Pandemic Data Quality Modelling: A Bayesian Approach

Modellazione della qualità dei dati pandemici: un approccio bayesiano

Luisa Ferrari, Giancarlo Manzi, Alessandra Micheletti, Federica Nicolussi and Silvia Salini

Abstract When dealing with pandemics like COVID-19, it is crucial for policymakers to constantly monitor the emergency. Correct data reporting is a hard task during pandemics, and errors affect the overall mortality, resulting in excess deaths in official statistics. In this work, we provide tools for evaluating the quality of pandemic mortality data. We accomplish this through a spatio-temporal Bayesian approach accounting for the bias implicitly contained in the data.

Abstract *Quando si affrontano pandemie come il COVID-19, è fondamentale che si monitori costantemente lo stato della pandemia. Tuttavia, una corretta raccolta dei dati è un compito difficile in questi casi e gli errori influiscono sulla valutazione della mortalità complessiva, traducendosi in un eccesso di mortalità nelle statistiche ufficiali. In questo lavoro, si forniscono strumenti per valutare la qualità dei dati sulla mortalità pandemica attraverso un approccio spazio-temporale bayesiano.*

Key words: Pandemics, Bayesian analysis, variance models, time-space models

Luisa Ferrari
Department of Statistical Sciences ”Paolo Fortunati”, University of Bologna, e-mail: luisa.ferrari5@unibo.it

Giancarlo Manzi
Department of Economics, Management and Quantitative Methods, University of Milan, e-mail: Giancarlo.Manzi@unimi.it

Alessandra Micheletti
Department of Environmental Science and Policy, University of Milan, e-mail: alessandra.micheletti@unimi.it

Federica Nicolussi
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy, e-mail: federica.nicolussi@polimi.it

Silvia Salini
Department of Economics, Management and Quantitative Methods, University of Milan, e-mail: Silvia.Salini@unimi.it

1 Introduction

The COVID-19 pandemic brought the world close to a halt in 2020 and 2021 and killed almost seven million people as of early 2023. In similar contexts, actions of surveillance with limited error is crucial, especially in mortality monitoring [3, 6, 9]. In this paper, we consider the bias between excess mortality and the official Italian COVID-19 data in the first 2020 outbreak for evaluating data quality in a space-time context. To model this bias we use a Bayesian framework where two different quality measures ought to be evaluated: (i) the share in the population dying because of a particular infectious disease without being officially reported, and (ii) the coverage of the epidemic by the health systems, which can be considered an adequate indicator of their quality and a proxy for the efficacy of the crisis response.

2 Data

In order to evaluate the data quality on COVID-19 mortality we considered two data sources: (i) official data series on pandemic mortality (ii) national or supranational statistical institute data on population mortality. Data (i) are weekly provincial (EU NUTS-3 level) COVID-19 deaths from February 24th to May 11th, 2020. While data about new cases were regularly published, the number of weekly COVID-19 new deaths was not officially available at this level. Nevertheless, we reconstructed the time series of COVID-19 at a NUTS-3 level indirectly, using other official sources like regional authorities' daily bulletins on provincial deaths and other information sources [4]. Bulletins were in general in a ”pdf” format so we were able to scrape data from these documents and to retrieve the data of interest for the majority of the Italian provinces. Regarding data (ii), each year the Italian National Statistical Istitute (ISTAT) provides a weekly record of deaths reported in each municipality in Italy. Here we use a 5-year window consisting of the period 2015-2019 to represent the stable mortality level. The excess mortality is then found by subtracting this stable level from the COVID-19 2020 deaths data, in each province and week of the year.

3 Proposed metrics

The aim of the metrics to be defined is to provide an estimate for the under-reporting mortality bias that the official data has been subject to. Two different metrics with different interpretations for the policy-makers are proposed.

Let D_{ij} be the officially reported total number of deaths in province i and week j which exceeds the average of the previous 5 years, assuming that they can all be imputed to the COVID-19 emergency. Let $\hat{D}_{ij}$ be an estimate for D_{ij}. Let Y_{ij} be the officially reported number of COVID-19-related deaths in province i and week j.

Finally, let POP_i be the average population in province i along the considered period. The additive bias is built as the difference between the *actual mortality* D_{ij}/POP_i and the *official mortality* Y_{ij}/POP_i. This bias is defined as "additive" because it must be added to the official mortality to get the unbiased value:

$$b^A_{ij} = 1000 \cdot \frac{\hat{D}_{ij} - Y_{ij}}{POP_i}$$

In terms of interpretation, $b^{(A)}$ defines the share in the population that died because of COVID-19 without being officially reported, so large values represent a negative scenario. Its trend over time and space (but not its magnitude) is a rough proxy for the part of the pandemic that was concealed and undetected by the public administration.

The ratio between Y_{ij} and D_{ij} assesses the probability of a COVID-19-related death being officially reported. In order to transform it into a bias metric, the complementary probability of not being reported is considered instead and called $b^{(M)}_{ij}$. This is equivalent to the additive bias divided by the excess mortality rate.

$$b^M_{ij} = 1000(1 - \frac{Y_{ij}}{\hat{D}_{ij}})$$

A large bias indicates a bad situation. Regarding its meaning, b^M measures the promptness of the health system to react to the pandemic, thus it is an adequate indicator of its quality and a proxy for the efficacy of the crisis response.

4 Model

Our spatio-temporal model for the two above metrics model resembles the one by Franco-Villoria et al. [5] with the temporal component being a Gaussian random walk model of order 1. For the spatial component we adopted an ICAR model [1] [2], but instead of considering the traditional adjacency matrix $\mathbf{M}$ with only non-negative entries and a null diagonal, and a corresponding diagonal matrix $\mathbf{D}$ where $d_{p,q} = \sum_q m_{p,q}$ to take into account the geographical boundaries, we considered an adjacency matrix with smartphone location data, which actually estimates the average commuting of individuals between two provinces, no matter their actual geographical location:

$$\mathbf{u} \sim N_I\left(\mathbf{0}; \Sigma_u = (\mathbf{I_J} - \mathbf{D^{-1}M})^{-1}\mathbf{D}^{-1}\right).$$

The covariance matrix of the interaction term is defined as the Kronecker product of the covariance matrix of the two main effects, following the work of Knorr-Held [7]. Finally, as for the prior specification we reparametrize (as in Franco-Villoria et al.

[5]) the original variances $\sigma_u^2, \sigma_v^2, \sigma_w^2$ into a total residual variance V, the proportion ψ of this V given by the interaction term, and the proportion ϕ of main effects variance imputable to the spatial effect. The prior specification is then chosen on this new set of parameters. Specifically, the INLA default prior on variance parameters on σ_ε^2, a Uniform on ϕ, and a Penalized Complexity (PC) prior [8] on ψ with base model $\phi_0 = 0$, and a PC prior on V with base model $V_0 = 0$.

5 Results

Figure 1 shows the posterior mean of the spatial random effects over the provinces of Italy. For b_A on the left, provinces in the North of Italy experienced a larger share of underreported deaths with respect to the overall population. However, the spatial distribution completely changes for the b_M metric, as most of the Northern provinces show small values, while the highest effects are found in the Southern and North-Eastern provinces. These figures display how the two indices measure very different quantities, with b_M being much more consistent with the literature on the spatial distribution of health system quality indicators in Italy. With respect to the temporal pattern, the two metrics also show differences. The average temporal trend for b_A, shown in Figure 2, green curve, starts with an increasing part, up to the sixth week in the considered period, followed by a steady fall in the remaining weeks. This is a reasonable result, as it is expected that the indicator b_A performed the worst at the peak of the ”official” epidemic evolution, plus a delay due to the fact that deaths are considered instead of cases. Hence, this confirms the assumption that b_A is related to the level of stress of the health system, rather than to the quality of its response to a certain amount of stress.

The results for b_M are again completely different as the posterior means, shown in Figure 2, red curve, shows a steady decreasing trend. Finally, with a DTW clustering on b^M it was possible to detect four groups of provinces according to their performance in facing the emergency. Figure 3 shows the 4 different groups and their centroids, ordered by best (on the left) to worst (on the right) overall performance.

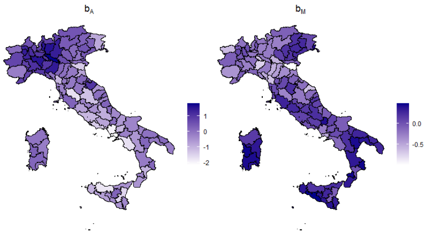

Fig. 1 Posterior mean of the spatial random effects on b_A and b_M

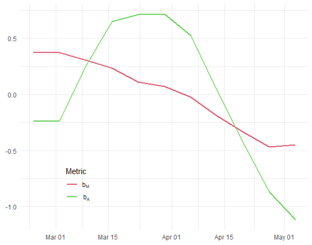

Fig. 2 Posterior mean of the temporal random effect on b_A and b_M

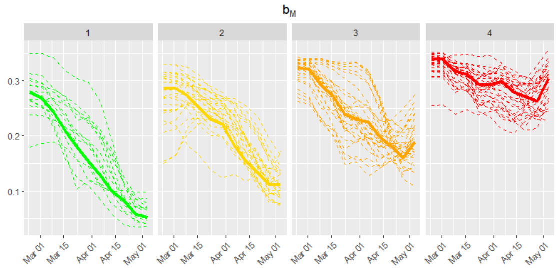

Fig. 3 Posterior mean of the fitted values divided in 4 clusters with corresponding centroids in the provinces of Aosta, Rimini, Catanzaro, and Cosenza

References

1. Besag J.: Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Stat. Soc. B. 36, 192–225 (1974)
2. Besag, J., York, J., Mollié, A.: Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math. 43(1), 1–20 (1991)
3. Colombo, R.M., Garavello, M., Marcellini, F., Rossi, E.: An age and space structured SIR model describing the Covid-19 pandemic. J. Math. Ind. (2020) doi: 10.1186/s13362-020-00090-4
4. Ferrari, L., Gerardi, G., Manzi, G., Micheletti, A., Nicolussi, F., Biganzoli, E., Salini, S.: Modeling Provincial COVID-19 Epidemic Data Using an Adjusted Time-Dependent SIRD Model. Int. J. Env. Res. Pub. He. (2021) doi: 10.3390/ijerph18126563
5. Franco-Villoria, M., Ventrucci, M., Rue, H.: Variance partitioning in spatio-temporal disease mapping models. Stat. Methods Med. Res. 31(8), 1566–1578 (2022)
6. Kantner, M., Koprucki, T.: Beyond just "flattening the curve": Optimal control of epidemics with purely non-pharmaceutical interventions. J. Math. Ind. (2020) doi: 10.1186/s13362-020-00091-3
7. Knorr-Held, L.: Bayesian modelling of inseparable space-time variation in disease risk. Stat. Med. 19(17-18), 2555–2567 (2000)
8. Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H.: Penalising model component complexity: A principled, practical approach to constructing priors. Stat. Sci. 32, 1–28 (2017)
9. Wu, J., Tang, B., Bragazzi, N.L., Nah, K., McCarthy, Z.: Quantifying the role of social distancing, personal protection and case detection in mitigating COVID-19 outbreak in Ontario, Canada. J.Math.Ind. (2020) doi: 10.1186/s13362-020-00083-3