medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Modeling the Dynamics of the COVID-19 Population in Australia: A Probabilistic Analysis

Ali Eshragh* Saed Alizamir† Peter Howley‡ Elizabeth Stojanovski‡

Highlights

• This work applies a novel and effective approach using a partially-observable stochastic process to study the dynamics of the COVID-19 population in Australia over the 1 March–22 May 2020 period.

• The key contributions of this work include (but are not limited to):

(i) identifying two structural break points in the numbers of new cases coinciding with where the dynamics of the COVID-19 population are altered: the *first*, a major break point, on 27 March 2020, is one week after implementing the "lockdown restrictions", and the *second* minor point on 18 April 2020, is one week after the "Easter break";

(ii) forecasting the future daily numbers of new cases up to 28 days in advance with extremely low mean absolute percentage errors (MAPEs) using a relative paucity of data, namely, MAPE of 1.53% using 20 days of data to predict the number of new cases for the following 6 days, MAPE of 0.43% using 34 days of data to predict the number of new cases for the following 14 days, and MAPE of 0.20% using 55 days of data to predict the number of new cases for the following 28 days;

(iii) estimating approximately 33% of COVID-19 cases as unobserved by 26 March 2020, reducing to less than 5% after implementing the Government's constructive restrictions;

(iv) predicting that the growth rate, prior to the Government's implementation of restrictions, was on a trajectory to infect numbers equal to Australia's entire population by 24 April 2020;

(v) estimating the dynamics of the growth rate of the COVID-19 population to slow down to a rate of 0.820 after the first break point, with a slight rise to 0.979 after the second break point;

(vi) Advocating the outlined stochastic model as practically beneficial for policy makers when considering implementation and easing of virus restrictions due to the demonstrated sensitivity of the dynamics of the COVID-19 population in Australia to both major and minor system changes.

• The model developed in this work may further assist policy makers to consider the impact of several potential scenarios in their decision-making processes.

*School of Mathematical and Physical Sciences, University of Newcastle, NSW, Australia, and International Computer Science Institute, Berkeley, CA, USA. Email: `ali.eshragh@newcastle.edu.au`

†School of Management, Yale University, CT, USA.

‡School of Mathematical and Physical Sciences, University of Newcastle, NSW, Australia.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

The novel Corona Virus COVID-19 arrived on Australian shores around 25 January 2020. This paper presents a novel method of dynamically modeling and forecasting the COVID-19 pandemic in Australia with a high degree of accuracy and in a timely manner using limited data; a valuable resource that can be used to guide government decision-making on societal restrictions on a daily and/or weekly basis. The "partially-observable stochastic process" used in this study predicts not only the future actual values with extremely low error, but also the percentage of unobserved COVID-19 cases in the population. The model can further assist policy makers to assess the effectiveness of several possible alternative scenarios in their decision-making processes.

1 Introduction: COVID-19 Pandemic

The novel beta-coronavirus, later named "COVID-19", was first reported in late December 2019 in Wuhan City, China [13]. Early reports indicated a wet market in Wuhan to be the origin of the outbreak, affecting approximately 66% of market staff, and comprising symptoms resembling pneumonia of fever, dry cough, and fatigue [25]. The market closed 1 January 2020, following an epidemiologic alert announced by the local health authority in China on 31 December 2019. The infection was reported to have spread to many cities across China over January 2020, with thousands in China becoming infected by the disease, while also spreading rapidly globally, affecting countries including Thailand, Japan, Korea, Vietnam, Singapore, United States and Germany [25]. The World Health Organization (WHO) declared the outbreak a pandemic on 11 March 2020 and, as of 22 May 2020, a total of $4,993,470$ confirmed cases of COVID-19 globally were reported by WHO with $327,738$ related deaths across at least 216 countries[1].

Pandemic in Australia. According to official reports, the novel Corona Virus COVID-19 arrived on Australian shores on around 25 January 2020. From 5 March 2020, the number of new cases grew rapidly and reached over 300 cases daily in late March[1]. Following lockdown restrictions by the Australian Government from mid-March, the daily number of new cases started declining from early April, reaching approximately 20 cases daily by late April[1].

Preventative measures to minimize transmission were increasingly imposed by the Australian Government from 1 February 2020 with foreign nationals from mainland China banned entry to Australia, and 14 days of self-quarantining imposed for returning citizens from China[2]. Travel restrictions were subsequently imposed with all travelers arriving to Australia required to self-isolate for 14 days from 15 March 2020, with fines of up to AUD$50,000 for non-compliance[3]. A general travel ban was imposed from 20 March 2020 with Australia closing its borders to all non-residents[4].

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/
[2] https://www.pm.gov.au/media/extension-travel-ban-protect-australians-coronavirus
[3] https://www.bbc.com/news/world-australia-51894322
[4] https://www.australia.gov.au/coronavirus-updates

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A human bio-security emergency was declared in Australia on 18 March 2020 with a social distancing rule of 4 square meters per person imposed from 20 March 2020. From 22 March 2020, a mandatory closure of non-essential services was imposed with some states closing their borders allowing only the state's residents to return[3] and from 23 March 2020, all places of social gathering were closed with cafes and restaurants limited to takeaway[5]. From 29 March 2020, public gatherings were limited to two people if they were not from the same household and there were only four acceptable reasons for leaving homes, comprising shopping for essentials, medical or compassionate needs, exercise and work or education purposes[5].

Some subtle variation in the timing of the implementation of these measures occurred between States/Territories with State Governments/Territory officials also imposing additional restrictions in response to State/Territory-specific data. For example, some states introduced social distancing measures in schools from 15 March 2020, preventing students and staff from congregating in large numbers with several university graduations, conferences, events, classes and student organized events also canceled[6].

At the time of writing, the Government had a three-stage plan to reopen Australia by July 2020[7]. The three stages reflect increasing the numbers of permissible visitors in homes and public places, whilst still maintaining noted hygiene and social spacing, along with the opening of various places of employment and social interaction (restaurants, community centers in Stage 1). Accompanying this are the lifting of travel restrictions: local and regional in Stage 1, interstate in Stage 2 and partial international, principally Pacific region, in Stage 3. The seven States and Territories invoke these at slightly differing times to reflect their local experiences and numbers of infected people.

Vital need for modeling. The outbreak of COVID-19 and its accompanying pandemic has created an unprecedented challenge and unilateral response worldwide, and urged every nation to deploy its utmost resources toward combating the disease whilst managing the economic and social impacts. Tracking the epidemic and estimating the size of the infected population and effects of potential guidelines and restrictions has become a critical priority for most governments around the globe as it has immediate ramifications on all subsequent policy interventions (e.g. see [7, 8, 17, 19, 22]).

Stochastic processes are designed to deal with change that involves randomness and uncertainty, both aspects that are paramount to the COVID-19 outbreak. In particular, *partially-observable stochastic processes* specifically account for incomplete knowledge of a system that arises from knowing only partially about a given situation, without knowledge of the complete situation. A common form of partial observation is one whereby the state of the system, or each component of the system, can be observed with only a certain degree of certainty. An example is the case of biological invasions, whereby an invasive species or

[5] https://www.australia.gov.au/coronavirus-updates

[6] https://www.smh.com.au/national/nsw/school-assemblies-excursions-and-events-to-be-cancelled-2020031 html

[7] https://www.pm.gov.au/media/update-coronavirus-measures-08may20

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

individual of the species can be detected with only a certain probability upon each survey (e.g. see [20]). Another application is in medical testing, where a test can provide a false negative, so that the infection can then only be detected with a certain probability for an infected individual upon administering the test (e.g. see [12, 16]). Further applications of partially-observable stochastic processes include (but are not limited to) recognizing patterns [11], analyzing digital signals [24], and understanding biological processes [1, 2].

In this paper, we focus on modeling the early stages of the COVID-19 outbreak in Australia, and provide an epidemic model that complements others in use by providing extremely accurate estimates of COVID-19 transmission in Australia, including estimates of hidden cases, in timeframes relevant to policy implementation. More precisely, we utilize a special class of stochastic processes, the *partially-observable pure birth process*, to model the dynamics of the COVID-19 population in Australia. In the present epidemiological context of modeling the COVID-19 outbreak, the main source of uncertainty comes from the stochastic dynamics of the system as well as the structure of sampling in which each infected individual can be tested with only a certain degree of certainty. Our model particularly suits situations where the number of infected citizens is relatively minimal compared to the total at-risk population, which is the case for the majority of regional and national jurisdictions. This is a critical phase of disease spread, and requires policy measures that effectively control growth. The effectiveness of these policies, in turn, depends heavily on the quality of the models used and the precision of the estimates that they generate. The following two features of our model establish its benefits relative to other models (such as Susceptible, Infected and Recovered (`SIR`) model):

- The *robust predictive nature* where, with only small amounts of data, the (subsequently released) future actual values are forecasted very well;
- The capability to estimate not only the growth rate of the COVID-19 cases, but also the percentage of *unobserved cases*, which represent those in the population who have not been officially diagnosed.

The Highlights identify the main contributions of this work. In summary, the novel modeling identifies key break points associated with social restrictions imposed by the Government; estimates the percentage of unobserved cases; accurately predicts future numbers of COVID-19 cases pre- and post-implementation of restrictions; demonstrates how growth rates in cases changes in response to major and minor break points; and provides guidance for policy makers in terms of the sensitivity of the dynamics of COVID-19 in Australia.

The structure of this paper is as follows: Section 2 applies a partially-observable pure birth process to model the dynamics of the COVID-19 population in Australia and predicts future values as well as the percentage of unobserved hidden cases. Section 3 discusses the strengths and limitations of the model used in this study. Section 4 provides the final discussions on policy implications and concluding remarks. Appendix A presents an overview of partially-observable continuous-time Markov population processes along with their theoretical and applied properties.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2 Modeling: Dynamics of the COVID-19 Population in Australia

A *Continuous-time Markov population Process* (CTMPP) is a class of stochastic processes often used to model biological phenomena (e.g. see [5, 18, 21]). The study of a CTMPP under partial observations, referred to as a "partially-observable continuous-time Markov population process" (PO-CTMPP) is of interest for the present study. A special class of PO-CTMPPs is the *partially-observable pure birth process* (PO-PBP) whereby, while the underlying model is a stochastic "pure birth process", observations are made partially according to a binomial distribution.

Bean et al. [3] extensively studied the theoretical properties of a PO-CTMPP and a PO-PBP, and derived the conditional probability distribution of the true state of the system and future values of partial observations, given the history of partial observations. Furthermore, they showed that, unlike a pure birth process, a PO-PBP is not Markovian of any order. Bean et al. [4] applied these results to find the Fisher Information for a PO-PBP and derived the optimal experimental design. The details of these results are summarized in Appendix A.

We utilize these approaches here to model and analyze the dynamics of the COVID-19 population in Australia. Due to practical limitations described in Appendix A.1, not all infected cases may be observed each day, implying that the confirmed cases reported officially are only partial representations of actual cases. Therefore, a PO-PBP can be considered a superior model to explain the complex dynamics of the COVID-19 population.

Remark 1. We assume that there is no shortage of "COVID-19 test kits" in Australia. If this assumption were false, the model would lose applicability once the required sampling rate reached the limit imposed by kit shortages. However, this has not happened to date in Australia, according to the Australian Prime Minister's statements during the pandemic[a].

[a] https://www.linkedin.com/in/scottmorrisonmp/

The data for this study are obtained from "daily WHO reports"[8] provided on their website[9]. All algorithms are coded in C, and the outputs are analyzed in MATLAB R2019a.

The first step in data analysis is visualization. Figure 1 displays the cumulative new COVID-19 cases in Australia from 1 March to 22 May 2020 and demonstrates two *structural break points* where the population dynamics have been altered, mainly attributed to new polices and exogenous factors:

(i) The first break point occurs on 27 March 2020. This point corresponds to one week after implementation of the "lockdown restrictions". As shown in Figure 1, this is a crucial break

[8] It should be noted that there are a few official resources for the COVID-19 data with minor differences in their reports. Although the structure of our model and main output stay consistent for the data from different resources, the numerical results might slightly vary.

[9] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

point where the curvature of the cumulative new cases dramatically changes from a convex exponential growth to a concave stable pattern. Furthermore, it is observed that the growth rate starts declining after this break point.

(ii) The second break point occurs on 18 April 2020 which corresponds to a week after the "Easter break" in Australia. Unlike the first break point, the second one does not transform the curvature or stability of the graph, but instead shifts it up slightly and slows down the speed at which the growth rate parameter is declining.

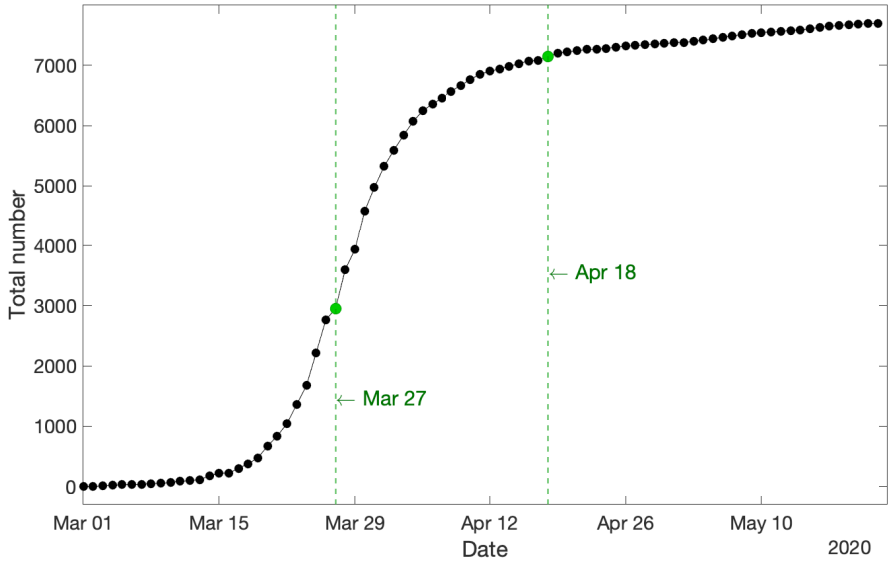

Figure 1: Cumulative new COVID-19 cases in Australia for the whole span of the data. Two structural break points in the dynamics of the COVID-19 population are evident: a major break point on 27 March 2020, and a minor break point on 18 April 2020.

Remark 2. Figure 1 along with those two structural break points indicate that the dynamics of the COVID-19 population in Australia appear very sensitive towards major/minor system changes, which should be a serious consideration for policy makers while easing out virus restrictions.

To gain insights into the complex nature of the population for modeling its dynamics, we carry out our analysis in three nested real time steps:

- Step 1: $1-26$ March 2020, until the first break point (cf. Section 2.1);
- Step 2: 1 March–17 April 2020, until the second break point (cf. Section 2.2);
- Step 3: 1 March–22 May 2020, the whole span of the data (cf. Section 2.3).

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The general scope of our modeling for each step is as follows: we fit a PO-PBP to the data to model the dynamics of the COVID-19 population over the designated period. A PO-PBP possesses two major parameters comprising the growth rate λ_t and the observation probability p_t (cf. A.2). We construct the likelihood function of partial observations according to (1) in conjunction with Corollary 1, and truncate the involved infinite sums by utilizing (2) and Proposition 1. Then, the logarithm of the likelihood function is maximized over the range of parameters to find their maximum likelihood estimates (MLEs). Finally, Proposition 2 is applied to predict the future values of partial observations.

Furthermore, in order to evaluate the accuracy of predictions generated by the estimated models, the dataset for each nested real time step is partitioned into two mutually exclusive segments consisting of the *training data* to estimate the parameters and forecast future values, and the *test data* to evaluate the accuracy of predictions. To measure the latter, we utilize *mean absolute percentage error*, introduced in Definition 1.

Definition 1 ([14]). *The mean absolute percentage error (MAPE) is defined as:*

$$\mathsf{MAPE} = \frac{1}{h}\sum_{t=1}^{h}\left|\frac{f_t - x_t}{x_t}\right| \times 100\%,$$

where f_t, x_t and h are the forecasted values, actual values, and prediction horizon, respectively.

Remark 3. It should be noted that the quality of predictions reported in Sections 2.1 to 2.3 is robust to the choice of training and test data, provided there are enough observations in the former set (cf. Figure 5).

2.1 Step 1: 1–26 March 2020

This step involves the beginning of the pandemic in Australia where the COVID-19 population is growing exponentially fast. The data from $1-20$ March 2020 are used as the *training data* to estimate the parameters of the model, and the data from $21-26$ March 2020 are used as the *test data* to evaluate the accuracy of predictions. For this period of modeling, we consider the flowing dynamics for the two parameters growth rate λ_t and observation probability p_t for the underlying PO-PBP:

$$\lambda_t = \alpha_1^{(t-t_0)}\lambda \qquad \text{for } t \text{ from } 1-26 \text{ March 2020},$$

and

$$p_t = \min\{\beta_1^{(t-t_0)}p, 1\} \qquad \text{for } t \text{ from } 1-26 \text{ March 2020},$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where, $\alpha_1 > 0$ and $\beta_1 > 0$ are constant coefficients, λ and p are unknown initial values of the parameters, and t_0 is the date of the first observation (i.e., 1 March 2020). After constructing the likelihood function and maximizing over the parameters, the MLEs in Table 1 are derived. Since MLE of α_1 and β_1 turn out equal to 1, both $\mathsf{MLE}(\lambda_t) = 0.235$ and $\mathsf{MLE}(p_t) = 0.67$ are fixed for all t in the range.

Parameters	λ	α_1	p	β_1
MLE	0.235	1.000	0.67	1.00

Table 1: MLE of parameters

By applying Proposition 2, the expected values of partial observations over the span of test data (i.e., $21 - 26$ March 2020) are predicted. The results are shown in Figure 2, where the training data, test data, and predictions are displayed by the black solid plot, red dot-dash plot, and blue solid plot, respectively. It is readily seen that the predicted values are so well fitted to the actual test data. This observation is numerically confirmed with $\mathsf{MAPE} = 1.53\%$.

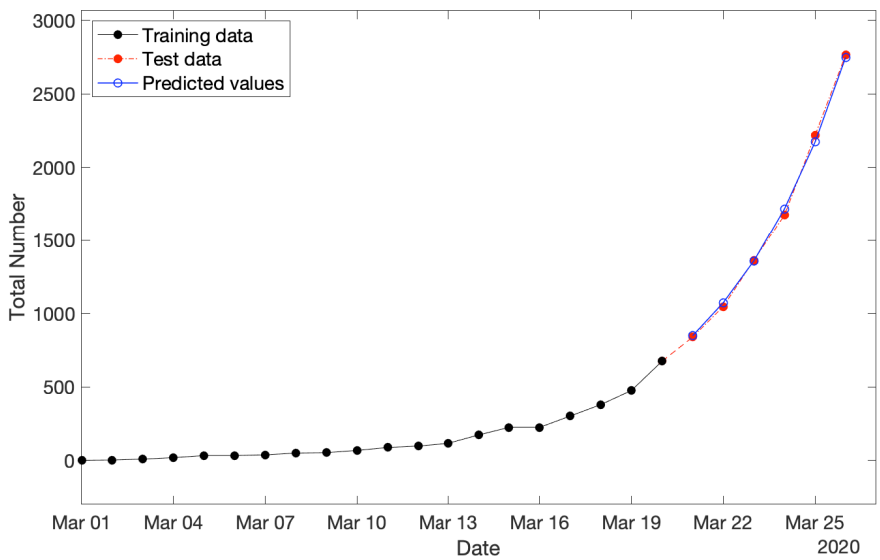

Figure 2: Cumulative new COVID-19 cases in Australia for three categories consisting of: (i) training data spanning from $1 - 20$ March 2020 (in black), test data spanning from $21 - 26$ March 2020 (in red), and (iii) predicted values over the period of test data (in blue) with $\mathsf{MAPE} = 1.53\%$.

The model additionally suggests that prior to the government's implementation of restrictions, the growth rate was on a trajectory to hit infection numbers equal to Australia's entire population by 24 April 2020, a prediction which would have probably been softened only somewhat by limiting factors such as our island status. This asserts the effectiveness

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

of Government's policies and restrictions. Figure 3 displays the semi-log plot (where the y-axis is scaled logarithmically) of the cumulative new COVID-19 cases in Australia (from 1 March–22 May 2020) and the model predictions (from 21 March–24 April 2020) in black and blue, respectively. This figure reveals the exponential growth of the COVID-19 population before the impact of lockdown restrictions on 27 March 2020 (marked in green).

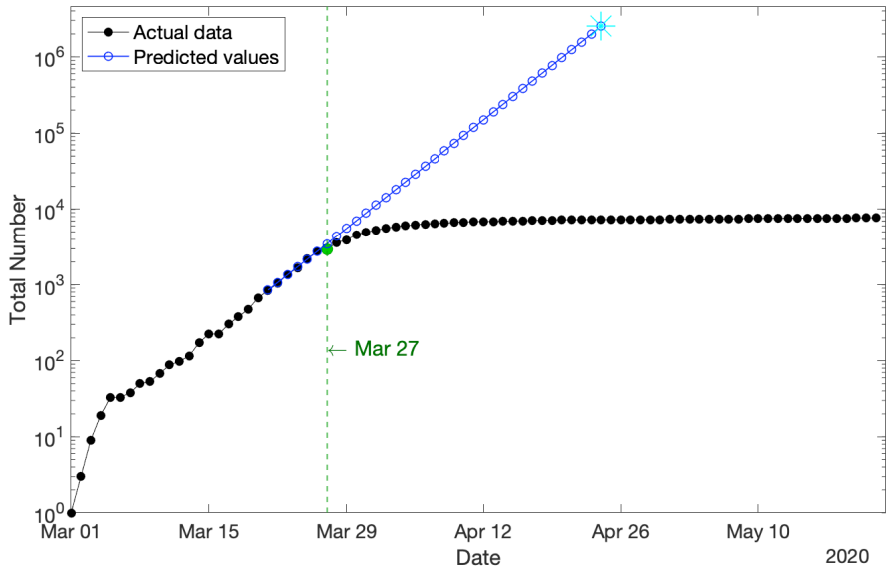

Figure 3: Cumulative new COVID-19 cases in Australia for the whole span of the data (in logarithmic scale) compared with the predicted values for the scenario in which the lockdown restrictions had not been implemented. It shows the exponential growth of the COVID-19 population before the impact of lockdown restrictions on 27 March 2020 (marked in green).

Remark 4. Figure 3 shows an initial break point on 6 March 2020. As it is located at the outset of the pandemic in Australia and the number of confirmed cases is still very small during that short period, we disregard it as a break point in our analysis. It does not have a significant influence on the results.

Finally, the MLE of the observation probability p_t estimates that only 67% of COVID-19 cases in Australia had been tested by 26 March 2020, and the hidden 33% cases had not been recorded/diagnosed officially, by this date.

Identifiability analysis. In statistical inference, there are several tools to measure the quality of estimates, including *identifiability*.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Definition 2 ([15]). *A statistical inference is called "identifiable", if different values of the model parameters generate different probability distributions of the observable variables.*

If an inference is truly not identifiable, then mathematically, the value of the likelihood function will be a constant at all values of the parameters which are equivalent. Therefore, in this case, one would expect to see a ridge on the likelihood surface of roughly constant values as the parameters change.

In order to see the identifiability of the MLEs of the main parameters λ and p given in Table 1, we plot the log-likelihood function of partial observations in terms of the two parameters. As depicted in Figure 4, it is clearly observed that there exists a curvature in the log-likelihood function, illustrating these estimates are identifiable.

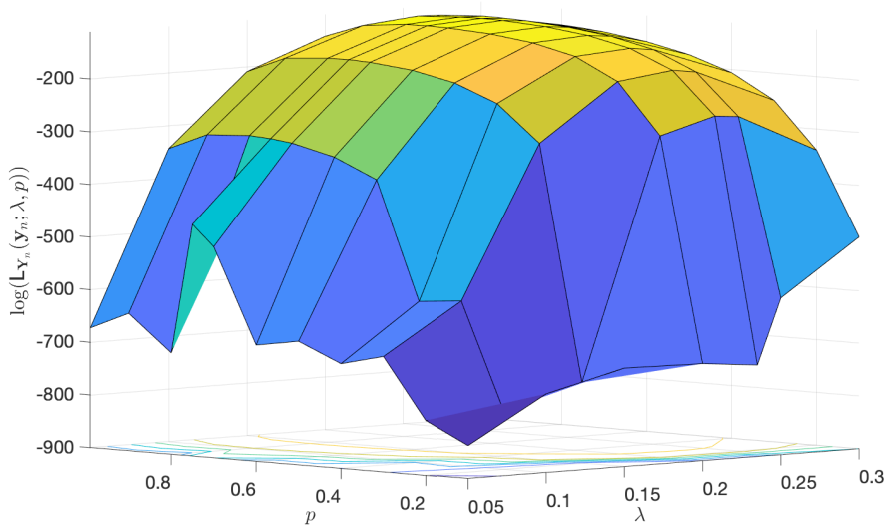

Figure 4: The log-likelihood function of partial observations in terms of the two parameters λ and p. This plot illustrates the identifiability of the estimates provided in Table 1.

Remark 5. There is a heuristic that so long as the log-likelihood function changes by at least 2 units, then it is regarded as a worthwhile change, implying the identifiability of estimates. So, by considering the locus of points in (λ, p) that remain within 2 units of the log-likelihood function at $(\mathsf{MLE}(\lambda) = 0.235, \mathsf{MLE}(p) = 0.67)$, the locus allows for $\mathsf{MLE}(p)$ to range within $[0.55, 0.75]$ and for $\mathsf{MLE}(\lambda)$ to range within $[0.225, 0.245]$. A few other points within those two ranges are chosen as the MLEs of λ and p, but no significant difference in MAPE of predictions is observed.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Sensitivity analysis. Due to the very small amount of available data (26 observations in this step, and 76 data in total), there is limited opportunity to investigate the robustness of estimates. In spite of this, the MLE of parameters for size training data, varying from 20 to 25, are estimated and the future values over the span of the corresponding test data (where their sizes varying downward from $6-1$) are predicted. The MAPE of those predictions versus the size of training data are displayed in Figure 5. The largest MAPE is 3.18%, which is still very low, illustrating high quality forecasts, while demonstrating the robustness of our estimates to the choice of training data.

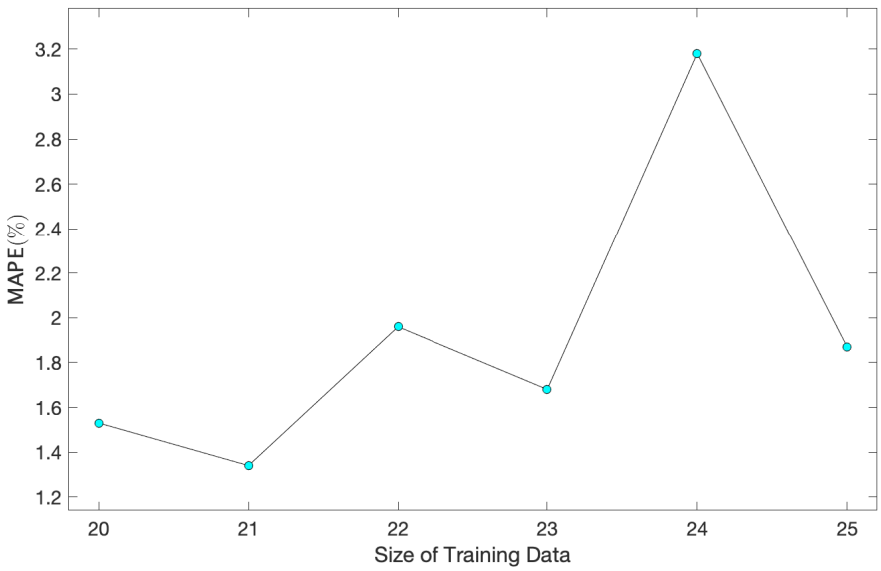

Figure 5: MAPE of predictions vs. size of training data for a range of size of training data (test data) varying from $20-25$ $(6-1)$.

2.2 Step 2: 1 March–17 April 2020

This step involves the first break point when the influence of lockdown restrictions appears in the COVID-19 population growth. The data from 1 March–3 April 2020 are used as the *training data* to estimate the model parameters, using data from $4-17$ April 2020 as the *test data* to evaluate the accuracy of predictions. Figure 1 indicates changes to the dynamics of the growth rate after the first break point. Accordingly, we consider the following dynamics for the growth rate and the observation probability parameter for the underlying `PO-PBP`:

$$\lambda_t = \begin{cases} \alpha_1^{t-t_0}\lambda & \text{for } t \text{ from } 1-26 \text{ March 2020,} \\ \alpha_2^{t-bp_1+1}\alpha_1^{bp_1-t_0}\lambda & \text{for } t \text{ from 27 March–17 April 2020,} \end{cases}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

and

$$p_t = \begin{cases} \min\{\beta_1^{t-t_0} p, 1\} & \text{for } t \text{ from } 1-26 \text{ March 2020,} \\ \min\{\beta_2^{t-bp_1+1} \beta_1^{bp_1-t_0} p, 1\} & \text{for } t \text{ from 27 March–17 April 2020,} \end{cases}$$

where $\alpha_2 > 0$ and $\beta_2 > 0$ are new constant coefficients, and bp_1 is the date of the first break point (i.e., 27 March 2020). The two new parameters α_2 and β_2 control the impact of the first break point on the growth rate and observation probability, respectively. Table 2 provides the MLE of the parameters.

Parameters	λ	α_1	α_2	p	β_1	β_2
MLE	0.235	1.000	0.814	0.67	1.00	1.06

Table 2: MLE of parameters

By applying Proposition 2, the expected values of partial observations over 14 days, that is the span of test data from $4-17$ April 2020, are predicted. The results are shown in Figure 6, where the training data, test data, and predictions are displayed by the black solid plot, red dot-dash plot, and blue solid plot, respectively, and the first break point is marked in green. Clearly, the predicted values are remarkably fitted to the actual test data with a significantly small MAPE $= 0.43\%$, which is notably less than one percent error.

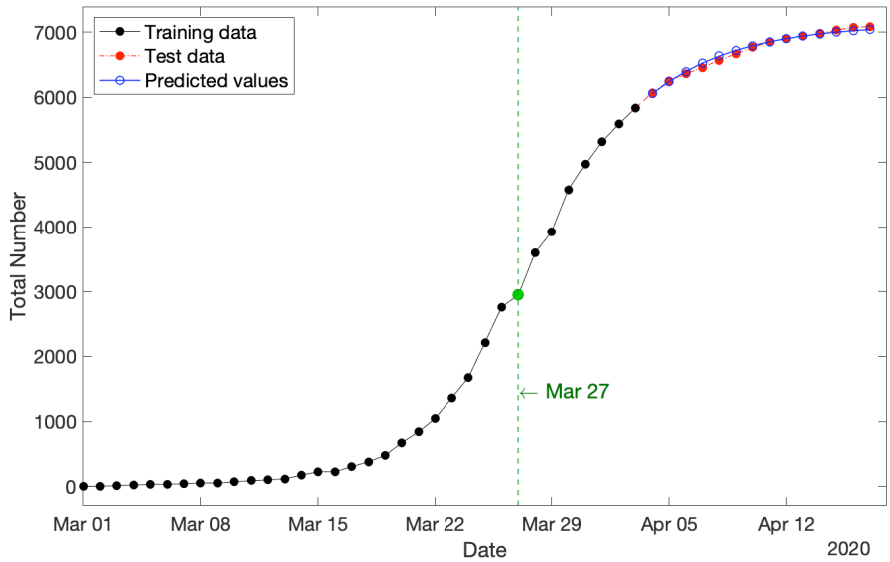

Figure 6: Cumulative new COVID-19 cases in Australia for three categories: (i) training data spanning 1 March–3 April 2020 (in black), test data spanning $4-17$ April 2020 (in red), and (iii) predicted values over the period of test data (in blue) with MAPE $= 0.43\%$. The first break point on 27 March 2020 is marked in green.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Remark 6. The MLEs provided in Table 2 imply that the observation probability p_t starts increasing after the first break point with the estimated boosting factor of $\mathsf{MLE}(\beta_2) = 1.06$ such that after one week, it reaches the upper bound of 1. However, by considering the identifiability of these point estimations as well as Remark 5, it is observed that the locus allows for $\mathsf{MLE}(p_t)$ to range within $[0.95, 1.00]$ over the test period of $4-17$ April 2020. Hence, the model estimates that the percentage of observed COVID-19 cases from $2-17$ April 2020 lies within the range $[95\%, 100\%]$. Furthermore, analogous to Figure 5, the robustness of estimates is confirmed.

2.3 Step 3: 1 March–22 May 2020

The last step is for the whole span of the data, involving both structural break points. The data from 1 March–24 April 2020 are used as the *training data* to estimate the parameters of the model, and the data from 25 April–22 May 2020 are used as the *test data* to evaluate the quality of predictions. Motivated from Figure 1 along with Steps 1–2, we define the following dynamics for the growth rate and the observation probability parameter for the underlying PO-PBP:

$$\lambda_t = \begin{cases} \alpha_1^{t-t_0}\lambda & \text{for } t \text{ from } 1-26 \text{ March 2020,} \\ \alpha_2^{t-bp_1+1}\alpha_1^{bp_1-t_0}\lambda & \text{for } t \text{ from 27 March–17 April 2020,} \\ \alpha_3^{t-bp_2+1}\alpha_2^{bp_2-bp_1}\alpha_1^{bp_1-t_0}\lambda & \text{for } t \text{ from 18 April–22 May 2020,} \end{cases}$$

and

$$p_t = \begin{cases} \min\{\beta_1^{t-t_0}p, 1\} & \text{for } t \text{ from } 1-26 \text{ March 2020,} \\ \min\{\beta_2^{t-bp_1+1}\beta_1^{bp_1-t_0}p, 1\} & \text{for } t \text{ from 27 March–17 April 2020,} \\ \min\{\beta_3^{t-bp_1+1}\beta_2^{bp_2-bp_1}\beta_1^{bp_1-t_0}p, 1\} & \text{for } t \text{ from 18 April–22 May 2020,} \end{cases}$$

where $\gamma > 0$ is the new inflation parameter on the growth rate after the second break point, and bp_2 is the date of the second break point (i.e., 18 April 2020). Table 3 provides the MLE of the parameters.

Parameters	λ	α_1	α_2	α_3	p	β_1	β_2	β_3
MLE	0.235	1.000	0.820	0.979	0.67	1.00	1.06	1.00

Table 3: MLE of parameters

By applying Proposition 2, the expected values of partial observations over the span of test data (i.e., 28 days from 25 April–22 May 2020, inclusive) are predicted. Results are shown in Figure 7, where the training data, test data, and predictions are displayed by the black solid plot, red dot-dash plot, and blue solid plot, respectively, and the two break points

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

are marked in green. It is clearly evident that the predicted values are exceptionally closely fitted to the actual test data with MAPE = 0.20%, which is notably much less than one percent.

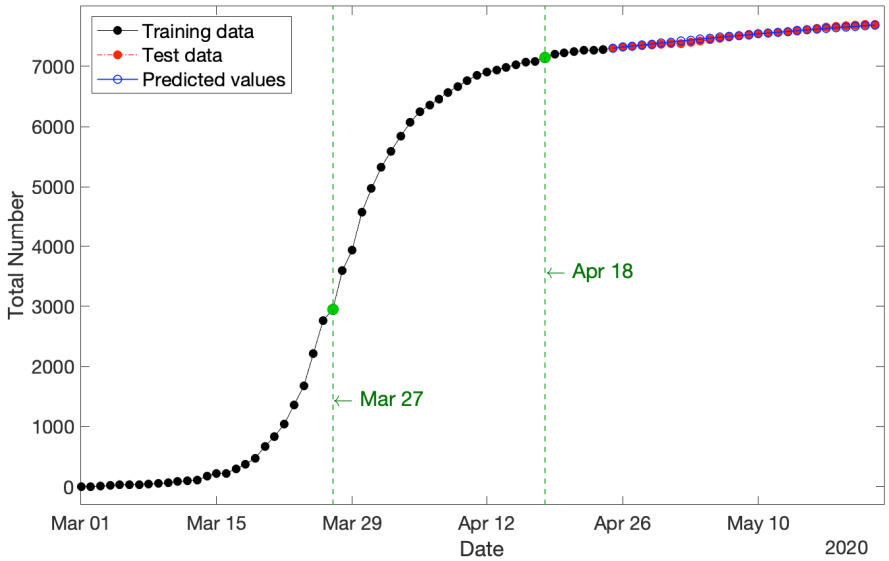

Figure 7: Cumulative new COVID-19 cases in Australia for three categories: (i) training data spanning 1 March–24 April 2020 (in black), test data spanning 25 April–22 May 2020 (in red), and (iii) predicted values over the period of test data (in blue) with MAPE = 0.20%. The two break points are marked in green.

Remark 7. The MLEs provided in Table 3 show that the declining parameter on the growth rate is inflated from $\mathsf{MLE}(\alpha_2) = 0.820$ to $\mathsf{MLE}(\alpha_3) = 0.979$. Although it is still less than one (indicating that the population is stable and not exploding), but such an increase as a consequence of people's interactions during the Easter break as well as releasing a few restrictions on 2 May 2020 should be taken into account by policy makers for easing out the COVID-19 restrictions. Furthermore, by considering the identifiability of estimations as well as Remark 5, it is observed that the locus allows for $\mathsf{MLE}(\alpha_3)$ to range within $[0.940, 1.020]$. Hence, there is a chance that the parameter $\mathsf{MLE}(\alpha_3)$ could be greater than one, implying that the population starts growing again. If this takes place, the population size will quickly resume exponential growth (cf. Figure 3). Furthermore, analogous to Figure 5, the robustness of estimates is confirmed.

Remark 8. One can easily see that the MLEs of parameters provided in Tables 1 to 3, alter after each per-identified structural break point. These results confirm our choice of break points, and also motivate us to suggest an algorithmic way to detect the break

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

points. In that case, the whole time frame can be partitioned into τ mutually exclusive time intervals (t_{k-1}, t_k) for $k = 1, \ldots, \tau$, and the following dynamics for the growth rate parameter λ_t and the observation probability p_t can be constructed:

$$\lambda_t = \begin{cases} \alpha_1^{t-t_0}\lambda & \text{for } t \in [t_0, t_1), \\ \Pi_{j=1}^{k-1}\alpha_j^{t_j-t_{j-1}}\alpha_k^{t-t_{k-1}+1}\lambda & \text{for } t \in [t_{k-1}, t_k),\ j = 2, \ldots, \tau, \end{cases}$$

and

$$p_t = \begin{cases} \min\{\beta_1^{t-t_0}p, 1\} & \text{for } t \in [t_0, t_1), \\ \min\{\Pi_{j=1}^{k-1}\beta_j^{t_j-t_{j-1}}\beta_k^{t-t_{k-1}+1}p, 1\} & \text{for } t \in [t_{k-1}, t_k),\ j = 2, \ldots, \tau. \end{cases}$$

After estimating all parameters, those consecutive intervals showing distinct MLEs for α_i and β_i could be an indication of a "structural break point". If one wants to trim estimating the location of each break point, one should merge those consecutive intervals first, then partition them into some more sub-intervals with new parameters, and re-estimate all parameters together. Such trimming procedure can be repeated until there is no change in the break points.

3 PO-PBP: Strengths and Limitations

The main strength of the PO-PBP model presented in this paper is the high accuracy of predictions within a timescale in the order of 4 weeks. It is timescales of this order that governments are considering to adjust restrictions and modify adjustments to restrictions[10]. Such high accuracy on these timescales makes the model not only applicable but also highly appealing to base policy decisions on.

A natural question to ask is whether comparable accuracy can be obtained in this work by directly fitting a pure birth process to the data-segments that lie between break points. We have made this calculation and conducted the comparison with our PO-PBP model and found that although the PBP performs reasonably well, it is not as accurate as our PO-PBP. For instance, the MAPE for predictions before the lockdown restrictions obtained by the former model is in the order of 6.03%, which is not as good as the MAPE for our PO-PBP, which is only 1.53%.

Would the accuracy of predictions of this model extend into the long-range future? This model, as with any models, should be used within its range of validity, with care to take into account the assumptions built into it. Due to some of the assumptions in the model that are described below, long-range predictions will lose accuracy unless they are corrected for by updated real-life data.

The model in this paper is based on a PBP, which is a special case of a more general Continuous-time Markov population process that includes both births and deaths. In this

[10] https://www.pm.gov.au/media/update-coronavirus-measures-08may20

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

case, a "birth" is a new infection, and a "death" represents removal from the infected population by means of either recovery from illness or death. The real-life process of COVID-19 infection includes both infection and potential recovery or death. Thus in that sense a richer model that includes both birth and death would be appropriate. However, there is a trade-off. The theory for the richer model exists and is sound, but is computationally much less tractable.

The trade-off between representational accuracy and computability is a common one in mathematical modeling, and well understood in the literature (e.g., see [23]). In the case of birth and death processes, if births happen much faster than deaths, it is legitimate to disregard deaths within an appropriate timescale. In the case of COVID-19, infection can happen very quickly – in a matter of hours, whereas recovery or potentially death takes much longer: weeks potentially stretching into a month or more. Thus on timescales of a few weeks, much insight can be gained from a `PBP`.

Furthermore, the trade-off between the tractability of a `PBP` versus the longer range accuracy of one utilizing both births and deaths can be further tipped in the balance towards pure birth modeling by a method which to some extent accounts for both processes. Consider a growing birth and death process with the birth and death rate of λ and μ, respectively, such that $\lambda > \mu$. Then to some extent, both may be accounted for in a `PBP` in which birth rate is modeled by the difference $\lambda - \mu$. Our `PO-PBP` model employs this strategy: the λ in our model is really the difference between an underlying birth and death rate. This extends slightly the timescales within which the model is useful.

The considerations above are relevant to the prediction in Figure 3, in which the blue line indicates that if growth rates had continued in the same pattern as pre-lockdown, all Australians would have been infected by 24 April 2020. In reality that pattern would have been somewhat mitigated by factors not currently present in our `PO-PBP`. One such factor is recovery/death rates discussed above. Another is the finite size of the Australian population and our island status.

Due to Australia being an island along with recently closing its borders, any circulating virus has only a finite collection of approximately 25 million people to potentially infect. The more people infected, the greater the chance that when an infected person comes into contact with another person, that other person is also already infected, so that no new infection can occur. This saturation effect is not built into our `PO-PBP`, which means that the blue line in Figure 3 would shoot straight off the page had we not truncated it. In alternative models that account for the limiting effect of finite population size, that blue line would have curved down slightly as the infected proportion of the population became comparable with the overall population size. In other words, the domain of validity of the present `PO-PBP` is limited to the situation in which the overall proportion of infected individuals is still relatively low, as is presently the case, as of May 2020.

At this point, it is insightful to compare with another form of disease modeling, the use of `SIR` models utilizing differential equations. The acronym "SIR" stands for three categories, respectively: Susceptible, Infected and Removed. A *susceptible* person has not yet caught the infection, but potentially could; an *infected* person actively has the infection, and is

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

assumed capable of passing it on; and a *removed* person is removed from the population of susceptible or infected people by either death or immunity. In basic SIR modeling, a flow diagram is drawn between the three categories, and the flows between these are modeled by differential equations in which parameters are introduced to represent the chances of transitioning between the states. In this way the inter-relationships between the categories are captured. Solutions to the equations predict how the size of the three categories will develop over time. These solutions may be obtained either exactly or numerically, and sometimes depend critically on accurate values of the parameters in the SIR model.

An advantage of the SIR model over our PO-PBP is that the SIR accounts for the finite population size, and removes persons who have become immune or died from the susceptible/infected population. The disadvantage, however, is that the SIR model does not explicitly account for the "partially observed" nature of COVID-19 cases. Policy makers and Scientists only have access to reported data, yet there may well be infections in the population which are driving the growth of the pandemic, but which are not directly diagnosed and hence observed.

The most remarkable advantage of the PO-PBP is that it provides means of estimating and incorporating the proportion of hidden cases. More precisely, our model employs a new "observation probability" parameter to the underlying PBP to construct the new PO-PBP. Then, by maximizing the complicated likelihood function of the PO-PBP, all parameters including the observation probability at each time t are estimated. Due to the invariant property of MLEs, one minus the MLE of that observation probability will estimate the proportion of hidden cases in the population.

Any model, however, is only as valid as its assumptions. An assumption of the PO-PBP is that sampling is "uniform and random", implying that the model assumes any infected person as likely to be tested and identified as any other person. The assumption of randomness is almost never 100% satisfied for any realistic scenario – some biases will always be present. What matters is how impactful these are. It is worth considering the impact of model assumptions in this modeling.

One feature of relevance is the availability of test kits. If shortage of test kits were to severely curtail sampling, this would undermine the validity of the model. In Australia, initial testing was largely limited to people considered "high risk", and include those who recently returned from overseas, had contact with a confirmed COVID-19 case, or in hospital with severe symptoms matching the disease, while other population members were considered "low risk". If perchance significant infection had established in the "low risk" part of the population, and if a low death rate had allowed this hidden population to remain undetected, the proportion of undetected cases reported by our model would be an underestimate. However, the predicted pattern of confirmed infections would remain valid.

Recently, the Australian government has substantially expanded testing opportunities and let testing for COVID-19 be available to every Australian with mild respiratory symptoms including a cough and sore throat[11]. This makes the "random sampling" assumption of the PO-PPB considerably more robust. We should soon be able to determine whether there

[11] https://www.pm.gov.au/media/update-coronavirus-measures-24april20

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

has been a significant reservoir of undetected COVID-19 cases in Australia.

It should be noted that we are implementing a continuous-time model, whilst observations are reported just once daily. Use of a continuous-time model is still valid, however, since it is general enough to take account of the discrete data structure. This continuous-time model is utilized due to the power of the theory that underlines the model and the relevance of that theory to this modeling situation.

We conclude this section by stating that the `PO-PBP` models only the impact of COVID-19 with respect to the numbers of infections – it does not model other impacts on society (negative or positive) of policy control measures, as it is recognized that restrictions on gatherings affect people's lives in different ways. A positive example is reduced air pollution due to reduced travel [9], while a negative example is increased risk of harms like domestic violence [6]. As yet there is no single model which incorporates all of these factors.

4 Discussion and Conclusion: Policy Implications

In this paper, we apply a class of continuous-time Markov population stochastic processes, namely the *partially-observable pure birth process*, to model and analyze the dynamics of the COVID-19 population in Australia. Specifically, we use the theoretical properties of this stochastic process to construct the likelihood function of cumulative confirmed cases to find the maximum likelihood estimates of its parameters. These estimations are used to predict future values of the population along with the number of unobserved hidden COVID-19 cases.

The Markovian stochastic process model that we develop is based on a partially observable pure birth process, and its predictions fit the actual observations at the Australian national level surprisingly well. Aside from its simplicity and high accuracy, there are several other advantages of this model from a policy perspective.

The stochastic process in our model revolves around only two parameters, both of which have clear and communicable practical interpretations: the former represents the speed of the spread, while the latter provides a measure of detection likelihood. We postulate that in the absence of any policy interventions, both parameters follow an evolution pattern that resembles a geometric decay. A shift in policy, however, may ratchet up or down the decay rate, thereby inducing a new infection trajectory.

As demonstrated, the model captures the complex dynamics of the detected/infected ratio, which is a critical component in the design of any containment policy. Furthermore, policymakers gain access to a coherent and insightful representation of the situation to informatively contemplate the consequences of action/inaction over the span of a few weeks (i.e., how the epidemic unfolds in the absence of any interventions).

Because of its efficacy, this model equips policymakers with a powerful tool to conduct scenario-based analyses, and enables predictions with a high degree of precision, of how a particular decision drifts the evolution trajectory of the disease, and enables such a prediction only a week after it is enacted. This supports early identification and reinforcement of effective policies, and timely scale back or discontinuation of others.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The model empowers decision-makers to evaluate and compare the implications of the two fundamental hallmarks of the model: lowering the infection rate versus increasing the detection likelihood. Depending on which one of these two avenues should be pursued, the government resources should be directed accordingly, and the corresponding message should be conveyed to the public.

If extensive community screening were undertaken particularly for asymptomatic citizens, then this model would be expected to give very accurate predictive power throughout the full range of possible policy implementation with regard to social distancing restrictions. The model further lends itself to crafting hybrid policies that utilize a combination of these two approaches and the delicate division of available resources between them.

Availability of test kits is a practical consideration in interpreting this model. The current model assumes an adequate supply of test-kits so that sampling is not restricted by a shortage. If this assumption was not fully met during the early days of the pandemic in Australia, it would mean we would have potentially underestimated the "hidden fraction" of undetected COVID-19 cases. Statements by the government[12] indicate that there is no current shortage of test kits in Australia. It would be interesting but quite difficult future work to try to explicitly incorporate test-kit shortage in the model. Such future work would have particular relevance in other countries where test-kit shortages are a pressing issue.

Additionally, our epidemic model treats the entire nation as a single pool of homogeneous agents with equal exposure to risks and similar contact behaviors. While this assumption may sound fairly restrictive, we believe that it has not impacted the quality of our findings in a profound way. Nevertheless, accounting for inherent heterogeneities and local characteristics of smaller regions/communities would further enhance the richness of the model and strengthen its outcomes. This modification particularly lends itself to countries such as U.S. where there is substantial variation in the extent and timing of the epidemic across states.

This model is likely to be applicable to many other countries and circumstances beyond Australia, since there are not a large number of location-specific assumptions built in, the main one being that testing involves a reasonably uniform sampling of the population. This strength partly derives from the analytical nature of the model, rather than being one which is simulation-based and within that highly customized to local factors. Ideally, different models are used in conjunction, and this model could profitably be used in conjunction with other types of models, in understanding the spread of COVID-19 in the future both in Australia and elsewhere.

A benefit of this model is that the 4 weeks prediction horizon allows officials to fine-tune their short-term actions and contingency planning in light of reasonable confidence in the immediately expected upcoming pattern in the number of cases.

This model suggests several possibilities for further research, that would enhance its applicability across a broader range of circumstances. For example, in the Australian data, the structural break-points were identified visually. There is potential to develop a purely computational method of doing these identifications, according to Remark 8. This would enhance applicability to more complex and long-term data sets, as may be expected in the

[12] https://www.linkedin.com/in/scottmorrisonmp/

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

future across the world.

All in all, the PO-PBP is a useful model for understanding and predicting the trajectory of COVID-19 in Australia under various policy choices. On a short timescale, relevant to government actions, predictions have been shown to be very accurate. The model also appears to be sensitive to subtle shifts in population behavior, allowing it to be useful in considering the impact of social events such as the Easter break in April 2020.

Acknowledgment The authors are very grateful to Dr Judy-anne Osborn from the University of Newcastle in Australia for extensive useful conversations about modeling assumptions, implications and presentation thereof, during preparation of this paper. The authors also thank Prof. Nigel Bean from the University of Adelaide in Australia for suggesting some improvements.

References

[1] Aggoun, L. and Elliott, R.J. (1998). Recursive Estimation in Capture-Recapture Methods. *Sultan Qaboos University, Oman, Science and Technology* **3,** 67–75.

[2] Baldi, P. and Brunak, S. (2001). *Bioinformatics*. MIT Press, Cambridge.

[3] Bean, N.G., Elliott, R., Eshragh, A. and Ross, J.V. (2015). On Binomial Observations of Continuous-time Markovian Population Models, *Journal of Applied Probability* **52(2),** 457–472.

[4] Bean, N.G., Eshragh, A. and Ross, J.V. (2016). Fisher Information for a Partially-Observable Simple Birth Process. *Communications in Statistics-Theory and Methods* **45(24),** 7161–7183.

[5] Black, A.J. and McKane, A.J. (2012). Stochastic Formulation of Ecological Models and Their Applications. *Trends in Ecology and Evolution* **27,** 337–345.

[6] Bradbury-Jones C. and Isham L. (2020). The Pandemic Paradox: The Consequences of COVID-19 on Domestic Violence. *Journal of Clinical Nursing*, https://doi.org/10.1111/jocn.15296.

[7] Chen, T.M., Rui, J., Wang, Q.P. Zhao, Z.Y., Cui, J.A. and Yin, L. (2020). A Mathematical Model For Simulating the Phase-based Transmissibility of a Novel Coronavirus. *Infectious Diseases of Poverty* **9(24),** https://doi.org/10.1186/s40249-020-00640-3.

[8] Dandekar, R.A., Henderson, S.G., Jansen, M., Moka, S., Nazarathy, Y., Rackauckas, C., Taylor, P.G. and Vuorinen, A. (2020). Safe Blues: A Method for Estimation and Control in the Fight Against COVID-19. https://doi.org/10.1101/2020.05.04.20090258.

[9] Dutheil, F., Baker, J.S. and Navel, V. (2020). COVID-19 As a Factor Influencing Air Pollution? *Environmental Pollution* **263(A),** 114466.

[10] Elliott, R.J., Aggoun, L. and Moore, J.B. (1994). *Hidden Markov Models: Estimation and Control*. Springer, New York.

[11] Fink, G.A. (2008). *Markov Models for Pattern Recognition: From Theory to Application*. Springer, Berlin-Heidelberg-New York.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[12] Gerberry, D.J. (2009). Trade-off Between BCG Vaccination and the Ability to Detect and Treat Latent Tuberculosis. *Journal of Theoretical Biology* **261,** 548–560.

[13] Huang, H, Wang, Y., Li, X., Ren, L., Zha, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J. and Cao, B. (2020). Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *Lancet* **395,** 497–506.

[14] Hyndman, R.J. and Koehler, A.B. (2006). Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, **22(4),** 679–688.

[15] Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag New York, Inc.

[16] Kao, R.R. (2003). The Impact of Local Heterogeneity on Alternative Control Strategies for Foot-and-Mouth Disease. *Proceedings of the Royal Society B: Biological Sciences*, **270,** 2557–2564.

[17] Koo, K.R., Cook, A.R., Park, M., Sun, Y., Sun, H., Lim, J.T., Tam, C. and Dickens, B.L. (2020). Interventions to Mitigate Early Spread of SARS-CoV-2 in Singapore: A Modelling Study. *Infectious Diseases of Poverty*, **9(24),** https://doi.org/10.1186/s40249-020-00640-3.

[18] Keeling, M.J. and Ross, J.V. (2008). On Methods for Studying Stochastic Disease Dynamics. *Journal of the Royal Society Interface* **5,** 171–181.

[19] Moss, R., Wood, J., Brown, D., Shearer, F., Black, A.J., Cheng, A.C., McCaw, J.M. and McVernon, J. (2020). Modelling the Impact of COVID-19 in Australia to Inform Transmission Reducing Measures and Health System Preparedness. https://doi.org/10.1101/2020.04.07.20056184.

[20] Olson, C.A., Beard, K.H., Koons, D.N. and Pitt, W.C. (2012). Detection Probabilities of Two Introduced Frogs in Hawaii: Implications for Assessing non-Native Species Distributions. *Biological Invasions* **14,** 889–900.

[21] Renshaw, E. (1993). *Modelling Biological Populations in Space and Time*. Cambridge University Press, Cambridge.

[22] Small, M. and Cavanagh, D. (2020). Modelling Strong Control Measures for Epidemic Propagation with Networks – A COVID-19 Case Study. arXiv:2004.10396v2.

[23] Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society* **58(1),** 267–288.

[24] Vaseghi, S.V. (2009). *Advanced Digital Signal Processing and Noise Reduction*. 4th Edition, Wiley, United Kingdom.

[25] Wu, Y., Chen, C. and Chan, Y. (2020). The Outbreak of COVID-19: An Overview. *Journal of the Chinese Medical Association* **83,** 217–220.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A Appendix: Background

In this appendix, we present a brief overview of the underlying stochastic process used for modeling in this paper. An extensive discussion can be found in [3, 4].

A.1 Partially-observable Continuous-time Markov Population Process

Suppose $\{X_t,\, t \geq 0\}$ is a continuous-time Markov population process with the unknown parameter vector $\boldsymbol{\theta}_t$. The vector $\boldsymbol{\theta}_t$ parameterizes the q-matrix (generator) $Q(\boldsymbol{\theta}_t)$ of the model. We restrict our attention to CTMPPs where the range of the random variable X_t includes non-negative integers, and the initial value of this process, x_0, is known. Moreover, we suppose that the process is time-homogeneous, that is the conditional probability $\mathsf{P}_{(X_{t_2}|X_{t_1})}(x_{t_2}|x_{t_1})$ for any values of $t_2 > t_1 \geq 0$ depends only on x_{t_1}, x_{t_2} and $t_2 - t_1$.

In order to estimate the unknown parameter vector $\boldsymbol{\theta}_t$, we take n observations of $\{X_t,\, t \geq 0\}$ at times $0 < t_1 \leq t_2 \leq \cdots \leq t_n$. Suppose that at each observation time t_i, we do not observe X_{t_i} directly, but rather only a random sample. This may be due to practical restrictions such as time or budget constraints which limit the ability to survey comprehensively, or might be because of an implicit component of the data collection process. A common model for the sampling is binomial, where the state of the system, or each component of the system, is observed with a probability p_t at observation time t. Definition 3 provides a formal definition of a *partially-observable continuous-time Markov population process*.

Definition 3 ([3]). *Consider the* CTMPP *$\{X_t,\, t \geq 0\}$ with the parameter vector $\boldsymbol{\theta}_t$. Suppose the random variables Y_t are defined such that the conditional random variable $(Y_t|X_t = x_t)$ follows the* Binomial*(x_t, p_t) distribution, that is*

$$\mathsf{P}_{(Y_t|X_t)}(y_t|x_t) = \binom{x_t}{y_t} p_t^{y_t}(1-p_t)^{x_t - y_t} \quad \text{for } y_t = 0, 1, \ldots, x_t\,.$$

Then the stochastic process $\{Y_t,\, t \geq 0\}$ is called a PO-CTMPP *with the parameter vector $(\boldsymbol{\theta}_t, p_t)$.*

Remark 9. It is readily seen that a PO-CTMP model with parameter vector $(\boldsymbol{\theta}_t, 1)$ reduces to a CTMP model with parameter vector $\boldsymbol{\theta}_t$.

In order to find the MLE of the unknown parameter vector $(\boldsymbol{\theta}_t, p_t)$, we first need to construct the likelihood function of partial observations, that is,

$$\mathsf{L}_{\boldsymbol{Y}_n}(\boldsymbol{y}_n; \boldsymbol{\theta}_t, \boldsymbol{p}_n) = \mathbf{Pr}(\boldsymbol{Y}_n = \boldsymbol{y}_n),$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where the random vector $\boldsymbol{Y}_n := (Y_0, Y_{t_1}, Y_{t_2}, \cdots, Y_{t_n})$, the realization vector $\boldsymbol{y}_n := (x_0, y_{t_1}, y_{t_2}, \cdots, y_{t_n})$, the probability vector $\boldsymbol{p}_n := (1, p_{t_1}, p_{t_2}, \cdots, p_{t_n})$, and $\mathbf{Pr}(Y_0 = x_0) = 1$.

Bean et al. [3] utilized the Conditional Bayes' Theorem [10] and derived the following analytical results:

Theorem 1 ([3]). *Consider a* `PO-CTMP` *process with the parameter vector* $(\boldsymbol{\theta}_t, p_t)$.

(i) The conditional p.m.f. of the true value of the underlying process given the partial observations is

$$\mathsf{P}_{(X_{t_n}|\boldsymbol{Y}_n)}(x_{t_n}|\boldsymbol{y}_n) = \frac{\varrho_n^{x_{t_n}}}{\sum_{\ell=y_{t_n}}^{\infty} \varrho_n^{\ell}} \quad \text{for } x_{t_n} = y_{t_n}, y_{t_n}+1, \ldots,$$

where,

$$\varrho_n^{\ell} := e y_{t_n}! \binom{\ell}{y_{t_n}} p_{t_n}^{y_{t_n}} (1-p_{t_n})^{\ell - y_{t_n}} \sum_{j=y_{t_{n-1}}}^{\infty} \mathsf{P}_{(X_{t_n}|X_{t_{n-1}})}(\ell|j)\varrho_{n-1}^{j},$$

for $\ell = y_{t_n}, y_{t_n}+1, \ldots$, $n = 1, 2, \ldots$, *and the initial conditions* $\varrho_0^{x_0} = 1$ *and* $\varrho_0^{\ell} = 0$ *for* $\ell \neq x_0$.

(ii) The conditional p.m.f. $\mathsf{P}_{(Y_{t_{n+1}}|\boldsymbol{Y}_n)}(y_{t_{n+1}}|\boldsymbol{y}_n)$ *for* $y_{t_{n+1}} = 0, 1, 2, \ldots$ *equals to*

$$\frac{\sum_{x_{t_{n+1}}=y_{t_{n+1}}}^{\infty} \sum_{x_{t_n}=y_{t_n}}^{\infty} \binom{x_{t_{n+1}}}{y_{t_{n+1}}} p_{t_{n+1}}^{y_{t_{n+1}}} (1-p_{t_{n+1}})^{x_{t_{n+1}}-y_{t_{n+1}}} \mathsf{P}_{(X_{t_n}|X_{t_{n-1}})}(x_{t_n}|x_{t_{n-1}}) \varrho_n^{x_{t_n}}}{\sum_{\ell=y_{t_n}}^{\infty} \varrho_n^{\ell}},$$

for $n = 1, 2, \ldots$.

A.2 Partially-observable Pure Birth Process

A popular model in the class of `CTMPP` is the stochastic *pure birth process* (`PBP`). Let $\{X_t, t \geq 0\}$ be a time-homogeneous `PBP`, with the parameter λ_t (known as the birth/growth rate) at time t, and known initial population size of x_0. If $X_t = x_t$, then the *transition rate* equals to $\lambda_t x_t$. It can be shown [21] that if the birth rate over a given time interval $[t_1, t_2]$ does not vary and equals to λ_{t_1}, then the transition probability at times $0 \leq t_1 \leq t_2$ is given by

$$\mathsf{P}_{(X_{t_2}|X_{t_1})}(x_{t_2}|x_{t_1}) = \binom{x_{t_2}-1}{x_{t_1}-1} e^{-\lambda_{t_1}(t_2-t_1)x_{t_1}} (1-e^{-\lambda_{t_1}(t_2-t_1)})^{x_{t_2}-x_{t_1}} \quad \text{for } x_{t_2} = x_{t_1}, x_{t_1}+1, \ldots.$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Let the stochastic process $\{Y_t,\ t \geq 0\}$ be the corresponding *partially-observable pure birth process*(PO-PBP), with the parameter vector (λ_t, p_t). Bean et al. [3] simplified Theorem Theorem 1 for a PO-PBP, as provided in Corollary 1.

Corollary 1 ([3]). *Consider a PO-PBP $\{Y_t,\ t \geq 0\}$ with the parameter vector (λ_t, p_t), and the underlying PBP $\{X_t,\ t \geq 0\}$ with the known initial population size of x_0.*

(i) The quantity ϱ_n^ℓ for $\ell = y_{t_n}, y_{t_n}+1, \ldots,$ and $n = 1, 2, \ldots,$ is given by

$$e y_n! \binom{\ell}{y_n} p_{t_n}^{y_n} (1-p_{t_n})^{\ell - y_n} \sum_{j=\overline{x}_{t_{n-1}}}^{\ell} \binom{\ell-1}{j-1} e^{-\lambda_{t_{n-1}}(t_n - t_{n-1})j} \left(1 - e^{-\lambda_{t_{n-1}}(t_n - t_{n-1})}\right)^{\ell - j} \varrho_{n-1}^j,$$

where $\overline{x}_{t_n} := \max\{x_0, y_{t_1}, \cdots, y_{t_n}\}$. The initial conditions are as provided in Theorem 1.

(ii) The conditional p.m.f. of the next partial observation, given all past n partial observations equals to

$$\mathsf{P}_{(Y_{t_{n+1}}|\boldsymbol{Y}_n)}(y_{t_{n+1}}|\boldsymbol{y}_n) = \frac{1}{\sum_{\ell=\overline{x}_{t_n}}^{\infty} \varrho_n^\ell} \left(\sum_{x_{t_{n+1}}=\overline{x}_{t_{n+1}}}^{\infty} \sum_{x_{t_n}=\overline{x}_{t_n}}^{x_{t_{n+1}}} \binom{x_{t_{n+1}}}{y_{t_{n+1}}} p_{t_{n+1}}^{y_{t_{n+1}}} (1-p_{t_{n+1}})^{x_{t_{n+1}} - y_{t_{n+1}}} \right.$$
$$\left. \times \binom{x_{t_{n+1}}-1}{x_{t_n}-1} e^{-\lambda_{t_n}(t_{n+1}-t_n)x_{t_n}} \left(1 - e^{-\lambda_{t_n}(t_{n+1}-t_n)}\right)^{x_{t_{n+1}} - x_{t_n}} \varrho_n^{x_{t_n}} \right),$$

for $y_{t_{n+1}} = 0, 1, 2, \ldots,$ and $n = 1, 2, \ldots$.

An important question that may arise here is the dependency structure of the stochastic process $\{Y_t,\ t \geq 0\}$ which is addressed in $t \in (0, \infty)$. Theorem 2.

Theorem 2 ([3]). *The PO-CTMP process is not Markovian of any order. That is, for any fixed value of $k = 1, 2, \ldots,$ there exist $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$, $y_1, y_2, \cdots, y_n$, and $n > k$, such that,*

$$\mathbf{Pr}(Y_{t_n} = y_{t_n} | Y_{t_1} = y_{t_1}, \cdots, Y_{t_{n-1}} = y_{t_{n-1}}) \neq \mathbf{Pr}(Y_{t_n} = y_{t_n} | Y_{t_{n-k}} = y_{t_{n-k}}, \cdots, Y_{t_{n-1}} = y_{t_{n-1}}).$$

Likelihood function. Although, Theorem 2 makes finding the likelihood function of a PO-PBP more challenging and complicated, one can use the chain rule along with Corollary 1 to construct the likelihood function:

$$\mathsf{L}_{\boldsymbol{Y}_n}(\boldsymbol{y}_n; \boldsymbol{\lambda}_n, \boldsymbol{p}_n) = \prod_{k=1}^{n} \mathsf{P}_{(Y_{t_k}|\boldsymbol{Y}_{k-1})}(y_{t_k}|\boldsymbol{y}_{k-1}), \tag{1}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.05.25.20112482; this version posted May 26, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where $\boldsymbol{\lambda_n} := (\lambda_0, \lambda_{t_1}, \cdots, \lambda_{t_n})$. Now, by having the likelihood function at hand, one can find the MLE of unknown parameters for a PO-PBP. However, there are some infinite sums involved with the likelihood function which should be handled carefully in numerical computations. One approach to deal with those infinite sums is to truncate them by exploiting Chebyshev's inequality. More precisely, Chebyshev's inequality prescribes to truncate the infinite sum over the realizations of the conditional random variable $(X_{t_n} \mid \boldsymbol{Y}_n = \boldsymbol{y}_n)$ at

$$\mathbb{E}\left[X_{t_n} \mid \boldsymbol{Y}_n = \boldsymbol{y}_n\right] + 20\sqrt{\mathsf{Var}(X_{t_n} \mid \boldsymbol{Y}_n = \boldsymbol{y}_n)}\,, \tag{2}$$

to guarantee that at least 99.75% of the corresponding probability distribution is covered. Been et al. [4] derived those expected values involved in the truncation point (2) analytically.

Proposition 1 ([4]). *Consider a PO-PBP $\{Y_t,\ t \geq 0\}$ with the parameter vector (λ_t, p_t), and the underlying PBP $\{X_t,\ t \geq 0\}$. We have,*

$$\mathbb{E}\left[X_{t_n} \mid \boldsymbol{Y}_n = \boldsymbol{y}_n\right] = \frac{\overline{x}_{t_n} + (1-p_{t_n})(1-e^{-\lambda_{t_n} t_n})}{p_{t_n} + (1-p_{t_n})e^{-\lambda_{t_n} t_n}},$$

$$\mathsf{Var}(X_{t_n} \mid \boldsymbol{Y}_n = \boldsymbol{y}_n) = \frac{(\overline{x}_{t_n} + 1)(1-p_{t_n})(1-e^{-\lambda_{t_n} t_n})}{(p_{t_n} + (1-p_{t_n})e^{-\lambda_{t_n} t_n})^2},$$

where $\overline{x}_{t_n}$ is as defined in Corollary 1.

Prediction. In order to predict the future values of the process given the past partial observations, we use the MLE of the conditional expected value $\mathbb{E}\left[Y_{t_{n+1}} \mid \boldsymbol{Y}_n = \boldsymbol{y}_n\right]$. Due to the invariant property of MLEs, we only need to find the MLE of the unknown parameters λ_t and p_t and replace them in the equation provided in Proposition 2.

Proposition 2 ([4]). *Consider a PO-PBP $\{Y_t,\ t \geq 0\}$ with the parameter vector (λ_t, p_t), and the underlying PBP $\{X_t,\ t \geq 0\}$. We have,*

$$\mathbb{E}\left[Y_{t_{n+1}} \mid \boldsymbol{Y}_n = \boldsymbol{y}_n\right] = p_{t_{n+1}} e^{\lambda_{t_n}(t_{n+1}-t_n)} \mathbb{E}\left[X_{t_n} \mid \boldsymbol{Y}_n = \boldsymbol{y}_n\right],$$

where $\mathbb{E}\left[X_{t_n} \mid \boldsymbol{Y}_n = \boldsymbol{y}_n\right]$ is as given in Proposition 1.