Noname manuscript No.
(will be inserted by the editor)

A simulation is worth a thousand worlds

Importance sampling Monte Carlo applied to COVID-19 contingency

Andrea Rapuzzi · Tomaso Vairo

Received: date / Accepted: date

Abstract COVID-19 outbreak has become a global pandemic that affected more than 200 countries worldwide. Predicting the behavior of this outbreak has a crucial role in organizing preventive and protective actions, and in improving the decision making process.
The aim of predicting the number of people who contract the virus has so far been pursued with regression models (exponential, logistics, ...) [5], [6], [7], but regressions can integrate the variable context only a posteriori. The regression models are all dependent on their own history, thus, they can not display anything which did not happen before.
The pandemic infection of COVID-19 presents a transmission behaviour that is widely changing over time. This is due to the growth of the efficiency in the detection of infected, for the changes in social distancing measures and for the widespread use of individual protection devices.
The approach presented in this paper, starting from the definition of simplified risk assessment framework, aims at designing a probabilistic model for the virus transmission and detection, keeping into account this context changes, binding the correct set of variables to them, and at inferring the distribution for the underlying stochastic variables. This is a key to unlock innovative and valuable insights from the current events. The model has been built in Gen [10], a probabilistic programming system, built at MIT and embedded in Julia.

Keywords COVID-19 · Julia Language · Predictive model · Inference · Gen

Andrea Rapuzzi
A-SIGN S.r.l - via XXV Aprile 10/3a - 16121 Genoa, Italy
E-mail: andrea.rapuzzi@a-sign.it

Tomaso Vairo
DICCA, Civil, Chemical and Environmental Engineering Dept. Genoa University, via Opera Pia 15, 16145,Genoa, Italy

1 Introduction

In December 2019, Wuhan, a city in China, became the center of the COVID-19 outbreak. A few months later, the world health organization, declared a global pandemic contingency. From the beginning of the contingency, more than 40 Mln confirmed cases and 1 Mln deaths worldwide have been officially reported [1], [8].
Covid-19 has been considered as the most significant planetary crisis since the World War-II [4]. COVID-19 is a highly contagious disease with moderate fatality rate. It transmitted among humans via touching contaminated bodies with viral particles or contacting infected patients [2]. The incubation period of the disease ranges from 2 to 21 days, and one of the main issues is that it may transmit from infected people which doesn't have any symptoms. This fact poses a crucial attention to the detection efficiency. Severe complications may occur in elderly persons with other debilitating diseases [4].
Accumulating evidence suggests that various policies on the reduction of social interactions, and on the massive use of facial masks, slowed down the growth of COVID-19 infections. All of those rules affect virus spread by changing people's behavior (e.g., stay-at-home order) [3].
In the present paper, we applied the risk assessment techniques (bow-tie analysis), for describing, at high level, the COVID-19 transmission risk, and for identifying barriers and escalation factors, then a probabilistic model is developed, in order to predict the system behaviour, and quantitatively assess the impact of various policies from the beginning of the outbreak. The predictive model relies on a generative, contextual and nonparametric approach. **Generative**, because it is designed as a model to 'generate' (or mimic) the observed behaviour (and not the other way round). **Contextual** because it tries to correctly collocate phenomena which have a very different context binding. **Nonparametric** because, even if we have introduced a few high-level model parameters, the model can dynamically and unboundedly infer its complexity and the number of parameters it needs to fit.

The results have shown that the model is capable of inferring posterior distributions, as shown by its generative capabilities. This allows for more ambitious and valuable goals to be achieved in the coming second study phase.

2 The risk model

2.1 Bow-Tie analysis

A 'Bow-Tie' is a diagram that visualizes the risk you are dealing with in just one, easy to understand the picture. The diagram is shaped like a bow-tie, creating a clear differentiation between proactive and reactive risk management. It gives an overview of multiple plausible scenarios, in a single picture and provides a simple, visual explanation of a risk that would be much more difficult to explain otherwise.

There are two things that the Bow-Tie does. First, the Bow-Tie analyses chains of events, or possible accident scenarios. The way it does can be decomposed in 2 different methods. The first method is the fault tree (FT) which covers the left side of the Bow-Tie, the second is the event tree (ET) which can be seen on the right side of the Bow-Tie.
Fault Tree Analysis (FTA) and Event Tree Analysis (ETA) are the quantitative part of the risk assessment process.
Fig. 1 depict the developed Bow-Tie.
The two inference models, built with FT and ET techniques, deriving from

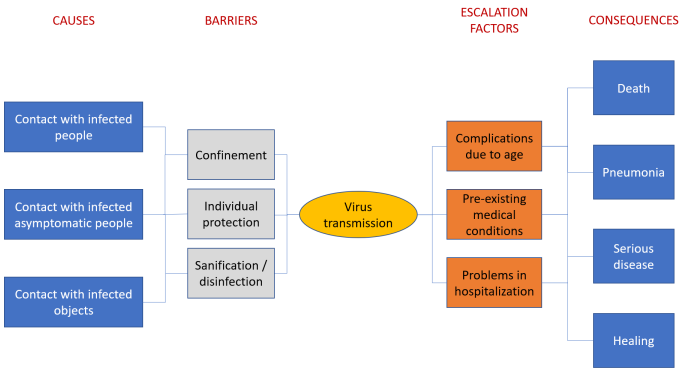

Fig. 1 Bow-Tie model for COVID-19 transmission

the Bow-Tie model structure, are described below.

Fault Tree Analysis .
Any sufficiently complex system is subject to failure as a result of one or more subsystems failing. The likelihood of failure, however, can often be reduced through improved system design. Fault tree analysis maps the relationship between faults, subsystems, and redundant safety design elements by creating a logic diagram of the overall system.
The barriers on the FT side of the Bow-Tie are:

- Confinement. Limiting the social contacts;
- Individual protection. Wearing suitable protection systems;
- Sanification and disinfection of shared areas.

The FTA is depicted in Fig. 2.

Event Tree Analysis .
The overall goal of event tree analysis is to determine the probability of possible negative outcomes that can cause harm and result from the chosen initiating event. The escalation factors on the ET side of the Bow-Tie are:

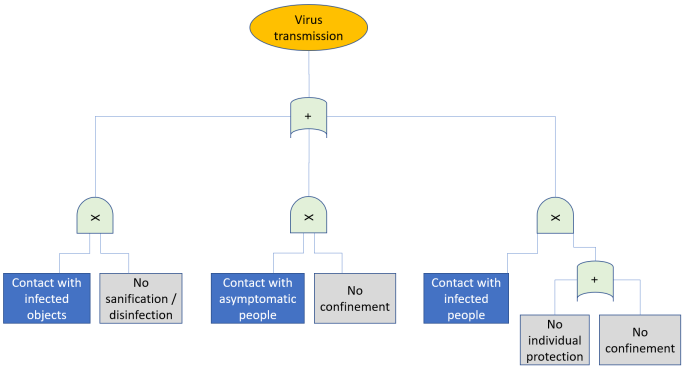

Fig. 2 FT for COVID-19 transmission

– Complication related to the age of the infected person;
– pre-existing medical condition;
– Problems in hospitalization.

The ETA is depicted in Fig. 3.

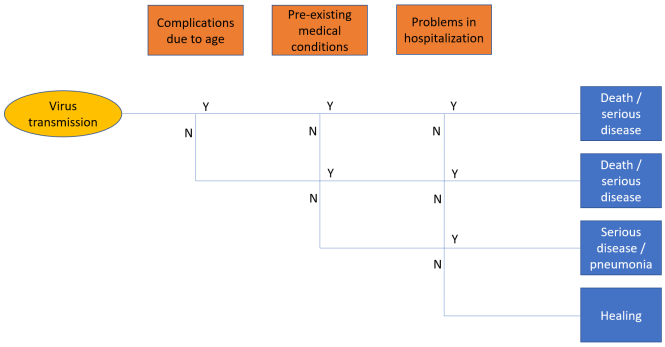

Fig. 3 ET for COVID-19 transmission

2.2 The transmission model

The possible sources of virus transmission are represented by contact with infected people (or infected objects, but for objects, the criterion of disinfection is a sufficient barrier). The contact with infected people is a context strictly dependent factor. In fact there are 2 circumstances that define the transmission context:

- infected people can show the symptoms of the disease (and therefore, from the transmission point of view, they are no longer so relevant, because in observation, at home or in the hospital), or they can be asymptomatic, therefore in the incubation period. The incubation period is very large (ranges from 2 to 21 days, approximately, with an average value of 5.2 days, and an interval between 4 and7 days with 95% confidence), therefore the transmission possibilities are very different.
- The government intervened with various gradually increasing restrictive measures: the closure of schools on February 29, the definition of the first red zone on March 7, the closure of the nation, as the only red zone, on March 11, and the complete lock-down on March 21 [9]. THe lock-down lasted 69 days, and subsequently the restrictions decreased, until October 18, where partial restrictions were reintroduced, to cope with a new increase in infections, with the substantial difference linked to an increased detection and tracking capability of the infected. Therefore the chances of coming into contact with infected people are very different depending on the changing conditions.

The variables at play are:

- not directly observable: we can observe the number of detected infected people (in the following just 'detected people'), but we don't know the number of infected ones
- correlated to the observed variables, but with a variable and unknown time shift: it is difficult, once detected an infected person, to know the time at which the infection occurred (in any case, this data is not registered and publicly available)
- highly dependant on a changing context: the cultural, social and normative contexts, to name a few, influence the underlying statistics. Some contexts are slowly changing, or not at all in the timeframe of an experiment, like the cultural one. Others present abrupt changes (for example, in the case of a lockdown).
- noisy: even the observed variables present a high level of noise (e.g. from their registration process)

3 The predictive model

We have designed our model as a generative one, that is a stochastic process aimed at generating the same variables we have observed and based on common sense to model a simplified version of the natural phenomena.

3.1 Population

Our model does not account for total population count, like in an infinite population or at an early pandemic phase. Our model accounts for three different

and mutually exclusive states for a person: healthy, infected and detected. Since we model daily people cohorts, not single people, and our model does not include model counters for healthy people.

In our abstraction, the region of interest is considered closed (there are no people going out of it or coming into it) with the exception of the initialization phase (see below).

3.2 Context

We introduce the concept of **context**, to model in a simplified way, and with a single abstraction, the sum of all the relevant contexts that can influence the observed phenomena (like cultural, social and normative context). A single context is in place at any given simulation time. Changes in context, or context switches, can occur during a single simulation altering the behaviour of the model. We don't make apriori assumption about how may context switches occur (or when) in the course of our simulations, excluding an apriori distribution for the probability of a context switch occurring at any given day.

The variables we have modelled can be classified on the base of their contextual binding as:

- steady: a single distribution of the variable is generated for a single simulation run (they don't change inside a simulation run)
- contextual: a different distribution is generated for any context of a single simulation run (they change for each context)
- daily: a different distribution is generated for any day of a single simulation run (they change for each day)

3.3 Influence windows

In a pandemic model like ours, variables at a given time are related to other variables from the past. Or, to provide a more generative perspective, a variable at a given time impacts the generation of another variable in the future. For example, the number of people that have been infected today, influences, directly or indirectly, the number of people that will be infected in the coming days. But how to model a direct correlation? To account for this, we have designed weighted influence windows. They are a mechanism to weight the impact to a variable coming from a past variable. Let's say that in a given, unweighted, model, the number of infected people at day n (in) is sampled from the following distribution:

$$i_{n+1} \sim Poisson(f \cdot \sum_{d=1}^{n} i_d)$$

In this case (see Fig. 4), the value at day n is affected by the sum of all the values from the previous days.

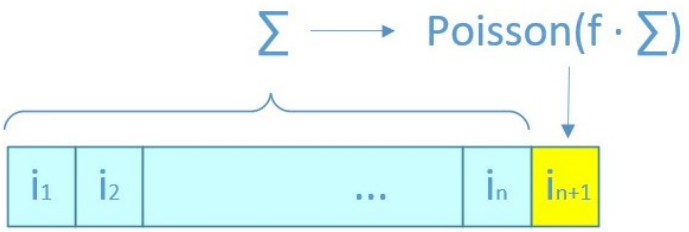

Fig. 4 Unweighted approach

With weighted influence windows (see Fig. 5), we allow the model to account differently for the values coming from the previous days. The weights of influence are another variable of the model (a steady one, as we will see).

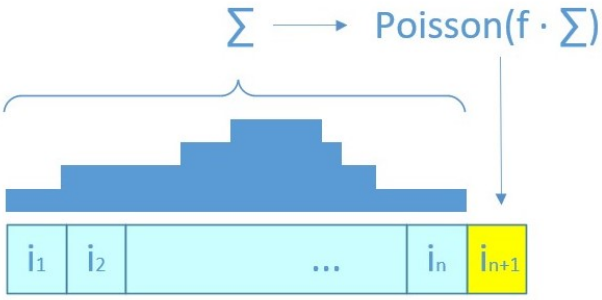

Fig. 5 Weighted windows approach

3.4 Model variables

The main model variables are discussed in the following sections, grouped by their contextual binding: steady, contextual and daily.

3.4.1 Steady variables

Influence windows weights for transmission: a set of influence windows weights at the model level (different for any simulation runs) to account for the capability of people infected in the past days to infect people on a present day. This variables set models a characteristic of the virus itself that it is, in statistical terms, constant in time and across different pandemic scenarios. It is a sort of virus signature in its ability to transmit from one person to another.

Influence windows weights for detection: a set of influence windows weights at the model level (different for any simulation runs) to account for the likeliness of detecting an infected person. This variables set accounts for a miscellaneous of different factors; some from the virus itself (like its likeliness of manifesting symptoms driving to medical tests), some from the social context (like the ability of the health system to proactively detect infected people). These latter factors are indeed less steady and more contextual, and a simplification has been made, by considering it steady, to build a simpler model.

3.4.2 Contextual variables

Transmission factor: a contextual factor to model the ability of the virus to spread under a certain context (e.g. lockdown, social unawareness, etc.).

Detection factor: a contextual factor to model the likeliness of an infected person detection under a certain context.

3.4.3 Daily variables

Context switch: a daily variable indicating a change in context.

Infected count: a daily variable to model the number of people that have been infected on a given day. It is the daily amount of newly infected people, not the cumulative figure. Once inferred for a given day, on a single simulation run, it remains constant.

Detected count: a daily variable to model the number of people that have been detected on a given day. It is the daily amount of newly detected people, not the cumulative figure. Once inferred for a given day, on a single simulation run, it remains constant. These are also the model **observed variables**.

Exposed count: a daily variable to model the number of people that have been infected on a given day and have not yet been detected. It is continuously updated in the course of the simulation rollout.

Other ancillary variables: for stability and convergence reasons, contextual variables are not directly modelled at the context level, but at the daily level and then averaged for each context.

3.5 Transmission and detection model

Given a transmission factor f_t for a given context, the transmission influence function $TI()$, the count of exposed people at days $n-1$, $n-2$, .., $n-k$ (E_{n-1}, E_{n-2}, .., E_{n-k} where k is the extension of the transmission influence

window), the number of people infected at day n is sampled from the following distribution:

$$I_n \sim Poisson(f_t \cdot TI(E_{n-1}, E_{n-2}, .., E_{n-k}))$$

Given a detection factor f_d for a given context, the detection influence function $DI()$, the count of exposed people at days $n-1$, $n-2$, .., $n-k$ (E_{n-1}, E_{n-2}, .., E_{n-k} where k is the extension of the detection influence window), the number of people detected at day n is sampled from the following distribution:

$$D_n \sim Binomial(f_d \cdot DI(E_{n-1}, E_{n-2}, .., E_{n-k}))$$

Once the number of detected people has been inferred for a given day, the model assigns the contribution of any past day to the detection (answering the question: when have, the people detected today, been infected?) inverting the logic of the detection influence function.

3.6 Initialization

A lot of attention has been dedicated to the correct initialization of the model state; that is to the value to assign to the infected population on the days preceding the first observation. Since the model has a recursive nature (i.e. the data inferred for a given date are one of the inputs for the inference at later days), having initial populations that are distributed like they have been generated by the model itself, is key to have a good model convergence. The solution we choose is to use the model itself to generate initial data. For the initialization phase, we accounted for the possibility of infected people coming daily into our region of interest from outside (otherwise no contagion is possible) in the for of:

$$\sim Poisson(\lambda_{in})$$

where λ_{in} is the daily average number of people entering in the region.

3.7 Inference and generation

Once the model is in place, in the form of a stochastic generative function, Gen performs posterior inference with importance sampling Monte Carlo.

It runs and evaluates several thousand (millions in our experiments) simulation runs. Given a generated trace, the model also allows the generation of new artificial scenarios using the inferred (posterior) distributions in a sort of what-if or possible worlds.

4 Results and Discussion

The experimental results are based on the dataset made publicly available from the Italian Presidenza del Consiglio dei Ministri - Dipartimento della Protezione Civile. To provide for a region both homogeneous and relevant in size, the data from the Lombardia region, have been used (data points between the 24th of February, 2020 and the 25th of October 2020). In order to perform an evaluation of the capability of the model to find posterior distributions that are coherent with the observed data, we have used the distribution learned in the traces and used the model as a generative function. Figure 6 shows 5 simulations (coloured lines) generated from a single trace, in comparison with the observed data (red dots).

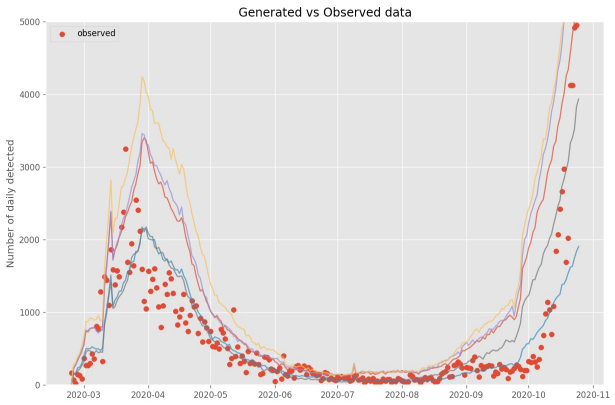

Fig. 6 Generated data

Figure 7 shows over 2.000 simulations (footprinted in the grey cloud) generated by 50 traces, in comparison with the observed data (red dots).

4.1 Coming activities

Even if encouraging, these results, that clearly show the fitness of the model generative capabilities, are just the beginning of our study. As of this writing, we are evolving the model (and the experiments) to allow it to produce:

- stable and coherent influence windows for transmission; this will allow us to fully characterize the capability of the virus to transmit at a given time distance from the infection.
- stable and coherent context switches for a given scenario; this will allow to signal a change in context occurring days before a change in observable data trends. This does not mean that the model will predict future

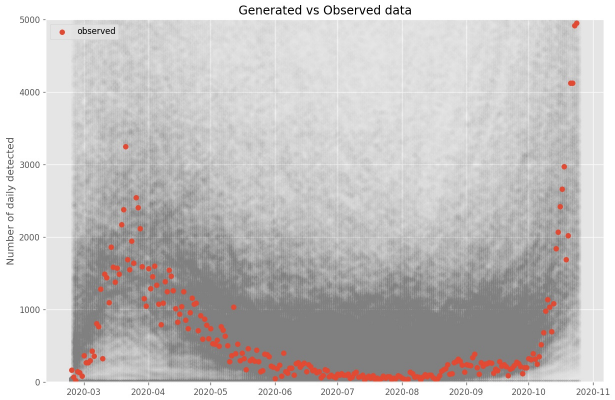

Fig. 7 Generated data cloud

context changes but that it will assign the correct timing information to a context change. In a scenario, like the one introduced by the current pandemic, where the impact of a current decision is delayed (e.g. a couple of weeks) and noisy, this model could provide a valuable credit assignment mechanism.

5 Conclusion

In this paper, we have proposed a probabilistic model for COVID-19 transmission and detection. The model different contextual binding levels (steady, contextual, daily) and its ability to freely and unboundedly allocating context switches make it capable of simulating a scenario as challenging as the one presented by the recent pandemic. The current work has demonstrated the fitness of the model generative capabilities and the coming phase in the study will concentrate on some important and innovative inference features.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. European Centre for Disease Prevention and Control, COVID-19 situation update worldwide, as of 19 October 2020, EU (2020, oct. 19th)
2. Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M., Liu, M., Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries, Science of The Total Environment 729, 139051 (2020)

3. Chernozhukov, V., Kasahara, H., Schrimpf, P., Causal impact of masks, policies, behavior on early covid-19 pandemic in the U.S., Journal of Econometrics, (2020)
4. F. Wu, S. Zhao, B. Yu, Y. Chen, W. Wang, Z. Song, et al., A new coronavirus associated with human respiratory disease in China, Nature., 579, 265-269 (2020)
5. Lamiaa A. Amar, Ashraf A. Taha, Marwa Y. Mohamed, Prediction of the final size for COVID-19 epidemic using machine learning: A case study of Egypt, Infectious Disease Modelling 5, 622-634 (2020)
6. Chaurasia, V., Pal, S. COVID-19 Pandemic: ARIMA and Regression Model-Based Worldwide Death Cases Predictions. SN COMPUT. SCI. 1, 288 (2020)
7. Almeshal, A.M.; Almazrouee, A.I.; Alenizi, M.R.; Alhajeri, S.N., Forecasting the Spread of COVID-19 in Kuwait Using Compartmental and Logistic Regression Models. Appl. Sci., 10, 3402 (2020)
8. Worldometer, Coronavirus cases worldwide (2020, Oct. 19th)
9. Italian Civil Protection Dept. Covid-19, (2020, Oct. 19th)
10. Cusumano-Towner, Marco F. and Saad, Feras A. and Lew, Alexander K. and Mansinghka, Vikash K., Gen: A General-Purpose Probabilistic Programming System with Programmable Inference. Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation. 2019.