A Heavily Trained Time-Dependent SIRD Model for Local Covid-19 Data in Italy

Luisa Ferrari[1], *Giuseppe Gerardi*[2], *Giancarlo Manzi*[2], *Alessandra Micheletti*[3], *Federica Nicolussi*[2], *Elia Biganzoli*[4], *and Silvia Salini*[2]

1 UCL, London, UK – luisa.ferrari.20@ucl.ac.uk
2 DEMM, University of Milan, Milan, Italy – giuseppe.gerardi@unimi.it, giancarlo.manzi@unimi.it, federica.nicolussi@unimi.it, silvia.salini@unimi.it
3 DESP, University of Milan, Milan, Italy – alessandra.micheletti@unimi.it
4 DCSCH, University of Milan, Milan, Italy – elia.biganzoli@unimi.it

We present a time-dependent SIRD model for the spread of COVID-19 infection at a provincial (i.e. EU NUTS-3) level in Italy, using official data from the Italian Ministry of Health, integrated with data extracted from daily official press conferences of regional authorities and from local newspaper websites. This integration concerns COVID-19 death data which are not available at NUTS-3 level from open official data channels. The model is trained for improved forecasting performance with similarity techniques putting together data from time series most similar to that for which the forecast is performed.

1 Introduction

The outbreak of the Covid-19 epidemics in early 2020 has caused an unprecedented effort of the scientific community to produce models that could monitor and predict the evolution of the epidemics in a reliable way, also to promptly advice governments in order to take actions which could mitigate the burden on hospitals in treating the affected patients and reducing the mortality rate of the infection.

The first reported Italian Covid-19 case dates back to February 20th, 2020,[1] in the city of Codogno, southern Lombardy, and the epidemics spread particularly in Italian northern regions, that is, those most commercially connected with China, where the epidemics had its origin. It was immediately clear that the initial Covid-19 outbreak in Italy was not homogeneously spreading within regions, but there were many differences from province to province in the same region. Therefore, we decided to model epidemic data at a provincial rather than at a regional level, contrary to the majority of the analyses about the virus spread in Italy published recently.

DOI: 10.54103/milanoup.73.41

2 Methods and Results

We consider here the classical approach of a model consisting of 4 compartments: susceptible (S), infected (I), recovered (R) and deaths (D), therefore adopting a classical epidemiological approach to model epidemic data. However, in adopting this approach we faced a twofold problem: (i) data at provincial level were available only for susceptible and infected people, not for recovered and dead people; (ii) parameters of epidemics, particularly the Covid-19 epidemic, evolve in time depending implicitly on external factors like the limitation of the mobility of the citizens, the measures of protection of the healthcare personnel and of the workers who kept on doing jobs which were considered essential services to the community, and on the number of swab tests performed locally to detect the infected subjects, in order to put them in strict quarantine as soon as possible.

To solve (i) we extracted the number of daily Covid-19 deaths from daily official press conferences of regional authorities and from local newspaper websites; also, we computed the number of the cumulative recovered individuals at a provincial level using the recovery rate at the regional level. The reason for estimating this number proportionally to that of the region is that patient treatment for the illness due to Covid-19 could be considered sufficiently homogeneous across the provinces (with almost the same recovery rate across provinces within the region), contrary to the number of deaths which depends more on the local level of the infection.

The model for forecasting all the parameters involved in the SIRD model, i.e. β, the transmission rate γ_R and γ_D, the mortality rate. From these parameters we also predicted the values for the reproduction number, i.e. the expected number of people potentially infected by an infected individual. For each of these parameters we form an autoregressive model combined with a cost function to be minimized, expressed in terms of a ridge regression. The ridge penalizing parameter λ is obtained via cross-validation. The optimal maximum number for J concerning β is used also in the other two autoregressive models concerning γ_R and γ_D, as to have three homogeneous time series.

However, the model features described above might struggle in producing reliable estimates in contexts where the number of cases is very low and there is considerable fluctuation or inconsistency in the data, as it happened in some provinces where the outbreak was not so intense, whilst it seems to give more robust results when data are aggregated at a higher level, as in the entire country's time series. So, we decided to implement a boosting of the model training in the following way:

- (i) *Regional-based training*. For each province, the corresponding region is selected. In order to choose the regions whose situation most resembles the one in the selected region, a time series correlation measure between all of the regions was computed: this correlation was based on the new cases' daily time series divided by the regions' population and was time-weighted, so that the most recent days have a much bigger impact than the days at the beginning of the epidemic. Weights were then normalized. The regions showing a high correlation with the region containing the province of interest were selected and their data were aggregated to compose the training set for the province of interest. The threshold used to select the regions was the median value of the correlation coefficients.

- (i) *Provincial-based training*. The training set was made up by the provinces that are highly correlated with the province of interest and the threshold chosen is again the median value of the correlation coefficient, which is computed exactly the same way as in the regional training approach. The provinces that have a correlation higher than the median value are selected and aggregated with the province of interest: this composed data set is used to train the model.

Model coefficients resulting from this training process were then applied to the local data in order to make predictions for each province: given the predictions about the future values of the parameters β, γ_R and γ_D, the future values of the variables S, I, R, and D are computed using the customary discrete time difference equations of the classic SIRD model.[2]

In Figures 1 and 2 an example on applying this model to the Piacenza province is shown. Piacenza is one of the provinces in northern Italy with the highest cumulative infection rate, since its main hospital is located very close to Codogno, the town with the first cluster of observed Covid-19 cases in Italy. In this case, the model is trained using the regional-based training. The vertical dashed lines represent the dates in which the Italian government's containment measures were implemented.

3 Discussion

The COVID-19 outbreak has made it necessary strong containment measures. The spread of this disease in Italy involved heterogeneously the whole Italian territory, see.[3] The full lock down of industrial and other productive activities (March, 22th) was legitimized by few municipalities with ever increasing number of cases (and deaths), by the inability to stop the contagion and by the emergence of new outbreaks. It is also true that the lockdown stopped municipalities barely touched from the COVID-19.

The analysis presented in this paper can be applied at the NUTS-3 region level in Italy in order to predict the future development of the epidemic in the specific provincial context. This choice is aimed at tackling the heterogeneity of the

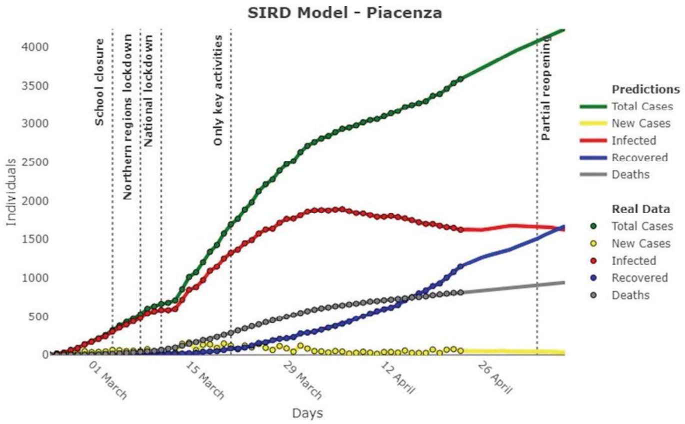

Figure 1: Heavily trained SIRD model for Piacenza province - April 23rd, 2020. The training set includes all of the data up to April 23rd; the model is then used to make predictions iteratively for the following 15 days. In this example, the regional aggregation training approach is implemented, and the resulting optimal number of lags is equal to 7.

COVID-19 spread on the territory. Furthermore, as expected in an emergency situation, some information may be temporarily incomplete.

The main issue found during the building of our model was the lack of detailed and consistent data about the epidemic at the provincial level, although

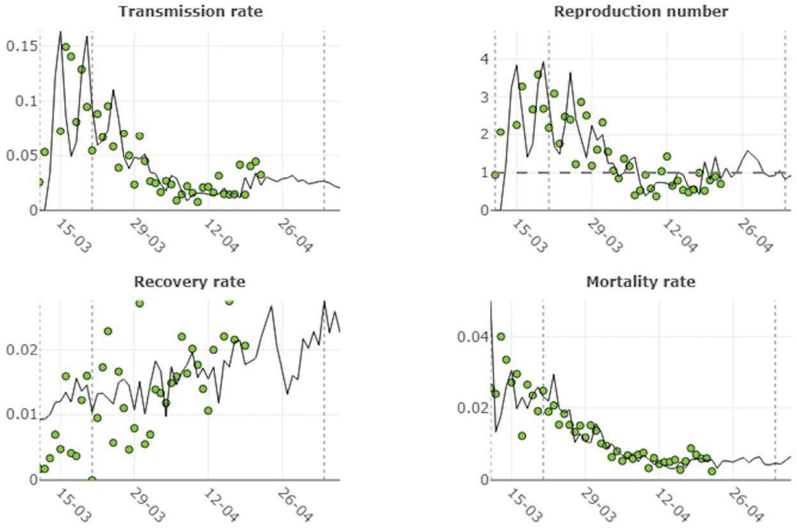

Figure 2: Heavily trained SIRD parameters predictions for Piacenza province - April 23rd, 2020. The green dots represent the daily values of the parameters up to April 23rd, while the solid black line represents the fitted model and the predictions for the following days.

the procedure adopted by the central government and the regional authorities was standardized in order to get similar aggregated data across regions. A full trustfulness of these aggregated data is put in doubt by the fact that some of them (deaths, recovered and number of tests) are not publicly available at a provincial level in the official daily releases by the Italian Civil Protection Agency and therefore, as in our case, they must be estimated or found in some ways. If more variables had been available, the model could have been extended to include other compartments, such as the hospitalized cases or the number of tested individuals or, again, the possibility to consider a percentage of recovered people as susceptible, or the asymptomatic cases, or the under- reporting of deaths. Indeed, we choose to implement the simplest (with the lowest number of parameters) model, which allows us to overlook strong and unlikely assumptions.

A dashboard built on this model regarding the spread of the Covid-19 virus in Italy's provinces has been implemented and can be found at http://demm.ceeds.unimi.it/covidModelling/.

References

[1] Guzzetta G., Poletti P., Ajelli M., Trentini, F., Marziano, V., Cereda, D., Tirani, M., Diurno, G., Bodina, A., Barone, A., Crottogini, L. et al. (2020). Potential short-term outcome of an uncontrolled COVID-19 epidemic in Lombardy, Italy, February to March 2020. *Eurosurveillance 2020*, 25(12): 1–4.

[2] Kermack, W.O., McKendrick, A.G. (1927). A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A*, 115 (772): 700–721.

[3] Bertuzzo E., Mari L., Pasetto D., Miccoli, S., Casagrandi, R., Gatto, M., Rinaldo, A. (2020). The geography of COVID-19 spread in Italy and implications for the relaxation of confinement measures. *Nature Communications*, 11: 1-11.