Estimation of the final size of the second phase of the coronavirus COVID 19 epidemic by the logistic model

Milan Batista
University of Ljubljana, Slovenia
milan.batista@fpp.uni-lj.si

(Mar 2020)

Abstract

In the note, the logistic growth regression model is used for the estimation of the final size and its peak time of the coronavirus epidemic in China, South Korea, and the rest of the World.

1 Introduction

In the previous article [1], we try to estimate the final size of the epidemic for the whole World using the logistic model and SIR model. The estimation was about 83000 cases. Both models show that the outbreak is moderating; however, new data showed a linear upward trend. It turns out that the epidemy in China was slowing but is begin to spread elsewhere in the World.

In this note, we will give forecasting epidemic size for China, South Korea, and the rest of the World and daily predictions using the logistic model. Full daily reports for China, Germany, Iran, Italy, Slovenia, South Korea, Spain and counties outside of China generated are available at

https://www.researchgate.net/publication/339912313_Forecasting_of_final_COVID-19_epidemic_size

The MATLAB program *fitVirus* used for calculations is freely available from

https://www.mathworks.com/matlabcentral/fileexchange/74411-fitvirus

We note that logistic models give similar results as the SIR model (at least for the case of China and South Korea). However, the logistic model is given by explicit formula and is thus much simpler for regression analysis than the SIR model, where one must on each optimization step solve a system of ordinary differential equations. (One may,

however, use approximate solution and thus obtain four-parameter problem which can be very sensitive to initial guess). Yet, the logistics model has its drawbacks as the epidemic approaches its final stage: the actual number of cases may be slightly larger than that predicted by the logistics model. If the actual number of cases begins to exceed the predicted end-state systematically, then a second phase of the epidemic is likely to occur, and the model will no longer be applicable.

2 Logistic grow model

In mathematical epidemiology, when one uses a phenomenological approach, the epidemic dynamics can be described by the following variant of logistic growth model [2-5]

$$\frac{dC}{dt} = rC\left(1 - \frac{C}{K}\right), \tag{1}$$

where C is an accumulated number of cases, $r > 0$ infection rate, and $K > 0$ is the final epidemic size. If $C\left(0\right) = C_0 > 0$ is the initial number of cases then the solution of (1) is

$$C = \frac{K}{1 + A\exp\left(-rt\right)}, \tag{2}$$

where $A = \dfrac{K - C_0}{C_0}$ When $t \ll 1$, assuming $K \gg C_0$, and therefore $A \gg 1$ we have the natural growth

$$C = \frac{Ke^{rt}}{e^{rt} + A} = \frac{C_0}{1 - C_0/K}\frac{e^{rt}}{1 + e^{rt}/A} \approx C_0 e^{rt} \tag{3}$$

When $t \to \infty$ the number of cases follows the Weibull function

$$C = K\left(1 - Ae^{-rt} + \cdots\right) \approx K\left[1 - e^{-r\left(t - t_0\right)}\right] \tag{4}$$

The growth rate $\dfrac{dC}{dt}$ reaches its maximum when $\dfrac{d^2C}{dt^2} = 0$. From this condition, we obtain that the growth rate peak occurs in time time

$$t_p = \frac{\ln A}{r} \tag{5}$$

At this time the number of cases is

$$C_p = \frac{K}{2}, \tag{6}$$

and the growth rate is

$$\left(\frac{dC}{dt}\right)_p = \frac{rK}{4} \tag{7}$$

To answer the question about doubling time Δt, i.e., the time takes to double the number of cases we solve $C\left(t+\Delta t\right) = 2C\left(t\right)$ for Δt. Result is

$$\Delta t = \frac{\ln 2}{r} - \frac{1}{r}\ln\left(\frac{1}{A} - e^{-rt}\right) - t \tag{8}$$

The first term represents initial exponential growth, then Δt increases with t. When $t \to t_p = \frac{\ln A}{r}$, i.e., when $C \to K/2$, then $\Delta t \to \infty$. For $C \geq K/2$ doubling time lost its meaning.

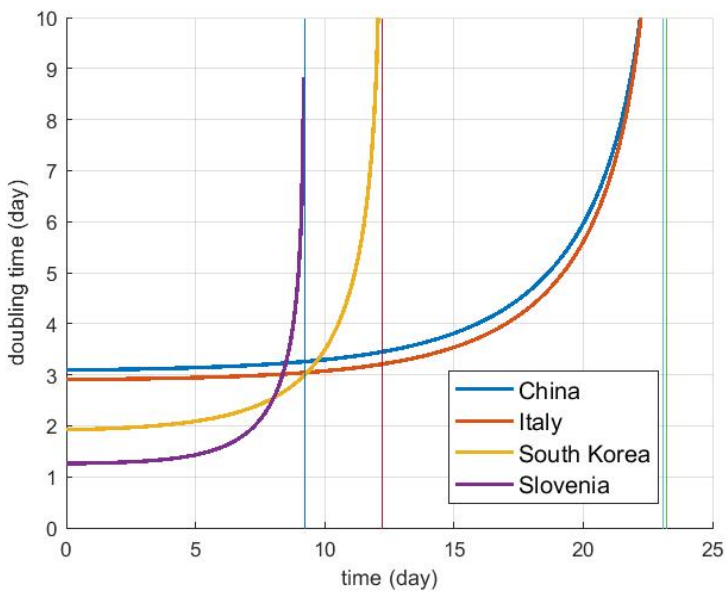

Figure 1. Doubling time (data up to 14 mar 2020)

Now, if $C_1, C_2, \ldots, C_n$ are the number of cases at times $t_1, t_2, \ldots, t_n$, then the final size predictions of the epidemic based on these data are $K_1, K_2, \ldots, K_n$. When convergence is achieved, then one may try to predict the final epidemic size by iterated Shanks transformation [6]

$$K = \frac{K_{n+1}K_{n-1} - K_n^2}{K_{n+1} - 2K_n + K_{n-1}}. \tag{9}$$

There is no natural law or process behind this transformation; therefore, it must be used with some care. In particular, the calculated limit is useless if $K < C_n$, i.e., it is below the current data.

The logistic model (2) contains three parameters: K, r, and A, which should be determined by regression analysis. Because the model is nonlinear, some care should be taken for initial guess. First of all, in the early stage, the logistic curve follows an exponential growth curve (3) , so the estimation of K is practically impossible. With enough data, the initial guess can be obtain in the following way. Expressing t from (2) and use three equidistant data point yield the following system of three equations:

$$t_k - 2m = \frac{1}{r}\ln\left(\frac{AC_{k-2m}}{K - C_{k-2m}}\right), \quad t_k - m = \frac{1}{r}\ln\left(\frac{AC_{k-m}}{K - C_{k-m}}\right), \quad .t_k = \frac{1}{r}\ln\left(\frac{AC_k}{K - C_k}\right) \tag{10}$$

This system has a solution [7]

$$K = \frac{C_{k-m}\left(C_{k-2m}C_{k-m} - 2C_{k-2m}C_k + C_{k-m}C_k\right)}{C_{k-m}^2 - C_k C_{k-2m}}. \tag{11}$$

$$r = \frac{1}{m}\ln\frac{C_k\left(C_{k-m} - C_{k-2m}\right)}{C_{k-2m}\left(C_k - C_{k-m}\right)}. \tag{12}$$

$$A = \frac{\left(C_k - C_{k-m}\right)\left(C_{k-m} - C_{k-2m}\right)}{\left(C_{k-m}^2 - C_k C_{k-2m}\right)}\left[\frac{C_k\left(C_{k-m} - C_{k-2m}\right)}{C_{k-2m}\left(C_k - C_{k-m}\right)}\right]^{\frac{t_k - m}{m}}. \tag{13}$$

The solution is acceptable when all the unknowns are positive.

Formulas (10),(11),(12) are used to calculate the initial approximation in the *fitVirus* program. For practical calculation, we take the first, the middle, and the last data point. If this calculation fails, we consider regression analysis as questionable. Final

values of the parameters K, r, and A are then calculated by least-square fit using the MATLAB functions *lsqcurvefit* and *fitnlm*.

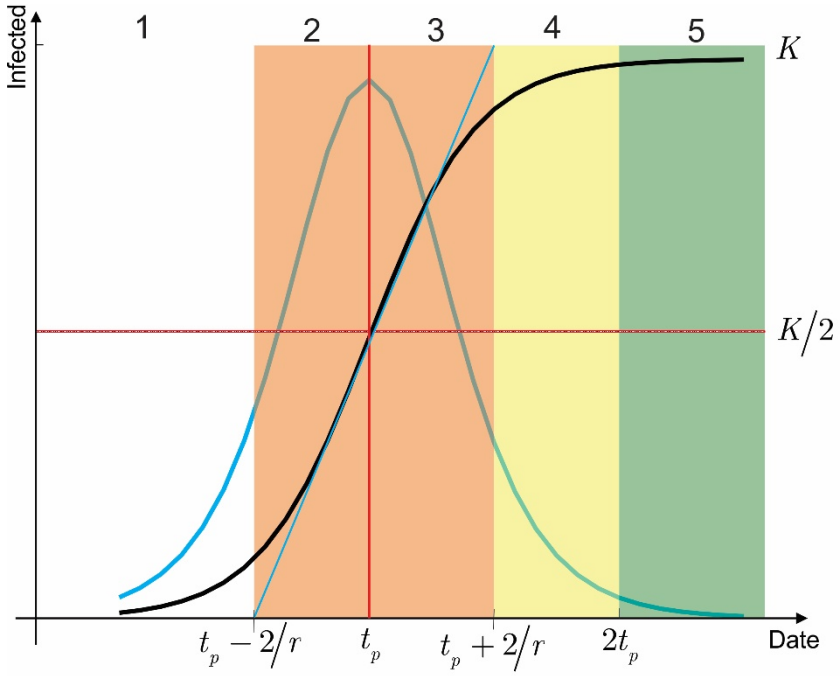

Figure 2. Epidemic phases

Before we proceed, we for convenience, introduce the following epidemy phases (see Fig 2):

1. Phase 1 – exponential growth (lag phase, slow growth) : $t < t_p - 2/r$
2. Phase 2 – fast growth (positive growth phase, acceleration phase) to an epidemy turning point: $t_p - 2/r < t < t_p$
3. Phase 3 – fast growth to steady-state (negative growth phase, deceleration phase): $t_p < t < t_p + 2/r$
4. Phase 4- steady-state (transition phase, slow growth, asymptotic): $t_p + 2/r < t < 2t_p$
5. Phase 5 – steady) ending phase (plateau stage): $t > 2t_p$

The duration of the fast-growing period is thus equal to $\tau = 4/r$. We note that the names of the phases are not standard, and are arbitrarily chosen.

3 Results

3.1 China (11.Mar 2020)

On the base of available data, one can predict that the final size of coronavirus epidemy in China using the logistic model will be approximately 81 000 ± 500 cases (Table 1) and that the peak of the epidemic was on 8 Feb 2020 (Table 2). It seems that the epidemic in China is in the ending stage (Fig 3, Fig 4).

The short-term forecasting is given in Table 3 where we see that the discrepancy of actual and forecasted number of cases is within 2%. However, actual and predicted daily new cases are scattered and vary between 13% to 300%. On 7 Mar 2020, the actual number of cases was 80695, and the daily number of cases was 44. Prediction in Table 3 is cumulative 80588 cases and 39 daily cases. The errors are 0.1% and 11%, respectively.

Table 1. Estimated logistic model parameters for China (data up to 11.Mar 2020)

	Estimate	SE	tStat	pValue
K	**80772**	489.04	165.16	2.0205e-72
r	0.22568	0.0062941	35.855	2.4094e-38
A	191.23	26.978	7.0882	3.5795e-09

```
Number of observations: 55, Error degrees of freedom: 52
Root Mean Squared Error: 1.9e+03
R-Squared: 0.997,  Adjusted R-Squared 0.997
F-statistic vs. zero model: 1.61e+04, p-value = 4.02e-77
```

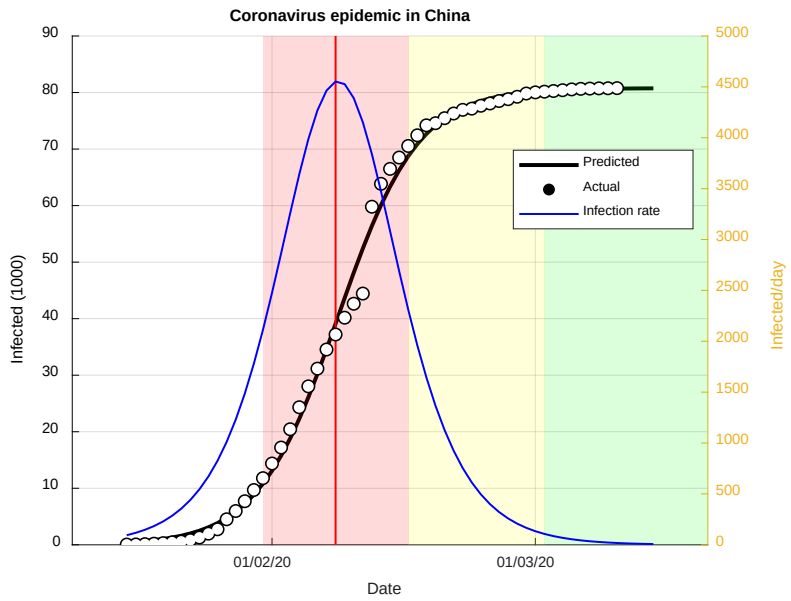

Figure 3. Predicted evaluation of coronavirus epidemic in China

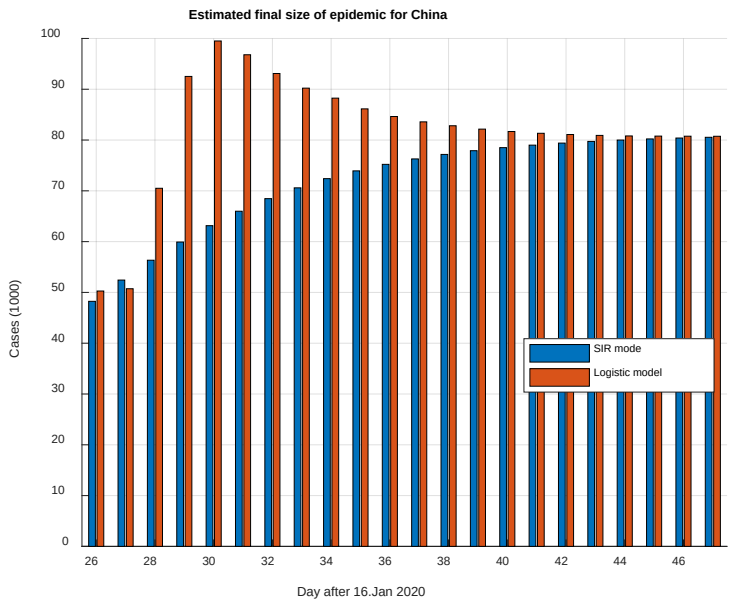

Figure 4. Predicted final size of coronavirus epidemic in China (prediction from 2. Mar 2020)

Table 2. Results of daily logistic regression for China (data from 4.Mar 2020)

day	date	C (cases)	K (cases)	r (1/day)	A	t_peak (day)	dC_peak (cases/day)	date_peak
16	31.jan.20	11789	17884	0.452	460.109	13	2019	30.jan.20
17	1.feb.20	14378	21578	0.415	398.528	14	2238	31.jan.20
18	2.feb.20	17203	25879	0.383	353.422	15	2476	1.feb.20
19	3.feb.20	20438	31161	0.354	319.58	16	2753	2.feb.20
20	4.feb.20	24332	38706	0.324	294.507	17	3136	3.feb.20
21	5.feb.20	28028	44255	0.308	283.606	18	3406	4.feb.20
22	6.feb.20	31161	45828	0.303	279.609	18	3476	4.feb.20
23	7.feb.20	34546	47586	0.298	272.671	18	3545	4.feb.20
24	8.feb.20	37198	48066	0.296	269.981	18	3561	4.feb.20
25	9.feb.20	40171	49272	0.292	261.091	19	3594	5.feb.20
26	10.feb.20	42638	50275	0.288	252.341	19	3614	5.feb.20
27	11.feb.20	44438	50724	0.286	247.719	19	3621	5.feb.20
28	12.feb.20	59800	70500	0.226	154.092	22	3989	8.feb.20
29	13.feb.20	63850	92522	0.198	138.603	24	4585	10.feb.20
30	14.feb.20	66492	99501	0.192	137.601	25	4785	11.feb.20
31	15.feb.20	68500	96788	0.195	138.393	25	4712	11.feb.20
32	16.feb.20	70548	93100	0.199	141.105	24	4625	10.feb.20
33	17.feb.20	72436	90217	0.203	144.985	24	4569	10.feb.20
34	18.feb.20	74199	88248	0.206	149.059	24	4540	10.feb.20
35	19.feb.20	74576	86134	0.21	155.327	24	4519	10.feb.20
36	20.feb.20	75464	84630	0.213	161.403	23	4512	9.feb.20
37	21.feb.20	76288	83577	0.216	166.902	23	4513	9.feb.20
38	22.feb.20	76936	82810	0.218	171.816	23	4518	9.feb.20
39	23.feb.20	77150	82149	0.22	176.855	23	4525	9.feb.20
40	24.feb.20	77658	81677	0.222	181.054	23	4533	9.feb.20
41	25.feb.20	78064	81329	0.223	184.574	23	4541	9.feb.20
42	26.feb.20	78495	81087	0.224	187.292	23	4547	9.feb.20
43	27.feb.20	78824	80912	0.225	189.429	23	4552	9.feb.20
44	28.feb.20	79251	80805	0.226	190.85	23	4556	9.feb.20
45	29.feb.20	79824	80774	0.226	191.288	23	4557	9.feb.20
46	1.mar.20	80026	80755	0.226	191.566	23	4558	9.feb.20
47	2.mar.20	80151	80741	0.226	191.792	23	4558	9.feb.20

Table 3. Short-term forecasting for China

Day	Date	Actual	Predicted	Error %	Daily actual	Daily predicted	Error %
44	28.feb.20	79251	79814	0.945	427	232	45.667
45	29.feb.20	79824	80000	0.407	573	186	67.539
46	1.mar.20	80026	80149	0.302	202	149	26.238
47	2.mar.20	80151	80268	0.265	125	119	4.8
48	3.mar.20	-	80363	-	-	95	
49	4.mar.20	-	80439	-	-	76	
50	5.mar.20	-	80500	-	-	61	
51	6.mar.20	-	80549	-	-	49	
52	7.mar.20	-	80588	-	-	39	

3.2 South Korea (11.Mar 2020)

On the base of available data, one can predict that the final size of coronavirus epidemy in of South Korea using the logistic model will be approximately 8050 ±70 cases (Fig 5, Table 4) and that the peak of the epidemic was on 1 Mar 2020. The epidemic in South Korea appears to be in the steady-state transition phase. These figures were already predicted on 4. Mar 2020 (Table 5), i.e., the prediction was approximately 7500 to 8500 cases and that the peak will be around 2 Mar.

On 7 Mar 2020, the actual number of cases was 7134, and the daily number of cases was 367. Prediction in Table 5 is cumulative 6572 cases and 259 daily cases. The errors are 8 % and 30%, respectively.

Table 4. Estimated logistic model parameters for South Korea up to 13.Mar 2020

	Estimate	SE	tStat	pValue
K	8046.5	71.784	112.09	7.8279e-32
r	0.36657	0.010259	35.733	5.581e-21
A	87.97	9.8747	8.9087	9.4612e-09

```
Number of observations: 25, Error degrees of freedom: 22
Root Mean Squared Error: 129
R-Squared: 0.998,  Adjusted R-Squared 0.998
F-statistic vs. zero model: 1.26e+04, p-value = 1.03e-35
```

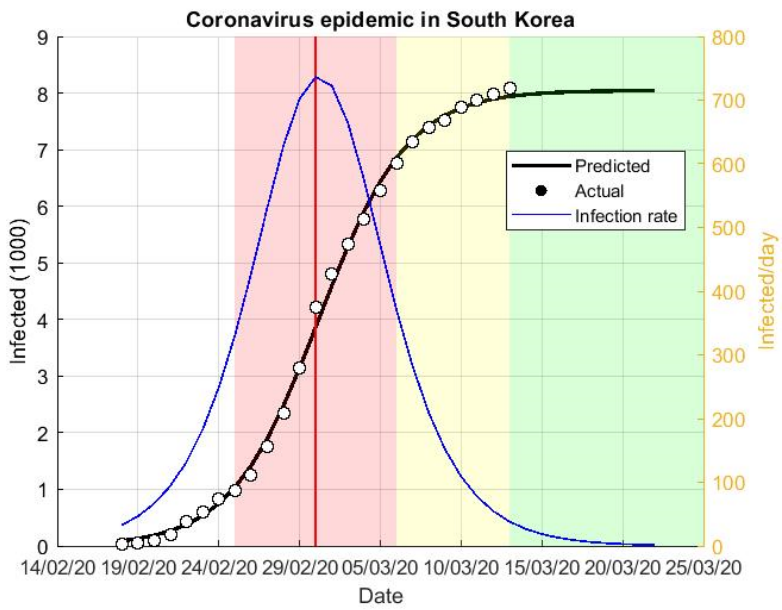

Figure 5. Predicted evaluation of coronavirus epidemic in South Korea (13.Mar 2020)

Table 5. Results of daily logistic regression for South Korea

day	date	C (cases)	K (cases)	r (1/day)	A	t_peak (day)	dC_peak (cases/day)	date_peak
7	24.feb.20	833	1105	0.824	47.332	4	227	23.feb.20
8	25.feb.20	977	1133	0.808	46.526	4	228	23.feb.20
9	26.feb.20	1261	1508	0.633	37.083	5	238	24.feb.20
10	27.feb.20	1766	3355	0.426	44.537	8	357	27.feb.20
11	28.feb.20	2337	7021	0.363	76.656	11	636	1.mar.20
12	29.feb.20	3150	17145	0.328	165.531	15	1405	5.mar.20
13	1.mar.20	3736	8639	0.359	95.534	12	775	2.mar.20
14	2.mar.20	4335	7214	0.378	88.434	11	682	1.mar.20

Table 6. Short-term forecasting for South Korea

day	date	actual	predict	Error %	Daily actual	Daily predict.	Error %
11	28.feb.20	2337	2393	29.7	571	562	1.6
12	29.feb.20	3150	3031	17.7	813	638	21.5
13	1.mar.20	3736	3708	17.2	586	677	15.5
14	2.mar.20	4335	4378	15.3	599	670	11.9
15	3.mar.20	-	4997	-	-	619	
16	4.mar.20	-	5532	-	-	535	
17	5.mar.20	-	5971	-	-	439	
18	6.mar.20	-	6313	-	-	342	
19	7.mar.20	-	6572	-	-	259	
20	8.mar.20	-	6761	-	-	189	

3.3. Rest of the World

The comparison of the predicted final sizes is shown in the graph in Figure 6. Based on the data from 11. Mar 2020, a very rough estimate indicates that the number of cases will be about 90000 (Fig 7); data from 13. Mar 2020 rise this number to 380 000. However, it is an early-stage epidemic, so these estimates are very questionable and will be changed with new data.

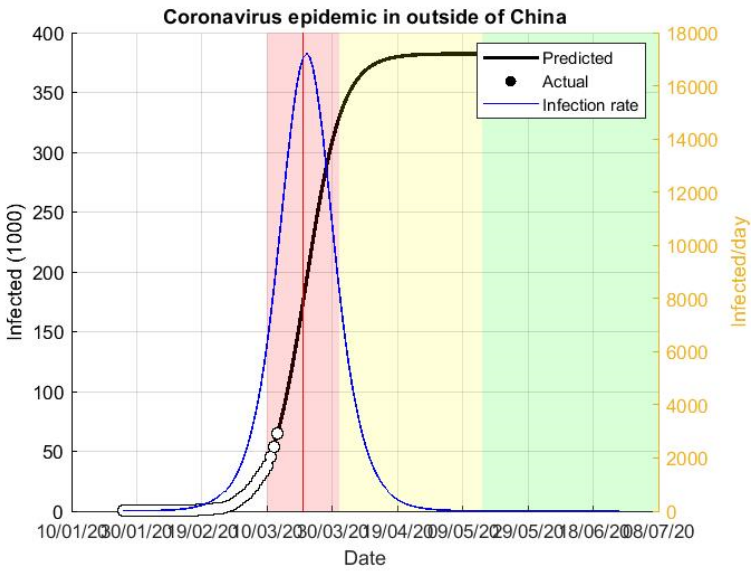

Figure 6. Predicted evaluation of coronavirus epidemic outside of China

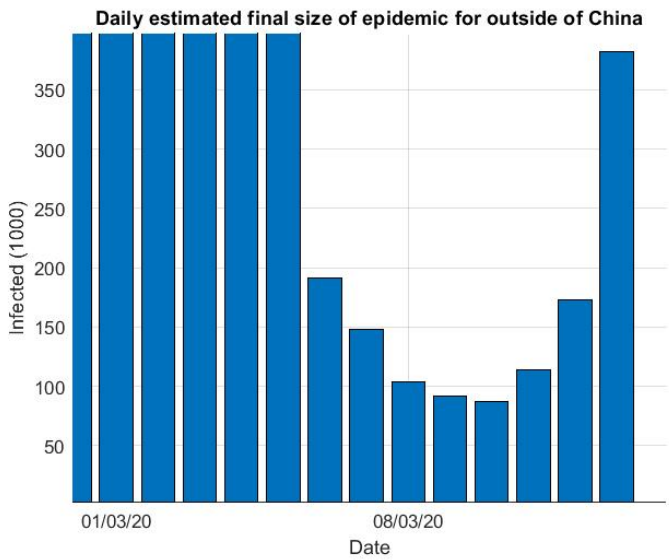

Figure 7. Predicted final size of coronavirus epidemic outside of China

4 Conclusion

On the base of available data, one can predict that the final size of coronavirus epidemy in China will be around 81 000 cases. For South Korea, a prediction is about 8000 cases. For the rest of the World, the forecasts are still very unreliable in is now approximately 380 000 cases.

We emphasize that the logistics model is a phenomenological, data-driven model. Thus, its forecasts are as reliable as useful are data and as good as the model can capture the dynamics of the epidemic. As daily epidemic size forecasts begin to converge, it can be said that the outbreak is under control. However, any systematic deviation from the forecast curve may indicate that the epidemic is escaping control. An example is China. By 25. Feb., the data follows a logistic curve and then begins to deviate from it. We now know that this was the beginning of the second stage of the epidemic, which is now spreading around the World. A similar linear trend can now be observed for South Korea (Fig 5); we hope this does not mark the beginning of the second stage of the epidemic in this country.

References

[1] M. Batista, Estimation of the final size of the COVID-19 epidemic, medRxiv, (2020) 2020.2002.2016.20023606.

[2] X.S. Wang, J.H. Wu, Y. Yang, Richards model revisited: Validation by and application to infection dynamics, J Theor Biol, 313 (2012) 12-19.

[3] B. Pell, Y. Kuang, C. Viboud, G. Chowell, Using phenomenological models for forecasting the 2015 Ebola challenge, Epidemics, 22 (2018) 62-70.

[4] F. Brauer, Mathematical epidemiology: Past, present, and future, Infectious Disease Modelling, 2 (2017) 113-127.

[5] S.L. Chowell G, Viboud C, Kuang Y., West Africa Approaching a Catastrophic Phase, or is the 2014 Ebola Epidemic Slowing Down? Different Models Yield Different Answers for Liberia. , PLOS Currents Outbreaks., (2014).

[6] C.M. Bender, S.A. Orszag, Advanced mathematical methods for scientists and engineers I asymptotic methods and perturbation theory, Springer, New York, 1999.

[7] P.F. Verhulst, Notice sur la loi que la population poursuit dans son accroissement, Correspondance mathématique et physique, 10 (1838) 113-121.