 doi:10.20944/preprints202005.0319.v1

STATISTICAL MODELING OF COVID-19 PANDEMIC STAGES WORLDWIDE

Ranadeep Datta[1], Prabhat Kumar Trivedi[1], Ayush Kumawat[1], Rahul Kumar[1], Isha Bhradwaj[1], Neha Kumari[1] Varun Agiwal[1,2], Saurabh Kumar[1,3], Ashok Kumar[1,4] and Ashutosh Shukla[5], and Jitendra Kumar[1,*]

[1]*Central University of Rajasthan, Bandarsindri, Kishangarh, Ajmer, Rajasthan, India*
[2]*Department of Community Medicine, Jawaharlal Nehru Medical College, Ajmer, Rajasthan*
[3]*Invertis University Bareilly, Uttar Pradesh, India*
[4]*SHKM Government Medical College, Nuh, Hariyana*
[5]*National Statistical Office, Field and Operationn Division, Agra, Uttar Pradesh, India*

ABSTRACT

COVID-19 is an infectious disease, growth of which depends upon the linked stages of the epidemic, the average number of people one person can infect and the time it takes for those people to become infectious themselves. We have studied the COVID-19 time series to understand the growth behaviour of COVID-19 cases series. A structural break occurs in the COVID-19 series at the change time form one stage to another. We have performed the structural break analysis of data available for 207 countries till April 20, 2020. There are 42 countries which have recorded five breaks in COVID cases series. This means that these countries are in the sixth stage of growth transmission and show a downward pattern in reporting in the daily cases, whereas countries with two and three breaks, record the rapid growth pattern in the daily cases. From this study, we conclude that the more the breaks in the series, there is more possibility to determine the constant or decreasing rate of daily cases. It is well fitted using lognormal distribution as this distribution is archived at its highest peak after some period and then suddenly it decreases at a longer time period. This can be seen in various countries like China, Australia, New Zealand and so on.

Keywords: COVID-19, Structural Break, Growth Stage

*Corresponding author at: Department of Statistics, Central University of Rajasthan, Bandersindri, Ajmer, India. Email addresses: vjitendrav@gmail.com

 doi:10.20944/preprints202005.0319.v1

INTRODUCTION:

COVID-19 is a novel coronavirus that has travelled from Wuhan, China, to the most of the countries (207) within 100 days. COVID-19 is an infectious disease, growth of which depends upon the stage of the epidemic, the average number of people one person can infect and the time it takes for those people to become infectious themselves. However, the growth rate of coronavirus cases is different in every country depends upon its health infrastructure, persons living life, environmental conditions and many other factors. So, the stages of disease transmission depends upon the changing pattern in the series of the number of COVID cases based on total population, total land area, medical facilities, etc. These changing patterns may be analyzed by a structural break model where shifting in the series can be determined using a change in the growth of spreading of COVID-19. Hence, the structural break can easily explain the shifting on the COVID-19 time series from one breakpoint to another by changing the model parameters. For each break interval, it follows a well-known growth model, and the growth rate of the series is different. Significant contribution in the study of a structural break in time series includes the work of Chow (1960), Nelson and Plosser (1982), Andrews (1993), Bai and Perron (1998), Chaturvedi and Kumar (2007), Bai (2010), Meligkotsidou et al. (2011, 2017) and Agiwal et al. (2018).

In COVID-19 cases, a structural break might occur when most of the population and land area are affected, sudden increments in corona patients on daily basis, population not following the government guidelines or other factors. It also depends upon the administrative model which has the main goal to slow down the growth of COVID-19 cases by various measures like social distancing, lockdown etc. Health services may be provided in a more structural and significant way when such services try to increase the recovery of COVID-19 cases. Most of the countries have managed the disease in reasonable time. However, some countries record the change in the COVID series, i.e., there is a possibility in the series that structural break(s) have happened. So, the present study has analyzed the changing pattern of COVID-19 cases and identified various suitable breakpoints. For that, we have determined the breakpoints using statistical methodology and then examined the changing trend in each break interval. We have analyzed every stage separately to understand the behaviour of transmission on infections. Various distribution models have been fitted in each break interval based on the number of infected days and the number of total cases. Based on the results, we determine the best-fitted model in the overall break interval.

In the analysis of break intervals, the Lognormal distribution fitted better among all the five discrete and continuous distributions under consideration. Similarly, in the analysis of the number of cases at breaks, the Lognormal distribution fitted better than the rest of the distributions, both discrete and continuous, under consideration.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

MATERIAL & METHOD

The data has been collected from our World COVID-19 databases of daily updates of confirmed cases. The data covers the total number of people infected with the COVID-19 virus from December 31, 2019, to April 20, 2020. As of April 20, 2020, there were 2350993 cases of infections worldwide. On April 20, there were a total of 205 countries/provinces infected with the virus. This data was processed for further analysis.

In the present paper, we apply the Chow F-test statistic to determine the potential breaks at all change points in the COVID series. This methodology is well discussed in R package "strucchange", developed by Zeileis et al. (2002). The number of change points (m) is determined by breakpoint (s) function in strucchange that employs the location of the break(s) using the minimum value of the residual sum of squares (RSS) and Bayesian information criterion (BIC) whereas. Various discrete and continuous distributions are fitted for all break intervals to show the development of the COVID19 cases feasibly. The selection of the best-fitted distribution is based on the minimum value of the Akaike information criterion (AIC) and Bayesian information criterion (BIC).

We will do the following analysis step by step:

Step-1: Determine the number of the structural break(s) and its locations.

Step-2: Classify the countries by taking the break interval of World COVID cases.

Step-3: Fit a distribution in each break interval based on total infected days and the total number of cases.

Step-4: Display the histogram and density plot based on the results obtained from the previous steps to conclude the COVID-19 cases.

RESULTS

Out of the total 205 countries infected till April 20, 2020, 10 countries do not have any breaks in their data, 3 countries have only 1 break, 23 countries have 2 breaks, 48 countries have 3 breaks, 79 countries have 4 breaks and remaining 42 countries have five breaks. All breaks identification depends upon the size of the series as well as how much times there is a shift in the series. We can interpret the number of breaks as various growth stages of COVID-19 cases. First breakpoint shows the change in the growth stage of COVID-19 cases at the initial time period. Those countries having a single breakpoint can be considered in the first stage because in these countries the growth of corona cases is increasing slowly. A country can be regarded as in the second and third stages of growth in COVID cases where there is a rapid change in the trend of corona cases. These countries have recorded two or three breakpoints. Some of the countries that are having three breaks are getting a higher peak and record a decreasing trend in the growth of cases. During the analysis, some countries like Australia, Austria, China, Greece, New Zealand and many more record a constant or decreasing growth rate in the COVID-19 cases. These countries show the fourth and fifth stage of growth and have recorded at higher numbers of breaks (four, five) in the series. These breakpoints are identified

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

based on the trend pattern of corona cases. So, we can consider break intervals as the growth stages of COVID cases in our analysis.

Analysis the stages based on duration time:

All infectious disease always depends majorly on two factors, the first number of carriers and the second time of infections. So, in the present section, we study the number of days in various COVID19 infection stages. First, the structural breaks are found for each country over the data of total cumulative cases. Maximum five breaks (B1, B2, B3, B4 and B5) are identified in each country series. Then intervals are formed between all the breakpoints, namely zero stage (S1: 1st day to B1), the first stage (S2: (B2+1) to B3), second stage (S3: B2+1 to B3), the third stage (S4: B3+1 to B4), fourth stage (S5: B4+1 to B5) and fifth stage (S6: (B5+1 to April 20). Here, each country has different number of breaks, and its locations are also different. So, we study the total number of days where there is a change in the series that means infection growth is from one stage to another stage. This means that each break interval for every county records the time to shift from one to another point and it is termed as I-stage, II-Stage, III-Stage, IV-Stage, Stage-V and Stage-VI. Table 1 records the descriptive statistics of the occurrence period of different stages.

Table 1 shows that the shortest duration between any two breaks in the series is four days. The average number of days for any interval is approximately eight days, with an exception in the first interval, which is highest about 20 days. The quartiles for the different intervals show that 75% of the countries have 6 to 25 days before coming into the first stage of the outbreak.

The common dataset of the number of days in the break intervals seems to be positively skewed as recorded in above Figure 1. The later intervals are found to have a smaller number of days in break intervals as compared to the others. The individual break interval datasets have similar asymmetrical plot like the collective interval data, shown in Figure 2. The plots are skewed to the right, and there are fewer countries that have larger break intervals.

Similarly, for the rest of the intervals, most of the countries have about 4 to 9 days before the next breakpoint occurs. The difference in the shape of the first interval and rest of the intervals is that there are more countries in the S1 than the rest, i.e., almost all the countries have observed their first breakpoint based on total number of cases within the country. Likewise, very few countries have observed their sixth stage in the number from the total cases, which leads to a distorted histogram for S6 and less skewed than the other intervals. This also shows a control situation of these countries that observed the sixth stage of COVID-19 infection.

The Cullen and Frey graphs were fitted for each of the intervals and used to find possible candidates for fitting various discrete and continuous distributions to the data. The following distributions are considered: Normal, Negative Binomial, Poisson, Lognormal and Gamma distribution. Table-2 describes the best-fitted distribution based on various assumed distribution. This table shows that lognormal distribution is fitted well in all break intervals except S4.

Tracing the theoretical density and distribution function over the empirical density and distribution are displayed in Figure 3 and 4. These figures give a nice view of the shape of each stage

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

interval of COVID-19 infection and fits well. Points the on P-P plots lie on the straight line representing a better fit, whereas in Q-Q plot departures from the straight line can be observed which increase at the later stages (due to fewer data in later stages). Deviations from the straight line are a hint of lack of fit.

Analysis the number of cases at each stage:

In the analysis of different COVID-19 infection, we also need to understand the COVID-19 cases as the infection in equally crucial for the number of carriers. So far, we modelled COVID-19 cases in different stages of spreading of COVID-19. First, the data consisted of distinctively very high values contributed by countries with a high number of infected cases. These are the countries like the United States, Italy, China, Spain, Germany, Iran, etc., these data points consisted of about 15-19% of the data and has been removed considering outliers which are recorded in Table 3.

The outlier countries are the countries that are worse affected by the disease than the countries in the corresponding stages. It is also recorded that countries which are severely affected by COVID19 (United States, China, etc.), are not having reached in last stages like fifth and sixth. Table 4 records the descriptive statistics of the break interval based on total cases, and Figure 5 displays its density plot.

The aggregate data, after removing outliers, have a positively skewed shape and mildly peaked in the first half of the data. The average number of cases at B0 is 74, which is reasonable as these are moderately affected (after removing outlier) countries and therefore have fewer cases till the first break. This average increased rapidly till B4, reaching approximately 580 cases, where about 75% of the countries have less than 802 cases in total. At break point B5, this average falls. The reason behind this is a possibility to decrease the growth pattern of the COVID-19 cases in these countries. Most of these countries have a high number of infected cases that were pulling the average up at the earlier breaks.

The suitable candidates for distribution fitting were found using Cullen and Frey graphs in Table 5. The lognormal distribution fitted the best among exponential, normal, Gamma and other discrete distributions. The trace of density & distribution function and the P-P plots are shown in Figures 7-8. These figures show that a better fitting is observed using the lognormal distribution as it has achieved its peak and after that, there is a sudden decreasing trend in the extended time period. This can be easily seen in China series. Departures from the straight line in the Q-Q plot at the higher quantiles form the distribution is fitted due to the highly skewed countries data. Hence the methods other than lognormal distribution are not able to thoroughly explain the skewness of the data accurately in the total cases.

DISCUSSION

This paper studies and analyzes the various stages in the growth of COVID-19 cases using the structural break methodology. We considered the cumulative cases of coronavirus for each country as a data series and identified the breaking point when the structure of the series is shifted suddenly. We

 doi:10.20944/preprints202005.0319.v1

have examined the inference based on each break period. This break period provides the length of the duration and number of cases in various growth stages. For each break interval, we fit a distribution to explain the growth pattern of COVID-19 cases. Based on the results, we observed that lognormal distribution is better fitted to the number of days and number of cases in each break interval of COVID19 in the whole world. This is so because lognormal distribution has a long right tail with an exponential growth pattern.

REFERENCES

1. Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. Econometrica, 61(4):821–856.
2. Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. Econometrica, 66(1):47–78.
3. Chaturvedi, A. and Kumar, J. (2007). Bayesian unit root test for time series models with structural breaks. American Journal of Mathematical and Management Sciences, 27(1-2):243–268.
4. Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. Econometrica, 28:591–605.
5. Meligkotsidou, L., Tzavalis, E., and Vrontos, I. (2017). On Bayesian analysis and unit root testing for autoregressive models in the presence of multiple structural breaks. Econometrics and Statistics, 4:70–90.
6. Meligkotsidou, L., Tzavalis, E., and Vrontos, I. D. (2011). A Bayesian analysis of unit roots and structural breaks in the level, trend, and error variance of autoregressive models of economic series. Econometric Reviews, 30(2):208–249.
7. Nelson, C. R. and Plosser, C. R. (1982). Trends and random walks in macroeconomic time series: some evidence and implications. Journal of Monetary Economics, 10(2):139–162.
8. Bai, J. (2010). Common breaks in means and variances for panel data. Journal of Econometrics, 157(1):78–92.
9. Agiwal V., Kumar J. and Shangodoyin D. K. (2018): A Bayesian Inference of Multiple Structural Breaks in Mean and Error Variance in Panel AR(1) Model. Statistics in Transition, 19(1):7-23.
10. Singh, A. K., Singh, A. and Engelhardt, M. (1997): The lognormal distribution in environmental applications. In Technology Support Center Issue Paper.
11. Delignette-Muller, M. L. and Dutang, C. (2015): fitdistrplus: An R package for fitting distributions. Journal of Statistical Software, 64(4):1-34.
12. Stedinger, J. R. (1980): Fitting log normal distributions to hydrologic data. Water Resources Research, 16(3):481-490.

 doi:10.20944/preprints202005.0319.v1

13. Sánchez, S., Ancheyta, J., and McCaffrey, W. C. (2007): Comparison of probability distribution functions for fitting distillation curves of petroleum. Energy & Fuels, 21(5): 2955-2963.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

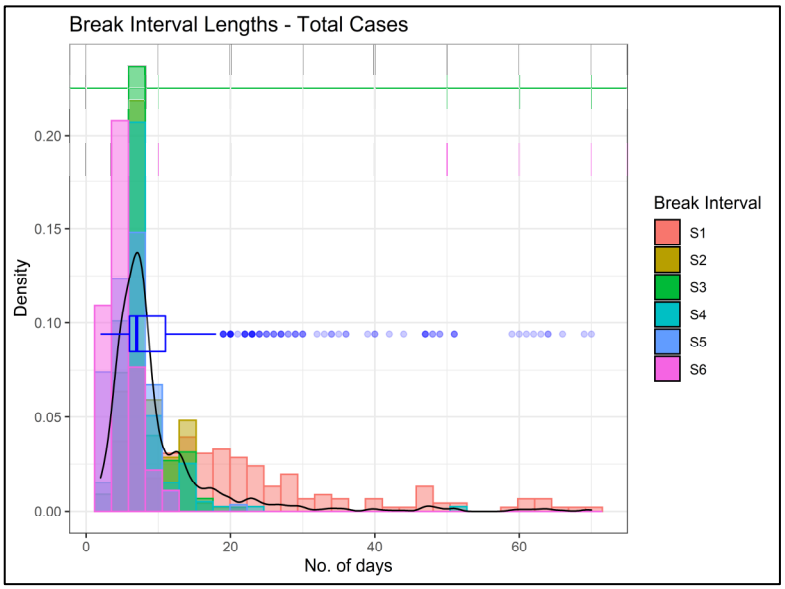

Figure 1: Density plot on overall break intervals

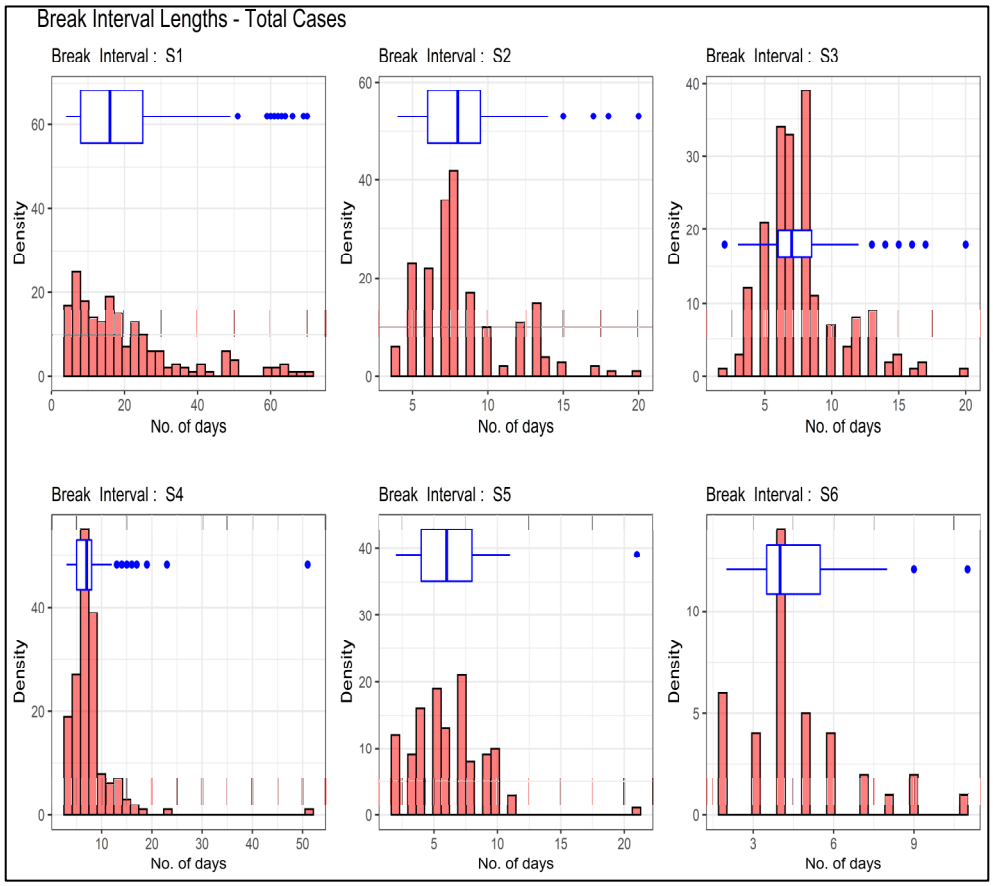

Figure 2: Histogram and box plot in each break interval

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

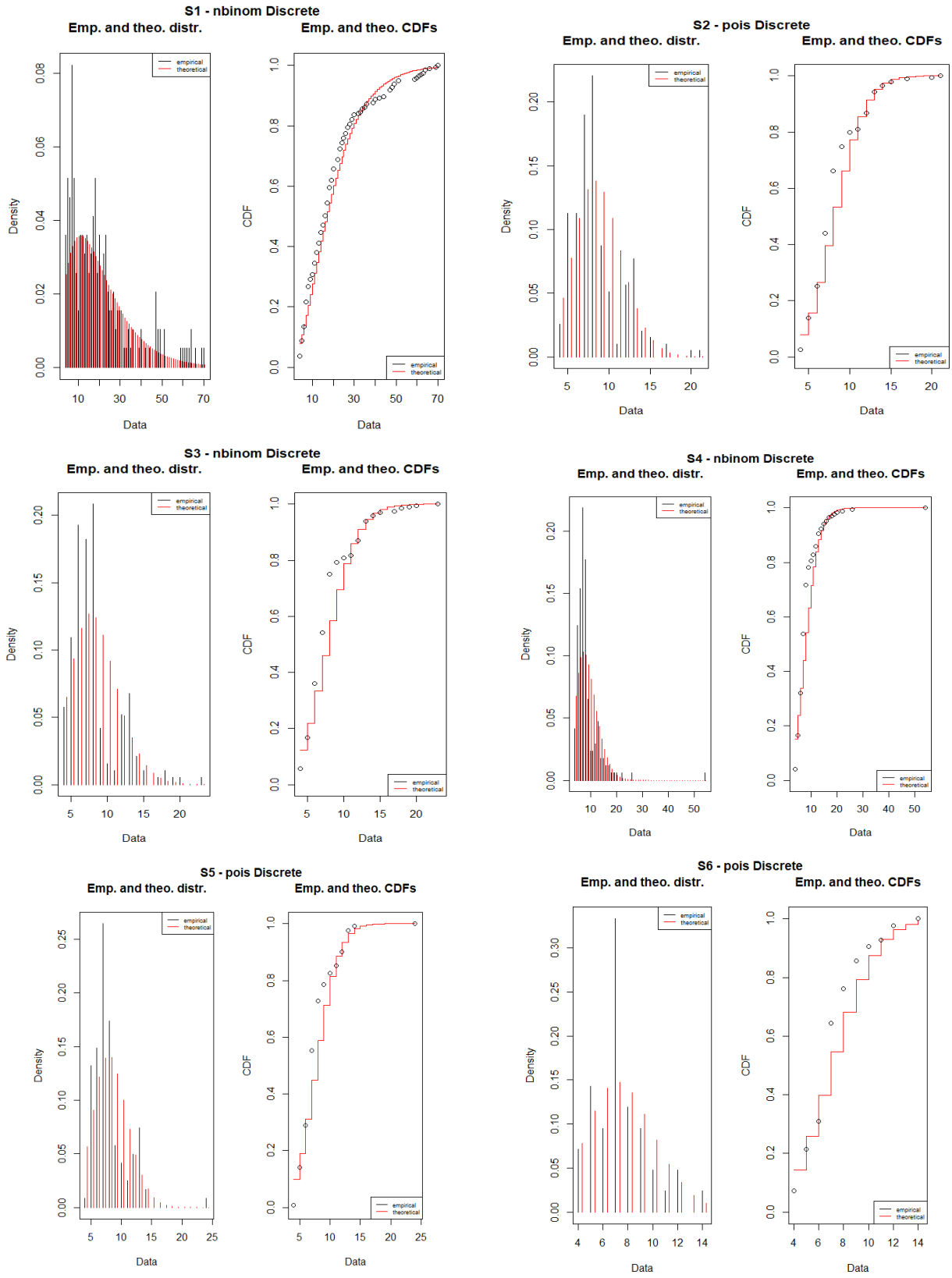

Figure 3: Best fitted discrete distributions in Break Intervals

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

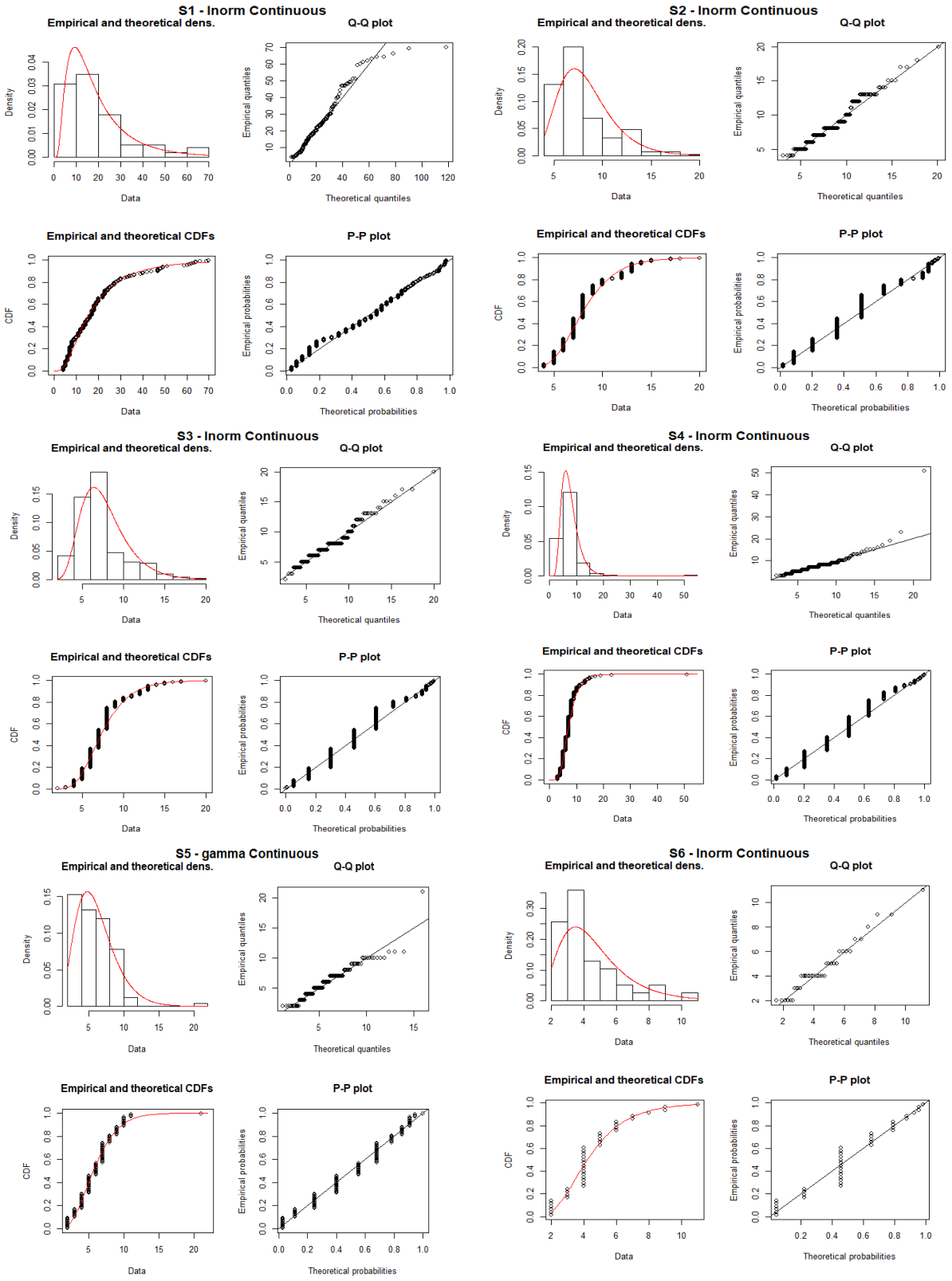

Figure 4: Best fitted continuous distributions in break intervals

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

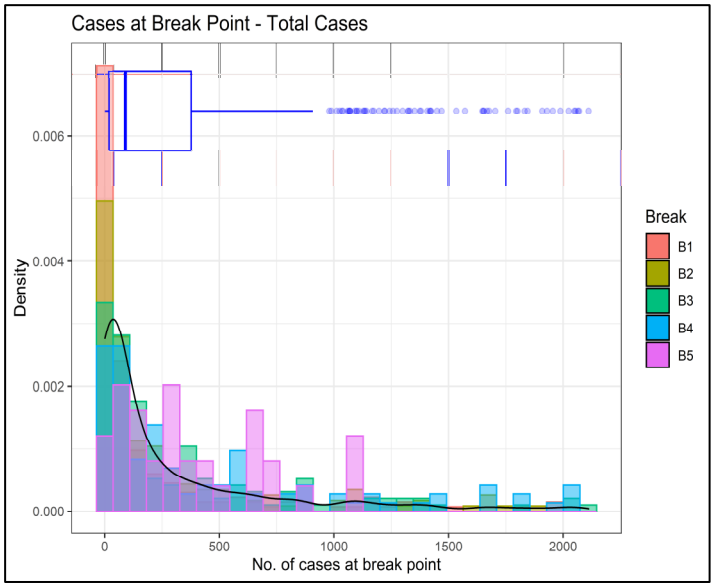

Figure 5: Density plot on overall break intervals

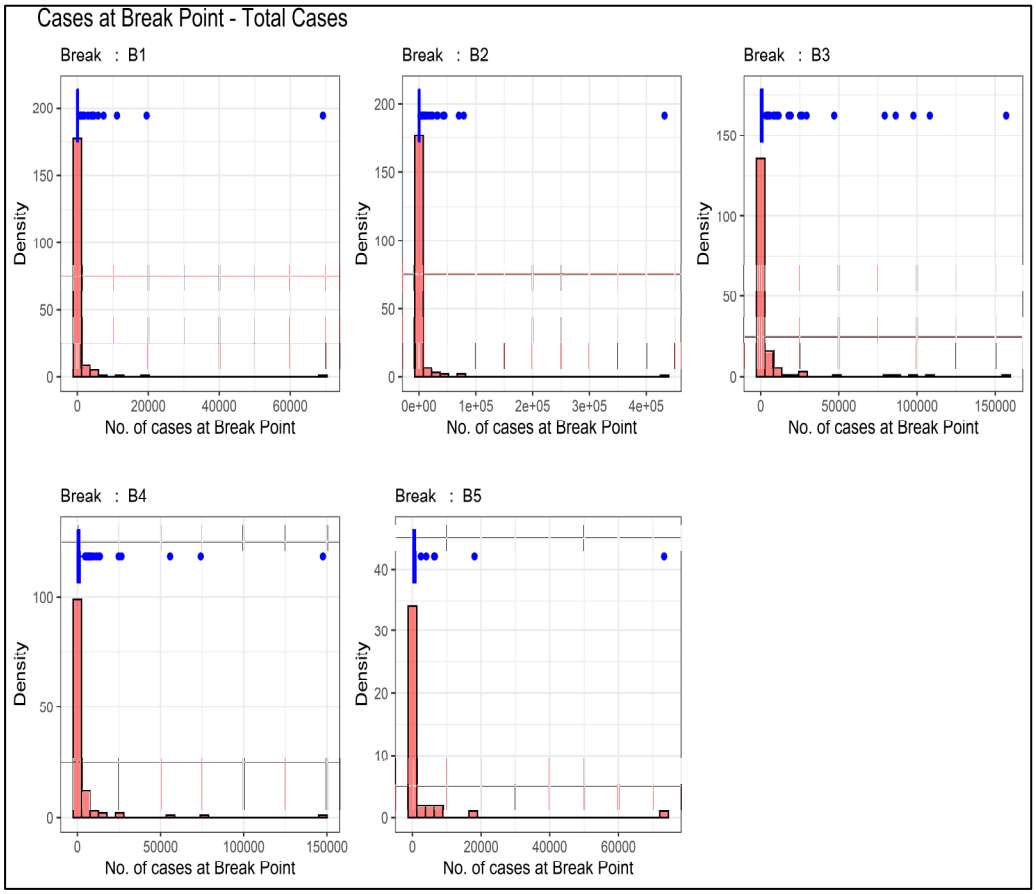

Figure 6: Histogram and box plot in each break interval

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

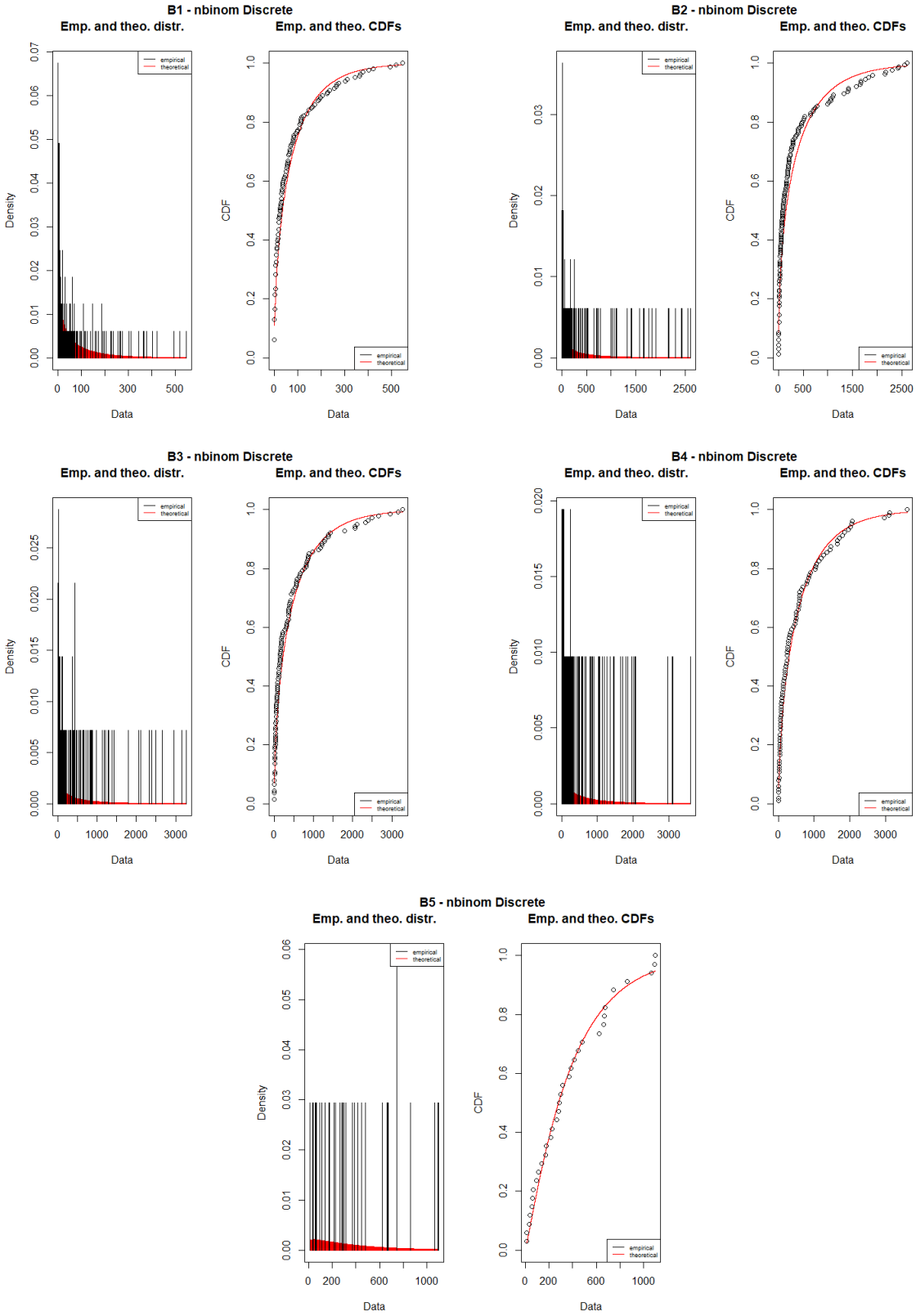

Figure 7: Best fitted discrete distribution of total cases at breakpoints

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

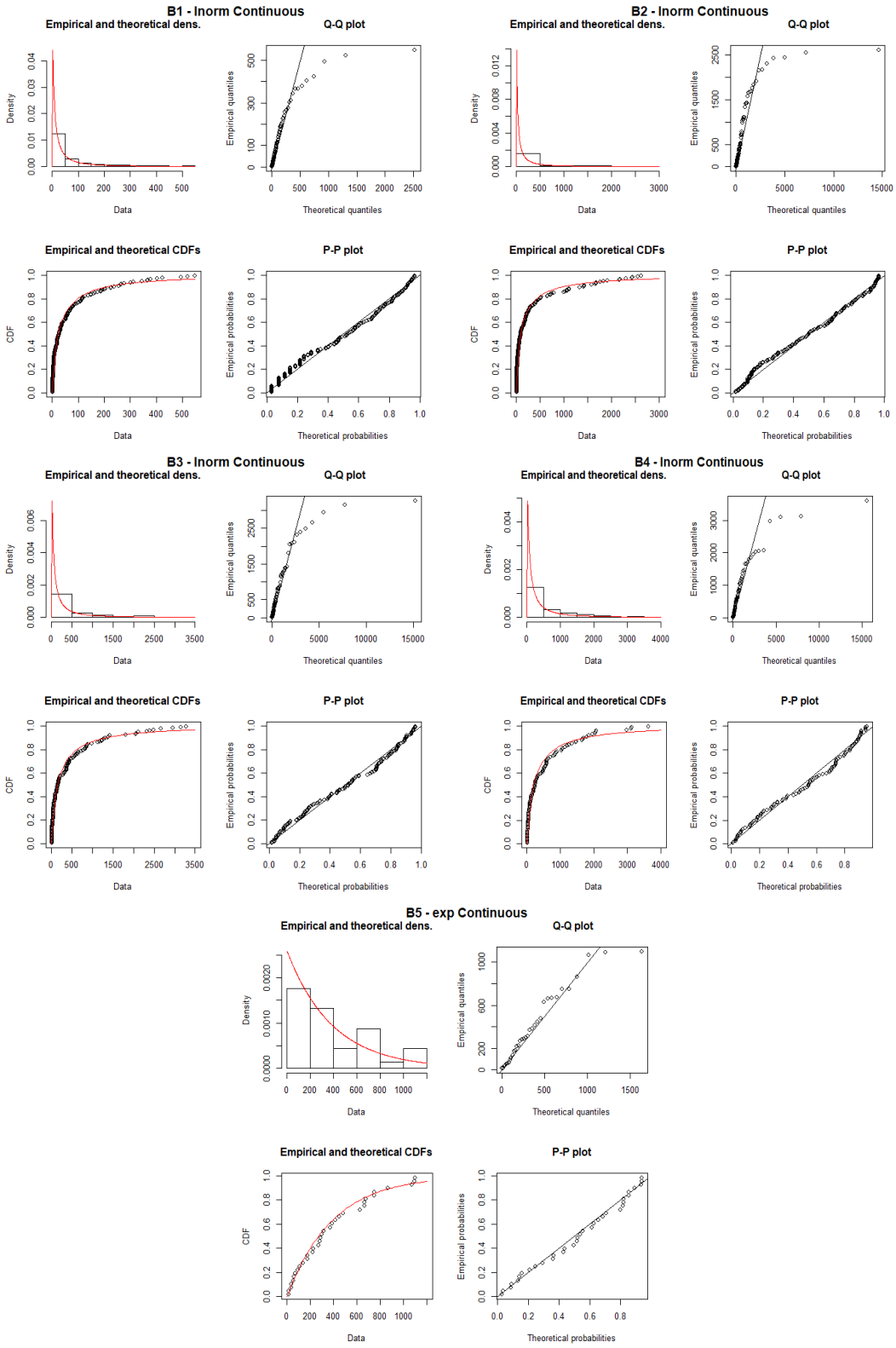

Figure 8: Best fitted continuous distributions of cases at breakpoints

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

Table 1: Descriptive statistics of break interval based on number of days						
Break Interval	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
S_1	4	8	16	20.04	25	70
S_2	4	6.5	8	8.426	9.5	21
S_3	4	6	7	8.073	8.25	23
S_4	4	6	7	8.562	9	54
S_5	4	6	7	8.025	9	24
S_6	4	6	7	7.357	8	14

Table 2: Distribution fitting on Break Intervals				
Distribution		AIC	BIC	Interval
Normal	mean= 20.041, sd= 15.193	1618.504	1625.050	S1
NegBinom	size= 2.289, mu= 20.036	1519.761	1526.307	
Poisson	lambda= 20.041	2845.370	2848.643	
LogNormal	meanlog= 2.737, sdlog= 0.728	1500.993	1507.539	
Gamma	shape= 1.74, rate= 0	1517.354	1523.900	
Normal	mean= 8.426, sd= 3.004	986.363	992.909	S2
NegBinom	size= 143.893, mu= 8.426	963.021	969.567	
Poisson	lambda= 8.426	961.411	964.684	
LogNormal	meanlog= 2.069, sdlog= 0.334	936.440	942.986	
Gamma	shape= 7.952, rate= 1	946.518	953.064	
Normal	mean= 8.073, sd= 3.294	1006.612	1013.127	S3
NegBinom	size= 30.527, mu= 8.073	969.217	975.732	
Poisson	lambda= 8.073	974.439	977.696	
LogNormal	meanlog= 1.981, sdlog= 0.364	917.121	923.625	
Gamma	shape= 6.884, rate= 1	923.273	929.777	
Normal	mean= 8.562, sd= 5.031	1029.682	1035.942	S4
NegBinom	size= 9.199, mu= 8.562	925.052	931.312	
Poisson	lambda= 8.562	999.098	1002.228	
LogNormal	meanlog= 1.946, sdlog= 0.408	838.067	844.326	
Gamma	shape= 2.865, rate= 0	897.692	903.951	
Normal	mean= 8.025, sd= 2.827	598.861	604.453	S5
NegBinom	size= 182946.75, mu= 8.025	580.531	586.123	
Poisson	lambda= 8.025	578.531	581.327	
LogNormal	meanlog= 1.693, sdlog= 0.488	583.292	588.883	
Gamma	shape= 4.689, rate= 1	578.328	583.919	
Normal	mean= 7.357, sd= 2.202	189.496	192.971	S6
NegBinom	size= 2332625.794, mu= 7.357	190.746	194.221	
Poisson	lambda= 7.357	188.746	190.484	
LogNormal	meanlog= 1.434, sdlog= 0.438	162.131	165.458	
Gamma	shape= 4.968, rate= 1	163.179	166.506	

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

Table 3: Outlier identification in each break interval

Break	No. of countries	No. of outliers	Outlier (%)	Country-Cases
B1	195	32	16%	Austria - 1332, Belarus - 700, Belgium - 1243, Brazil - 4256 , Canada - 4675, Chile - 1142, China - 1988, Ecuador - 789, France - 4499, Germany - 3062, India - 1071, Indonesia - 579, Iran - 5823, Ireland - 785, Israel - 1930, Italy - 3858, Japan - 1128, Mexico - 848, Netherlands - 2994, Peru - 1323, Poland - 901, Portugal - 642, Qatar - 634, Russia - 1534, Saudi Arabia - 1012, Singapore - 594, Spain - 11178, Sweden - 1167, Turkey - 7402, Switzerland - 3010, United Kingdom - 19522, United States - 69194
B2	192	27	14%	Austria - 5888, Belarus - 3281, Belgium - 10836, , Canada - 20748, Chile - 3031, China - 44724, , Germany - 42288, India - 7447, Iran - 16169, , Italy - 31506, Japan - 3906, Netherlands - 9762, , Portugal - 2995, Russia - 11917, Saudi Arabia - 2932, , Spain - 78797, Sweden - 3700, Switzerland - 10714, , United Arab Emirates - 2990, United Kingdom - 70272, United States - 432132
B3	169	30	18%	Australia - 5844, Austria - 11129, Belgium - 26667, Chile - 5546, China - 79355, Czech Republic - 3589, Ecuador - 4965, Germany - 108202, Iran - 29406, Ireland - 5364, Israel - 10743, Malaysia - 4228, Mexico - 5399, Netherlands - 17851, Peru - 11475, Poland - 5205, Portugal - 8251, Qatar - 3428, Romania - 4057, South Korea - 8162, Spain - 157022, Sweden - 9685, Switzerland - 19227, Ukraine - 4161
B4	121	18	15%	Austria - 13560, Chile - 8273, Czech Republic - 5312, Indonesia - 5136, Iran - 55743, Ireland - 11479, Italy - 147577, Norway - 5208, Pakistan - 5988, Poland - 7582, Portugal - 13141, Serbia - 4873, South Korea - 9661, Switzerland - 24820, Turkey - 74193
B5	42	8	19%	Argentina - 2432, Czech Republic - 6303, Denmark - 6511, Iran - 73303, Norway - 6415, Panama - 4016, Portugal - 18091, South Africa - 2506

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 May 2020 doi:10.20944/preprints202005.0319.v1

Table 4: Descriptive statistics of break interval based on number of cases						
Break	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
B_1	1	6	26	74.398	84.5	549
B_2	2	18	87	364.115	339	2615
B_3	4	39.5	158	469.136	571.5	3277
B_4	5	51.5	241	579.805	802	3614
B_5	12	117.2	293.5	388.500	653	1100

Table 5: Distribution fitting on number of cases at breaks				
Distribution		AIC	BIC	Break
Exponential	rate= 0.013	1732.877	1735.971	B1
LogNormal	meanlog= 3.133 ,sdlog= 1.716	1664.021	1670.208	
Gamma	shape= 0.446 ,rate= 0	1682.891	1689.078	
Normal	mean= 74.399 ,sd= 111.387	2003.014	2009.202	
NegBinom	size= 0.535 ,mu= 74.419	1686.026	1692.213	
Poisson	lambda= 74.399	20783.846	20786.94	
Exponential	rate= 0.003	2278.165	2281.271	B2
LogNormal	meanlog= 4.473 ,sdlog= 1.866	2154.149	2160.361	
Gamma	shape= 0.362 ,rate= 0	2187.982	2194.194	
Normal	mean= 364.115 ,sd= 605.353	2586.168	2592.38	
NegBinom	size= 0.452 ,mu= 364.262	2184.135	2190.347	
Poisson	lambda= 364.115	118204.847	118207.953	
Exponential	rate= 0.002	1989.949	1992.883	B3
LogNormal	meanlog= 4.967 ,sdlog= 1.735	1932.466	1938.335	
Gamma	shape= 0.451 ,rate= 0	1946.092	1951.961	
Normal	mean= 469.137 ,sd= 698.406	2219.032	2224.9	
NegBinom	size= 0.531 ,mu= 469.156	1944.489	1950.358	
Poisson	lambda= 469.137	107695.872	107698.807	
Exponential	rate= 0.002	1518.715	1521.35	B4
LogNormal	meanlog= 5.296 ,sdlog= 1.685	1494.836	1500.105	
Gamma	shape= 0.566 ,rate= 0	1495.598	1500.868	
Normal	mean= 579.806 ,sd= 770.847	1665.684	1670.954	
NegBinom	size= 0.582 ,mu= 579.74	1496.071	1501.34	
Poisson	lambda= 579.806	85921.07	85923.705	
Exponential	rate= 0.003	475.436	476.962	B5
LogNormal	meanlog= 5.452 ,sdlog= 1.197	483.439	486.492	
Gamma	shape= 1.457 ,rate= 0	478.832	481.885	
Normal	mean= 388.5 ,sd= 321.866	493.129	496.182	
NegBinom	size= 1.123 ,mu= 388.62	477.247	480.3	
Poisson	lambda= 388.5	9466.545	9468.072	