 doi:10.20944/preprints202003.0077.v3

The ongoing COVID-19 epidemic curves indicate initial point spread in China with log-normal distribution of new cases per day with a predictable last date of the outbreak version 3: Test for when after a peak in daily cases the predicted equation becomes reliable and use of the derivative of the equation to detect time of key changes determining the length of the outbreak.

Stefan Olsson[1,2*] and Jing Zhang[1]

[1]State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, College of Plant Protection
[2]Plant Immunity Center, Haixia Institute of Science and Technology, College of Life Science
Fujian Agriculture and Forestry University, No.15 Shangxiadian Road, Cangshan District, Fuzhou City, Fujian Province, China. P.C. 350002

* E-mail stefan@olssonstefan.com or stefan.olsson@fafu.edu.cn

Abstract

During an epidemic outbreak it is useful for planners and responsible authorities to be able to plan ahead to estimate when an outbreak of an epidemic is likely to ease and when the last case can be predicted in their area of responsibility. Theoretically this could be done for a point source epidemic using epidemic curve forecasting. The extensive data now coming out of China makes it possible to test if this can be done using MS Excel a standard spreadsheet program available to most offices. The available data is divided up for whole China and the different provinces. This and the high number of cases makes the analysis possible. Data for new confirmed infections for Hubei, Hubei outside Wuhan, China excluding Hubei as well as Zhejiang and Fujian provinces all follow a log-normal distribution that can be used to make a rough estimate for the date of the last new confirmed cases in respective areas. In the version 2 continuation work, 9 additional days were added for the Chinese data to evaluate the previous predictions. The extra data then available from China follows the previous predicted trend supporting the usefulness of this simple technique. In the version 2 we also tested the feasibility for a non-specialist to make similar predictions using additional data from S Korea now available. In this third continuation the predictions for Version 2 are evaluated for S Korea and fits well the beginning of the decline but it seems to be difficult to bring down numbers of cases per day under about 100 new cases per day, potential reasons for this is discussed. To further evaluate when in a prediction becomes reliable the Chinese data was used to evaluate to make predictions for each day around the peak in number of cases and after2-3 consecutive days of decreasing new cases per day the prediction becomes reliable. In version 3 data for Italy just reaching this point was used to make further predictions for that country. A second new analysis was also added to use the fitted equation to detect when the acceleration of new cases per day stopped increasing exponentially. In the Chinese case this measured point coincides with the date of the complete Hubei lockdown and in the new Italian analysis it coincides with the mandatory Italian lockdown. Predicted dates for the end of the Italian outbreak is also added.

Introduction

In epidemics starting as a point source the number of new cases often follows a log-normal distribution or more precisely a Poisson-Gamma distribution. How this distribution will develop over time can theoretically be determined by fitting a log-normal distribution equation to the data for new cases per day are reported. The estimate will of cause be more accurate the further into the outbreak. A literally "breaking point" for the accuracy of the estimate for the end of the outbreak comes after the number of

 doi:10.20944/preprints202003.0077.v3

new cases per day have reached its peak. From there on the estimate should be better and better. Here a simple method that could be used without access to special resources for getting such estimates after the peak has been reached is presented using data from the ongoing COVID-19 epidemic in China.

Results and discussion

A log normal distribution can be relatively nicely fitted all data sets (Fig 1&2). When using a log scale for the Y-axis it is apparent there are deviations in the early dates especially for Hubei (Fig 1A). This could be caused by a lag in detection of new cases in the beginning of the outbreak. The deviations in the latest dates can have many different causes like changing criteria for new cases, or simply a backlog in cases confirmation due to highly stressed health care system in the worst hit city Wuhan. Both the data from Hubei outside Wuhan (Fig 1B) and China outside Hubei (Fig 1C) on the other hand closely follows a log normal distribution.

To see if the same relationships holds also outside Hubei, two provinces with quite different number of cases, Zhejiang with many cases and Fujian with few cases, was also tested (Fig 2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0077.v3

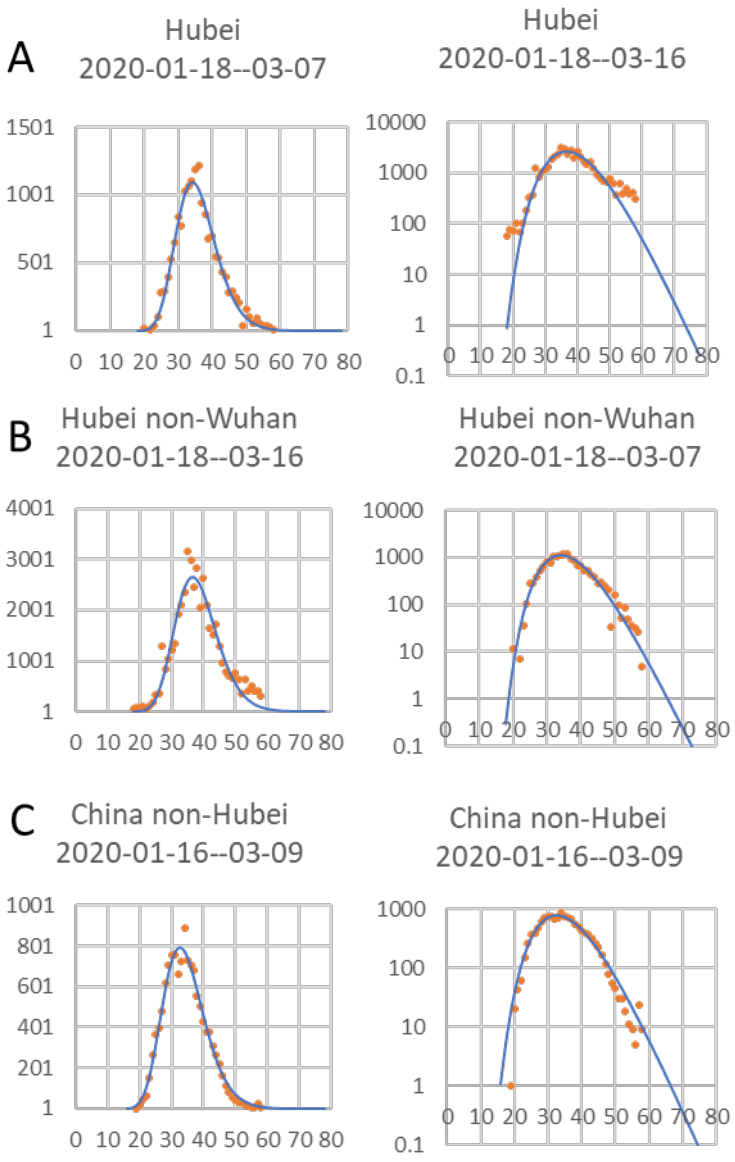

Figure 1. Log normal distribution of new confirmed cases for each day since 1 Jan 2020 Hubei, Hubei-nonWuhan and in reest of China. The Log of day values with start on the first day a case could have been confirmed was used curve fitting although here in the plot the actual number of days since 1st January was used as X-axis. Number of new confirmed cases per day and fitted curve (left) and Log number of new cases per day to show start and stop days (right). Headings shows estimated dates for 1st and last confirmed case. Y axes both to the left and right start at 1 to highlight the first and last case.

In Zheijiang the outbreak followed the general pattern very closely (Fig 2A) but for the much smaller outbreak in Fujian (Fig 2B) the number of cases per day dropped more than the model for the last days.This is caused by the approximation to log-normal distribution instead of a Poisson distribution that is more correct for data with few cases (Gonzales-Barron and Butler, 2011) but more difficult to handle using standard Excel curve fitting. This discrepancy mean that the last new infection date will be overestimated especially for limited outbreakes like the one in Fujian province. From planning point of view it should however be safer to oversestimate the length of the outbreak than underestimate it. A

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0077.v3

fairly good estimate of the last data could be done as soon as the number of new confirmed cases per day started to decrease.

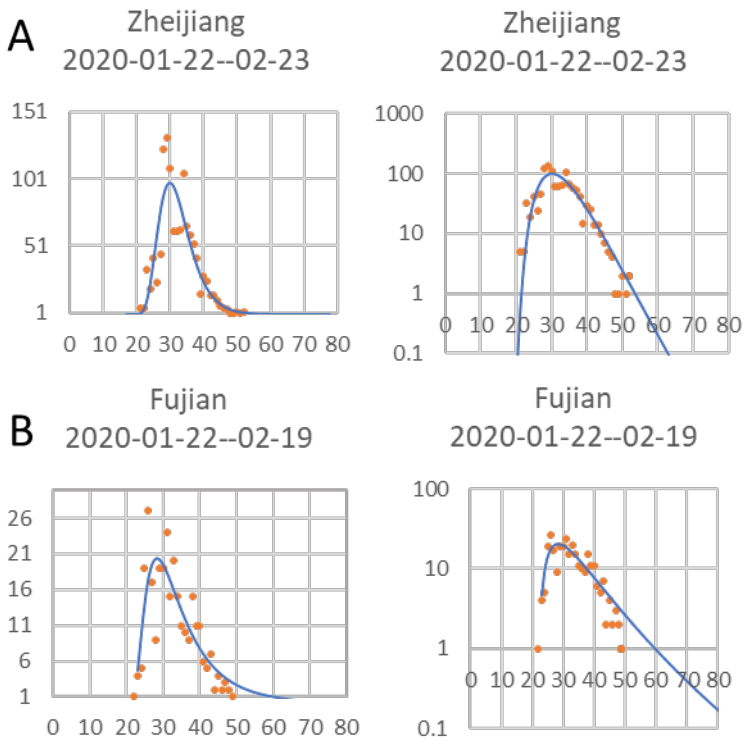

Figure 2. Log normal distribution of new confirmed cases for each day since 1 Jan 2020 in two provinces with relatively high numers of cases, Zeijiang with high numbers and Fujian with low numbers. The Log of day values with start on the first day a case could have been confirmed was used curve fitting although here in the plot the actual number of days since 1st January was used as X-axis. Number of new confirmed cases per day and fitted curve (left) and Log number of new cases per day to show start and stop days (right). Headings shows estimated dates for 1st and last confirmed case. Y-axes both to the left and right start at 1 to highlight the first and last case.

The estimated start date for when new cases could have been confirmed caused by community spread was for Hubei and Wuhan the 18th January while outside Hubei the data indicate a 2 day earlier start if the disease behaved similarly. This is a bit surprising but could indicate that the disease was brought to Wuhan city and Hubei province from a less populated area and found good conditions for spread in Wuhan. The estimated start dates for when new cases could be confirmed in the two provinces Zheijiang and Fujian were both the 22nd January a few days later than in the epicenter for the outbreak.

Test 9 days later if predictions were reasonable

In the follow up test of the original prediction the new data for the next 9 day follow the prediction (Fig. 1) surprisingly well (Fig. 1 continued). This applies for all three cases but especially good was the prediction for Hubei non-Wuhan (Fig. 1B continued). Interestingly, for China non-Hubei that previously seemed to predict a later end date than the data indicated (Fig. 1C), now with the new data it is apparent that this is not the case (Fig. 1C continued). Finally, for Hubei the decrease in new cases for the additional dates in principle follow the shape of the fitted curve but with a slight lag (Fig. 1A continued)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0077.v3

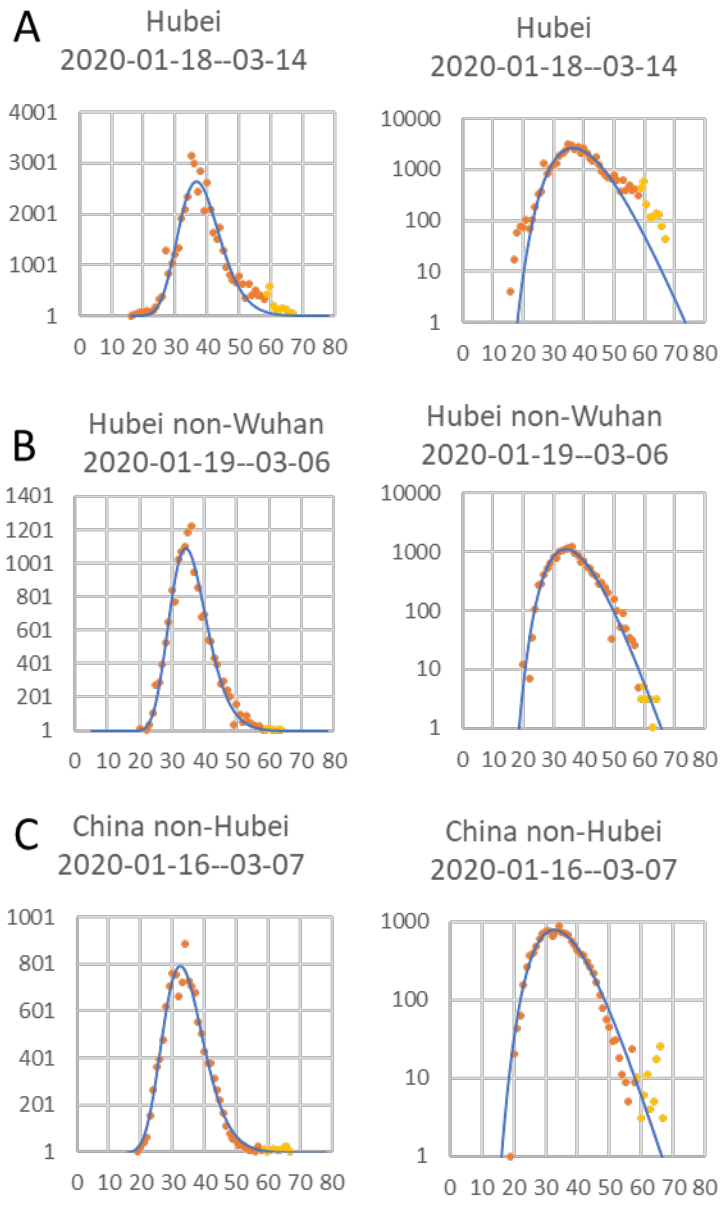

Figure 1 Continued. Follow up of the development seen in Figure to evaluate the predictions made previously. Same data and same data-fitting as in Figure 1 in manuscript V1 but with new data from February 27 to March 07 added (yellow dots). The Log of day values with start on the first day a case could have been confirmed was used curve fitting although here in the plot the actual number of days since 1st January was used as X-axis. Number of new confirmed cases per day and fitted curve (left) and Log number of new cases per day to show start and stop days (right). Headings shows estimated dates for 1st and last confirmed case. Y axes both to the left and right start at 1 to highlight the first and last predicted case.

Test if the MsExcel sheets with the instructions can be used by a non-bioinformatician

The Excel sheet was sent to a previous master student now living in another city (now also co-author) to test the feasability of using the sheets to do curve-fitting and predictions using the MsExcel file. After some initial problems finding out how to find the Solver Add-In for an iMac version of MsExcel things

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0077.v3

went smoothly. The problem was solved by the master student through an internet search for how to find and add the Solver Add-in to the iMac version. Also the S Korea data can be efficiently modelled using the same approach (Fig. 3).

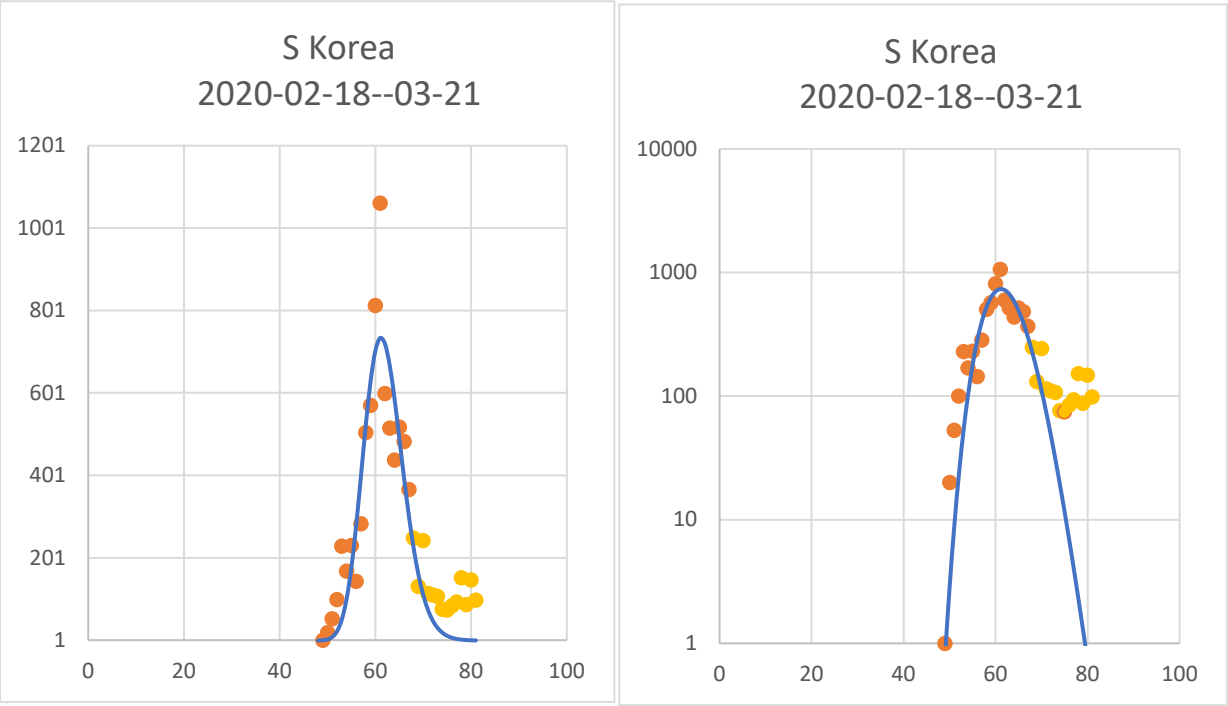

Figure 3 Continued. Follow up of the development seen in Figure3 to evaluate the predictions made previously. The Log of day values with start on the first day a case could have been confirmed was used for curve fitting although here in the plot the actual number of days since 1st January was used as X-axis. Same data and same data-fitting as in Figure 1 in manuscript V2 but with new data from Mach7 to March 24 added (yellow dots). Number of new confirmed cases per day and fitted curve (left) and Log number of new cases per day to show start and stop days (right). Headings shows estimated dates for 1st and last confirmed case. Y axes both to the left and right start at 1 to highlight the first and last predicted case.

Test 9 days later if predictions for S Korea were reasonable

The first 3-4 days follows the predicted curve very clos but the latest points stop declining at the level of about 100 new cqases per day there can be many explanations for this. One could be that the restrictions in S Korea was not a complete lockdown as in China alowiinfg a low level spread that maintains the levels of new infection. It could also be that in these cases are new infecteion leaking in from abroad. S Korea is not in the same situation as China was with basically no cases outside China or of cause a combination of both.

Test when in an outbreak the predictions becomes reliable

We have stated in the previos versions of the manuscript that one need to wait until at or after the peak in numbers per day to for the predictions to be reliable. Now we use the data for whole China to test this notion. We thus made predictions for consecutive day just before and after the peak in numbers per

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0077.v3

day. Thus we can plot curves showing these predictions and compare with where in the curve the predictions where made (Fig. 4). It is apparent that for the China data 2-3 days of decrease in numbers was needed to be able to reliable predict the magnitude and end of the outbreak (Fig. 4).

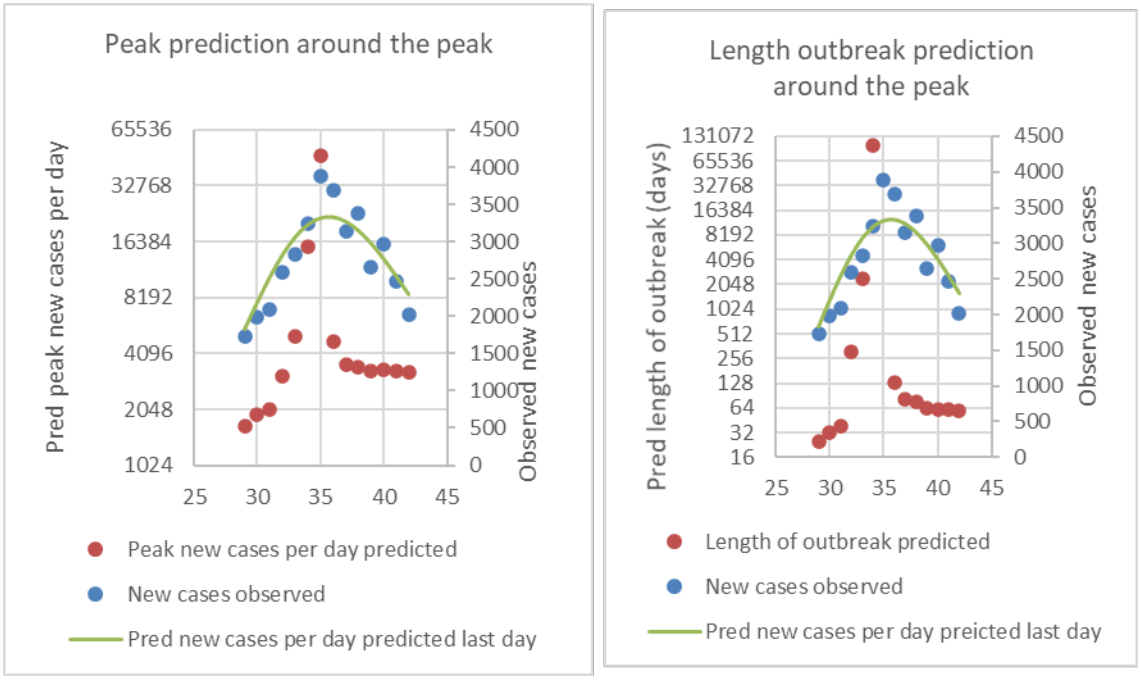

Figure 4. Analysis for when in an outbreak the analysis becomes reliable

Left: Prediction of peak hight using the data around the peak starting some days before the peak and finishing some days after the peak. Left Y axis shows a log scale for the predictions (red dots) and the right Y-axis show unlogged values for the observed values (blue dots) as well as the predicted fitted equation from the last day (day 42).

Right: Same as left figure but now with predicted length of the outbreak from predicted first to last case instead of peak height.

Use of the fitted equation to determine when the outbreak starts to slow down for then later to reach the peak new infections per day: Use on China data and on Italy data

An equation like the normal distribution has an acceleration phase, a slowdown of increase of acceleration then maximum acceleration before going into a deceleration towards the peak. Thus, the point where the acceleration of this acceleration starts to break is the point where something could have happened that determined the whole outbreak size and duration. To determine this the change in predicted new cases from one day to next (in principle the derivative of the equation) was plotted together with the predicted number of cases for the whole outbreak (Fig.5 Left). As can be seen in the figure the acceleration of the acceleration (the red curve) starts to slow down at day 28 (January 28) and the grid have been adjusted so that can be seen easier in the figure. A similar analysis was performed for the Italian data (Fig 5 right where it can be seen that the acceleration of the acceleration stops at day 70

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0077.v3

(March 10). The Italian prediction data as of March 24 is also presented together with observed data (Fig 6).

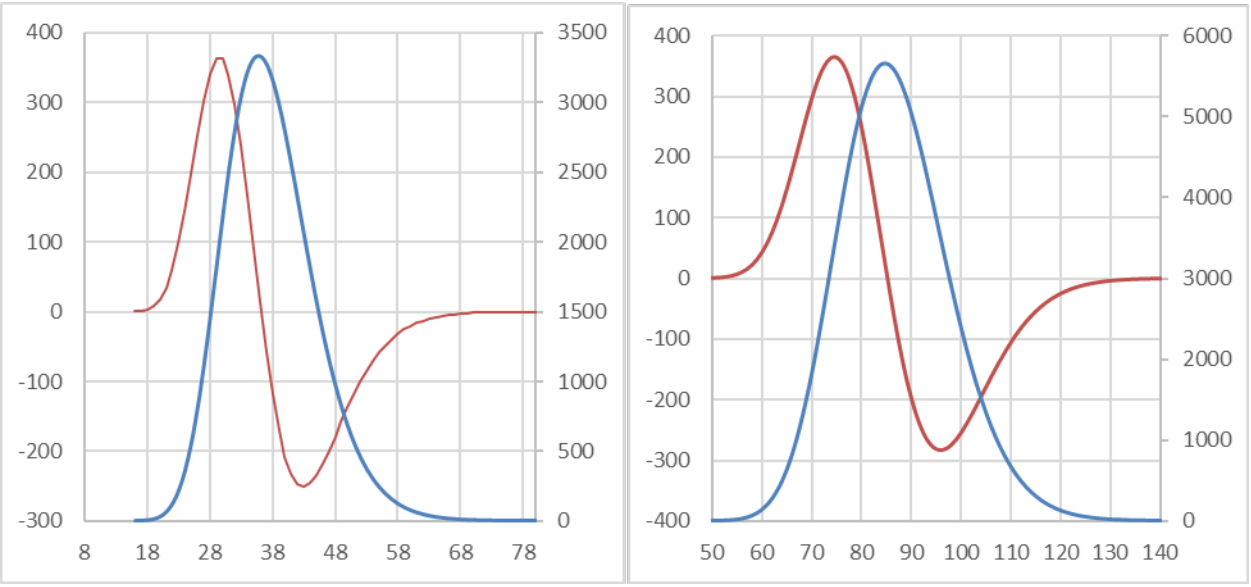

Figure 5 Plot of acceleration in new cases per day (red lines) together with cases per day predicted by the fitted equation (blue lines).
Values for whole China (left) shows that the Acceleration of the acceleration in new cases per day started to slow down on day 28 (January 28) (thus the unusual X axis to show that point with grid lines). Values predicted for Italy (right) shows a similar change on day 70 (March 10) Left Y axes are increase in numbers of new cases per day and right Y axes shows new cases per day. X-axes is days since start of year 2020.

As can be seen for Italy new cases are predicted to start falling at around March 24-25 and the outbreak is predicted to get its last case on May 23 if present measures by the Italian government is kept. With even stronger measures this could maybe be shortened and total cases lower. If measures are relaxed the whole outbreak becomes longer and total number of cases will increase.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0077.v3

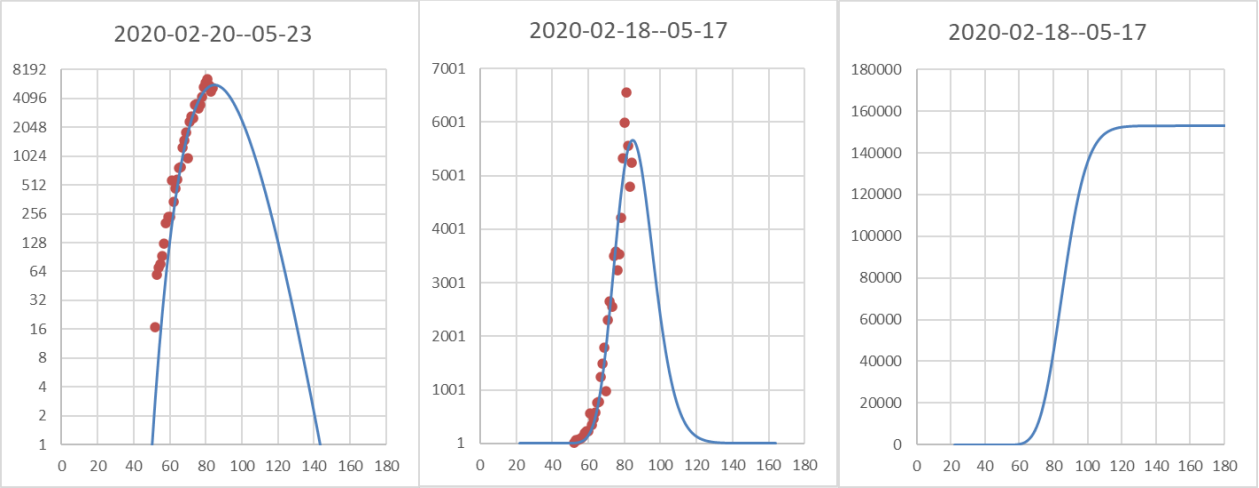

Figure 6 Predictions for Italy the 24th March 2020

The Log of day values with start on the first day a case could have been confirmed was used for curve fitting although here in the plot the actual number of days since 1st January was used as X-axis. Log number of new cases per day to show start and stop days (left), number of new confirmed cases per day and fitted curve (Middle) and cumulative predicted curve (right). Headings shows estimated dates for 1st and last confirmed case. Y axes both to the left and middle start at 1 to highlight the first and last predicted case.

The day the acceleration of cases stopped was the day lockdowns were enforced.

Since January 28 for China and March 10 for Italy marks the day when the outbreaks took a new direction according to our analysis and the acceleration of the acceleration of new cases started to slow down we decided to look up that day in the WiKi-pages (Anonymous, 2020a) that also records the decisions taken by officials to see if anything unusual happened those days. For China the mandatory lockdown of Hubei was announced the January 28 and the lockdown was in effect the 29th. For Italy the lockdown was similarly introduced on March 10. This may point to the importance of an early lockdown as in China since it took effect when there were around 1000-1500 new cases per day. In Italy the same did not happen until there were more than 2000 new cases per day only 2-3 days later in the outbreak. This difference seem to double the length of the outbreak and triple the total number of cases compared to China (Fig. 6).

Conclusion

Plotting new confirmed cases per day against time can be used during a large point source epidemic outbreak to relatively early after the peak in new cases determine a likely last date for new cases. Such information should be useful to people in charge for planning how to allocate resources. The information will also be available when resources are as most stretched with a large number of active cases just after the peak in number of new cases per day, In addition, if the data continue to fit the curve for a point source outbreak in one area there has most likely been no new introduction of cases or any change to the virus or the likelihood that a person becomes infected within that area. The latter seems to be the case for the COVID-19 outbreak in China 2019-2020 pointing to that the quarantining

measures stopping further spread between provinces and cities after the first few days of person-to-person transfer have worked efficiently.

In this extended work we tested the predictions for the 9 following days in the previous preprint paper (or version) against the new data and we found that the technique managed to predict the new data very well. In addition, we have now also found that it is feasible to put the Excel file in the hands of a non-bioinformatician and get useful results as can be seen for the S Korea newly added figures (Fig. 3).

In this further extended work, we evaluated the predictions previously made for S Korea and found that predictions were valid but inflow from other countries and some minor outbreaks and/or not strong enough measures might make it difficult get the outbreak to completely disappear. We have in V3 added an analysis for when the prediction using our method becomes reliable and that happens a few days after the peak where number of infections per day start a reliable decline. We also in this version added an analysis of when new infections pre day stopped accelerating with the same rate. We then found that it was rather early in the China case and only some days later for Italy. To our big surprise Both these dates coincide with the days mandatory lockdown took effect in both countries. The few days later in the outbreak lockdown in Italy is then also a probable cause of the higher peak and is predicted to result in 2 times longer outbreak with 3 times higher numbers of total cases than was the case for China. Thus a few days hesitation taking the lockdown decisions appears to have had huge effects.

Methods

Official referred to data for the COVID-19 outbreak in China is collected at a Wikipedia page (Anonymous, 2020b). Since the kind of analysis here presented is a relatively simple analysis it should be possible to do for anyone using a standard program Microsoft Excel with the standard available Solver plugin for data handling and curve fitting. The logarithm of number of days since the estimated start of the epidemic outbreak were used for fitting a normal distribution equation to the data but in the figures the data was plotted against the non-logged day number with day 1 on the 1st January to ease in determining the actual dates from readings on the X-axis and the values in the spreadsheet files.

The MS Excel file used for this analysis is available as Supplementary file and can easily be modified to be used with other data to relatively early after the peak in new confirmed cases be able to predict the end of an epidemic outbreak with a definite starting point having a “first case”.

Acknowledgement

When back in my home country Sweden I had to decide when to return to China after the winter break for the Chinese New Year (Spring Festival), I decided to look at the epidemiology data since I have been working with biological control trying to cause epidemics in fungal pathogens attacking plants. I thought of looking for data about the COVID-19 outbreak to be able to determine a time and a route back that limit the chances for me to catch the infection and bring it to my workplace. I found the very good Wikipedia entry I refer to in the methods and would like to thank everyone that has contributed to edit

 doi:10.20944/preprints202003.0077.v3

that site. Finally, I want to acknowledge my employer Fujian Agriculture and Forestry University that makes it possible for me to do research in China.

Supplemental file

"Corona model final.V3.xlsx" is a supplemental file containing all pervious data for China with an added Sheet for Whole of China. This sheet contains prediction reliability data and the data for calculating and showing the acceleration of number of cases per day. In addition, the file also contains instructions for how to use it to fit new data to make predictions.

"Corona model only S-Korea and Italy" is a supplemental file containing the previous S-Korea data sheet with the previous prediction so fit to new data can be evaluated. This file also contain data for Italy used for a fitting and prediction similar to what was done for Whole of China in the other Supplemental file.

References

Anonymous (2020a). 2019–20 coronavirus pandemic. *WikipediA*. Available at: https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic [Accessed March 25, 2020].

Anonymous (2020b). Timeline of the 2019–20 coronavirus outbreak. *WikipediA*. Available at: https://en.wikipedia.org/wiki/Timeline_of_the_2019%E2%80%9320_coronavirus_outbreak#Case_statistics [Accessed March 1, 2020].

Gonzales-Barron, U., and Butler, F. (2011). A comparison between the discrete Poisson-gamma and Poisson-lognormal distributions to characterise microbial counts in foods. *Food Control* 22, 1279–1286. doi:10.1016/j.foodcont.2011.01.029.