Applying Branching Process Theory to Model COVID-19 Spread

Zachary Moore
Roll: *Student*
Class: *Probability and Stochastic Processes*
Session: *VL2*
Email: *zmoore3@jh.edu*

Course: *625.721* Submission date: *12/13/2021*

Contents

Introduction and Branching Processes

Branching processes have many applications but are most typically used in the modeling of reproductive processes. In the simplest version of a branching process, starting at generation 0 there is 1 ancestor who reproduces according to a certain probability distribution [1]. Extensions to branching processes with multiple types of children have been made, namely in [2]. In this paper, the authors discuss the ability to predict the different birth distributions of the children as well as the parameters that define them distinctly. After studying this paper, I went down the path of potentially estimating the parameters of a generating distribution of COVID propagation throughout various U.S. counties. To do this I assumed that the spread of COVID could be modeled using a branching process where a "parent" infected and contagious thus can spread the virus and create children according to a probability distribution like in [3]. I will explain the data set and method used later in the paper.

We can define the branching process described above with random variables in turn define their probability generating functions. Doing this allows us to derive a probability of total extinction for the population. Put mathematically:

1. Let $X(n) = \#$ of children in generation n. $X(0) = 1$

2. Let Y be the random variable denoting the number of children obtained (infected).

3. Then, $X(1)$ is the number of children created by the ancestor and $X(1) \stackrel{d}{=} Y$

4. Because $X(2) = Y_1 + ... + Y_{X_{(1)}}$ we can define the generating function for the n^{th} generation to be:
$$g_n(t) = g_{n-1}(g(t)) \tag{1}$$

   where $g(t)$ is the probability generating function for the random variable Y.

Using $g(t)$, and according to Theorem 7.3 in [4], the probability of total extinction η is smallest non negative root of solution to $t = g(t)$.

With parents and children that generate offspring independently and identically (iid), according to (Gut CITATION) we can determine the expected value and variance of the number of children created at each generation only by understanding the probability distribution that they are generated with. From Theorem 3.7.1 in [4]:

Assuming $\mu = E[Y_1] < \infty$

$$E[X(n)] = \mu^n \tag{2}$$

and the $Var[X(n)]$:
$$= \sigma^2(\mu^{n-1} + \mu^n + ... + \mu^{2n-2}) \tag{3}$$

Because each generation reproduces independently of one another, we can also get the total expected population at generation n:

$$E[X(0) + ...X(n)] = E[X(0)] + ... + E[X(n)] = 1 + \mu + \mu^2 + ... + \mu^n \tag{4}$$

In the following sections I intend to test these theoretic moments for branching processes through simulation as well as attempt to estimate the parameters of the probability distribution that generated the case counts through an application of time series fitting. I originally thought of applying the MCMC method for parameter estimation but I do not believe that is a viable option for this case based on the way that counts are generated by multiple children, not a singular process like the daily text message data shown in the first chapters of the Bayesian Inference for Hackers book.

COVID-19

Modeling the spread of COVID-19 has been at the forefront of statistical application and the use of branching processes has been attempted by many ([5],[6]). Using the JHU daily COVID case reporting data for the U.S. [7], I intend to investigate the spread of COVID over 20 randomly selected counties. The data consists of dates ranging from January 2020 to current time, county and state names, among other attributes. Below is a screenshot of the cleaned data table used for analysis:

To make the work reproducible, I have set a random seed in the attached Jupyter notebook when randomly selecting the counties from the dataset. Below is a plot showing the case counts over the period of January 1st, 2020 - December 31st, 2020 for each county mentioned:

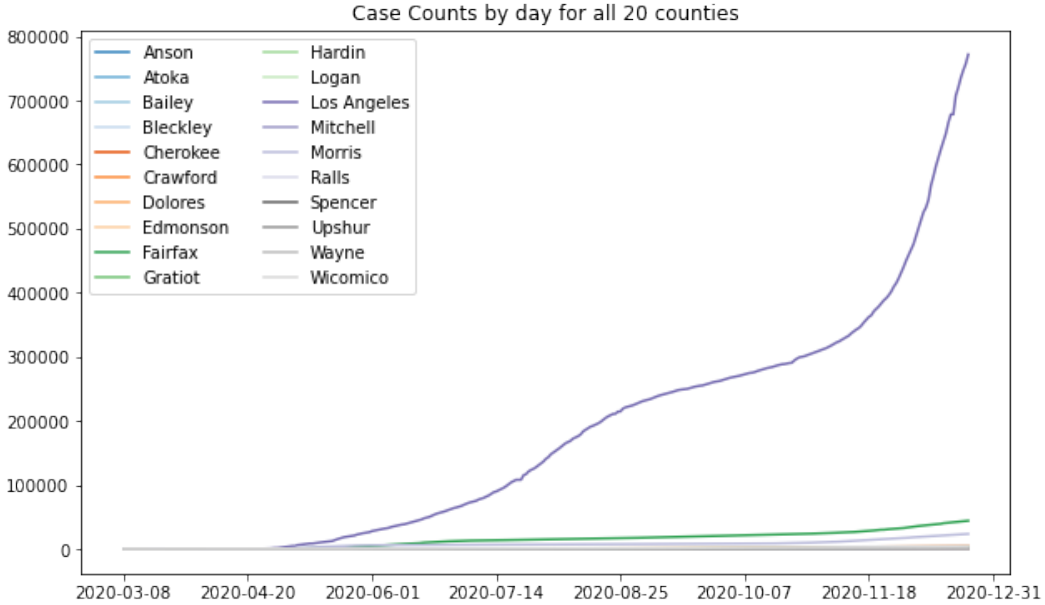
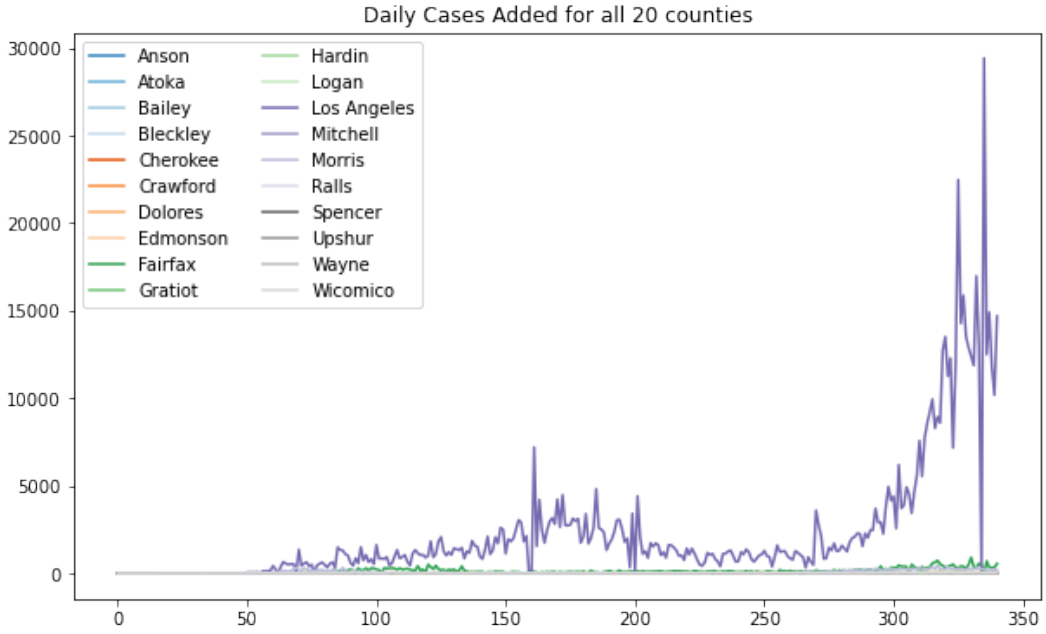


Figure 1 shows that a reasonable assumption can be made that the generating functions across counties are not identical, so the question then becomes, can a number for

the probability of infection be found for each county? In order to do this, we would need to have insight into the daily cases added by day for each county. That plot is shown below:



Looking above, the large majority of counties don't come close to the daily spread that we see in Los Angeles County (the most populous county in the data set). The maximum daily case count for LA is 29,423! The maximum case count for Dolores County, CO is only 5 cases. This plot shows us emphatically that the probability distribution across counties is not identical. Knowing this, we need a way to estimate the parameters for each distribution, more on that in the following sections.

Branching Process Simulation

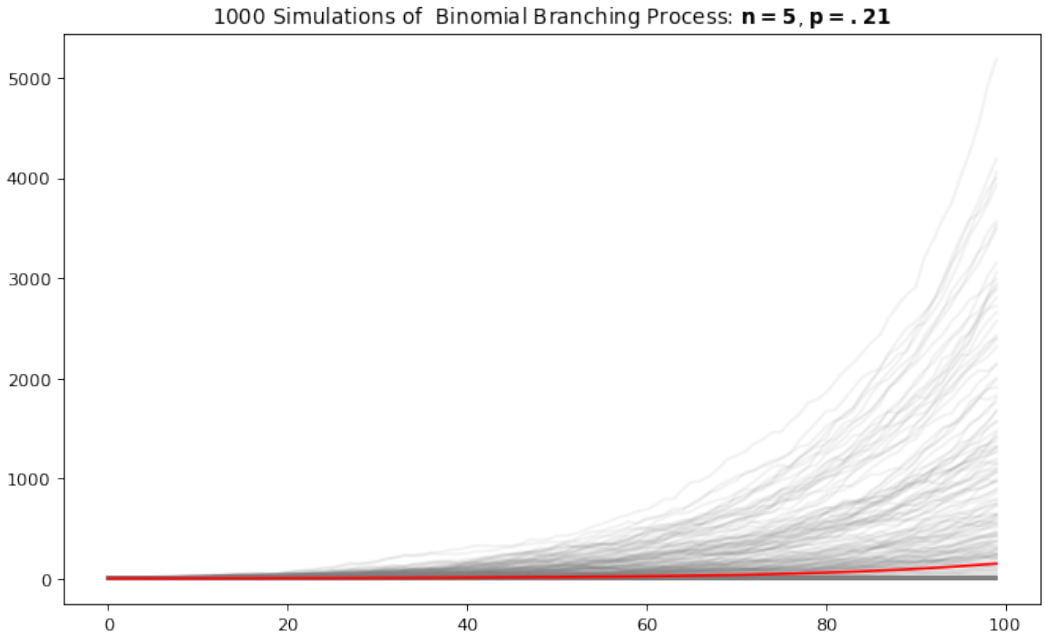
For this branching process, I assume that an infected individual passes COVID to uninfected people with binomial distribution. Here, n is the number of encounters a contagious individual has during that generation and p is probability of passing on the virus. Unlike in [3], I do not give an infected individual a time window.

Using the data above, it seems pretty clear that COVID spread is not identical throughout different counties. For example, in Dolores County, CO (Pop. 1,841) there were only 3 recorded cases through July whereas in Fairfax County (Pop. 1.1M) there were xxx cases. Is this a case of prolonged interactions increasing the probability of spread? Or is it more a case of higher interactions (trials) leading to more spread. For modeling purposes, I will use the first 100 days (generations) for each county's branching process simulation. Unfortunately, it is impossible know just how many people were infected at the time the first cases were reported for each county, so to start the branching process for each county at generation 0 I will use the first date that a case was reported as the initial ancestor and propagate from there. The branching process simulation code is in the supplied Jupyter Notebook [8].

Starting at generation 0, the ancestor produces k random children according to the binomial density [9]:

$$P(Y = k) = \binom{5}{k} (.21)^k (1 - .21)^{n-k}$$

I chose the values for n, p arbitrarily and with an eye to reasonably bound the $E[X(n)]$ given the observed case counts. At each subsequent generation, every child produces children identically. These counts are stored and then summed at each generation. There is a variance amongst the simulations and for 1000 simulations at a set n, p the results are shown below for the branching process defined with the probability function above:



For the figure we see the 1000 simulations in gray and the empirical mean at each generation n in dark red. Using this simulation structure, we can apply it to each county of interest and use the 100^{th} day's number of cases as the expected value of the branching process. This looks like:

$$\mu_{county} = \#\ cases\ recorded\ on\ 100^{th}\ day$$

Using that mu, and sticking with our binomial distribution with $n = 5$. We can find the expected value of the distribution for each county with:

$$E[X_{county}(n)] = (\mu_{county})^{1/100} \tag{5}$$

Now, using Eq.5n 5 we can find the probability of success, p for each county with:

$$E[X_{county}(n)] = 5 * p_{county}$$

$$p_{county} = \frac{E[X_{county}(n)]}{5} \tag{6}$$

Using the above, we can simulate the branching process for each county and compare that to the actual case recordings. The results of that are shown below:

It is clear from the figure that a major issue occurs when the 100th day has no recorded cases, not unexpected if μ_{county} is zero. Conversely, we can see that for some counties, both the theoretical and empirical spread do a decent job of fitting the data.

In the next section, I will attempt to experiment with fitting a time-series in hopes that I can arrive at a more clear parameter estimate for each county. Thus allowing for a "training" set using n generations that would allow decision-makers to predict spread $n+1, n+2..., n+m$ generations into the future.

Time Series Analysis

Can we instead, since we don't know the true spread and distribution for modeling, use [10] to show that if we can guess an expected value, the number of cases generated will be "independent of mixed distributions." Using this, we can dive into estimating the spread with an underlying branching process with expected value μ_{county}. The models to be tested will be:

$$\hat{y}_{n+1} = (\mu_{county})^{n+1} \tag{7}$$

where , in addition to just using the number of recorded cases on the nth day:

$$\mu_{county} = (\frac{y_1 + ... + y_n}{n})^{1/n} \tag{8}$$

and

$$\mu_{county} = (\frac{y_n + ... + y_{n-7}}{7})^{1/n} \tag{9}$$

which Eq. 8 is just the average of daily cases (and assuming a binomial distribution generates these cases this is the Maximum Likelihood Estimator used to generate p for the distribution) and Eq. 9 is the 7-day average for COVID-19 cases. To prove that the average daily cases method is the MLE for p, we need to show that in this method:

$$\hat{p} = \frac{\bar{y}}{N}$$

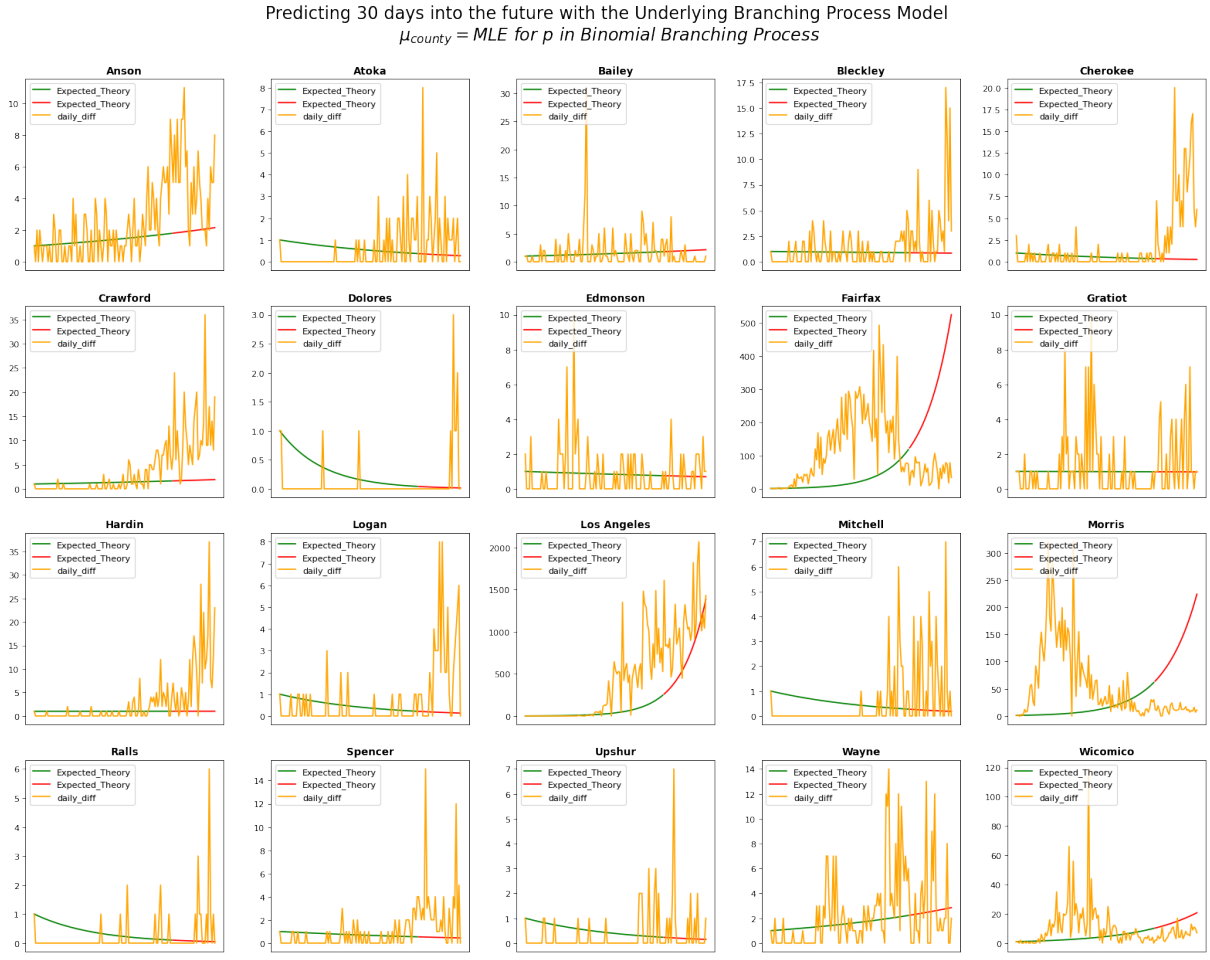
where $N = 5$.

This is shown with:

$$\mu^n_{county} = \frac{y_1 + ... + y_n}{n} = Np_{county}$$

$$p_{county} = \frac{y_1 + ... + y_n}{nN} = \frac{\bar{y}}{N} \tag{10}$$

Eq. 10 is just the MLE for p of a random variable that follows a binomial distribution. This proves that using the mean number of daily cases for μ_{county} is equivalent to using the MLE for p_{county}. Below, the figure shows the predictions for 30 days out using this MLE method for each p_{county} in the Binomial Branching Process:



While the above shows some decent fits for certain counties with the training data, this model form does not perform very well. The table below shows the results for each branching process model's RMSE for predictions 30 days into the future:

	nth_rmse	MLE_rmse	seven_rmse
Anson	18.164562	22.092220	21.807663
Atoka	11.789826	10.804891	9.427676
Bailey	9.380832	11.479057	11.130030
Bleckley	30.248967	27.604929	23.493957
Cherokee	46.270941	45.051350	44.708820
Crawford	39.658073	66.226406	38.257778
Dolores	4.000000	3.966754	4.000000
Edmonson	7.348469	5.682933	6.990172
Fairfax	257.452194	1426.873389	411.677448
Gratiot	11.747340	11.794476	13.855355
Hardin	43.593161	67.297108	55.581049
Logan	14.560220	17.337904	16.568952
Los Angeles	30776.804103	2722.850212	14251.755756
Mitchell	12.845233	12.141039	10.676935
Morris	48.197237	697.726547	109.523982
Ralls	7.141428	7.025707	6.968209
Spencer	21.546688	21.594244	17.692863
Upshur	8.185353	7.787499	7.399533
Wayne	19.475568	19.031131	37.355798
Wicomico	30.887625	51.067180	27.747033

In the table, there is no clear branching process model that outperforms the others when predicting new case counts. Specifically, for LA, the MLE model performs the best by far compared to the other 2 models but when looking at Wicomico county it is the worst performing model. This can most likely be attributed to measures being put in place to curb the spread of COVID-19 as well as the lack of types in the branching process model.

When looking to fit a time-series dataset and make predictions, it is typical to use basic methods before moving to more advanced models to baseline the measures of performance for the models. In [11], some methods are outlined and I will compare those to the BP models already explained in this paper. The 3 models I will look into are the Mean method, Naive method, and Drift method.

The Mean method just uses the average of all observations up to time t and predicts for time $t + h$ to be equal to the average. The Naive method just uses the observation at time t to make predictions for time $t + h$. The drift method takes the difference between the first observation and the t^{th} observation to make predictions for time $t + h$. Mathematically that looks like:

$$\hat{y}_{t+h} = h * (\frac{y_t - y_1}{t - 1})$$

The results of these methods and their RMSE have been summarised for each county in the table below:

	nth_rmse	MLE_rmse	seven_rmse	Mean_RMSE	Naive_RMSE	Drift_RMSE
Anson	18.164562	22.092220	21.807663	22.139828	13.856406	369.380563
Atoka	11.789826	10.804891	9.427676	10.688639	11.789826	103.556748
Bailey	9.380832	11.479057	11.130030	10.835266	9.380832	10.295630
Bleckley	30.248967	27.604929	23.493957	27.445072	30.099834	120.569482
Cherokee	46.270941	45.051350	44.708820	44.424937	45.880279	332.258935
Crawford	39.658073	66.226406	38.257778	65.004615	44.609416	528.151493
Dolores	4.000000	3.966754	4.000000	3.925303	4.000000	99.503769
Edmonson	7.348469	5.682933	6.990172	5.685772	7.280110	199.150697
Fairfax	257.452194	1426.873389	411.677448	457.860669	277.151583	651.369327
Gratiot	11.747340	11.794476	13.855355	11.788639	11.747340	14.696938
Hardin	43.593161	67.297108	55.581049	64.118445	47.696960	528.820385
Logan	14.560220	17.337904	16.568952	17.198837	14.525839	17.972201
Los Angeles	30776.804103	2722.850212	14251.755756	4859.523421	3604.751864	151066.160400
Mitchell	12.845233	12.141039	10.676935	11.998625	12.845233	105.099952
Morris	48.197237	697.726547	109.523982	305.144215	31.527766	1012.393698
Ralls	7.141428	7.025707	6.968209	6.949820	7.141428	101.123687
Spencer	21.546688	21.594244	17.692863	21.851773	18.275667	279.204226
Upshur	8.185353	7.787499	7.399533	7.701104	8.124038	99.784768
Wayne	19.475568	19.031131	37.355798	19.232135	19.026298	181.592951
Wicomico	30.887625	51.067180	27.747033	31.148788	31.575307	65.375837

Surprisingly, the BP models used to predict COVD spread with different μ_{county} fits do outperform the baseline time-series methods. Perhaps going forward these models would provide a better baseline for certain counties when looking to implement more sophisticated time-series modeling methods to predict COVID-19 outbreaks. These results can be interpreted as a way for some decision-maker to look at expected propagation if there was no attempt to stop the spread. Maybe in some cases there would be no need to impact local economies and businesses if the predicted spread is inconsequential with no impacts in a county with low population. Unfortunately, the models implemented do a pretty poor job overall of predicting the number of cases that were realized for the counties in question which leads me to the conclusion that more advanced time series models are required. This conclusion is unsurprising but it was still fun to investigate the performance of BP models compared to basic forecasting methods.

Future Work

My main recommendation for future work would be that of implementing the more advanced time series models to predict COVID-19 spread. There doesn't seem to be much seasonality in the dataset and we could look to apply auto-regressive techniques like ARMA and ARIMA models to see how they compare to the BP models as well. Additionally, multi-type branching processes could lead to better modeling of COVID-19 spreading in places where a single generating function is not sufficient or realistic.

References

[1] Wikipedia. *Branching Process*. URL: https://https://en.wikipedia.org/wiki/Branching_process.

[2] Benoît Azaïs Romain Henry. "Maximum likelihood estimation for spinal-structured trees". In: *Annalen der Physik* (2021). DOI: arXiv:2101.05099v1.

[3] Jérôme et al. Levesque. "A model of COVID-19 propagation based on a gamma subordinated negative binomial branching process." In: *Journal of theoretical biology* 512.110536 (2021). DOI: doi:10.1016/j.jtbi.2020.110536.

[4] Allan Gut. *An Intermediate Course in Probability*. Springer. Springer Science, 2009. ISBN: 9781489984463.

[5] Robin N Thompson. "Novel Coronavirus Outbreak in Wuhan, China, 2020: Intense Surveillance Is Vital for Preventing Sustained Transmission in New Locations." In: *Journal of clinical medicine* 9.498 (2020). DOI: doi:10.3390/jcm9020498.

[6] Andrea et al. Bertozzi. "The challenges of modeling and forecasting the spread of COVID-19." In: *Proceedings of the National Academy of Sciences of the United States of America* 1.498 (2020), p. 7. DOI: https://www.pnas.org/content/pnas/117/29/16732.full.pdf.

[7] JHU. *COVID-19 U.S. County JHU Data Demographics*. URL: https://www.kaggle.com/headsortails/covid19-us-county-jhu-data-demographics.

[8] Zachary Moore. *Modeling COVID-19 Propagation Using A Branching Process Model*. URL: https://https://github.com/zchmoore/COVID-Project.

[9] Wikipedia. *Binomial Distribution*. URL: https://en.wikipedia.org/wiki/Binomial_distribution.

[10] Moradi Mojtaba. Vajargah Behrouz Fathi. "Simulation Branching Processes by Mixing Distributions." In: *Journal of Applied Mathematics, Statistics, and Informatics* 8.1 (2012), p. 4.

[11] Athanasopoulos George Hyndman Rob J. *Forecasting: Principles and Practice*. O-Texts. O-Texts, 2018. ISBN: 9780987507112.