Statistical communication of COVID-19 epidemic using widely accessible interactive tools

Comunicazione statistica dell'epidemia di COVID-19 attraverso l'utilizzo di uno strumento interattivo

M. Mingione and P. Alaimo Di Loro

Abstract *High-quality data is crucial for guiding decision making. Data quality frailties have been exposed worldwide during the current COVID-19 pandemic. The latter complicates the prediction of its evolution and the assessment of both health and economic interventions. Indeed, the process of data collection of the main pandemic variables is murky and not intended for statistical analysis, favoring convenient narratives and only apparently supporting policy-making processes. We aim at providing proper communication to the general public and inform on the daily evolution of the epidemic. That is achieved by the interactive tool here introduced, along with some alerts highlighting the fallacy of indicators as poorly informative when considered alone. We discuss the utmost importance to consider simultaneously multiple indicators, cross-verifying their behavior in order to distinguish relevant information from harmful and dangerous misinterpretations. Information are summarized through easily readable and accessible graphs and interactive maps. Predictions are based on novel approaches and models and can be used as alerts to identify at-risk situations.*

Abstract *Dati di alta qualità sono cruciali per guidare il processo decisionale. Lacune nella qualità dei dati sono emerse in tutto il mondo durante l'attuale epidemia di COVID-19. Queste lacune complicano la previsione dell'evoluzione dell'epidema e la valutazione dei relativi interventi sanitari ed economici. Infatti, il processo di raccolta dati dei principali indicatori dell'epidemia è confuso e non progettato per l'analisi statistica, favorendo interpretazioni convenienti e soltanto apparentemente a supporto del processo legislativo. Il nostro scopo è quello di fornire al pubblico una corretta comunicazione statistica e di informare sull'andamento giornaliero dell'epidemia. Ciò si realizza attraverso uno strumento interattivo introdotto di seguito, in aggiunta ad alcune avvertenze che mostrano la scarsa informatività degli indicatori se considerati singolarmente. Si rileva la fondamentale importanza del considerare contemporaneamente più indicatori, mediante la verifica incrociata del loro comportamento, al fine di distinguere le informazioni rilevanti da interpretazioni errate, dannose e pericolose. Le informazioni sono riassunte in grafici e mappe interattivi. Le previsioni si basano su nuovi approcci e i modelli possono essere utilizzati come segnali per identificare le situazioni a rischio.*

Key words: COVID-19, Shiny, Data quality, Open data

Marco Mingione
University of Rome "La Sapienza", Statistical Science Department, e-mail: marco.mingione@uniroma1.it

Pierfrancesco Alaimo Di Loro
University of Rome "La Sapienza", Statistical Science Department, e-mail: pierfrancesco.alaimodiloro@uniroma1.it

1 Introduction

This work is the result of the joint 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 poor quantitative communication. In this regard, throughout the first year of the epidemic, the goal of our research group was manifold: (i) predict the evolution of the most relevant epidemic indicators and produce the forecast of the day of the peak for each curve; (ii) predict ICU occupancy by region to allow for an optimal allocation of health resources; (iii) sensitize the general public to the importance of correct statistical communication, allowing for a transparent and reproducible policy-making process.

COVID-19 public Italian data present several issues that severely affect their quality. Since the beginning of the epidemic, data have been collected for administrative and surveillance purposes mainly. Attention to the coherency, comparability and consistency of the collection process has been largely overlooked, hindering the inferential capability of any statistical analysis. To the best of our knowledge, data are and have always been gathered with very few shared standard guidelines. As a matter of fact, each regional healthcare department has its own different data collection and transmission system, which do not require compliance to any specific criteria. Measurement errors and errors in data entry are therefore expected to be often present, as well as substantial delays in reporting. Hence, any analysis of these data shall be limited to monitoring the *status quo* and produce scenarios projections rather than reliable medium to long-term predictions. In order to study and understand current and future states of the epidemic, higher quality and detailed information is of the utmost importance. Indeed, it is necessary that research groups are able to align the different indicators and follow the individual pathways of contagion and clinical evolution. Currently, the only recognized source of public data about the Italian COVID-19 pandemic is the Italian Protezione Civile (IPC) Github repository[1]. Data are aggregated and daily updated with the new flow of information coming from the regional system at around 6 p.m. . Despite all these limitations, StatGroup-19 believed that a more compelling and informative picture of the pandemic could be sketched using that data. This motivated the production of the *web application* described in Section 2.

2 A COVID-19 web app

The web application described here is built using `R Shiny` [8] and intends to provide the general public with a tool for accessing information about the Italian COVID-19 epidemic in an interactive and transparent way. The application is automatically updated at every user access with the most recent version available in the IPC Github repository and is accessible at https://statgroup19.shinyapps.io/Covid19App/[2]. It shows both descriptive and model-based analysis, allowing the user to customize several choices. In particular, it is composed of 4 main 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 daily incidence indicators, at national and regional level; (iii) *"ICU Nowcasting"*, which is specifically built to provide robust and trustworthy 1-day ahead intensive care unit (ICU) hospitalizations forecast; (iv) *"Vaccines"*, which includes some useful information about the vaccination campaign in Italy. Plots, data and all source codes are public and can be freely accessed at https://github.com/minmar94/StatGroup19, in the spirit of a completely *Open Data* community.

[1] https://github.com/pcm-dpc/COVID-19

[2] English version of the app is available at https://statgroup19.shinyapps.io/StatGroup19-Eng/

Fig. 1: Overview - daily report.

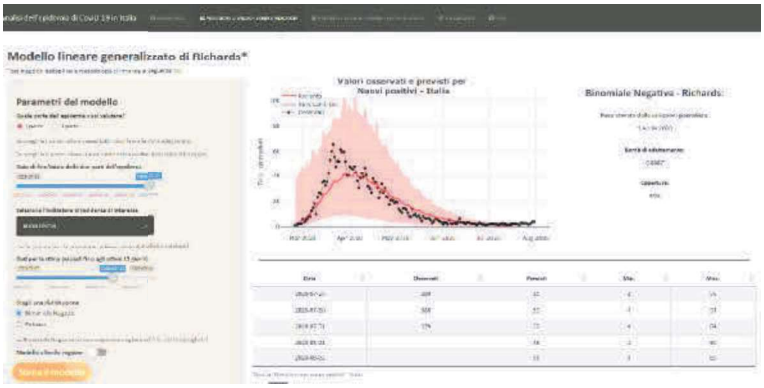
2.1 Overview

In the *Overview* page, the most relevant epidemic indicators 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. The page is organized in different sections and includes: (i) the IPC daily report (see Figure 1), enriched with some more interpretable relative measures; (ii) the *decrees* timeline, allowing for the understanding and quantification of the eventual effects yielded by the measures adopted to contain the spread of the contagion; (iii) time-series and maps, so that a detailed investigation of the temporal and the spatial distribution of the available indicators and their ratios (e.g. positivity rate, fatality rate, healing rate, etc.) is possible; (iv) the table containing the daily raw data (the user can go back to the day in which the systematic data gathering process started, i.e. February 24, 2020), together with some relative indices for comparison among regions.

2.2 Short-term forecast of incidence indicators

This section provides short to medium term forecast of incidence indicators at both national and regional level. Incidence indicators measure the number of individuals with a particular condition, related with the epidemic, recorded during a given period. These indicators can be considered, by analogy with the terminology used in econometrics, as flow data, quantifying the daily input (e.g. positives) and output (e.g. deceased and recovered/discharged) of the system. We propose a parametric regression model for the modeling of incidence indicators based on the use of the Richard's curve [6] as response function in place of the widely used exponential or polynomial trend. Furthermore, we replace the generally entrenched Gaussian assumption for the distribution of log-counts [5; 7] by the more appropriate Poisson or Negative Binomial distributions for counts.

Further details on the specific methodology are described in [2]. The current version of the model provided robust and accurate forecasts during the first wave, but it is able to describe only one pandemic



(a)

(b)

Fig. 2: Short-term forecasting of incidence indicators (a) and nowcasting of ICUs (b).

wave at a time. The user can decide either modeling the first or the second wave, but a more comprehensive extension is currently under development and will be added in the next future. This page shows a graphical representation of the fit, predicted values and the 95% confidence intervals (with related coverage) up to the next 15 days and reports the estimated day of the *true peak* and various goodness of fit measures (see Figure 2a).

2.3 Nowcasting of intensive care units

The overcrowding of hospital facilities and the consequent risk of a breakdown of the National Health Care System is the greatest challenge this pandemic has put Italy through. Hence, monitoring the available ICU capacity is critical in order to act timely and prevent this from happening. We dedicated a specific section of the application to the 1-day ahead prediction of ICU occupancy for each region. Specific details of the methodology are described in [3]. The model is based on an optimal ensemble of two simple methods. The terms are a generalized linear mixed regression model [1], that pools information

over different areas, and an area-specific non-stationary integer autoregressive methodology [4]. Both regional population and ICU capacity are used as offsets in the modeling efforts.

As soon as Protezione Civile updates the Github repository, the app updates predictions for ICU occupancy for the next day. Point predictions are provided together with 99% confidence intervals. Since the beginning of the epidemic, the forecasts have always been accurate up to $3-6$ beds at a regional level, with the confidence intervals containing the true future value in $\approx 100\%$ of the cases. The user can compare and download predicted and observed values for each day (see Figure 2b).

2.4 The vaccination campaign

Italy (and the rest of the world) started seeing the light at the end of the tunnel on December 27, 2020. On this day, also called *V-Day*, the vaccination campaign started and, ever since, most of the effort put in place from the health-care systems has been dedicated to this task. The goal is to complete the vaccination of the whole Italian population (or at least the 70% of it) by the end of 2021. For this reason, we decided to dedicate a section to the monitoring of the Italian vaccination campaign. Percentages of vaccinated people are available at both national and regional level by gender, category and age class. The user can also customize a regional map in which administered vaccine doses are reported either in absolute value, either as a fraction of the delivered doses or as a fraction the residents.

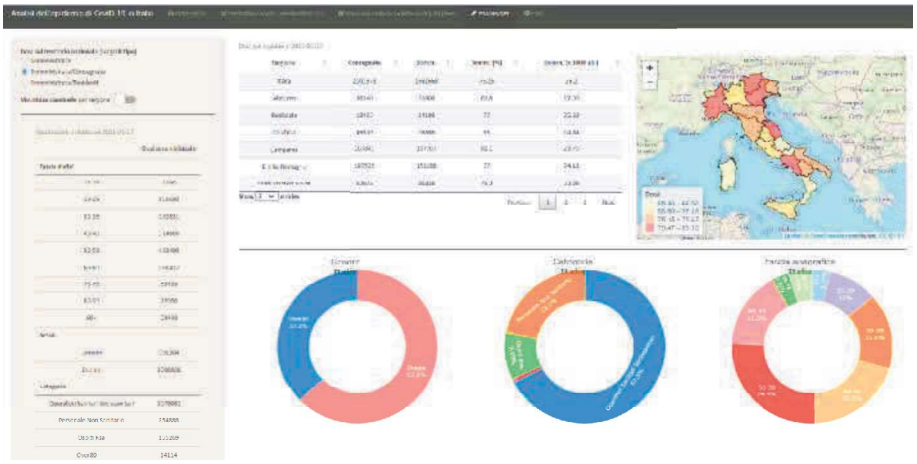


Fig. 3: Vaccination campaign in Italy.

3 Conclusions and further development

The web app described in Section 2 has been release on a https://www.shinyapps.io/ server since the beginning of May 2020. Many other web apps have been devised with the same purposes during the last year, both at national and international levels. Our group wanted to give its own contribution in monitoring what we believe are the key aspects of the epidemic. Even though far from perfection, it has

drawn considerable attention since its first releas. The app is particularly appreciated for its ease of usage and the interactive visualization tools that facilitates interpretation of the IPC data in a more friendly and perceptive way. The app is continuously under development, following the new possibilities and needs as well as the feedback and suggestions of the most zealous users.

Nevertheless, we must deal with the fact that the data necessary to construct more insightful and adequate information are currently in possession of government agencies and bodies, but not made available to the wide scientific community. We are perfectly aware that the guarantee of privacy and confidentiality are at stake, but we are concerned that further unknown considerations are limiting the proper pre-processing and masking that would turn the raw data into harmless accessible information. At this point in the evolution of the pandemic, the aggregated public data are no longer sufficient to make the government's decision-making mechanism transparent. More importantly, the scientific community has not been able to understand (and to replicate) some crucial quantities on which these decisions are taken.

Acknowledgments

We would like to thank our fellow colleagues from the StatGroup-19, Professors Alessio Farcomeni, Fabio Divino, Giovanna Jona Lasinio, Gianfranco Lovison and Antonello Maruotti for getting us involved in this research adventure.

References

[1] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.

[2] Pierfrancesco Alaimo Di Loro, Fabio Divino, Alessio Farcomeni, Giovanna Jona Lasinio, Gianfranco Lovison, Antonello Maruotti, and Marco Mingione. Nowcasting covid-19 incidence indicators during the italian first outbreak. *arXiv preprint arXiv:2010.12679*, 2020.

[3] Alessio Farcomeni, Antonello Maruotti, Fabio Divino, Giovanna Jona-Lasinio, and Gianfranco Lovison. An ensemble approach to short-term forecast of covid-19 intensive care occupancy in italian regions. *Biometrical Journal*, 2020.

[4] Konstantinos Fokianos, Anders Rahbek, and Dag Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439, 2009.

[5] G. Grasselli, A. Pesenti, and M. Cecconi. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: Early experience and forecast during an emergency response. *Journal of the American Medical Association*, 323: 1545–1546, 2020.

[6] FJ Richards. A flexible growth function for empirical use. *Journal of experimental Botany*, 10(2):290–301, 1959.

[7] G. Sebastiani, M. Massa, and E. Riboli. COVID-19 epidemic in Italy: evolution, projections and impact of government measures. *European Journal of Epidemiology*, 35:341–345, 2020.

[8] Hadley Wickham. *Mastering Shiny*. O' Reilly, 2020.