneous judgements and statements, especially when we are talking about life of people. As above explained it seems very simple how a comparison should be made, but journals didn't managed to inform people in the proper way especially at the beginning of the pandemic.

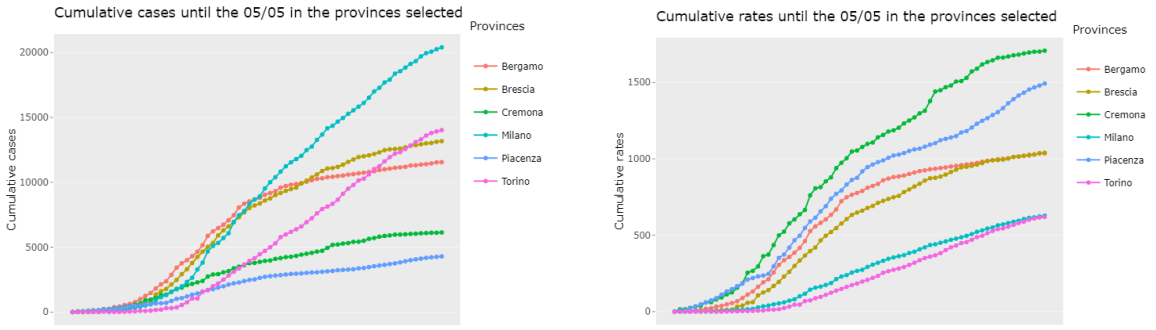


Figure 10 - cumulative cases and cumulative rates in Italy. A dashboard for province-based analysis[13]

5 Solutions for estimating deaths

When we start analysing numbers of people died for the *Covid-19* disease, in a first instance we should be aware of the words we use; as the magazine "Statistics and Society" reports[14], journals have made an abuse of the word *mortality* without really knowing the exact meaning. Mortality in general is an indicator that counts the number of people died out of the total population, and making a simple exercise of computing it, we can easily see how it is not meaningful since the denominator is very high relative to the dead people (population in Italy is about 60.000.000). What they were talking about is properly called *lethality* and it counts the deaths as a percentage of the people that has caught the virus. It is simply calculated by the ratio between the total number of deaths and the total cases up to the same day of reference (namely 't') and it explains as a matter of facts how much the virus is lethal. Remembering the variables[8] in the Civil Protection dataset, they are explained by the following formulas, assuming a fixed population of 60.000.000:

$$\boldsymbol{mortality_t = \frac{Dt}{pop}} \qquad\qquad \boldsymbol{lethality_t = \frac{Dt}{Nt}}$$

We can visualize the difference between the two indicators discussed above in the graph below (Figure 11), and in particular how it is meaningless to talk about mortality.

[14] Statistica e Società. (April 2020). website: http://www.rivista.sis-statistica.org

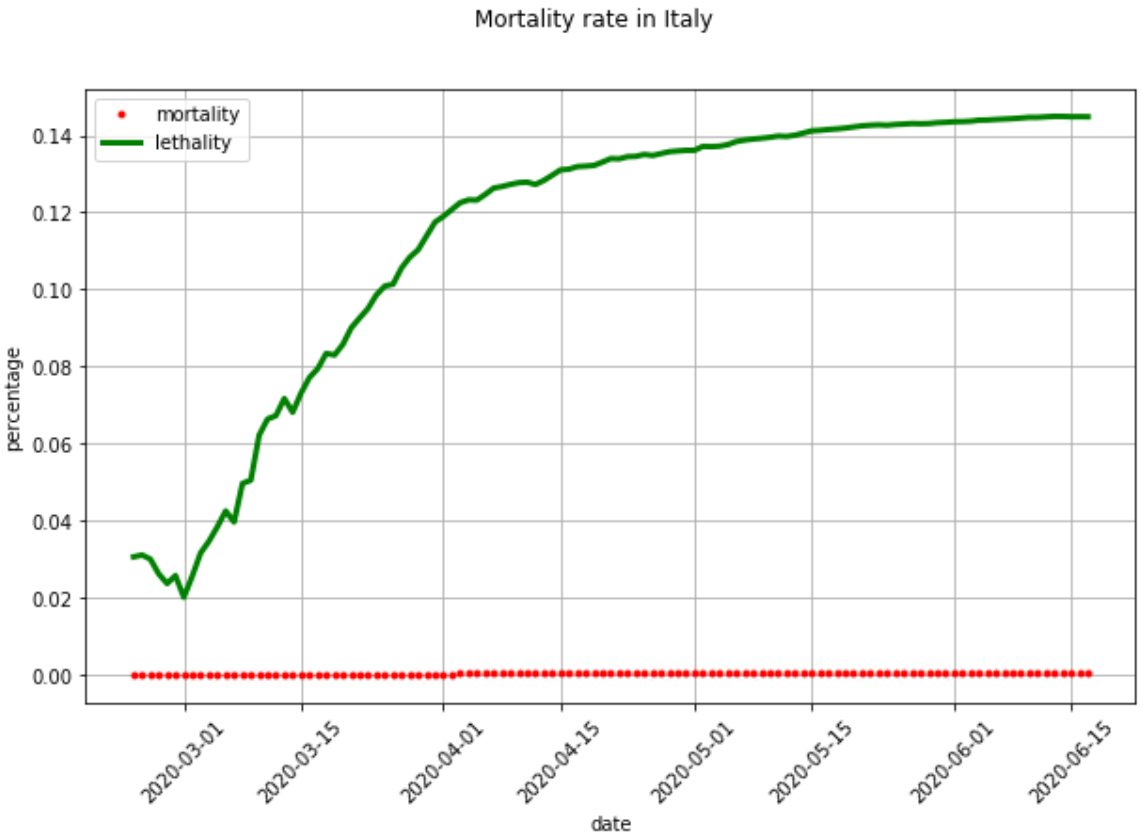


Figure 11 – Mortality and Lethality rates

But going deeper in the analysis of this indicator, Furno and Olivari in an article[15] warned that there is a procedural error: the denominator is maybe too small with respect to the *actual cases*. Between the argument supporting this thesis there is a study of the municipality of Vò (in Padua, Italy) that has proven a very high presence of asymptomatic cases, roughly 43% with a confidence level of 95% (32,2%-54,7%) (Lavezzo et al, 2020)[16]. What they have found is something very relevant because the samples of swabs, collected with 2 surveys at the beginning and the finish of lockdown, were respectively 2812 and

[15] la Voce. (April 2020). website: https://www.lavoce.info/archives/65205/decessi-da-covid-come-leggere-i-numeri/

[16] Lavezzo et al. (April 2020). *Suppression of COVID-19 outbreak in the municipality of Vo', Italy*. medRxiv preprint.

2343. These are very representative numbers since the city is very little: they represent 85,9% and 71,5% of the population, and these samples are also homogeneous in age. The strength of the results of a little city like Vò is also a weakness and it would be completely erroneous to make the inference to all the Italian cases, nonetheless the study gives a useful insight in how numbers we have are maybe not totally reliable.

Since now we have seen that maybe *lethality* rate has been overestimated by the underestimation of cases, but there is another cause that can affect our result and it is related to the deaths themselves. In the opposite way of the denominator, it could be possible an overestimation of the numerator.

An individual that is in the dead's compartment is only an individual who has died *while* having the virus, and we do not have the detail of whether the *SARS-CoV-2* was the primary cause or not (Menchetti et al, 2020)[6]. This means that people died for other primary causes or at least with Covid-19 as comorbidity are in the counts as individuals that has been struck down by the Virus. So even when we look at data on deaths, that could seem less apt to errors, we should be careful.

In the second part of the third chapter we have long discussed about the source of data of Istat; these data are still not enough, because they are representative respectively only for 14% and 21% of all the municipality in Italy and when they will cover a much more substantial part of them these data would be an extremely useful benchmark in order to see whether the deaths have been underestimated or not. The analysis that could lead us to see that is very simple: comparing data about weekly deaths in the four cities selected (Cremona – Milano – Bergamo – Piacenza)[13] it is evident how the trends in the 5 years 2015-2019 (the dotted lines) are pretty much the same. An evidence like this suggest us that, all other things being equal, the 2020's curve should be the same, and it is the case

until the first day of March. From that moment on the line start growing vertiginously in all the 4 cities represented. Assuming no other big changes in the health status of citizens from the last 5 years considered, the numbers of people died in excess can be reconducted to *SARS-CoV-2* infection. Further comparing the excess mortality found in Istat data with the Civil Protection reports can give us an idea of a possible incorrect estimation.

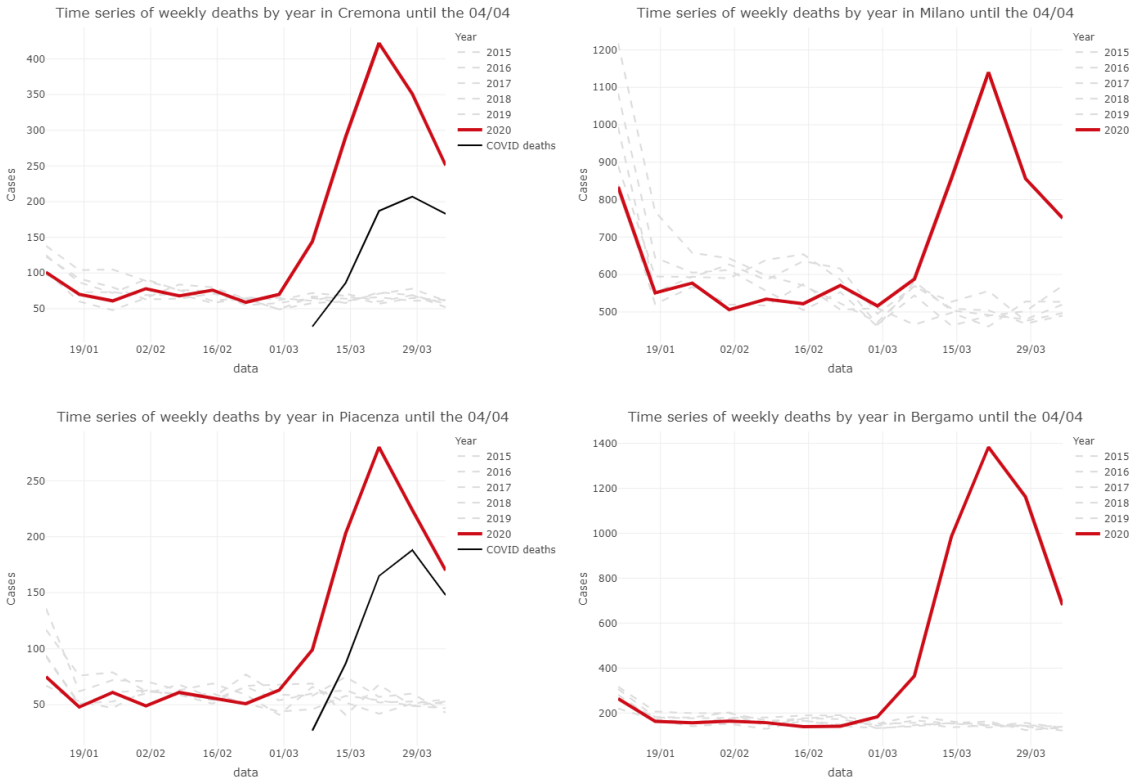


Figure 12 – Time series of weakly deaths by year in 4 representative cities (Cremona – Milano – Piacenza - Bergamo) – 'A dashboard for province based analysis'[13], dashboard: http://demm.ceeds.unimi.it/covid/.

In the quantitative study done by University of Florence[6] they report that in the period for which they have data the mean number of people died in 2015-2019 is 3300. Instead they are 6035 in the same weeks of the current year, almost doubled; and given that numbers of the five years are in the range between the minimum of 3214 in 2017 and the maximum of

3442 in 2019, we can be quite confident to say that the "*excess mortality is caused by epidemy of COVID-19"*[6] (Menchetti et al, 2020).

In this exercise of estimation of actual deaths from the source of Istat we should always remind that there have been many people died *with Covid-19*; ISS reports that 59,5%[17] of dead had 3 or more pre-existing diseases [Figure 13] and the mean age was 80 years old (in Figure 14 it is plotted the distribution of deaths between the ages in decades and it is very skewed to the right – towards the oldest individuals). The most frequent previous diseases found are hypertension, diabetes, kidney failure and ischemic cardiopathy (Figure 15). All these data suggest us that there have been at least 'some' individuals that have not died for *Covid-19* as a *primary cause*, and so that has happened what in epidemiology is called *harvesting*. It is a process that describes the anticipation of a mortality that would be observed just some months later if there was not the epidemy; this not to justify the high numbers between the oldest or less 'healthy' part of population, but it is useful to deeply understand how much lethal this virus in reality is. This phenomenon can also be explained by what SiSMG[18] reported about the last winter (2019): a decreased mortality maybe due to mild season. This delay in mortality has in a sense increased the number of people exposed to the subsequent Virus that would have arrive in 2020 (Menchetti et al, 2020).

[17] ISS. 2020. *Analysis done on a sample of 32938 dead patients while positive to Sars-CoV-2 infection*. Website: https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia.

[18] System of Surveillance of daily mortality in Italy (SiSMG, Italy)

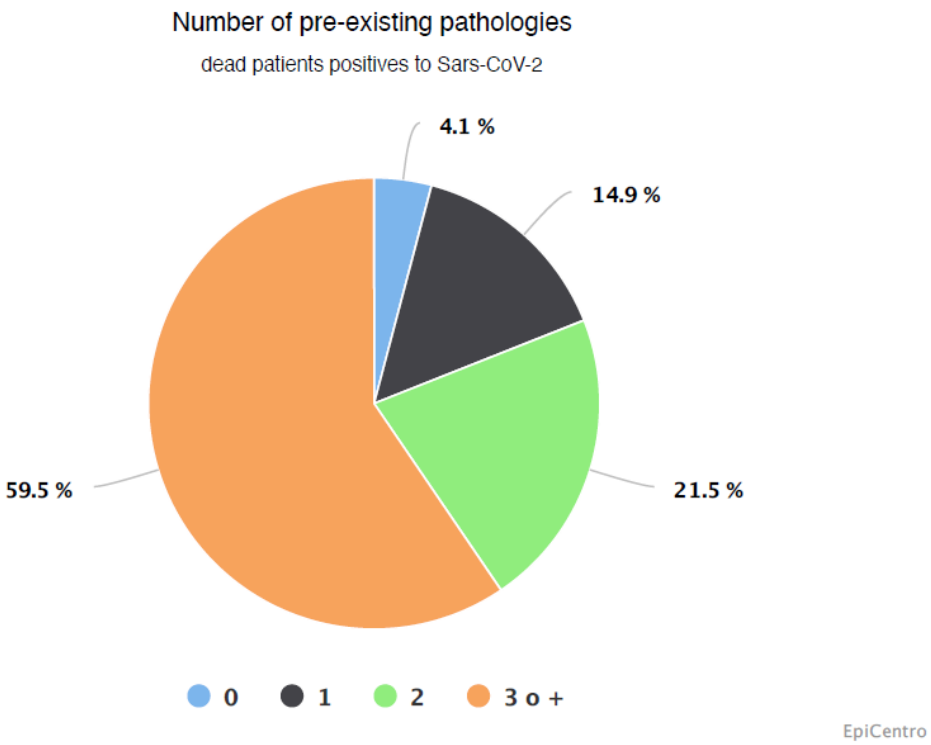


Figure 13 – number of pre-existing pathologies

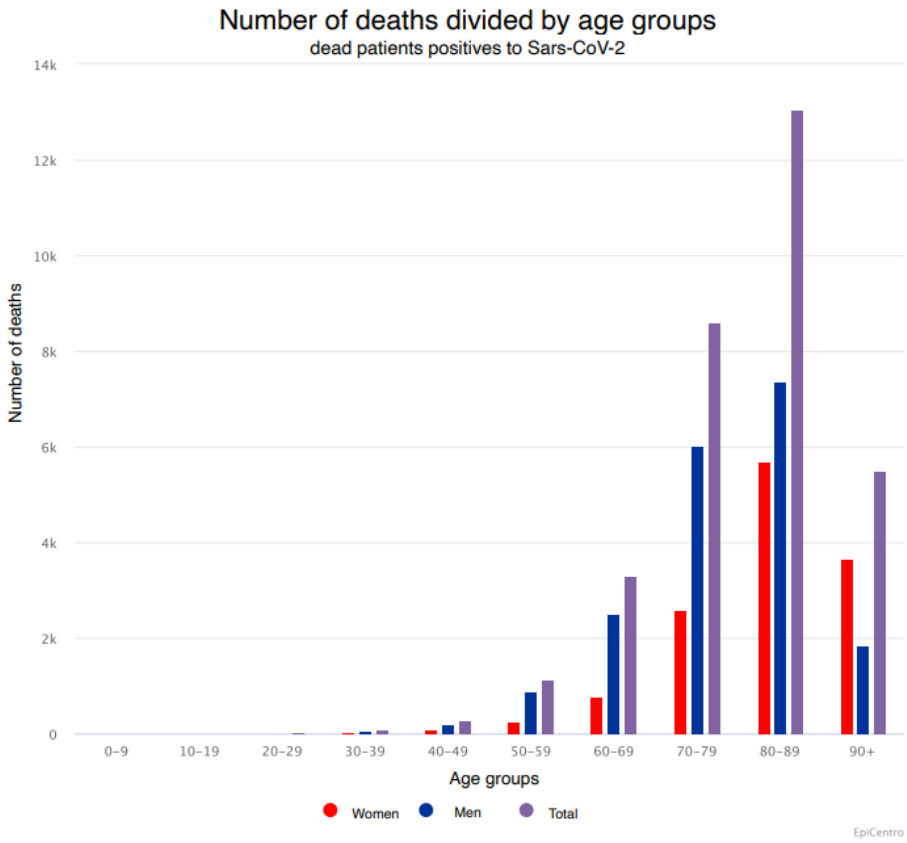


Figure 14 – number of deaths for decades-divided ages

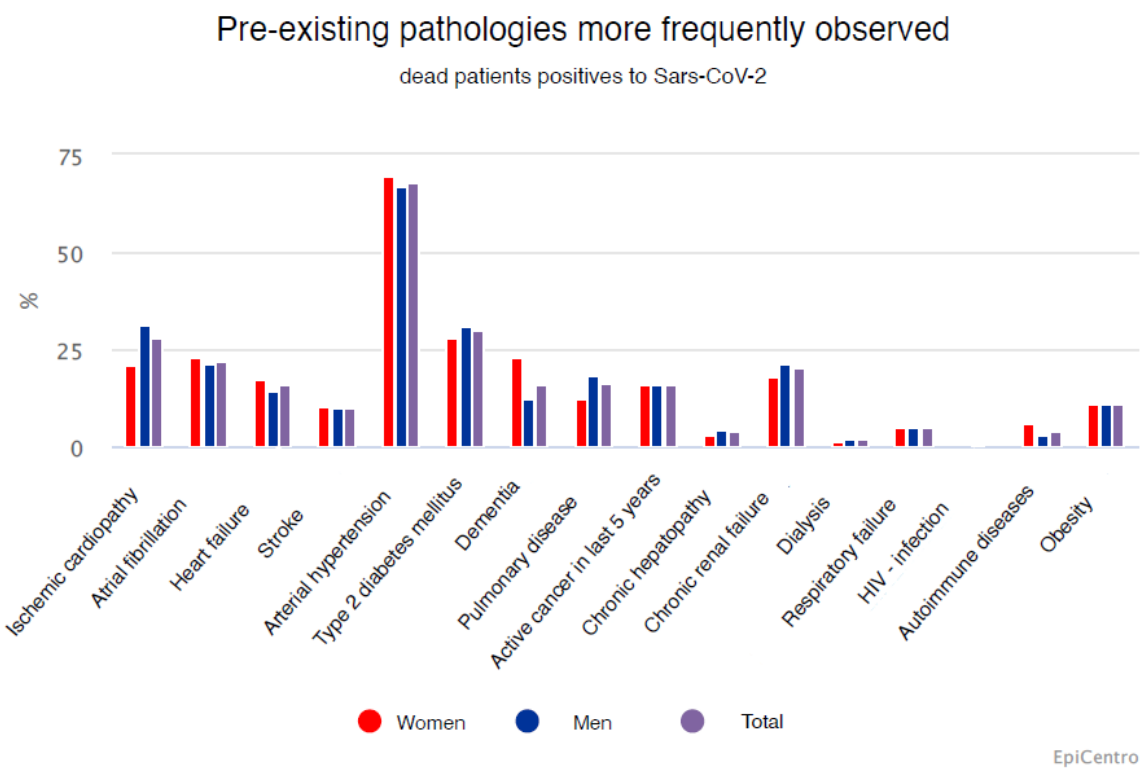


Figure 15 – frequencies of the more recurrent pre-existing diseases among dead

Estimation of death rate is not an easy task and still, even when we talk about deaths we should be careful because the data we have are not so reliable to directly say how much lethal and dangerous the virus has been.

6 Deterministic approach: *SIR* models

The transmission of a virus can be studied from a probabilistic point of view with the very well-known stochastic model "*SIR*", where the variables taken into account are supposed to be aleatory in order to explain which is the probability of infection, recovery and mortality, but for the sake of simplicity we are going to approach to that model from the deterministic point of view. "*SIR*" is the acronym of the compartments that are studied as functions of time *t,* with variables that can be retrieved from the Civil Protection Department. Individuals in a population '*N*', assumed to be constant in time, are divided between the three clusters or compartments:

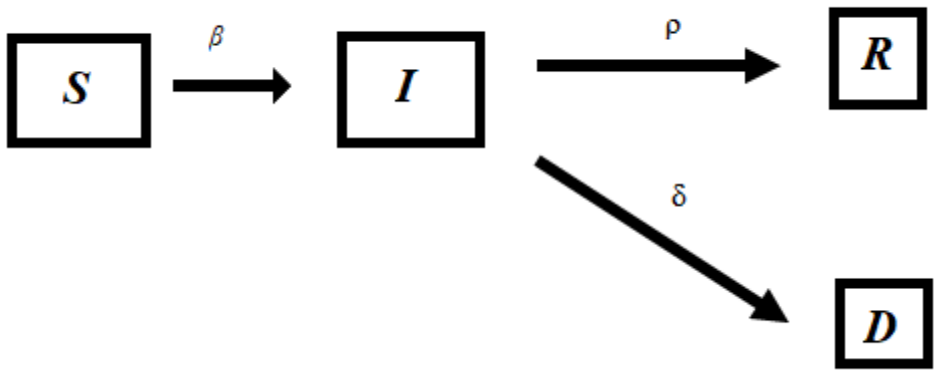
- Susceptible individuals ***'S'*** that can be infected:
- Infected ***'I'*** that can infect;
- Removed from the disease ***'R'***, that are out of the disease;

In this way, assuming also that individuals removed from the disease (whether died or recovered) can no more be infected (long-standing immunity in case of recovered), we can summarize it in the following formula:

$$\boldsymbol{N = S_t + I_t + R_t}$$

At time $t = 0$, when there is the so-called case-zero (the first infected) we have indeed only one infected in the population, and all the remaining people ($N - 1$), in particular the ones that are in contact with case-zero, can be considered susceptible (here an aleatory variable

would be more precise in order to estimate the probability to get infected); for being part of the removed it needs time that people start getting infected so at time zero no one populate the "*R*" compartment. We can further distinguish two sub-categories in this cluster of removed: "***D***" for the dead and "***R***" for the recovered.



As the model above synthetize, individuals pass from being Susceptible to Infected and then from infected to Dead or Recovered with trends that are described by the following parameters:

- *β (transmission rate)*
- ρ *(recovery rate)*
- δ *(death rate)*

Given that suppression of Covid-19 needed the avoiding of social contacts and close interactions between individuals, we can be pretty sure that transmission rate has been positively affected by the lockdown measures. When considering the reciprocal of it we can have the indicator of the average time between effective contagious contacts ($1/\beta$)

(Ferrari et al, 2020); instead the reciprocal of the sum of the other two parameters (1/ (ρ+ δ)) gives us the average time before removal from the infectious class[19].

Recovery and death rate instead depend also on the burden that hospitals had, especially how intensive care units were overcrowded. Parameters above mentioned can be retrieved at each time *t* from the typical differential equations of SIR models:

$$\frac{dS_t}{dt} = S_{t+1} - S_t = -\frac{\beta_t S_t I_t}{n}$$

$$\frac{dI_t}{dt} = I_{t+1} - I_t = \left(\frac{\beta_t S_t I_t}{n} - \rho_t - \delta_t\right) I_t$$

$$\frac{dR_t}{dt} = R_{t+1} - R_t = \rho_t I_t$$

$$\frac{dD_t}{dt} = D_{t+1} - D_t = \delta_t I_t$$

These differential equations express respectively the daily change in *susceptibles, infected, recovered* and dead at each time *t*. For the *susceptibles* we can see a minus in front of the transmission rate parameter because as infection spreads, the number of individuals that can catch the virus is obviously lower and lower; *Infected* compartment instead follows an increasing trend as a positive function of the transmission rate and a negative function of both recovery and death rates. Time series of the retrieved values of corresponding

[19] Luisa Ferrari, Giuseppe Gerardi, Giancarlo Manzi, Alessandra Micheletti, Federica Nicolussi, Elia Biganzoli and Silvia Salini. (2020). *Modelling provincial Covid-19 epidemic data in Italy using an adjusted time-dependent SIRD model.* Department of Economics, Management and Quantitative Methods, Department of Environmental Science and Policy, Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy.

parameters have been then used to make predictions with an elaborated model called Finite Impulse Response filter (FIR) by the study of Ferrari et al, 2020[19], but this goes beyond the purposes of this thesis.

Now, having explained all these parameters we are ready to better understand the concept of the *basic reproduction number* widely used as a benchmark for the understanding of whether the spreading of the virus was critical and not under control or the country was able to manage the situation.

7 Basic reproduction number as a crucial parameter

We have heard a lot that until R_0 parameter won't be below unity we can consider to be in danger, that something is going wrong in the managing of the critical situation and at least each individual is going to infect more than another one. This parameter should be taken below that threshold, and it has been the purpose of all the decisions of the *Council of Ministry of Italy* since March 2020. But what really means, and how it is computed is not an easy thing.

Now what we need to remind is the *SIR* model just presented in the last chapter, especially the parameters that explains how people get infected, die and recover. Ferrari et al in a dashboard developed make an estimation of the *basic reproduction parameter* as the ratio between the transmission rate β (numerator) and the sum of the recovery ρ and death rate δ (as denominator)[13]. What we should expect, in order to keep that value below 1 is a decreasing trend in the numerator (lowering the rate at which people transmit the infection) combined with an increase in the denominator; this could happen hopefully only by an increase in the rate at which individuals recover, but a decrease in R_0 is expected also with an increase in death rate (also in the denominator).

$$R_0 = \frac{\beta}{\delta + \rho}$$

Each of the single parameter (β - δ - ρ) have been estimated also in different ways by researchers.

We have seen that transmission rate is a parameter that explains how the virus is spreading, so it is dealing with infectious people. In the study[20] of Brogi and Guadabascio it is proposed a model, (based on Chenlin et al, 2020) that distinguishes infectious individuals $I_{(t)}$ in three categories:

- I_c = infectious found by swab;
- I_a = infectious not found by swab;
- I_q = infectious quarantined;

Total number of infectious is given by: $I(t) = I_c + I_a + I_q$; Under this assumption they describe the spreading process from the first infectious case as a function of the transmission rate β. Each individual who has, in a first instance, caught the virus, before being inspected, then found positive to the *Sars-CoV-2* and at least cared in hospitals or quarantined has a probability to transmit the virus that has been stylized with the Poisson distribution. So the expected infectious people in the next day E[I(t)] is equal to the Infectious individuals at time *t* to the power of the transmission rate β.

$$E[I(t)] = I^{\beta t}$$

Given that an individual is contagious in mean for 7.5 days and for manifesting its symptoms they are required 5-7 days Brogi and Guadabascio consider a time span of 14 days. So transmission rate can be expressed by this formula[20]:

[20] Brogi, Guadabascio. 2020. *Un modello per la stima dell'andamento del contagio da COVID-19 in Italia*. Research Gate

$$\beta = \frac{I_t}{\sum_{i=t-7}^{t-1} I_i}$$

At least many researchers between mathematicians and statistical scientists are trying to find out which is the best model to describe something that was meant to be only matter of epidemiologists.

Calling what we said in the introduction of this chapter each of these parameters has been and it is still essential for the advice of science to politics. Further many experts were called, so many models have arisen from research; we should be always aware that models are abstractions from reality that allow us to make predictions but testing them is essential in order to see if they are correct. In this case we need something that is not the mere or ludicrous testing of hypothesis, but we are retrieving parameters as describing factors of events (spreading, death and recovery) in order to understand whether we are doing the right thing in copying with the virus.

Below in figure 16 there is reported the estimation of the dashboard mentioned for one of the most afflicted cities (Cremona)[13], following the computation explained in the previous chapter. We can see the trends in the parameters of transmission, recovery and mortality and at least the $R_{0.}$

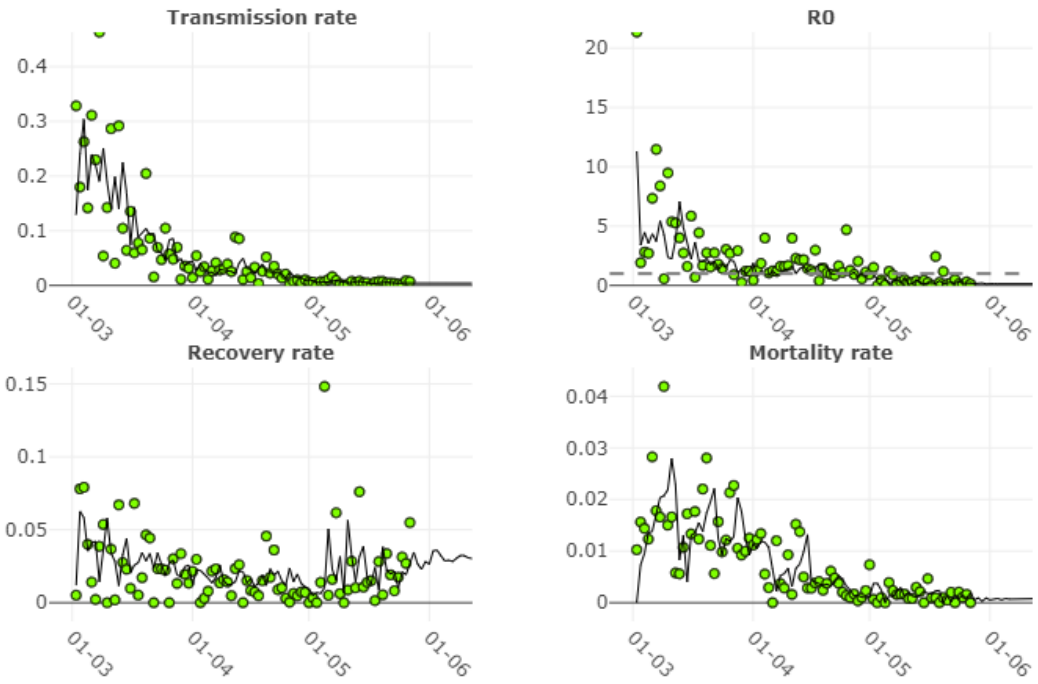


Figure 16 – Transmission rate, R_0, Recovery rate and Mortality rate in Cremona estimated by Ferrari et al dashboard[13]

8 Conclusions

With this short paper we have analysed in a quantitative way some facets of a phenomenon that changed our lives, and still will. Numbers are not just numbers that go up and down, but they are between the most powerful sources that allow to us to deeply understand with statistical tools many useful things for future-sightseers decisions.

Coordination with other Sciences is a key requirement in order to make solid assumptions for robust conclusions. So what I want to remark here is the relationship between Science and Society, and at least between scientists and citizens. Communication in this relationship plays an essential role, because what Science finds should be made accessible to everybody, otherwise remains something of the elites. Scientific research needs to be translated to a society with an insufficient data literacy that could make them understanding important findings, especially in a pandemic period like the one analysed, when "*facts are uncertain, stakes high, values in dispute and decisions urgent*" (Funtowicz and Ravetz, 1993). Then a new paradigm of science is what we require to copy with that problems, and as we have seen philosophers of science (Funtowicz and Ravetz) already given an answer to this question in the 90s with the concept of *post-normal science.* Again what is needed is an *extended peer community* in which everybody with a specific competence is responsible of their task. In this way the learning process from reality, then abstraction and modelling and at least the creation of a theory is more fluid, attached to reality and more sensitive to misrepresentation. It is the disruption of the elitist view of science, and it is a revolution that humankind is making. In the last decades science has become something more democratic in which the contribution of many (not everybody) has been essential. By the way University is more accessible to citizens, making science more a common source. This

is what we need, this is what we are; we should be ready in times like these to take part in this revolution in which science is in the centre, it is crystal-clear and of everybody. We should no more trust in the arrival of a great scientist once in the centuries to make a step forward in the future; if today we are going faster and further is because there exists a scientific community that has channels of communication that are working well.

Even Politics, that has the decision power for copying with that problem, asked the advice of many individuals from the world of Science, creating task forces every time they didn't know what to do. But why do they have not done the same for the communication of what they were exactly doing? The communication of whether the situation was under control or not? It is possible that scientists themselves are the communicator of their Science?

In this pandemic period, especially when Civil Protection Department stopped their press conferences, journals became the oligopolists of information, starting to exploit that source of power to capture the visualization of anxious and scared citizens. Even though no filter of understanding was put into that data reported, they still think to make analysis and insights without any competence. They remain open questions in which we should take time to consider the importance of not letting to the journals a power like this, but at the same time how much it could be useful the role of a mediator between Science and Society that makes citizens well aware of what is happening around without any speculation on it and that not creates infodemic worries in period of crisis.

Ringraziamenti

Desidero ringraziare la Professoressa Silvia Salini per il tempo prezioso che mi ha dedicato e per tutti gli spunti interessanti con cui ha saputo stimolarmi e spronarmi. Inoltre, ci tengo a ringraziarla per la sua disponibilità e gentilezza con cui mi ha accompagnato in

questi mesi.

Il mio ringraziamento più importante va a Margherita. La mia compagna di avventure, che da un senso a tutte le scelte della mia vita.

Un enorme grazie ai miei genitori, Michele e Lucia, per aver creduto in me da sempre. A mio fratello Alessio e alla nostra passione per l'informatica e la tecnologia che ci ha sempre uniti in un legame fortissimo.

Grazie a tutti voi, questa tesi è stata scritta con tanta passione e non potrei esserne più orgoglioso.

References

Bibliography and Sitography:

[1] Istituto Superiore di Sanità. (April 2020). website: https://www.iss.it/coronavirus

[2] Waltner-Toews, Biggeri, De Marchi, Funtowicz et al. (preprint - April 2020). *Post-normal Pandemics: Why Covid-19 requires a New Approach to Science*. ESRC STEPS Centre blog.

[3] Funtowicz and Ravets. (September 1993). *Science for the post normal age*. Futures, 25 (7), retrieved from: https://search-proquest-com.pros.lib.unimi.it:2050/docview/1292249765?accountid=12459

[4] Agosto et al. (March 2020). *Monitoring Covid-19 contagion growth in Europe*. CEPS Working Document, 1-2.

[5] Istat. (April 2020). website: https://www.istat.it/it/archivio/4216

[6] Menchetti et al. (2020). *Guida alla lettura e all'interpretazione dei dati Covid-19*. Università degli Studi di Firenze, Italy.

[7] Dipartimento della Protezione Civile. (27th February 2020). *Ulteriori interventi urgenti di protezione civile in relazione all'emergenza relativa al rischio sanitario connesso all'insorgenza di patologie derivanti da agenti virali trasmissibili*. Ocdpc n. 640.

[9] GitHub. (April 2020). website: https://github.com/pcm-dpc/COVID-19

[10] Istat. (April 2020). website: https://www.istat.it/it/archivio/4216

[11] Gruppo San Donato. (April 2020). *Tampone faringeo: come funziona il test per la diagnosi del coronavirus*, website: https://www.grupposandonato.it/news/2020/marzo/tampone-faringeo-coronavirus

[12] European Centre for Disease Prevention and Control. (2020). *Novel coronavirus (SARS-CoV-2) - Discharge criteria for confirmed COVID-19 cases*. technical report.

[13] Luisa Ferrari, Giuseppe Gerardi, Giancarlo Manzi, Alessandra Micheletti, Federica Nicolussi, Silvia Salini. (2020). *Covid-Pro in Italy: A dashboard for a province-based analysis*. Department of Economics, Management and Quantitative Methods – Department of Environmental Science and Policy, University of Milan

[14] Statistica e Società. (April 2020). website: http://www.rivista.sis-statistica.org

[15] la Voce. (April 2020). website: https://www.lavoce.info/archives/65205/decessi-da-covid-come-leggere-i-numeri/

[16] Lavezzo et al. (April 2020). *Suppression of COVID-19 outbreak in the municipality of Vo', Italy*. medRxiv preprint.

[17] ISS. 2020. Website: https://www.epicentro.iss.it/coronavirus/sars-cov-2-decessi-italia; *Analysis done on a sample of 32938 dead patients while positive to Sars-CoV-2 infection*.

[18] System of Surveillance of daily mortality in Italy (SiSMG, Italy)

[19] Luisa Ferrari, Giuseppe Gerardi, Giancarlo Manzi, Alessandra Micheletti, Federica Nicolussi, Elia Biganzoli and Silvia Salini. (2020). *Modelling provincial Covid-19 epidemic data in Italy using an adjusted time-dependent SIRD model.* Department of Economics, Management and Quantitative Methods, Department of

Environmental Science and Policy, Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy.

[20] Brogi, Guadabascio. 2020. *Un modello per la stima dell'andamento del contagio da COVID-19 in Italia*. Research Gate.