Supporting information file

Simulating and Forecasting the Cumulative Confirmed Cases of SARS-CoV-2 in China by Boltzmann Function-based Regression Analyses

Xinmiao Fu, Qi Ying, Tieyong Zeng, Tao Long and Yan Wang

Methods

Sources of data

We collected the daily cumulative number of confirmed cases (from Jan 21, 2020 to Feb 14, 2020) infected by SARS-CoV-2 from official websites of the National Health Commission of China and of health commissions of provinces, municipalities and major cities. Overseas data were not included in our simulation due to the small number of confirmed cases. The cumulative number of confirmed cases of 2003 SARS in China and worldwide were obtained from the official website of WHO.

Data fitting with Boltzmann function and estimation of critical dates

Data were organized in Microsoft Excel and then incorporated into Microcal Origin software (note: 2021 Jan 21 was set as day 1 and so on). The Boltzmann function was applied to data simulation for each set of data regarding different geographic regions (e.g.., China, Hubei Province and so on) and parameters of each function were obtained, with the potential total number of confirmed cases being directly given by parameter A_2. Estimation of critical dates was performed by predicting the cumulative number of confirmed cases in the coming days post Feb 14, 2020, and the key dates were provisionally set when the number of daily new confirmed cases is lower than 0.1% of the potential total number. The Boltzmann function for simulation is expressed as follows:

$$C(x) = \mathrm{A}_2 + \frac{\mathrm{A}_1 - \mathrm{A}_2}{1 + e^{(x-x0)/dx}} \quad (1)$$

where $C(x)$ is the cumulative number of confirmed cases at day x; A_1, A_2, $x0$, and dx are constants. In particular, A_2 represents the estimated potential total number of confirmed cases of SARS-CoV-2. Details of derivation of the Boltzmann function for epidemic analysis are described in the supporting information file.

Estimation of uncertainty in the non-linear regression

A Monte Carlo technique is applied to assess the uncertainty in the estimated total number of confirmed cases due to the uncertainty in the reported number cases. 1000 non-linear regressions were performed with the same time series data but each data point in the time series was perturbed by multiplying with a random scaling factor that represents the relative uncertainty. We assumed that the relative uncertainty follows a single-sided normal distribution with a mean of 1.0 and a standard deviation of 10%. This implies that all reported cases are positive but there is a tendency to miss-reporting some positive cases so that the reported numbers represent a lower limit. The resulting mean and 95% confidence interval (CI) were presented.

Derivation of Boltzmann function for data fitting

We are interested in the derivation of the Boltzmann function which can be written as,

$$N(t) = A_2 + \frac{A_1 - A_2}{1 + e^{(t-t_0)/\Delta t}}$$

where A_1, A_2, t_0, and Δt are some suitable constant.

This formula is related to the sigmoid function except for a linear transform. Let us explain why this Boltzmann function is reasonable for explaining the data for the number of virus-infected cases. Readily, we have, $\mathrm{N}(\mathrm{t}_0) = \frac{\mathrm{A}_1+\mathrm{A}_2}{2}$, $\mathrm{N}(-\infty) = \mathrm{A}_1$, $\mathrm{N}(+\infty) = \mathrm{A}_2$. We can prove a simple result.

Theorem 1. *Denote* $\mathrm{c} = \frac{1}{(\mathrm{A}_2-\mathrm{A}1)\Delta \mathrm{t}}$ *, then we have*

$$\frac{\mathrm{dN(t)}}{\mathrm{dt}} = \mathrm{c}\big(\mathrm{A}_2 - \mathrm{N(t)}\big)(\mathrm{N(t)} - \mathrm{A}_1).$$

Proof. Calculate directly, we have,

$$\begin{aligned}
\frac{\mathrm{dN(t)}}{\mathrm{dt}} &= -(\mathrm{A}_1 - \mathrm{A}_2) \cdot \frac{\mathrm{e}^{(\mathrm{t}-\mathrm{t}_0)/\Delta \mathrm{t}}}{\left(1 + \mathrm{e}^{(\mathrm{t}-\mathrm{t}_0)/\Delta \mathrm{t}}\right)^2} \frac{1}{\Delta \mathrm{t}} \\
&= \frac{1}{\Delta \mathrm{t}} \cdot \frac{\mathrm{A}_2 - \mathrm{A}_1}{1 + \mathrm{e}^{(\mathrm{t}-\mathrm{t}_0)/\Delta \mathrm{t}}} \cdot \left(1 - \frac{1}{1 + \mathrm{e}^{(\mathrm{t}-\mathrm{t}_0)/\Delta \mathrm{t}}}\right) \\
&= \frac{1}{\Delta \mathrm{t}} \cdot \frac{\mathrm{A}_2 - \mathrm{A}_1}{1 + \mathrm{e}^{(\mathrm{t}-\mathrm{t}_0)/\Delta \mathrm{t}}} \cdot \left(1 - \frac{1}{1 + \mathrm{e}^{(\mathrm{t}-\mathrm{t}_0)/\Delta \mathrm{t}}}\right) \\
&= \frac{1}{(\mathrm{A}_1 - \mathrm{A}_2)\Delta \mathrm{t}}\big(\mathrm{A}_2 - \mathrm{N(t)}\big)\big(\mathrm{A}_1 - \mathrm{N(t)}\big) \\
&= \mathrm{c}\big(\mathrm{A}_2 - \mathrm{N(t)}\big)(\mathrm{N(t)} - \mathrm{A}_1).
\end{aligned}$$

The above relation is very interesting. If we define the cumulative infected cases as *N(t)*, Boltzmann function states that the increment of the $N(t)$ is proportional to the number of infected cases (except some infected cases, represented by A_1, who may not have the potency to infect other healthy people) and the remaining healthy pool. A_2 represents the potential maximal size of infected cases.

Furthermore, we assume that the cumulative number of confirmed cases (designated as *C(t)*) is proportional to the number of infected cases (i.e., *N(t)*) under specific circumstances, although the former has a lag time behind the latter. Then, *C(t)* can also be expressed in the form of Boltzmann function.

Results

Table S1 cumulative numbers of confirmed cases of SARS-CoV-2 before and after adjustment

date	mainland China				Hubei				Wuhan			
	a	*b*	*c*	*d*	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*
1/21	440		440	604	375		375	539				
1/22	571		571	765	444		444	638				
1/23	830		830	1069	549		549	788	459		459	758
1/24	1287		1287	1605	729		729	1047	572		572	945
1/25	1975		1975	2434	1052		1052	1511	618		618	1020
1/26	2744		2744	3365	1423		1423	2044	698		698	1153
1/27	4515		4515	5699	2714		2714	3898	1590		1590	2625
1/28	5974		5974	7524	3554		3554	5104	1905		1905	3146
1/29	7711		7711	9711	4586		4586	6586	2261		2261	3733
1/30	9692		9692	12224	5806		5806	8338	2639		2639	4358
1/31	11791		11791	14911	7153		7153	10273	3215		3215	5309
2/1	14380		14380	18337	9074		9074	13031	4109		4109	6785
2/2	17205		17205	22080	11177		11177	16052	5142		5142	8491
2/3	20438		20438	26335	13522		13522	19419	6384		6384	10541
2/4	24324		24324	31598	16678		16678	23952	8351		8351	13789
2/5	28018		28018	36594	19665		19665	28241	10117		10117	16706
2/6	31161		31161	40805	22112		22112	31756	11618		11618	19184
2/7	34546		34546	45429	24953		24953	35836	13603		13603	22462
2/8	37198		37198	49017	27100		27100	38919	14982		14982	24739
2/9	40171		40171	53094	29631		29631	42554	16902		16902	27909
2/10	42638		42638	56475	31728		31728	45565	18454		18454	30472
2/11	44750		44750	59302	33366		33366	47918	19558		19558	32295
2/12	59804	13332	46472	61681	48206	13332	34874	50083	32994	12364	20630	34065
2/13	63851	15384	48467	64430	51986	15384	36602	52565	35991	14031	21960	36261
2/14	66496	16522	49974	66496	54406	16522	37884	54406	37914	14953	22961	37914

a: reported cumulative number of confirmed cases.

b: cumulative number of confirmed cases determined by clinical features.

c: cumulative cases without those determined by clinical features.

d: adjusted cumulative number of confirmed cases for fitting are colored in red.

Figure S1

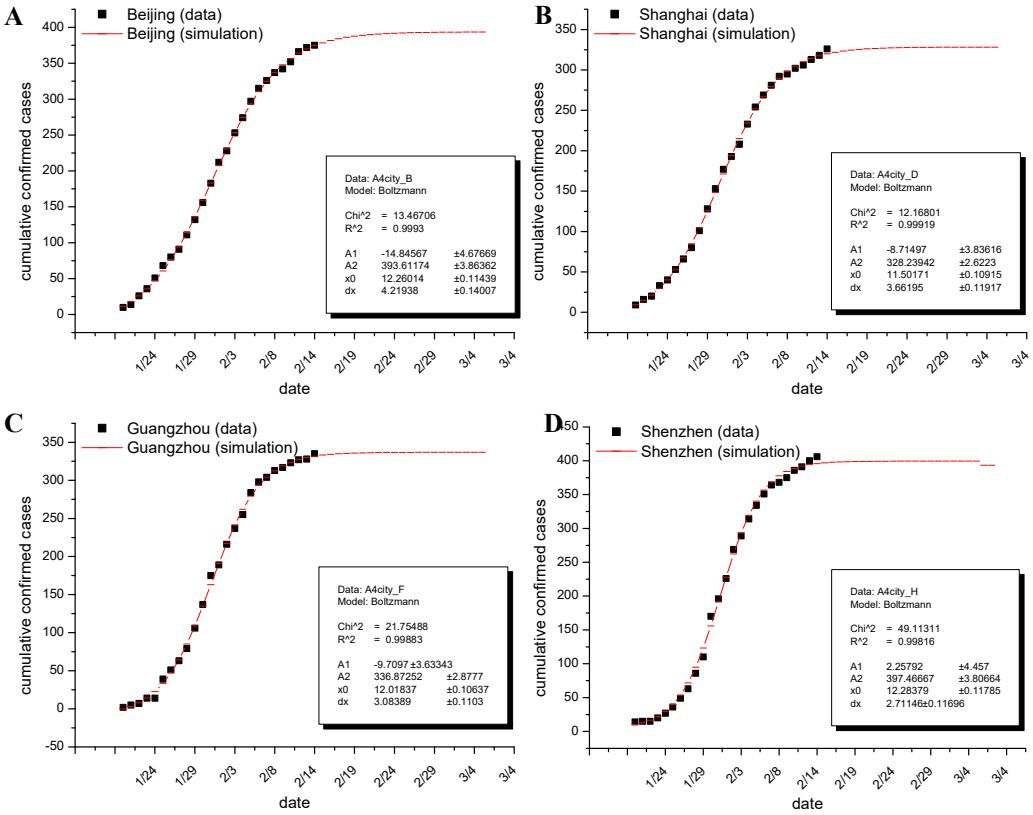

Figure S1 Fitting the cumulative number of confirmed cases from the top-4 major cities to Boltzmann function

Cumulative number of confirmed cases of SARS-CoV-2 as of Feb 14, 2020, in top4 major cities of mainland China (Beijing, panel A; Shanghai, panel B; Guangzhou, panel C; Shenzhen, panel D), are shown as black squares, and the simulation results from Boltzmann function are plotted as red short lines and parameters of each established function are shown in inserts.

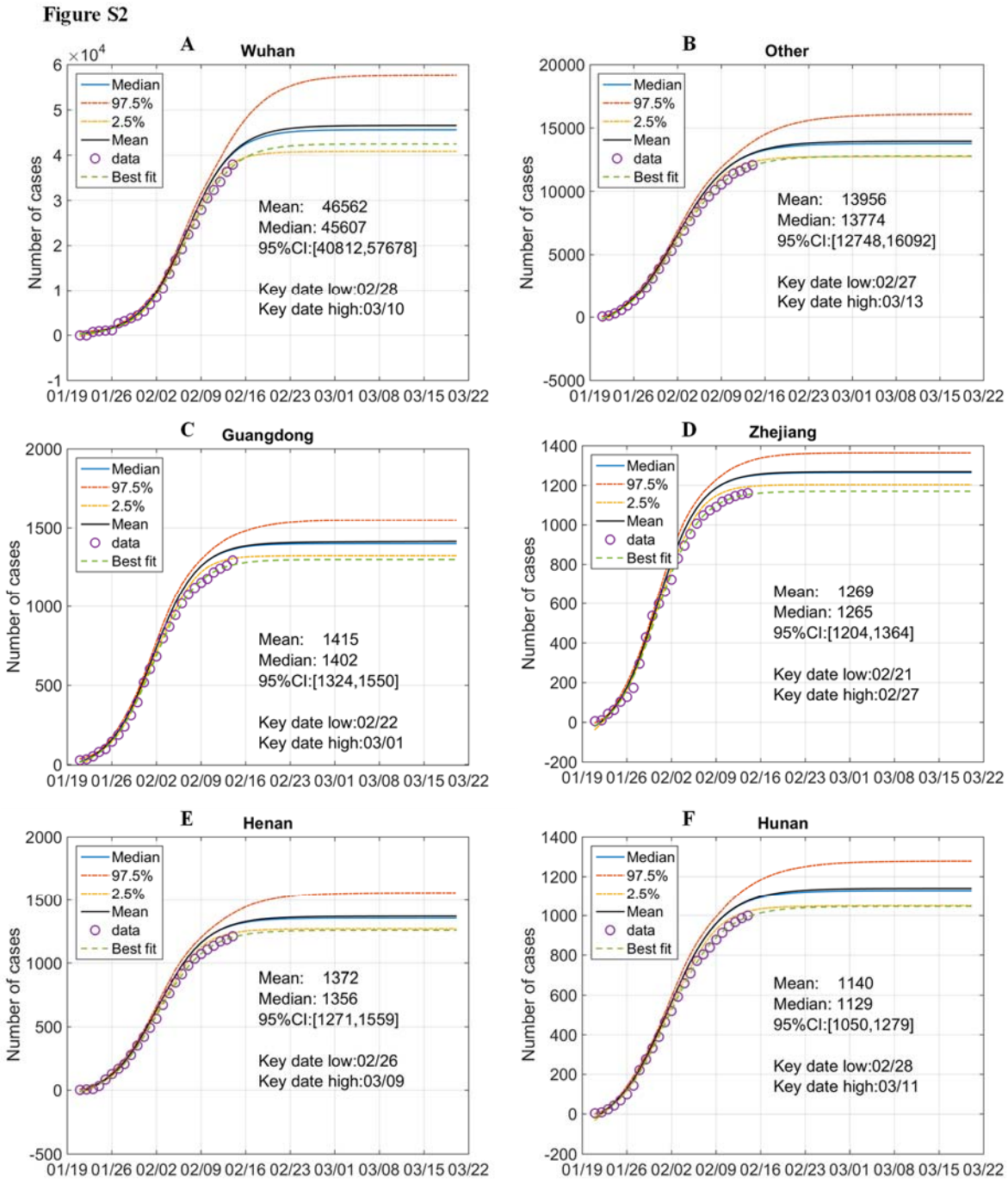

Figure S2 Analysis of the uncertainty of the estimated confirmed cases

Data of Wuhan City (panel A), of other provinces than Hubei (panel B), of Guangdong (panel C), of Zhejiang (panel D), of Henan (panel E) and of Hunan (panel F) were fitted to Boltzmann function assuming that the relative uncertainty of the data follows a single-sided normal distribution with a mean of 1.0 and a standard deviation of 10%. Original data are shown as circles; simulated results are presented as colored lines as indicated. Inserts show key statistics information. The key date is defined as the date when the number of daily new confirmed cases is less than 0.1% of the potential total number. The high and low key dates were determined by the simulated curve of CI at 2.5% and 97.5%, respectively.

Figure S3

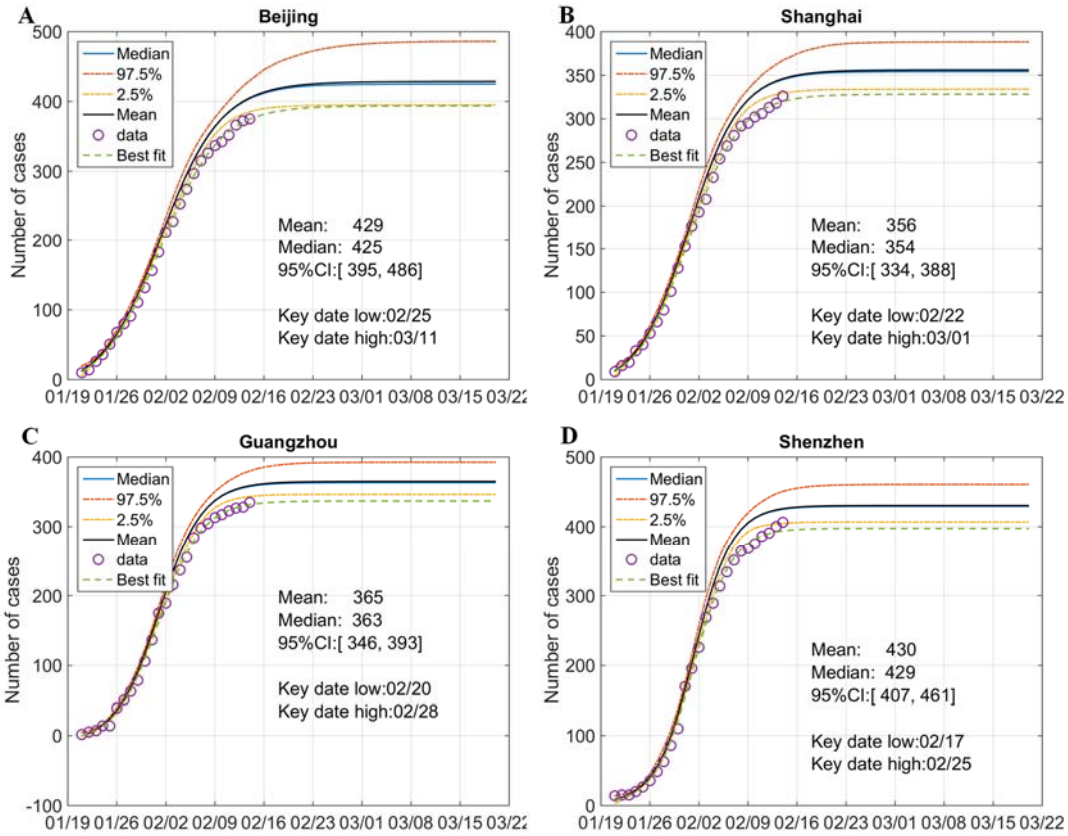

Figure S3 Analysis of the uncertainty of confirmed cases for the top 4 major cities of China

Data of Beijing (panel A), of Shanghai (panel B), Guangzhou (panel C) and Shenzhen (panel D) were fitted to Boltzmann function as described in **Fig. S2**.