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 2: Evaluation of previous prediction and testing the method for S Korea and the use of the method by an unexperienced person.

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 this continuation work 9 additional days were added for the Chinese data to evaluate the previous predictions. We also tested the feasibility for a non-specialist to make similar predictions using additional data from S Korea now available. The extra data now available from China follows the previous predicted trend supporting the usefulness of this simple technique.

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 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).

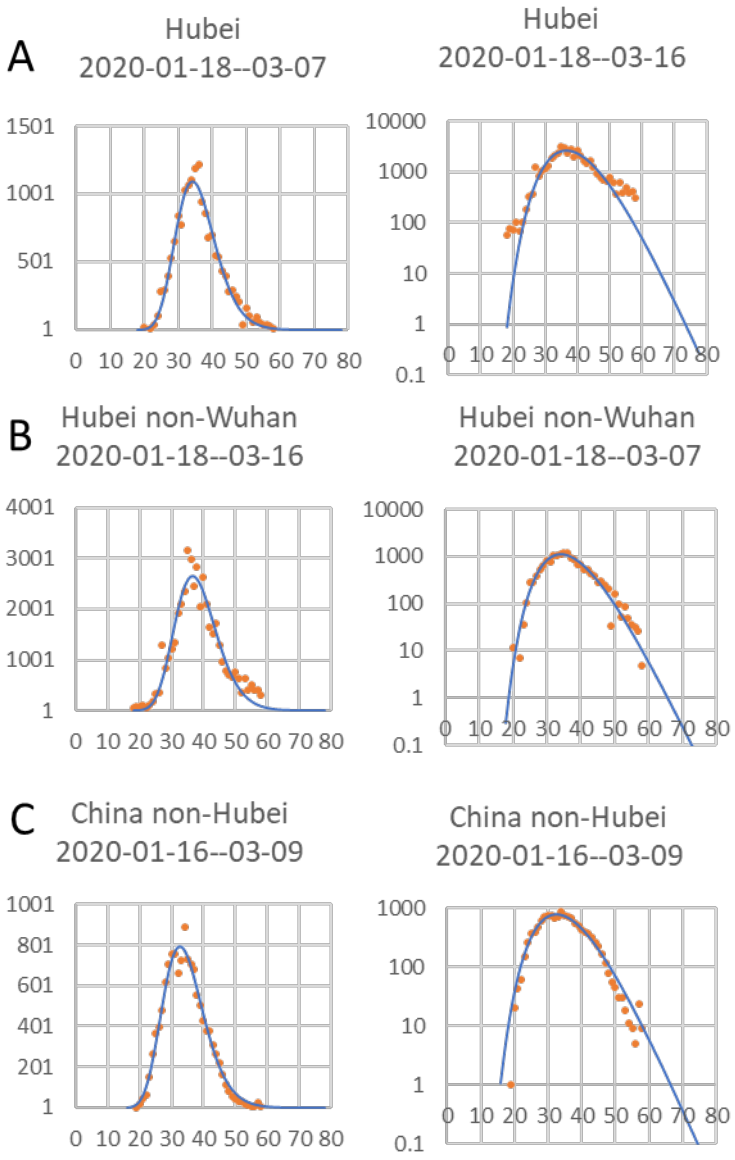

Figure 1. Log normal distribution of new confirmed cases for each day since 1 Jan 2020 Hubei, Hubei-nonWuhan and in reest of China. The Log of day values with start on the first day a case could have been confirmed was used curve fitting although here in the plot the actual number of days since 1st January 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 March 2020 doi:10.20944/preprints202003.0077.v2

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

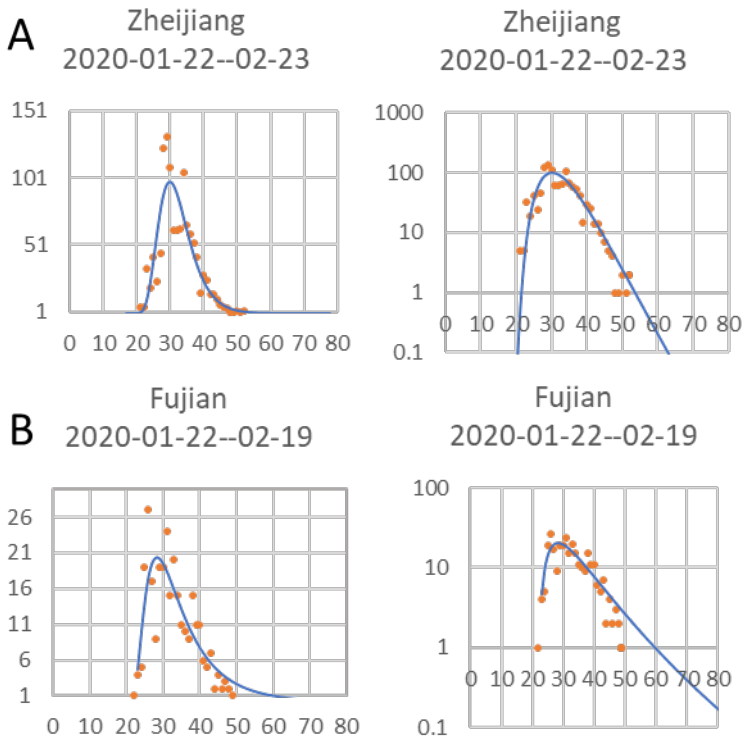

Figure 2. Log normal distribution of new confirmed cases for each day since 1 Jan 2020 in two provinces with relatively high numers of cases, Zeijiang with high numbers and Fujian with low numbers. The Log of day values with start on the first day a case could have been confirmed was used curve fitting 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 March 2020 doi:10.20944/preprints202003.0077.v2

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)

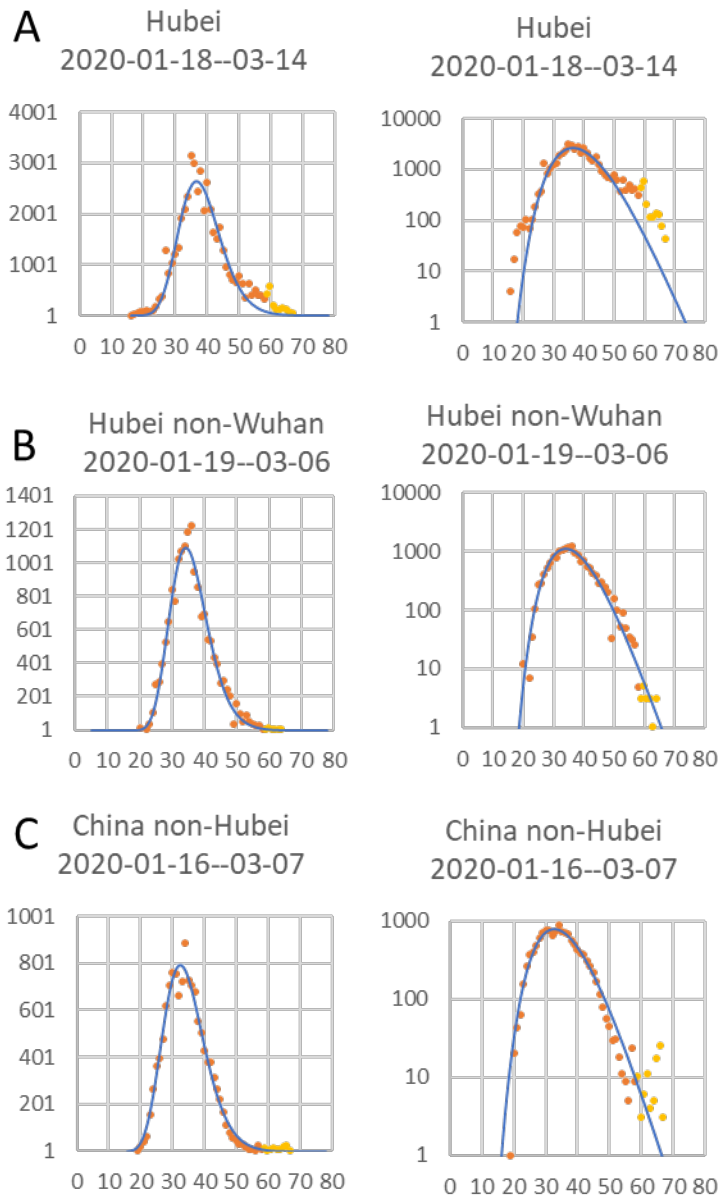

Figure 1 Continued. Follow up of the development seen in Figure to evaluate the predictions made previously. Same data and same data-fitting as in Figure 1 but with new data from February 27 to March 07 added (yellow dots).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 March 2020 doi:10.20944/preprints202003.0077.v2

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 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).

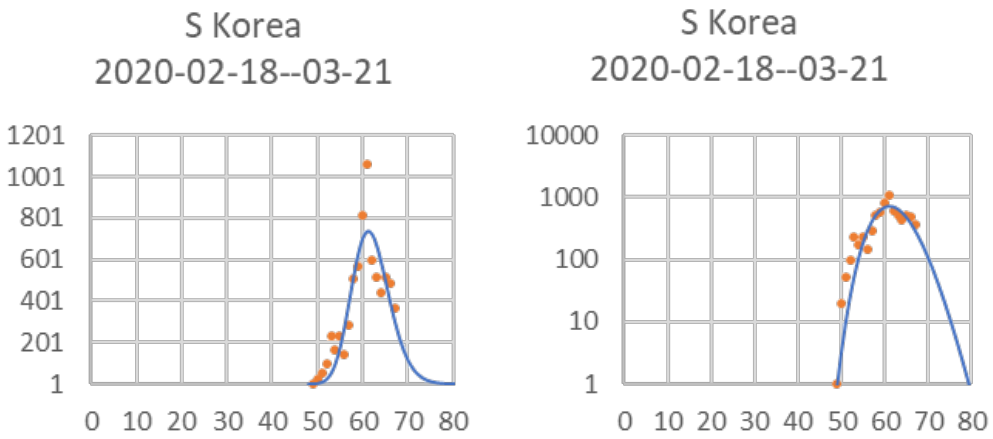

Figure 3. Log normal distribution of new confirmed cases for each day since 1 Jan 2020 South Korea. The Log of day values with start on the first day a case could have been confirmed was used curve fitting although here in the plot the actual number of days since 1st January was used as X-axis. Number 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.

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

 doi:10.20944/preprints202003.0077.v2

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).

Methods

Official referred to data for the COVID-19 outbreak in China is collected at a Wikipedia page (Anonymous, 2020). 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 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.V2.xlsx” is a supplemental file containing all pervious data from March 1 and calculations including the new data added. In addition, the file also contains instructions for how to use it to fit new data to make predictions.

References

Anonymous (2020). 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 and March 10, 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.