STATISTICAL APPROACH FOR MAKING PREDICTIONS OF CONFIRMED INFECTION AND DEATHS ON CORONA VIRUS

Edwin van den Heuvel, Marta Regis, Zhuozhao Zhan

Summary

The statistical approach that we are using to make short and long term predictions of the number of confirmed infections and deaths from the current Corona virus epidemic, consists of three different steps. The first two steps are based on the data of the provinces in China, while the third step is based on data of the specific country that is investigated for predictions. The first step is an investigation of the goodness-of-fit of the three-parameter logistic growth curve, that was used in the 19th century for the estimation of total population sizes in countries. The goal here is to obtain evidence that this curve is suitable for the accumulated data. The second step consists of determining the prediction bias when only limited data is available for fitting the full growth curve. This can be used to make corrections to our predictions when we only use limited data, in particular for the long-term predictions. The third step is an evaluation of the model fit to the data of a country (looking at convergence issues, influence of the choice of the starting values for the parameters, and consistency of predictions over time in relation to the bias correction). Our approach makes certain (untestable) assumptions on the representativeness of the data. It also assumes that the underlying processes that lead to the reported data (behavior of people, number of tests taken, procedure for testing new people) have no or minimal effect on the aggregated results.

Three parameter logistic growth curve

In the 19th century a three-parameter logistic growth curve was developed by Pierre Francois Verhulst (Ahuja and Nash, 1967) to provide prediction on the final population size within countries. The growth curve model is described as follows:

$$\mathbb{E}(Y_t) = M/(1 + \exp(-\beta(t - \alpha)),$$

with Y_t the accumulated observed number (of deaths or cases) at day t after a specific date, M the expected maximum number, α the number of days at which the expected number of counts is half way the maximum, and $\beta > 0$ the growth parameter. We will assume that the counts Y_t are approximately normally distributed with mean given by the three-parameter growth curve and variance σ^2. The parameters of the model are determined by the method of maximum likelihood. Analyses are conducted with procedure NLMIXED of the SAS software and with function nls of [R] for verification of numerical and convergence results. In the SAS package we estimate the logarithm of the standard deviation σ and in the [R] package we estimate the inverse slope $1/\beta$.

Choice of Data Set

Our intention is to use the official numbers from the countries.

Data comes from an online interactive dashboard, hosted by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, Baltimore, MD, USA, which tracks the reported cases of coronavirus disease 2019 (COVID-19) in real time and provides a visualization tool. The dashboard, first shared publicly on January 22, 2020, illustrates the location and number of confirmed COVID-19 cases, deaths, and recoveries for all affected countries.

The dashboard reports cases at province level in China, at city level in USA, Australia, and Canada, and at country level otherwise. The primary data source is DXY, an online platform run by members of the Chinese medical community. For countries and regions outside mainland China (that includes also Hong Kong, Macau, and Taiwan), data is manually updated from Twitter feeds, online news services, and direct communications sent through the dashboard. The information is then checked with regional and local health departments including the respective CDC of China, Taiwan, and Europe, the Hong Kong Department of Health, the Macau Government, and WHO, as well as city-level and state-level health authorities.

The data reported on the Johns Hopkins University dashboard aligns with the daily Chinese CDC and WHO situation reports for within and outside mainland China, respectively.

STEP 1: Analysis of Provinces in China and Mainland China

In the first step, we focus on the analysis of the data from the Chinese provinces. We select data from January 22, 2020 until March 09, 2020. We analyze the different provinces within mainland China separately. We omit Tibet since the number of confirmed infections and deaths was too limited.

The maximum likelihood estimates (MLE) of the parameters with their 95% confidence intervals are provided in **Table 1**. The fit of the models has been investigated through the R^2-statistic in a linear regression analysis of the predicted versus the observed numbers (**Table 1**).

Table 1: Parameter estimates of the growth curve fitting the number of confirmed cases, with their 95% confidence intervals and the estimated goodness-of-fit statistic (R^2).

Province	α			β			M			$\log \sigma$			R^2
	MLE	LCL	UCL	MLE	LCL	UCL	MLE	LCL	UCL	MLE	LCL	UCL	
Anhui	13.3	13.3	13.4	0.28	0.27	0.29	993	989	997	9.3	7.5	11.4	99.9
Beijing	11.9	11.7	12.2	0.22	0.21	0.23	411	407	414	7.5	6.1	9.2	99.7
Chongqing	11.2	11	11.5	0.24	0.23	0.26	573	568	579	11.8	9.6	14.6	99.7
Fujian	10.7	10.5	11	0.26	0.25	0.28	294	292	297	6.5	5.3	8.1	99.6
Gansu	11.7	10.8	12.6	0.26	0.21	0.31	96	93	99	7.4	6.0	9.1	95.4
Guangdong	11.7	11.5	11.8	0.28	0.27	0.29	1342	1334	1350	17.6	14.3	21.7	99.9
Guangxi	12.2	12	12.3	0.22	0.21	0.23	252	250	254	3.8	3.1	4.7	99.8
Guizhou	14.9	14.7	15.1	0.3	0.28	0.32	147	146	149	3.2	2.6	3.9	99.7
Hainan	12.6	12.3	12.9	0.23	0.22	0.24	170	168	171	3.6	3.0	4.5	99.6
Hebei	14.9	14.7	15.2	0.23	0.22	0.24	319	316	323	7.1	5.8	8.7	99.6
Heilongjiang	15.7	15.6	15.9	0.27	0.26	0.28	482	478	485	7.0	5.7	8.6	99.9
Henan	12.7	12.6	12.8	0.28	0.27	0.29	1270	1263	1277	15.6	12.7	19.2	99.9
Hubei	18.7	18.3	19.1	0.24	0.22	0.25	67625	66376	68875	7.7	7.5	7.9	99.3
Hunan	11.9	11.8	12.1	0.29	0.28	0.3	1018	1011	1024	14.9	12.1	18.3	99.9
Inner Mongolia	13.5	13.2	13.8	0.22	0.21	0.23	76	75	76	1.8	1.5	2.2	99.5
Jiangsu	13.5	13.3	13.7	0.26	0.25	0.27	635	630	640	11.4	9.2	14.0	99.8
Jiangxi	13.1	13	13.2	0.29	0.28	0.3	937	932	943	12.5	10.2	15.4	99.9
Jilin	13.6	13.4	13.8	0.34	0.32	0.36	92	91	93	1.8	1.4	2.2	99.8
Liaoning	10.3	10.1	10.5	0.26	0.25	0.28	122	121	124	2.6	2.1	3.2	99.6
Ningxia	14.4	13.9	14.8	0.21	0.19	0.23	74	73	75	2.4	1.9	2.9	99.2
Qinghai	9.1	8.7	9.4	0.38	0.34	0.43	18	18	18	0.8	0.6	1.0	98.4
Shaanxi	11.6	11.5	11.8	0.27	0.26	0.28	244	243	246	3.7	3.0	4.6	99.8
Shandong	17.5	16.4	18.6	0.15	0.13	0.17	781	748	813	41.5	33.8	51.1	97.6
Shanghai	10.6	10.4	10.7	0.28	0.27	0.29	336	334	338	4.9	4.0	6.0	99.8
Shanxi	11.6	11.4	11.8	0.31	0.29	0.32	133	131	134	2.6	2.1	3.2	99.7
Sichuan	13.2	12.8	13.5	0.21	0.2	0.22	535	529	542	13.3	10.8	16.4	99.5
Tianjin	13.9	13.7	14.2	0.21	0.2	0.22	137	136	138	2.3	1.9	2.8	99.8
Xinjiang	15.6	15.3	15.9	0.22	0.21	0.24	77	77	78	1.9	1.5	2.3	99.6
Yunnan	10.1	9.7	10.5	0.27	0.25	0.3	172	169	175	6.3	5.1	7.8	98.9
Zhejiang	10.5	10.2	10.7	0.32	0.3	0.34	1195	1184	1205	26	21.1	32	99.6

The R^2 values range from 95.4% to 99.9%. For 26 out of 30 investigated provinces the R^2 is larger than or equal to 99.0%, and 24 out of 30 provinces have an R^2 larger than or equal to 95.5%. Thus the three-parameter logistic curve seems to fit very well to the data on the virus in China.

Investigating the parameter estimates of the province-specific growth curves and their 95% confidence intervals, it is evident that there is variation between provinces on all three growth curve parameters. The parameter α ranges from 9.1 [8.7; 9.4] in Qinghai to 18.7 [18.3; 19.1] in Hubei (and to 17.5 [16.4; 18.6] in Shangdong when Hubei is excluded). The parameter β ranges from 0.15 [0.13; 0.17] in Shangdong to 0.38 [0.34; 0.43] in Qinghai. Finally, the parameter M ranges from 18 [18,18] in Qinghai to 67625 [66376; 68875] in Hubei (and to 1342 [1334; 1350] in Guangdong when excluding Hubei). The total number of confirmed infections in mainland China is given by 80556 (and without Hubei this number is 12931).

Investigating the number of deaths with the growth model is only possible for the provinces Heilongjiang, Henan and Hubei, since the other provinces have not enough counts on deaths to model them accurately (they had less than 10 deaths). The parameter estimates and the goodness-of-fit are given in **Table 2**.

The R^2 values for the three countries are larger than or equal to 98.9%, showing a very good fit and again the parameter estimates seem to vary between the three provinces.

Based on the estimated maximum number of cases and deaths at the last day included in our dataset (March 09, 2020), the expected percentage of deaths from COVD-19 can be estimated: Heilongjiang 2.70% (=13/481), Henan 1.73% (=22/1272) and Hubei 4.41% (=2986/67707). This shows relevant differences in the percentages of confirmed deaths with respect to the confirmed infections.

Since there is literature on the relation between the parameters of the growth curve (i.e. the parameter α is a function of the maximum value, the value at day zero and the parameter β), we investigate the correlation between the parameters. Kendall's tau correlation is equal to -0.30 for (α, β), -0.02 for (α, M), and 0.12 for (β, M).

Table 2: Parameter estimates of the growth curve fitting the number of deaths with their 95% confidence intervals and the estimated goodness-of-fit (R^2). Estimation is possible only for the provinces Heilongjiang, Henan and Hubei.

Province	α			β			M			$\log\sigma$			**R^2**
	MLE	**LCL**	**UCL**	**MLE**	**LCL**	**UCL**	**MLE**	**LCL**	**UCL**	**MLE**	**LCL**	**UCL**	
Heilongjiang	18.6	18.2	19.1	0.25	0.22	0.28	12.9	12.6	13.2	-0.6	-0.8	-0.4	98.9
Henan	22.5	22.0	23.0	0.25	0.22	0.28	21.7	21.1	22.3	-0.1	-0.3	0.1	99.0
Hubei	23.6	23.4	23.9	0.17	0.16	0.17	3007	2968	3045	3.7	3.5	3.9	99.9

Finally, we aggregate all the numbers of confirmed infections and deaths in mainland China and analyze them with the growth model. We take all provinces and all provinces without Hubei, since Hubei has a very strong influence on the overall curve. The parameter estimates and the goodness-of-fit are provided in **Table 3** and **Table 4** when all provinces are included and when we exclude Hubei, respectively. Here the goodness-of-fit is better for the number of confirmed infections than for the number of confirmed deaths, making it more difficult to predict the confirmed deaths. According to these estimates, the mortality rate of COVD-19 is 3.86% when all provinces are considered and it is 0.86% when Hubei is excluded.

Table 3: Parameter estimates of the growth curve fitting the number of confirmed cases and the number of deaths in mainland China with their 95% confidence intervals and the estimated goodness-of-fit.

Outcome	α			β			M			$\log\sigma$			**R^2**
	MLE	**LCL**	**UCL**	**MLE**	**LCL**	**UCL**	**MLE**	**LCL**	**UCL**	**MLE**	**LCL**	**UCL**	
Deaths	23.6	23.3	23.8	0.17	0.16	0.17	3117	3079	3154	3.7	3.4	3.9	92.3
Cases	17.7	17.4	18.1	0.22	0.21	0.24	80740	79432	82048	7.8	7.6	8.0	99.4

We have also plotted the estimated growth model (continuous black line) together with the observed accumulated numbers (black stars) for the confirmed infections and deaths in **Figure 1**. The dashed red lines in the same plots represent the estimated maximum values.

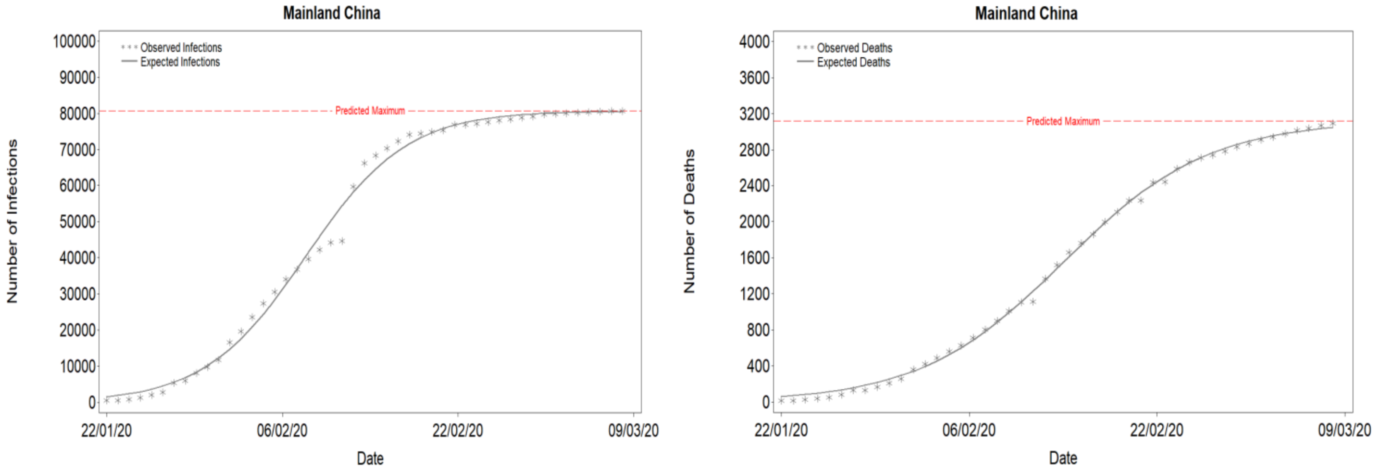

Figure 1: Estimated and observed number of confirmed deaths (top and bottom left) and number of confirmed cases (top and bottom right) in mainland China.

Table 4: Parameter estimates for the growth curve fitting the number of confirmed cases (Cases) and the number of deaths (Deaths) in mainland China (when Hubei is excluded) with their 95% confidence intervals and the estimated goodness-of-fit.

Outcome	α			β			M			$\log\sigma$			R²
	MLE	LCL	UCL	MLE	LCL	UCL	MLE	LCL	UCL	MLE	LCL	UCL	
Deaths	22.4	22.1	22.7	0.21	0.20	0.22	111	109	113	0.8	0.6	1.0	79.5
Cases	12.5	12.4	12.7	0.25	0.24	0.26	12873	12787	12958	5.2	5.0	5.4	99.9

STEP 2: Investigating bias in prediction based on early data

The growth curves of the provinces in China have almost reached the maximum value at 46 days after January 22, 2020. So, we could evaluate predictions of the values in the next few days based on an earlier period of observed data. We fit the growth curves after x days, with $x \in \{13, 14, \dots, 46\}$. We compare the estimates α_x, β_x, and M_x after x days with the final estimate from **Table 1**, expressed as a ratio. We also compare the predictions of the next three days after x days with the observed values after x days. Again we calculate the ratio of the predicted value with respect to the observed value. **Figure 2** shows the ratio of estimated parameters at day x with respect to the estimated parameters at day 46. Each line represents a province. The red line in the middle is the median line over all provinces. **Figure 3** presents the average bias over all province for one, two, and three day ahead predictions.

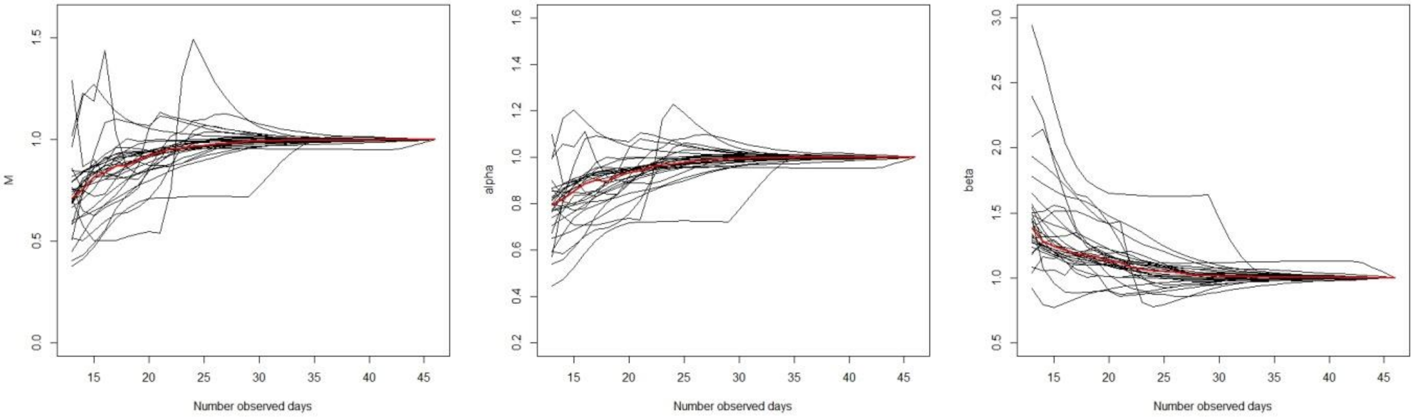

Figure 2: Parameter estimates of the fitted growth curve after x days of data are collected, compared to the parameter estimate at the final day expressed as ratio (left M, middle α, right β). Each line represents a province, while the red line is the day average across provinces.

The lines indicate that the parameters M and α on average are underestimated at 13 days and slowly converge to the final value if more data is used. The parameter β is overestimated, and it also slowly converges to the final estimate. From this we have created calibration curves (in particular for the maximum M) based on the data of the Chinese provinces. These can be used to correct early predictions for underestimation. However, the figures also show that there is variability and bias correction should be evaluated for each new country.

The figures on the predictions also show that the prediction underestimates what happens, but it seems to get closer to zero after 30 days of data, and almost vanishes when more data is available. It also shows that bias is larger when predictions are further ahead from the observed data. The prediction biases of the Chinese provinces are used to correct predictions for other countries.

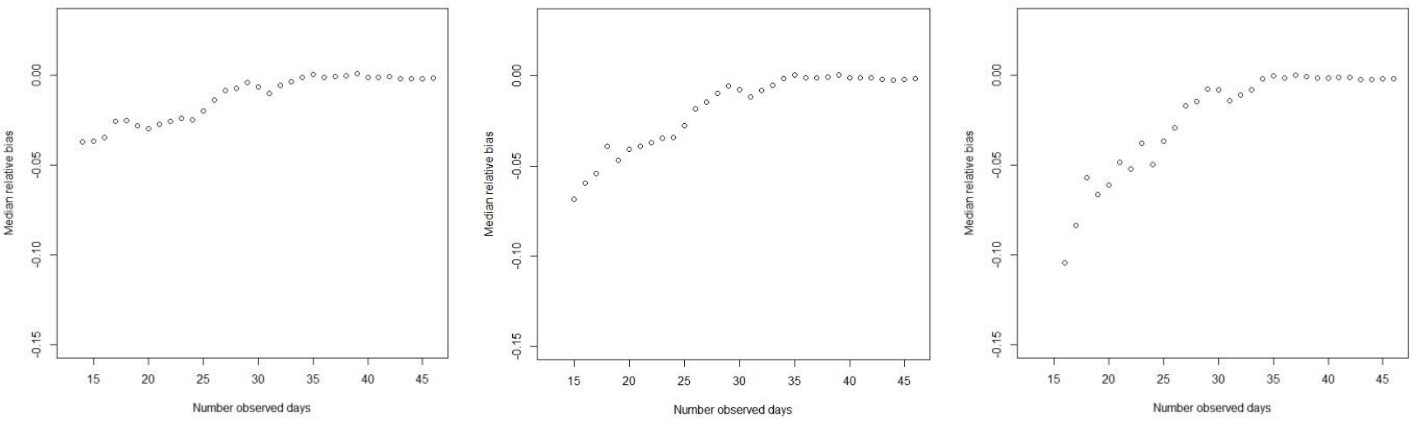

Figure 3: Bias in predictions for one (left), two (middle), and three (right) days ahead based on a fitted curve after x days of data are collected, compared to the observed values.

STEP 3: Evaluation of data of different countries

The analysis of the Chinese data has shown that it is not straightforward to make predictions. Thus data of each country first need to be investigated before we decide to make predictions. One issue is the dynamic of the virus diffusion, since some countries seem to have a longer period before a rapid growth occurs and other counties seem to grow faster in the beginning. It is therefore difficult to find the optimal date at which we start modelling the data. We have made the choice of taking as starting point for new countries the day at which a first death has been reported. We illustrate some of the evaluations we do for each country to decide if we have enough confidence to report our results to the public. We use the example of South Korea. **Table 5** shows the parameter estimate of M at 13, 14, 15, 16, and 17 days of data collected after the day of the first reported death. We also check whether the model converged without numerical issues, and provide the R^2 of the fitted model. Finally, in the last three columns, we display the one-, two- and three-days ahead predictions.

Table 5: Example. Curve growth estimation on the data from South Korea at 13, 14, 15, 16 and 17 after the first reported death. We report M, R^2, the status of convergence and one, two and three days ahead predictions.

Days	M (SE)	R^2	Converged	1-day	2-day	3-day
13	7642 (396)	99.8	Yes	6160	6552	6854
14	7504 (248)	99.9	Yes	6489	6775	6986
15	7646 (185)	99.9	Yes	6851	7077	7243
16	7853 (157)	99.9	Yes	7198	7384	7519
17	7953 (126)	99.9	Yes	7448	7592	7696

Looking at the estimates of the parameter M, we see a slow increase in the estimated value, which has a similar pattern as the one observed in the Chinese data. Based on these evaluations we decide if we produce calibrated estimates to go public on the TU/e website. For South Korea we reported the following predictions on March 11, 2020.

Table 6: Predictions that we reported on the TU/e website (on March 11, 2020).

Number of infections	Observed Infections Yesterday	Prediction (95% Lower and Upper Prediction Interval)			
		Today	Tomorrow	Next day	Final
South Korea	7513	7845 (7674; 8015)	7974 (7800; 8147)	8137 (7960; 8314)	8807 (8637; 8975)

Table 5 and **Table 6** show that the reported numbers deviate from the direct prediction of the three-parameter logistic curve.

Assumptions and Comments

The growth curve model essentially describes the accumulation of several different dynamic processes that are underneath the data. The first process is the spread of the virus among the population, and depends most likely on many different factors like human behavior. The second process is the sampling or testing approach to find infected people in the population. If serious changes in this strategy are being made, like a reduction of the number of tests, the estimated growth curve can seriously underestimate the severity of the virus spread. It is important that testing strategy does not change over time to be able to judge a change in the confirmed number of infections. The third process is determined by the set of policy measures that tries to influence or prohibit the spread of the corona virus. This may affect the growth curve and will then become visible in our projections. Thus when testing strategies are being maintained over time our growth curve shows the way the virus is behaving on confirmed infections and deaths, together with the effect of the policy.

It is important to note that the growth curves do not (necessarily) provide predictions of the true number of infections and deaths in the population due to the corona virus, but only describes the growth in confirmed number of infections and deaths. To obtain estimates on the true number of inhabitants in the population infected by the virus, we need to better understand the sampling strategy, the numbers of positive and negative tests among all tested inhabitants and the sensitivity and specificity of the corona test. With this information the growth curves may possibly be adapted to provide population numbers.