Journal Pre-proof

Using newspaper obituaries to “nowcast” daily mortality: evidence from the Italian COVID-19 hot-spots

Paolo Buonanno, Marcello Puca

PII: S0168-8510(21)00023-3

DOI: https://doi.org/10.1016/j.healthpol.2021.01.006

Reference: HEAP 4398

To appear in: *Health policy*

Received Date: 3 August 2020

Revised Date: 15 January 2021

Accepted Date: 21 January 2021

Please cite this article as: { doi: https://doi.org/

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 Published by Elsevier.

1. Introduction

Since the first suspected pneumonia cases observed on December 2019 in Wuhan (China), the novel coronavirus (SARS-CoV-2), causing a severe acute respiratory syndrome (COVID-19), has turned into a pandemic and a global challenge [c.f. 1]. In response, policymakers across the world are monitoring epidemiological indicators in real-time aiming to define successful containment measures [2, 3, 4]. Among such indicators, deaths caused by COVID-19, general mortality rates, and excess mortality are three standard epidemiological indicators used by the World Health Organization (WHO) to monitor the spread of an infectious disease [5]. However, comparing the current total mortality rate to the average mortality rate of previous years, which is the definition of excess mortality, is more reliable compared to "crude" or adjusted case fatality rates [c.f. 6, 7]. Indeed, the number of deaths caused by an infectious disease may either be underestimated if the health system fails to detect the infection in a dead patient (*censoring bias*), or overestimated if the health system focuses only on severe cases (*ascertainment bias*) [8, 9, 10, 11].

While excess mortality does not suffer from such biases, it might be difficult to estimate during a pandemic peak for two reasons. First, vested interests may incentivize some policymakers to hide deaths and cases caused by an infectious disease [12], which means that the reported number of deaths may under-represent the actual mortality rate [13, 14]. Second, even under the assumption that official statistics do not suffer from any bias, in almost all countries, official mortality records are published with a substantial delay. Indeed, the latest release of the WHO mortality database currently available, was updated on November 15, 2019 [15]. In this respect, the latest Italian National Statistical Institute (ISTAT) release of official mortality data, published on May 9, 2020, which covers the period between January 1, 2020 to April 15, 2020, should be seen as an "early release". Indeed, ISTAT usually publishes mortality data after at least one year (see Section 2 for further details). This release followed the concern raised by several Italian policymakers during the early stage of the COVID-19 outbreak that the official statistics were not accurately representing the actual number of deaths caused by COVID-19, which was believed to be much higher [16, 17].

In this paper we propose a novel approach to deal with the publication delay of mortality

Journal Pre-proof

Journal Pre-proof

data by "nowcasting" daily mortality using newspaper obituaries as an alternative source of information. Using standard forecasting techniques, we compare different models, and find that forecasts based on obituaries outperform autoregressive forecasts based on previous mortality rates according to several accuracy criteria. More specifically, we use obituaries published on the local newspapers of Bergamo (120,000 inhabitants) and Brescia (196,000 inhabitants), both in the Lombardy region, during the interval between February 24 to May 14, 2020. Lombardy is considered the European hot-spot of the first COVID-19 wave, with 88,183 reported cases and 15,974 deaths as of May 25, 2020, over a total population of approximately 10 million inhabitants [18, 19]. It therefore represents a relevant case study for our research purpose.
Our findings show that newspaper obituaries account for, roughly, half of the actual number of deaths and can be a valid, reliable, and timely alternative to official mortality statistics. Indeed, while being a non-uniform representation of a society, obituaries are used across the world and in different cultures [e.g. 20, 21]. These results may be relevant for many policymakers and health authorities for the surveillance of new infectious disease outbreaks. The next section presents the data used in our estimates, and the empirical methodology. In the third section we report our results. In the fourth and final section we discuss the limitations of our approach and provide possible policy recommendations.

2. Data and Methods

The basic principle of "nowcasting" is to exploit information which is published at a higher frequency than the variable of interest [22]. We thus use this approach to address the delay associated with the publication of official mortality data, and obtain a prediction of COVID-19 severity in almost real-time. We start by describing our dataset and then explore the accuracy of estimates based on newspaper obituaries published in two local newspapers using standard forecasting techniques.

Journal Pre-proof

Journal Pre-proof

2.1. Data

2.1.1. Newspapers obituaries

We digitalized obituaries published online by *L'Eco di Bergamo* and *Il Giornale di Brescia*, the two most circulated newspapers in the provinces of Bergamo and Brescia, respectively. In 2019, the daily number of readers of *L'Eco di Bergamo* was 402,000, while the daily number of readers of *Il Giornale di Brescia* was 427,000 (c.f. `http://audipress.it/quotidiani/`). Each obituary contains individual characteristics such as name, surname, age, gender, date of death and municipality of death. Our final dataset contains 4,054 and 3,874 unique individuals from February 24 to May 14, 2020 for the provinces of Bergamo and Brescia, respectively.

Newspaper obituaries contain a wide range of information about the victim such as name, surname, gender, age, date of death, and the municipality of death. Figure 1 displays the daily progression of mortality (solid line) and the number of published obituaries (dashed line) for each of the two municipalities in our sample. By comparing the number of obituaries to the number of actual registered deaths in our sample, we conclude that obituaries represent roughly 56% percent of the total number of deaths. Although obituaries represent only a subset of the officially registered deaths, with a gap increasing during the peak, the correlation between the two measures is glaring, and approximately equal to 97%.

2.1.2. Official mortality

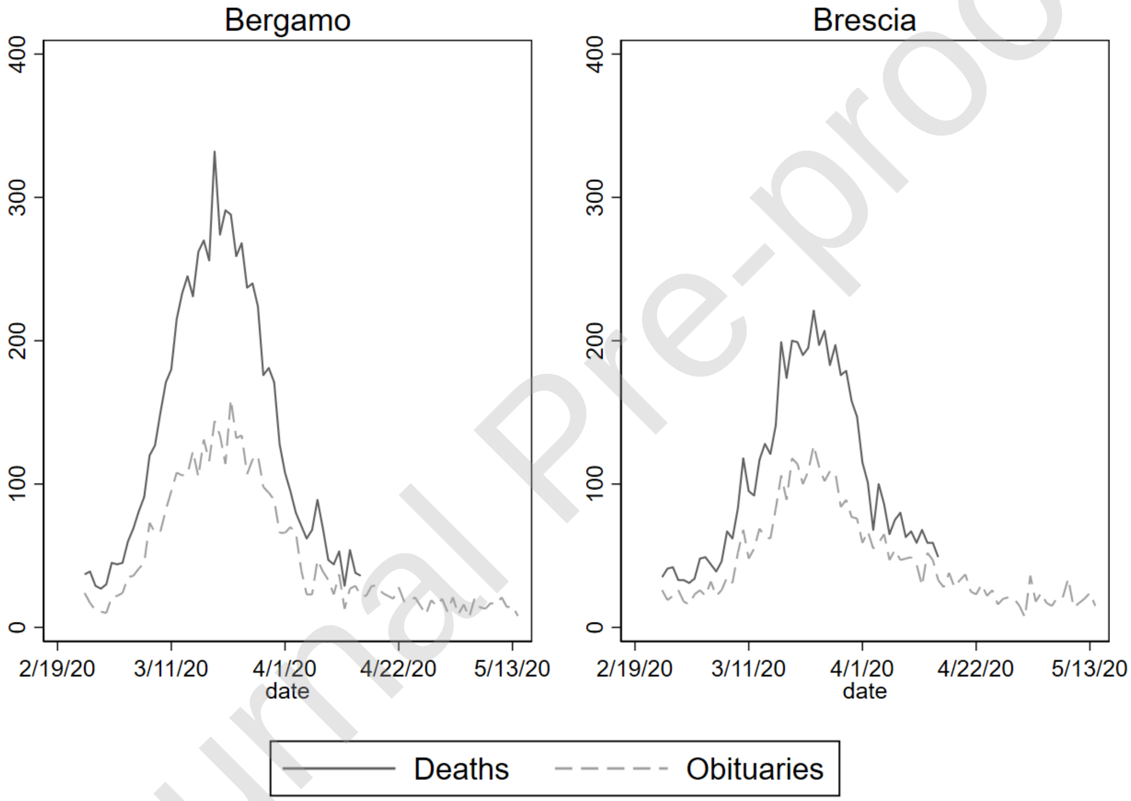
We combine obituaries data with mortality data at the municipality level released by ISTAT on May 9, 2020. ISTAT data are publicly available online at `https://www.istat.it/it/archivio/240401`. The ISTAT dataset contains daily deaths at the municipality level from January 1 to April 15, 2020 for a sample of 4,433 Italian municipalities. The ISTAT sample covers all municipalities belonging to the two provinces of our analysis (243 and 205 municipalities in the provinces of Bergamo and Brescia, respectively).

2.2. Methods

Formulation of the Augmented-ARMA(1,2) model. We use an auto-regressive moving-average model of mortality *augmented* with newspaper obituaries, which we label *Augmented-ARMA (AARMA)*. We select an AARMA(1,2) specification after inspectiing the autocorrelation and

Journal Pre-proof

Figure 1: Deaths vs Obituaries



Notes: Daily progression of deaths (solid line) and obituaries (dashed line) in our sample.

Journal Pre-proof

Journal Pre-proof

partial autocorrelation plots which display, respectively, a one-lag significant autocorrelation coefficient, and a two-lag partial autocorrelation coefficients (c.f. Fig4 and Fig5 in section 6). Formally, this writes as

$$\begin{aligned} y_t &= \mu_y + \sum_{i=1}^{n} \alpha y_{t-i} + \beta x_t + \varepsilon_t \\ \varepsilon_t &\xrightarrow{d} MA(2) \end{aligned} \tag{1}$$

where $y_t = \ln(deaths_t)$ is the log-transformed number of deaths observed on day t, μ_y is the unconditional mean of y, n is the number of model lags, and $x_t = \ln(obituaries_t)$ is the log-transformed number of newspaper obituaries referring to deaths occurred on day t, which is assumed to be exogenous with respect to the time series $\{y_t\}$ (i.e. $\mathbb{E}[\varepsilon_t|x_t] = 0$).

In what follows, we compare the accuracy of the AARMA(1,2) predictions to those we would obtain with: (i) an OLS model (i.e. $y_t = \mu_y + \beta x_t + \varepsilon_t$); (ii) an AR(1) (resp. AR(3)) model without information on obituaries, i.e. model (1) with $\beta = 0$ and $n = 1$ (resp. $n = 3$).

Accuracy metrics. The root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and Theil's U of the estimator $\hat{y}_t$ to the target mortality level y_t are defined, respectively, as

- $RMSE(\hat{y}_t, y_t) = [\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2]^{\frac{1}{2}}$
- $MAE(\hat{y}_t, y_t) = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t|$
- $MAPE(\hat{y}_t, y_t) = \frac{1}{n}\frac{\sum_{t=1}^{n}|\hat{y}_t - y_t|}{y_t}$
- $Theil(\hat{y}_t, y_t) = \frac{RMSE(\hat{y}_t, y_t)}{RMSE_{naive}}$

where $RMSE_{naive}$ refers to the RMSE of a naive forecast, i.e. $y_t = y_{t-1}$ [23]. The AIC and BIC are defined, respectively, as $AIC = 2k - 2\ln(\hat{L})$ and $BIC = k\ln(T) - 2\ln(\hat{L})$, where $\hat{L}$ maximizes the likelihood function of the estimated model, k is the number of estimated parameters, and T is the sample size.

Journal Pre-proof

3. Results

Table 1 reports the accuracy of different forecasting models of daily mortality from February 24, 2020 to May 15, 2020, with *Panel A* (resp. *Panel B*) reporting observations for the municipality of Bergamo (resp. Brescia). Specifically, we compare the estimated mortality level to the true mortality published by ISTAT on May 9, 2020 and computed different accuracy metrics described in Section 2. These measures include the root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), the Theil's U, the Akaike's information criterion (AIC), and the Bayesian Information Criterion (BIC). We compare these measures for (i) ordinary least squares (OLS) estimates; (ii) "augmented" autoregressive-moving-average (AARMA(1,2)) estimates with obituaries as exogenous variables; (iii) one-lag autoregressive estimates (AR(1)); (iv) three-lag autoregressive estimates (AR(3)). Lower values of each of these measures correspond to more accurate forecasts. Therefore, according to every performance metric, we report that AARMA(1,2) outperforms all other models, and for both provinces in our sample.

Table 1: Comparison of different forecasting models of mortality

	Model	**RMSE**	**MAE**	**MAPE**	**Theil's U**	**AIC**	**BIC**
Panel A: Bergamo	OLS	0.184	0.136	0.032	0.830	-24.299	-20.396
	AARMA(1,2)	0.137	0.113	0.026	0.581	-51.078	-39.370
	AR(1)	0.215	0.165	0.039	0.989	-8.131	-2.277
	AR(3)	0.210	0.163	0.039	0.961	-6.915	2.841
Panel B: Brescia	OLS	0.172	0.140	0.033	0.953	-31.734	-27.832
	AARMA(1,2)	0.158	0.122	0.029	0.863	-35.067	-23.359
	AR(1)	0.194	0.151	0.035	0.986	-23.034	-17.181
	AR(3)	0.191	0.148	0.034	0.981	-19.623	-9.866

Notes: Metrics of forecast accuracy. AARMA(1,2) refers to model (1). RMSE, MAE, MAPE, and Theil's U are for root mean squared error, mean absolute error, mean absolute percent error, and Theil's U statistic, respectively. AIC and BIC refer to the Akaike's Information Criterion and the Bayesian Information Criterion, respectively.

We also report in Figure 2 the forecasted mortality against the observed mortality, while Figure 3 displays the daily evolution of the estimated standard errors for each model. In Figure 3, note that the AR(1) and the AR(3) curves overlap. In Figure 2, note that the curves refer both to the in-sample (resp. out-of-sample) forecast, as they refer to the period before (resp. after) April 15, i.e. the last available date for official mortality data provided by ISTAT. Both the AARMA(1,2) and the OLS estimates have a much lower prediction error, and a close inspection of the estimates shows that both AARMA(1,2) and OLS estimates outperform models based

Journal Pre-proof

Journal Pre-proof

only on previously observed mortality data (i.e. AR(1) and AR(3)) over the entire period in our sample.

4. Discussion and Limits

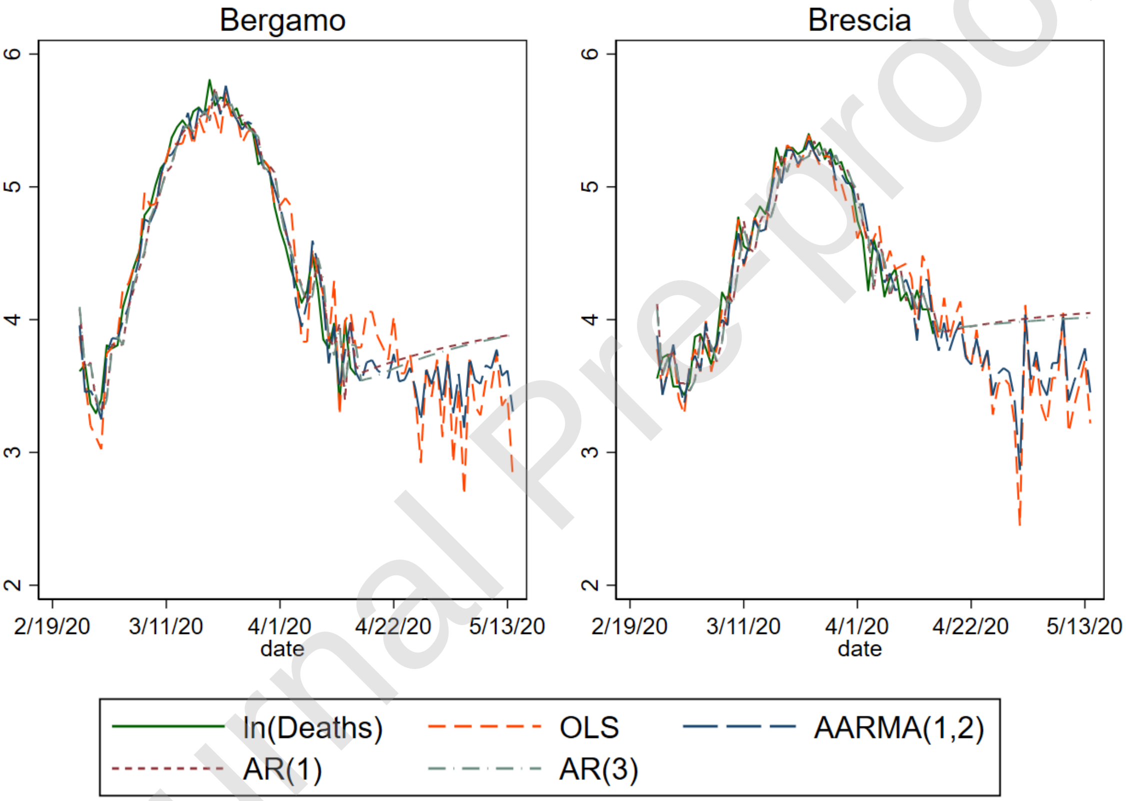
Our approach, though powerful, is not free from limitations. First, despite being concentrated in the most affected Italian region, our sample refers only to two provinces. Second, since they are costly, newspaper obituaries may under-represent the actual mortality level, in particular for lower income or marginalized communities. This issue becomes more severe during the epidemic peak (see Fig1). The effect of such under-representation on our estimates, however, is mixed. On the one hand, if obituaries were published for every COVID19 victim, our results would (obviously) be only marginally more accurate, as the sample would virtually coincide with the actual population. It is less obvious, on the other hand, what is the implication of the households-level heterogeneity towards the publication of obituaries on the accuracy of our forecasts. We are also agnostic about potentially heterogeneous behavioral attitudes towards such publications in other locations or cultures. In this respect, the large variance observed in civic attitudes and prosocial behavior across Italian municipalities may play a role in determining this attitude [cf. 24]. Understanding how such cultural heterogeneity may affect our conclusions, as well as increasing the sample size so as to include more municipalities, constitutes a valuable path for future research.

5. Conclusion and Policy Implications

We used newspaper obituaries to "nowcast" daily mortality observed in Italy during the first COVID-19 outbreak peak. We find that forecasting models which use obituaries outperform other models based on previously observed mortality, and we report that obituaries account, on average, for 56% of the actual deaths. Although the results should be interpreted cautiously in light of the assumptions and limitations inherent to our approach, their implications may help health authorities –possibly of any geographical entity – in the interpretation of observed surveillance data and provide indications for the magnitude and timing of possible future epidemic waves. More specifically, collecting newspaper obituaries in a centralized database would

Journal Pre-proof

Figure 2: Forecasts

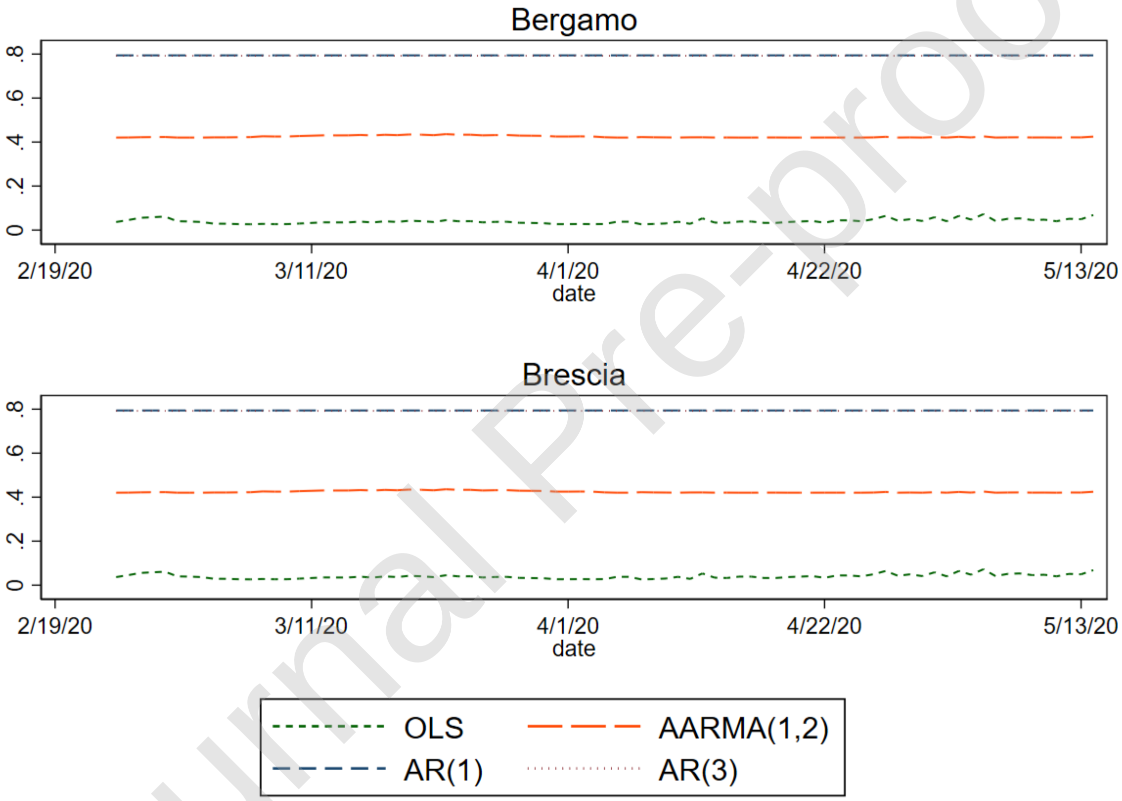


Notes: Daily progression of each forecasting model with the actual *ln*(*Deaths*) (solid green line) on the vertical axis and the corresponding *date* on the horizontal axis. The other lines refer to the forecasted *ln*(*Deaths*).

Journal Pre-proof

Journal Pre-proof

Figure 3: Prediction error



Notes: Daily evolution of the prediction error for each forecasting model. Note that the AR(1) and AR(3) curves overlap.

Journal Pre-proof

Journal Pre-proof

provide a valid alternative to official mortality figures, which could be used by health authorities to assess the actual severity of COVID-19 in almost real-time and, in turn, timely implement effective containment measures.

References

[1] Coronavirus disease (COVID-19) pandemic, author=World-Health-Organization, year=2020, institution=Available online at `https://www.who.int/emergencies/diseases/novel-coronavirus-2019`. Technical report.

[2] Richard J Hatchett, Carter E Mecher, and Marc Lipsitch. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proceedings of the National Academy of Sciences*, 104(18):7582–7587, 2007.

[3] Shihao Yang, Mauricio Santillana, and Samuel C Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, 2015.

[4] Shunqing Xu and Yuanyuan Li. Beware of the second wave of COVID-19. *The Lancet*, 395(10233):1321–1322, 2020.

[5] World-Health-Organization et al. Revealing the toll of COVID-19: a technical package for rapid mortality surveillance and epidemic response. Technical report, 2020.

[6] Paolo Buonanno, Sergio Galletta, and Marcello Puca. Estimating the severity of COVID-19: evidence from the Italian epicenter. *PLoS ONE*, 15(10): e0239569, 2020.

[7] Chirag Modi, Vanessa Boehm, Simone Ferraro, George Stein, and Uros Seljak. How deadly is COVID-19? A rigorous analysis of excess mortality and age-dependent fatality rates in Italy. *medRxiv*, 2020.

[8] Neil Ferguson, Daniel Laydon, and Nedjati Gilani *et al.* Under-reporting and case fatality estimates for emerging epidemics. *BMJ*, 350:h1115, 2015.

[9] Lipsitch Marc *et al.* Potential biases in estimating absolute and relative case-fatality risks during outbreaks. *PLoS neglected tropical diseases*, 9.7: e0003846, 2015.

[10] Andrew Atkeson. How deadly is COVID-19? Understanding the difficulties with estimation of its fatality rate. Technical report, National Bureau of Economic Research, 2020.

Journal Pre-proof

Journal Pre-proof

[11] James H Stock. Data gaps and the policy response to the novel coronavirus. Technical report, National Bureau of Economic Research, 2020.

[12] Guardian staff. Global report: Bolsonaro hides brazil's coronavirus death toll and case totals. *The Guardian. Available at https://www.theguardian.com/world/2020/jun/07/bolsonaro-strips-death-toll-and-case-totals-from-brazils-coronavirus-updates*, 2020.

[13] Jin Wu, Allison McCann, Josh Katz, and Elian Peltier. 161,000 Missing Deaths: Tracking the True Toll of the Coronavirus Outbreak. *The Guardian. Available online at https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html*, 2020.

[14] Carrie Arnold. What we'll need to find the true COVID-19 death toll. *National Geographic. Available online at https://www.nationalgeographic.com/science/2020/05/what-we-need-to-find-true-coronavirus-death-toll/*, 2020.

[15] World Health Organization. WHO Mortality Database. Technical report, World Health Organization, 2020.

[16] Mark Armstrong. Italian mayor claims the true death toll from COVID-19 likely to be much higher . *Euronews. Available online at https://www.euronews.com/2020/03/21/italian-mayor-claims-the-true-death-toll-from-covid-19-likely-to-be-much-higher*, 2020.

[17] Isaia Invernizzi. Coronavirus, the real death toll: 4.500 victims in one month in the province of Bergamo. *L'Eco di Bergamo. Available online at https://www.ecodibergamo.it/stories/bergamo-citta/coronavirus-the-real-death-tool-4500-victims-in-one-month-in-the-province-of_1347414_11/*, 2020.

[18] Marino Gatto, Enrico Bertuzzo, Lorenzo Mari, Stefano Miccoli, Luca Carraro, Renato Casagrandi, and Andrea Rinaldo. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences*, 117(19):10484–10491, 2020.

Journal Pre-proof

[19] Dino Gibertoni, Kadjo Yves Cedric Adja, Davide Golinelli, Chiara Reno, Luca Regazzi, and Maria Pia Fantini. Patterns of COVID-19 related excess mortality in the municipalities of Northern Italy. *medRxiv*, 2020.

[20] Marilyn Booth. The world of obituaries: Gender across cultures and over time. *Biography*, 26(3):453–456, 2003.

[21] Bridget Fowler and Esperança Bielsa. The lives we choose to remember: A quantitative analysis of newspaper obituaries. *The Sociological Review*, 55(2):203–226, 2007.

[22] Marta Bańbura, Domenico Giannone, Michele Modugno, and Lucrezia Reichlin. Now-casting and the real-time data flow. In *Handbook of economic forecasting*, volume 2, pages 195–237. Elsevier, 2013.

[23] H Theil. *Applied Economic Forecasting*. Rand McNally and Company, Chicago, 1966.

[24] Robert Putnam. The prosperous community: Social capital and public life. *The American prospect*, 13(Spring), Vol. 4. Available online: http://www. prospect. org/print/vol/13), 1993.

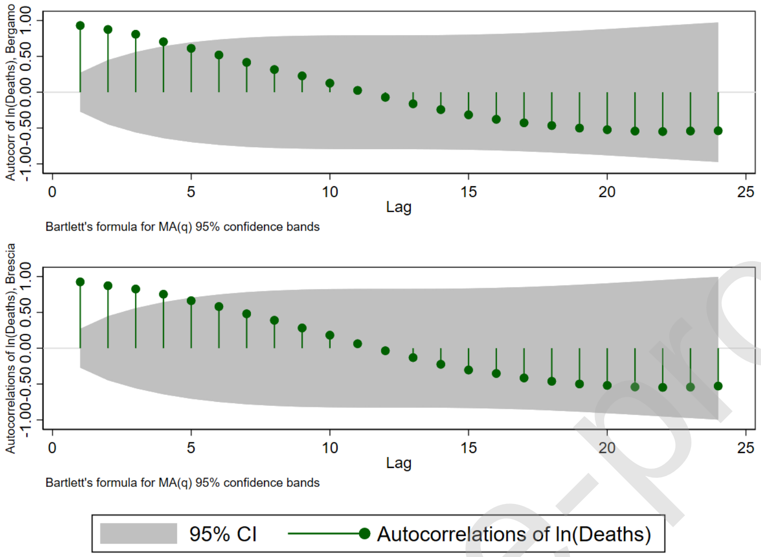
Journal Pre-proof

Journal Pre-proof

6. Appendix

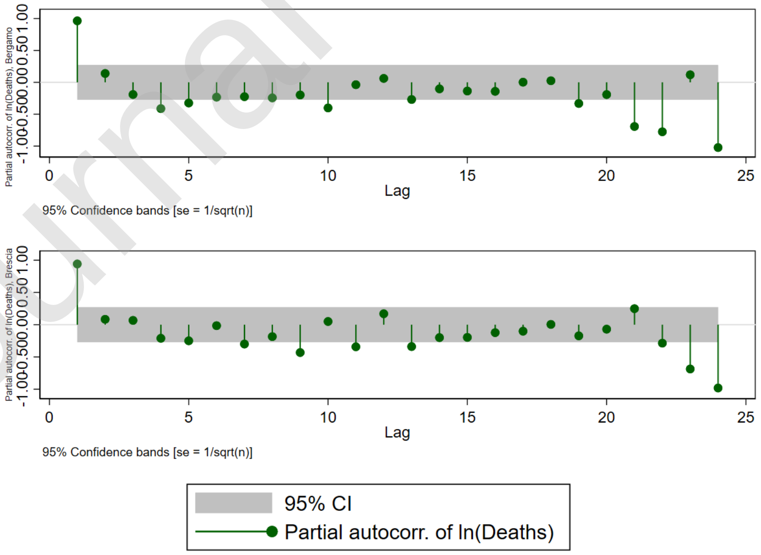
Figures

Figure 4: Autocorrelations



Notes: Autocorrelation of the natural logarithm of daily mortality for Bergamo (top) and Brescia (bottom).

Figure 5: Partial autocorrelations



Notes: Partial autocorrelation of the natural logarithm of daily mortality for Bergamo (top) and Brescia (bottom).

Journal Pre-proof