Nowcasting Italian Covid-19 epidemic
User guide

Alaimo Di Loro P.   Mingione M.   StatGroup19

May 21, 2020

CoviData19 - An app for all

The CoviData19 Shinyapp is the result of the common project of a group of statisticians who share the same commitment to the social role of statistics, but are aware of the pitfalls that can stem from its miscommunication. In this regard, the final aim of the app is to provide the general public with an accessible platform for consulting and monitoring information about the Italian COVID-19 epidemic in an interactive and transparent way. All data and graphics are automatically updated at every user access with the most recent version available in the Italian Protezione Civile (IPC) Github repository (`https://github.com/pcm-dpc/COVID-19`).

The web-app shows both descriptive and model-based analysis, allowing the user to customize the report as he deasires. In particular, our COVID-19 Shinyapp is composed of 4 panels: (i) *"Overview"*, which provides a general description of the Italian epidemic; (ii) *"Short-term forecast"*, which allows the modeling and short-term forecast of several indicators separately, at national and regional level; (iii) *"ICU Nowcasting"*, which is specifically built to provide robust and trustworty 1-day ahead intensive care unit (ICU) hospitalizations forecast; (iv) *"Info and Credits"*, which includes some useful information about the research team and the app itself.

Introductory notes and warnings

Vocabulary and definitions

First of all, it is important to explain how the indicators are built and what is their meaning. A first distinction must be made between **cumulative** and **daily** indicators. The formers are always positives and increasing over time, as they are reported as the cumulative sum of the daily values. Therefore, also their variations are always positives. However, for the reasons that will be explained in the next paragraph, some of these monotone indicators may decrease at regional level. The cumulative indicators provided by IPC are: cumulative positives, discharged recovered, deceased, swabs and tested cases[1]. Daily indicators are always positives too, but they could increase or decrease over time. As a result, their variations could be both positive or negative. Secondly, some of the indicators can be obtained by summing other basic indicators together. In particular:

- **Cumulative positives** = current positives + deceased + discharged recovered
- **Current positives** = hospitalized with symptoms + intensive care + home isolation
- **New positives** = cumulative positives of today - cumulative positives of yesterday
- **Current hospitalized** = hospitalized with symptoms + intensive care

Further important quantities that can be derived and should be monitored during a pandemic are the mortality rate, defined as the ratio between the total number of deaths and the population size; and the fatality rate, defined as the ratio between the total number of deaths and the total number of positives. Last, but not least, the user must also be aware that the data as they are reported, refer to the number of people that have been found positive to SARS-Cov-2. This does not imply that the same number of people developed COVID-19 disease.

Data issues

The user must be aware that the data present several issues which severely affect the data quality. The general problem of the database is that information are gathered and reported at a regional level and each regional healthcare organization has different transmission and data collection systems. Moreover, the national health data system was not meant to be updated daily. This implies that there is a severe measurement error issue, also including time-dependent propensity to diagnostic testing administration, delays in reporting, saturation of diagnostic and treatment facilities, transfer of patients without notification, and plain data entry errors. Least reliable indicators include positive status (which could have up to 30 days delay in reporting), recovery (which for hospitalized patients corresponded to discharge without ascertainment of negative status in certain regions, and which was severely undercounted in other regions where access to diagnostic test was given least priority to home isolated patients). In our opinion, the most reliable indicator is the count of ICU occupancy. Other issues are structural and mostly due to the suddenness and drama of the epidemic event that caught the country healthcare system mostly unprepared. For all these reasons, discordant data transmissions and lack of coordination are fairly common events; hence, when using the app, everyone must take into account the above and the following warnings:

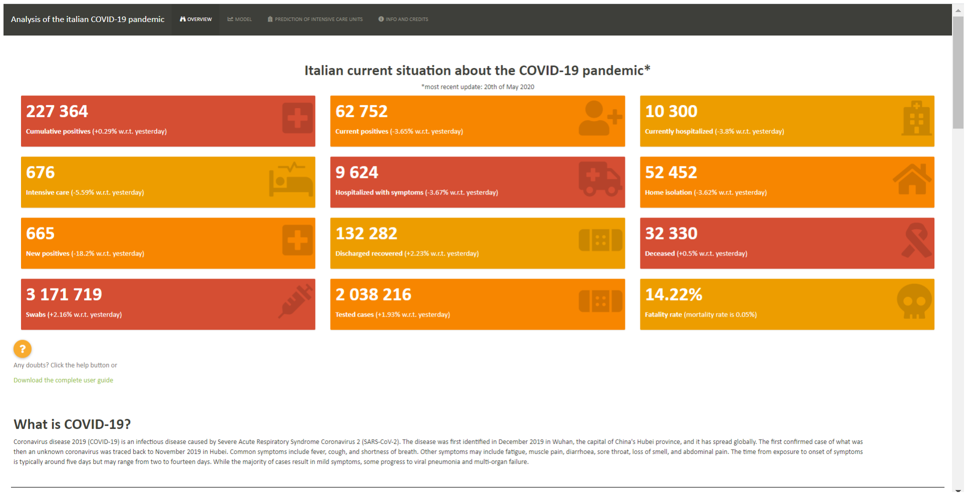
[1] The difference between swabs and tested cases corresponds to control swabs, which are done on the same individual to confirm the healing or for other necessities.

1. 1. Swabs and positive cases are not time aligned. For example, in countries like Singapore (`https://www.moh.gov.sg/covid-19`) daily data include information on total swabs tested, total unique persons swabbed as well as total swabs per 1,000,000 total population and total unique persons swabbed per 1,000,000 total population. In Italy only the total number of daily swabs is available and no linkage between swabs and tested individuals is kept in the data repository. Hence it is not possible to use the information on swabs in any modeling effort.

2. Deaths are aligned part to the date of death and part to the date of communication to the national health institute. Hence some care must be taken when daily data are considered, some variations maybe due to delayed data transmissions and not to a real peak in deaths.

3. ICU are relatively reliable although no official coordination at a national level exists on this point. Usually hospitals search for available beds in intensive care units by phone calling hospital offices directly. Then sometimes people have no access to ICU because of lack of information not because they do not need it.

1 Overview


Analysis of the italian COVID-19 pandemic
OVERVIEW
MODEL
PREDICTION OF INTENSIVE CARE UNITS
INFO AND CREDITS
?

Italian current situation about the COVID-19 pandemic*

*most recent update: 20th of May 2020

227 364
Cumulative positives (+0.29% w.r.t. yesterday)

62 752
Current positives (-3.65% w.r.t. yesterday)

10 300
Currently hospitalized (-3.8% w.r.t. yesterday)

676
Intensive care (-5.59% w.r.t. yesterday)

9 624
Hospitalized with symptoms (-3.67% w.r.t. yesterday)

52 452
Home isolation (-3.62% w.r.t. yesterday)

665
New positives (-18.2% w.r.t. yesterday)

132 282
Discharged recovered (+2.23% w.r.t. yesterday)

32 330
Deceased (+0.5% w.r.t. yesterday)

3 171 719
Swabs (+2.16% w.r.t. yesterday)

2 038 216
Tested cases (+1.93% w.r.t. yesterday)

14.22%
Fatality rate (mortality rate is 0.05%)

Any doubts? Click the help button or

Download the complete user guide

What is COVID-19?

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). The disease was first identified in December 2019 in Wuhan, the capital of China's Hubei province, and it has spread globally. The first confirmed case of what was then an unknown coronavirus was traced back to November 2019 in Hubei. Common symptoms include fever, cough, and shortness of breath. Other symptoms may include fatigue, muscle pain, diarrhoea, sore throat, loss of smell, and abdominal pain. The time from exposure to onset of symptoms is typically around five days but may range from two to fourteen days. While the majority of cases result in mild symptoms, some progress to viral pneumonia and multi-organ failure.

Figure 1: Overview - photograph of the Italian epidemic.

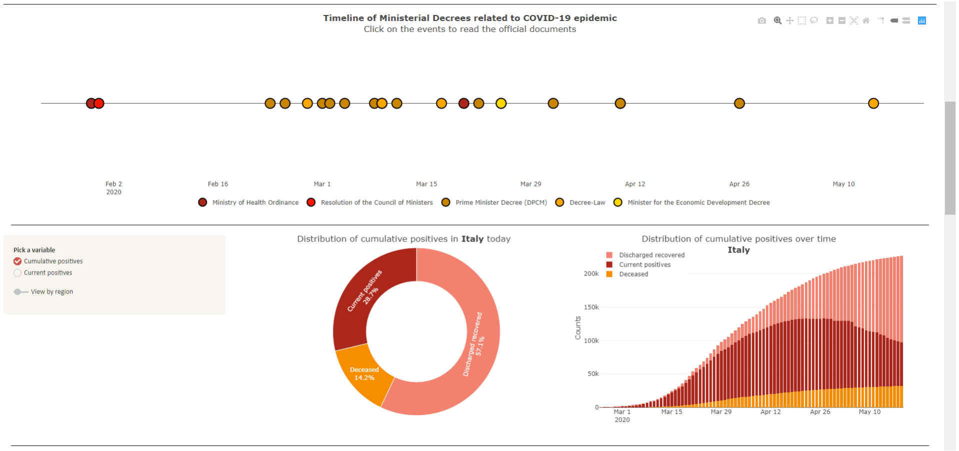
In the *overview* page, several variables are jointly recorded and visualized in order to provide an as accurate as possible picture of the Italian epidemic situation (both at national and regional level) based on data from the most recent update[2]. The page is organized in

[2] Note that every time a user starts the app, we check for updates and then, synchronise the data to the most recent version available on the Github repository of IPC.

six sections.

Section 1 The first section, positioned on top of the home page (see Figure 1), presents an introductory overview in which the user is provided with all the basic information (reported values and changes with respect to the previous day) about the aggregate Italian trend of the epidemic. Definitions, meaning and computing formulas of all the indicators are readily available by clicking the help button just below it; next to it, there is the link to download this user guide. Then, a brief explanation about the COVID-19 and the SARS-Cov2 is provided.

If the interest of the user goes further, the following panels investigate in more details the temporal, compositional and spatial aspects of the available indicators.


Timeline of Ministerial Decrees related to COVID-19 epidemic
Click on the events to read the official documents
Feb 2
2020
Feb 16
Mar 1
Mar 15
Mar 29
Apr 12
Apr 26
May 10
Ministry of Health Ordinance
Resolution of the Council of Ministers
Prime Minister Decree (DPCM)
Decree-Law
Minister for the Economic Development Decree
Distribution of cumulative positives in Italy today
Current positives 28.7%
Discharged recovered 57.1%
Deceased 14.2%
Distribution of cumulative positives over time
Italy
Discharged recovered
Current positives
Deceased
200k
150k
100k
50k
0
Counts
Mar 1
2020
Mar 15
Mar 29
Apr 12
Apr 26
May 10

Pick a variable

☑ Cumulative positives

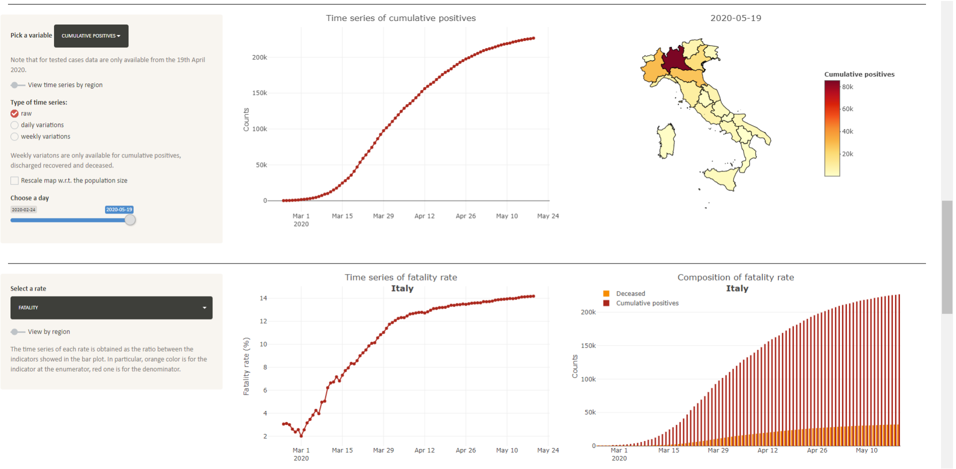
○ Current positives

View by region

Figure 2: Overview - decree timeline and composition of derived indicators.

Section 2 The second section (see the top part of Figure 2) shows a timeline of all the measures adopted by the government in order to keep under control the spread of the epidemic. This timeline is of the utmost importance in order to understand and appropriately interpret the temporal evolution of the indicators.

Section 3 This section, positioned just below the timeline (see Figure 2), presents the stacked barplot (over days) of the indicators' composition in terms of the counts contributing to their definition on the right-hand side. The user can decide to visualize either the *cumulative cases* or the *current positives*, both at national or regional level. On the left-hand side, a *Donut plot* shows the relative distribution (%) of the same indicator at the current date (corresponds to the last bar of the barplot).



Pick a variable CUMULATIVE POSITIVES

Note that for tested cases data are only available from the 19th April 2020.

View time series by region

Type of time series:
☑ raw
○ daily variations
○ weekly variations

Weekly variatons are only available for cumulative positives, discharged recovered and deceased.

□ Rescale map w.r.t. the population size

Choose a day: 2020-02-24 – 2020-05-19

Select a rate: FATALITY

View by region

The time series of each rate is obtained as the ratio between the indicators showed in the bar plot. In particular, orange color is for the indicator at the enumerator, red one is for the denominator.

Figure 3: Overview - spatial and temporal comparison.

Section 4 In this section (see top part of Figure 3), on the left-hand side, it is reported a time-series plot. The user can decide to visualize the temporal evolution of any of the available indicators by selecting the desired one from a drop-down menu. Since the user may want to compare the current situations of different regions, just next to it, a *Map* displays the regional distribution of the selected indicator at the current date. While the current date is the standard choice, the user can easily decide to visualize any other day through the control panel on the left. As the epidemic is known to spread out at different paces across regions, the user may want to focus on one or more specific area and compare the regional trends. With this purpose, it is possible to visualize the time series of the selected indicator at the regional level simply by checking a box on the control panel on the left and selecting the desired region. The possibility to include more than one region is available, so that trends of different regions can be easily compared on the same plot[3]. If the time series is visualized at the regional level, then the map on the right zooms on the selected region/regions and shows the distribution at the province level (only for cumulative positives, which is the only available indicator at the province level). The same control panel allows also to switch the visualization to the daily or weekly variations[4] (absolute or relative) just by checking a box, all others functionality being unchanged. As for the map, there is the possibility to rescale the distribution of the selected indicator with respect to the resident population (number of residents at 31st December 2020) so to get a relative measure the actual impact of the epidemic on the region.

Section 5 This panel (see Figure 3 below) includes functionalities very similar to the previous one. One the left, there is a time series plot in which the user can decide to

[3] Comparison is allowed for up to 5 regions for visualization clarity

[4] only available for cumulative positives, discharged recovered and deceased.

visualize the temporal evolution of 4 ratios between indicators. In particular, the user can pick: (i) the ratio between new positives and daily swabs, called *positivity rate*, and its alternative formulation with the daily tested cases[5] at the denominator; (ii) the ratio between deceased and cumulative positives, called *fatality rate*; (iii) the ratio between discharged recovered and cumulative positives, called *healing rate*; (iv) the ratio between ICUs and current hospitalized, called *severity rate*. Ratios are relative measures and the user may want to get information also on its two components, separately. Hence, on the right-hand side, we included the barplot of the raw indicators by day. The same information are available at the regional level, too.

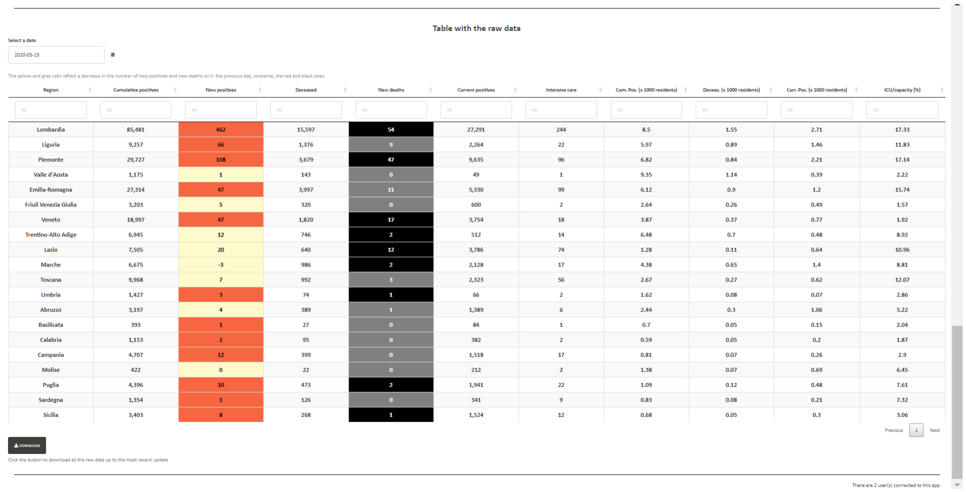


Table with the raw data

Select a date

2020-05-19

The yellow and grey cells reflect a decrease in the number of new positives and new deaths w.r.t. the previous day; viceversa, the red and black ones.

Region	Cumulative positives	New positives	Deceased	New deaths	Current positives	Intensive care	Cum. Pos. (x 1000 residents)	Deceas. (x 1000 residents)	Curr. Pos. (x 1000 residents)	ICU/capacity (%)
All	All	All	All	All	All	All	All	All	All	All
Lombardia	85,481	462	15,597	54	27,291	244	8.5	1.55	2.71	17.33
Liguria	9,257	66	1,376	9	2,264	22	5.97	0.89	1.46	11.83
Piemonte	29,727	108	3,679	47	9,635	96	6.82	0.84	2.21	17.14
Valle d'Aosta	1,175	1	143	0	49	1	9.35	1.14	0.39	2.22
Emilia-Romagna	27,314	47	3,997	11	5,330	99	6.12	0.9	1.2	15.74
Friuli Venezia Giulia	3,203	5	320	0	600	2	2.64	0.26	0.49	1.57
Veneto	18,997	47	1,820	17	3,754	18	3.87	0.37	0.77	1.92
Trentino-Alto Adige	6,945	12	746	2	512	14	6.48	0.7	0.48	8.92
Lazio	7,505	20	640	12	3,786	74	1.28	0.11	0.64	10.96
Marche	6,675	-3	986	2	2,128	17	4.38	0.65	1.4	8.81
Toscana	9,968	7	992	3	2,323	56	2.67	0.27	0.62	12.07
Umbria	1,427	3	74	1	66	2	1.62	0.08	0.07	2.86
Abruzzo	3,197	4	389	1	1,389	6	2.44	0.3	1.06	5.22
Basilicata	393	1	27	0	84	1	0.7	0.05	0.15	2.04
Calabria	1,153	2	95	0	382	2	0.59	0.05	0.2	1.87
Campania	4,707	12	399	0	1,518	17	0.81	0.07	0.26	2.9
Molise	422	0	22	0	212	2	1.38	0.07	0.69	6.45
Puglia	4,396	10	473	2	1,941	22	1.09	0.12	0.48	7.61
Sardegna	1,354	1	126	0	341	9	0.83	0.08	0.21	7.32
Sicilia	3,403	8	268	1	1,524	12	0.68	0.05	0.3	3.06

Click the button to download all the raw data up to the most recent update

There are 2 user(s) connected to this app

Figure 4: Overview - table with the raw data.

Section 6 In the last section (see Figure 4), at the bottom of the page, the user can visualize the table of the regional raw data of a specific day. The first column identify the region, while the following six columns contain: cumulative positives, new cases, total deaths, new deaths, current positives and ICU. The new cases and new deaths columns are coloured according to the trend with respect to the previous day: a pale tonality for a decreasing trend and a vivid tonality for an increasing trend (respectively on a scale from yellow to red and from grey to black). The next four columns contain relative measured of the same indicators, where total cases, total deaths and current positives are reported for every thousand inhabitants while ICU are divided by the region ICU capacity. The user can decide to order the table according to any of the columns and/or to swap positions between columns if he wants to facilitate the comparison between two or more indicators.

Finally, for those who wish to analyse the raw data by themselves, the whole table (including values for all the available days) can be freely downloaded in a convenient `.csv` format by pushing the button on the bottom-left corner of the page.

[5] only available starting from the 19th of April 2020.

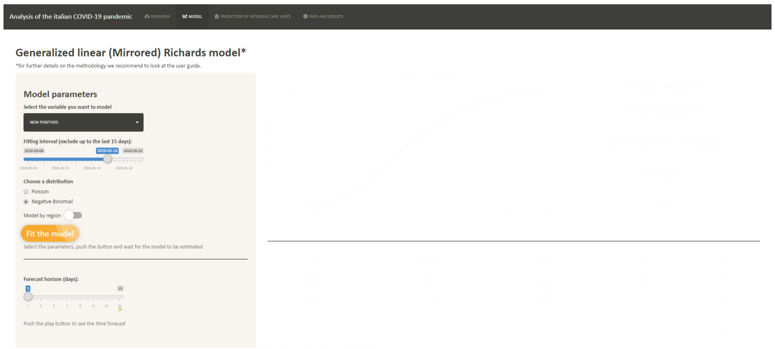
2 Short-term forecast

A simple description of the data at hand is sufficient for most users, but some other may be interested in "what is going to happen tomorrow?". Forecasting through appropriate statistical modelling is not available to everyone at a cost that is affordable. In the *Short-term forecast* section, we provide short-term (i.e. up to 15 days) forecasts of the main indicators, toghether with 99% confidence bounds to account for the estimation uncertainty. All the indicators currently provided by the *Protezione Civile Italiana* are counts, hence we considered as modeling options the Poisson $Poi(\theta_t)$ or the Negative binomial $NB(n, \theta_t)$ distribution.
The process mean θ_t is assumed to follow a pre-specified parametric form $\lambda_\gamma(t)$ over time and observations are considered independent fixed the mean function parameters γ.

In the case of cumulative counts (i.e. cumulative positives, deceased, ...), given the monotone increasing behaviour of the time series, $\lambda_\gamma(t)$ is assumed to follow the *Generalized Logistic Function* (Richard's Curve).

In the case of not cumulative counts (i.e. new positives, current hospitalized, ...), where the behavior is first increasing and then decreasing, the peak T is determined in advance through kernel smoothing. Afterwards, $\lambda_\gamma(t)$ is assumed to follow a function with an increasing logistic behaviour before the observed peak and a decreasing logistic behaviour after the observed peak. Gradient and hessian have been computed analytically and coded for all the *distribution-mean function* combinations and parameters are estimated by maximizing the log-likelihood through gradient based optimization routines (i.e. gradient descent). Robust estimates of their asymptotic variance-covariance matrix are obtained using the Huber Sandwich estimator.

Prediction intervals, unavailable analytically, are obtained through a double bootstrap procedure that accounts for both the uncertainty of parameter estimation and the randomness of the observations. Model estimation and the forecasts are provided in an interactive way, with the estimation routine repeated each time the user clicks on the *Fit the model* button.


Analysis of the italian COVID-19 pandemic
OVERVIEW
MODEL
PREDICTION OF INTENSIVE CARE UNITS
INFO AND CREDITS
Fit the model

Generalized linear (Mirrored) Richards model*

*for further details on the methodology we recommend to look at the user guide.

Model parameters

Select the variable you want to model: NEW POSITIVES

Fitting interval (exclude up to the last 15 days): 2020-05-06 2020-05-16 2020-05-20

Choose a distribution

○ Poisson

● Negative Binomial

Model by region

Select the parameters, push the button and wait for the model to be estimated

Forecast horizon (days): 1 15

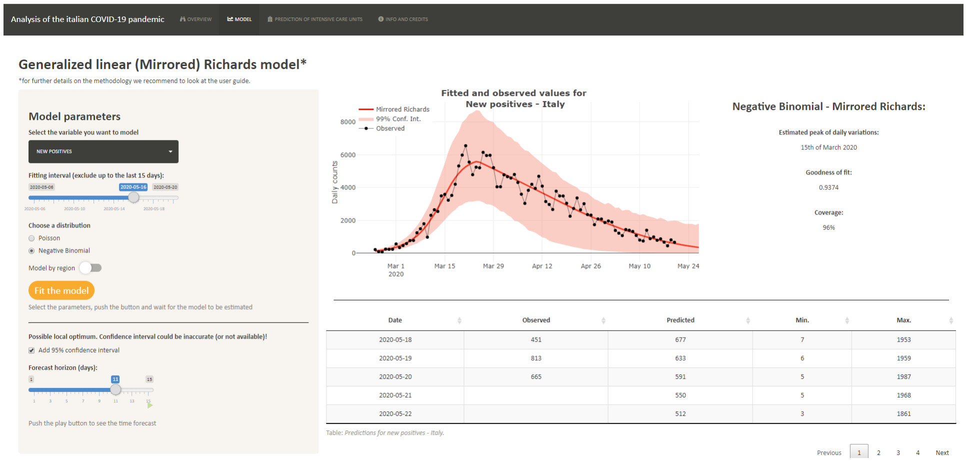
Push the play button to see the time forecast

Figure 5: Short-term forecast - input.

As shown in Figure 5, the user can also decide to vary the fitting interval (exclude up to the last 15 days) and evaluate the out-of-sample predictive performance on the days excluded from the fit. While we consider this functionality both useful (always uses up-to-date data and allows for model evaluation) and enjoyable for the user experience, this does come at a cost. Indeed, sometimes the optimization routine may fail to converge to the global optimum. Generally, this is not an issue in terms of the point estimation and forecasts, but can cause extremely large (and meaningless) confidence bounds. In such cases, the user is warned through a message on the control panel on the left, can decide to not plot the confidence bounds, and is kindly encouraged to re-try and fit the same model again. On the same control panel on the left, after the model fit (see Figure 6), the user can choose the *prediction horizon term* from 1 to up to 15 days. Alternatively, he can let the whole forecast appear day after day with an animation by clicking on the *play* button on the same panel.

Schematically, the app window reports (see Figure 6 on the right):

- **Top-left**. Plot with observed and predicted values for the selected variable
- **Top-right**. Summary for the results, which includes:
  - Day of the estimated peak for the daily variations
  - Estimated upper asymptote (only for cumulative indicators)
  - Goodness of fit
  - Coverage
- **Bottom**. Table containing observed and predicted values of the past 2 days and the forecasts of the next 15 days


Analysis of the italian COVID-19 pandemic
OVERVIEW
MODEL
PREDICTION OF INTENSIVE CARE UNITS
INFO AND CREDITS

Generalized linear (Mirrored) Richards model*

*for further details on the methodology we recommend to look at the user guide.

Model parameters

Select the variable you want to model: NEW POSITIVES

Fitting interval (exclude up to the last 15 days): 2020-05-06 – 2020-05-16 – 2020-05-20

Choose a distribution

○ Poisson

● Negative Binomial

Model by region

Select the parameters, push the button and wait for the model to be estimated

Possible local optimum. Confidence interval could be inaccurate (or not available)!

☑ Add 95% confidence interval

Forecast horizon (days): 1 – 11 – 15

Push the play button to see the time forecast

Negative Binomial - Mirrored Richards:

Estimated peak of daily variations:

15th of March 2020

Goodness of fit:

0.9374

Coverage:

96%

Date	Observed	Predicted	Min.	Max.
2020-05-18	451	677	7	1953
2020-05-19	813	633	6	1959
2020-05-20	665	591	5	1987
2020-05-21		550	5	1968
2020-05-22		512	3	1861

Table: *Predictions for new positives - Italy.*

Figure 6: Short-term forecast - output.

3 ICU Nowcasting

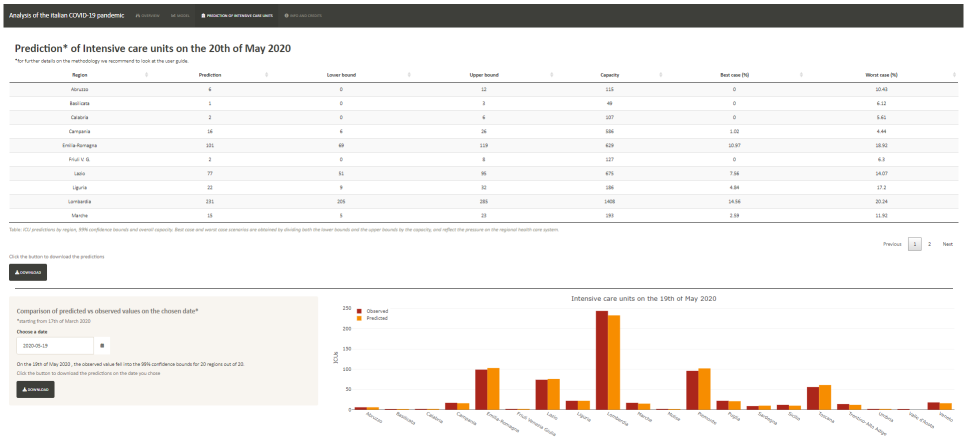
Since the beginning of the epidemic, the COVID-19 has placed a huge stress on the National Health System facilities of the most hardly hit regions. As a matter of fact, a percentage between 5% and 20% of the Italian severe cases required an ICU bed, pushing the critical care capacity to (and over) the limit for regions like Lombardia and Piemonte. The overcrowding of hospital facilities and the consequent risk of breakdown of the National Health System is the greatest challenge this pandemic has put Italy through, hence monitoring the available ICU capacity is key in order to act timely and prevent this to happen.

The *ICU Nowcasting* section provides forecasts of the number of ICU beds that are likely to be required in each region on the next day. These predictions are obtained through an ensemble method, working at regional level, based on the optimal weighting of two models trained on data from the previous two weeks.

The first model is a *Poisson Generalized Linear Mixed Model* (GLMM) with region-specific random intercepts and region-specific random linear and quadratic trends. The number of residents in the region as on the 31st December 2019 is included as an offset.

The second model is a non-stationary INAR(1) model with Poisson innovations, fitted separately on data from each region, including a polynomial trend up to the cubic degree (a BIC-based selection procedure is used to drop uninformative terms region-wise).

Predictions and 99% CIs resulting from the two models are then put together by taking a weighted average, where the optimal weights are determined by a leave-last-out validation.


Analysis of the italian COVID-19 pandemic
OVERVIEW
MODEL
PREDICTION OF INTENSIVE CARE UNITS
INFO AND CREDITS

Prediction* of Intensive care units on the 20th of May 2020

*for further details on the methodology we recommend to look at the user guide.

Region	Prediction	Lower bound	Upper bound	Capacity	Best case (%)	Worst case (%)
Abruzzo	6	0	12	115	0	10.43
Basilicata	1	0	3	49	0	6.12
Calabria	2	0	6	107	0	5.61
Campania	16	6	26	586	1.02	4.44
Emilia-Romagna	101	69	119	629	10.97	18.92
Friuli V. G.	2	0	8	127	0	6.3
Lazio	77	51	95	675	7.56	14.07
Liguria	22	9	32	186	4.84	17.2
Lombardia	231	205	285	1408	14.56	20.24
Marche	15	5	23	193	2.59	11.92

Table: *ICU predictions by region, 99% confidence bounds and overall capacity. Best case and worst case scenarios are obtained by dividing both the lower bounds and the upper bounds by the capacity, and reflect the pressure on the regional health care system.*

Click the button to download the predictions

Comparison of predicted vs observed values on the chosen date*

*starting from 17th of March 2020

Choose a date: 2020-05-19

On the 19th of May 2020 , the observed value fell into the 99% confidence bounds for 20 regions out of 20.

Click the button to download the predictions on the date you chose

Figure 7: Prediction of ICUs by region.

The page is organized in two panels, as it is shown in Figure 7.

Panel 1 In the first first panel, the user can observe a (downloadable) table containing the forecasts and the relative 99% confidence interval of the ICU beds to be dedicated to

COVID-19 patients on the next day, by region. The overall capacity of ICUs per region and the worst case/best case ratios are also reported.

Panel 2 The second panel provides the user with tools to evaluate the actual model performances on previous days. Indeed, the user can pick a date from the control panel on the left and generate a grouped bar-plot that compares predicted and observed values for the selected date. On the same control panel, just below the selected date, the user can read for how many of the regions the observed value was included in the forecast interval for the selected date. The user, if interested, can easily download the data generating the barplot by clicking the *download* button on the bottom-left corner of the page.

4 Info and Credits

In the last section, the user can find a brief description of StatGroup-19 and the assistants who collaborated with the research team for the web-app development. Links to the StatGroup-19's blog, facebook page and to the italian version of the ShinyApp are promptly available.

All data sources are referenced in the *Sources* paragraph, where an hyperlink connects directly to the IPC Github repository.

In the end, we wish to thank the *RStudio* software developers' team for providing such a flexible and powerful platform as *Shiny*, key tool in the production of this web-app. Moved by the aim of transparency and collaboration, all the codes used to generate the web-app are freely available in the Github repository `https://github.com/minmar94/StatGroup19`.