专题报告

SARS 流行预测分析

王建锋
（中国科学院力学研究所，北京 100080）

[摘要] 表面上突如其来的 SARS 本质上却有极规律的内在发展演化机制，遵从初始缓慢增长、加速、减速和稳定终止四个阶段总体道路，自然和社会生活领域众多事件演化都符合这一规律，因而可以运用广义的 Logistic 生长模型进行描述。基于先期流行的广东 SARS 感染病例数据，以及尚未结束的北京、全国 2003 年 SARS 流行统计数据，借助于最优化分析技术，运用广义的 Logistic 生长模型对该事件演化特征参量进行了辨识；在此基础上，又借助于广义生长模型的特例——Gompertz 函数进行了演化过程的预测，并与其他生长模型结果进行了比较。研究表明，生长模型模拟结果均与实际数据有很好的一致性，可以用来预测事件的发生演化过程，此次 SARS 事件堪称生长模型的经典实例。

[关键词] SARS；广义的 Logistic 生长模型；Gompertz 分布；预测；最优化
[中图分类号] R195.1；Q－332 [文献标识码] A [文章编号] 1009－1742（2003）08－0023－07

1 前言

在 SARS 流行初期，即从宏观上研究其发生、发展、演化规律并进行预测分析，对于防治传染性疾病大规模流行的战略决策、稳定社会秩序具有重要意义。笔者基于生物学领域生长预测模型，以广东、北京、全国 SARS 流行数据为例，初步分析了 SARS 流行的一般规律，并进行了预测分析，供有关方面参考、探讨。

2 特征参量辨识

特征参量主要指此次 SARS 感染累计人数、爆发峰值时间和终止时间、爆发峰值时刻的总感染人数及当日感染人数。这些都是系统发展演化的特征参量，对于局势判断极为重要。

广义的 Logistic 生长模型可以表示为[1]，

$$\frac{dN}{dt} = rN^{\alpha}\left[1 - \left(\frac{N}{K}\right)^{\beta}\right]^{\gamma}, \tag{1}$$

式中，r 是增殖系数；α，β，γ 是正的实常数；N 是状态变量，为时间的函数；K 是饱和值。此模型表现形式一般为 $N-t$ 空间中呈现不同形式的 S 型曲线，对应初始缓慢增长、加速、减速和稳定四个阶段，反映了一般生命历程，但是不存在解析解。

一般地，随系统演化内部驱动机制不同，曲线拐点位置会不同，各阶段持续时间、速率也不同。其特征参量中峰值爆发状态变量 $N(t)_{\inf}$（在拐点上）为

$$N_{\inf} = K\left(1 + \frac{\beta\gamma}{\alpha}\right)^{-(1/\beta)}, \tag{2}$$

式（2）只有在大于 $N^* = \left(1 + \frac{\beta\gamma}{\alpha - 1}\right)^{-(1/\beta)} K$ 时才有意义。

最大增长速率为 $(dN/dt)_{\max}$：

$$\left(\frac{dN}{dt}\right)_{\max} = \left(\frac{dN}{dt}\right)_{N_{\inf}} = rK^{\alpha}\left(\frac{\alpha}{\alpha + \beta\gamma}\right)^{\alpha/\beta}\left(\frac{\beta\gamma}{\alpha + \beta\gamma}\right)^{\gamma}, \tag{3}$$

爆发峰值时间（也在拐点上）$t_{\inf}$为

[收稿日期] 2003－05－30；修回日期 2003－06－17
[作者简介] 王建锋（1964－），男，河南荥阳市人，工学博士，中国科学院力学研究所副研究员

$$t_{\inf}=\frac{1}{rK^{\alpha-1}}\left[\frac{\left(1+\frac{\beta\gamma}{\alpha}\right)^{(\alpha-1)/\beta}}{1-\alpha}+\frac{\left(1+\frac{\beta\gamma}{\alpha}\right)^{(\alpha-1-\beta)/\beta}}{1-\alpha+\beta}+\frac{\gamma(\gamma+1)}{2!}\frac{\left(1+\frac{\beta\gamma}{\alpha}\right)^{(\alpha-1-2\beta)/\beta}}{1-\alpha+2\beta}+\cdots\right]-$$

$$\frac{1}{r}\left[\frac{N_0^{1-\alpha}}{1-\alpha}+\gamma\frac{N_0^{1-\alpha+\beta}}{K^{\beta}(1-\alpha+\beta)}+\frac{\gamma(\gamma+1)}{2!}\frac{N^{1-\alpha+2\beta}}{K^{2\beta}(1-\alpha+2\beta)}+\cdots\right],\alpha\neq 1 \quad (4)$$

针对不同的生命演化系统，相应于式（1）存在多个变种表达[2,3]。

首先进行数据预处理：

1）运用线性内插的方法，将北京4月19日至4月25日公布的截至每日晚8时的SARS累计感染人数，转换为截至每日上午10时的等效数据，以便于与以后所有公布的每日上午10时数据一致。

2）考虑到每日报告新增病例在随后公布的数据中又有排除，因此一律将排除病例数从最近的累计报告病例数中减掉，并由累计病例数推算每日实际新增病例数，以保证数据准确性和累计数据的非降性。

基于可以利用的并且首先进入稳定期的广东，以及将进入稳定期的北京、全国每日累计确诊SARS人数序列，运用优化算法，设定残差平方和最小为目标函数，特征参量 r，α，β，γ，K 为控制变量，约束条件为 $K>N(t)$，$N(t)$ 为最近一次累计感染人数，按式(1)进行参数识别。

模拟结果均表明，无论是地区性还是全国性的SARS流行都很好地遵从生长曲线（表1、图1、图2）。但是，由于广东和全国公布的数据为峰值爆发时间以后的序列，因此由这些短序列（特别对于广州）还不能预测到全部的特征参量，尤其是对应于式（2）、式（3）和式（4）的特征量。对于北京，采用已公布的所有数据序列（2003年4月19日—2003年6月12日），采用广义的Logistic生长模型进行拟合，所得特征参量为 $N(t)_{\inf}=953$，$(\mathrm{d}N/\mathrm{d}t)_{\max}=118$，$t_{\inf}=10.08$。结合表1，即说明最终的感染人数将为2 527人，峰值爆发时间是4月28日，峰值爆发时的累计感染人数为953人，其时每日最大感染人数为118人。

表1 广东、北京、全国SARS广义的Logistic生长模型特征参量辨识

Table 1 The predicted results of characteristic parameters of the generalized Logistic growth model for Guangdong, Beijing and Mainland China 2003 SARS

类别	特征参量					残差平方和	数据源
	r	α	β	γ	K^*		
广东	4.65E-10	3.2619	1 491.5992	115 614.4578	1 516	198.8312	2003-04-26～06-12
北京	56.0571	0.1137	3.6489	1.0612	2 527	7 668.0275	2003-04-19～06-12
全国	486.1013	-0.1255	12.8255	1.6362	5 355	4 482.8816	2003-04-26～06-12

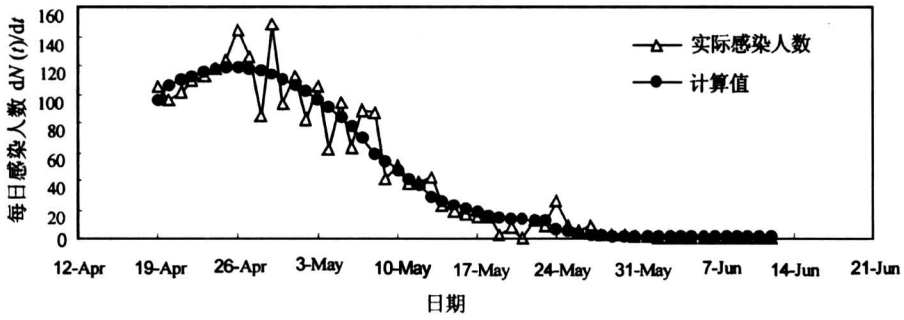
* 截至2003-06-12广东、北京、全国实际累计感染SRAS人数分别为：1511，2523，5328

应当说明的是，上述模拟结果都应当视为期望值，并且对于北京和全国SARS而言，其误差来源主要是由4月26日以前的最初几日统计数字准确度较差造成的。相对来说，由于笔者没有全部掌握广东前期SARS流行数据，其计算结果不够理想，但就饱和值来说，还是可以的。

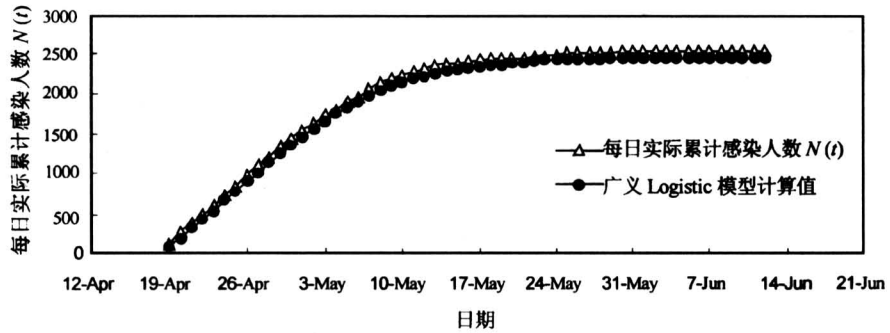
3 过程预测

上述广义的Logistic生长模型，由于不能通过积分给出 $N(t)$ 的解析解，因此尽管它能较精确描述与自然生命历程有关的系统演化特征，但是不能直接用于过程预测。

作为广义的Logistic生长模型的特例，历史上先后发展了多种预测模型，例如早期的指数模型、Verhulst-Pearl模型、Gompertz模型、Von Bertalanffy模型、Richards模型、Weibull模型以及Morgan-Mercer-Flodin模型等。对于一个事件演化的预测，最终运用或发展何种预测模型，最好的办法是首先对宏观趋势作估计，再经过各主要模型模拟，兼顾选择残差平方和最小者作为过程预测模型，但是始终不要认为残差平方和最小者一定是最好的预测模型。



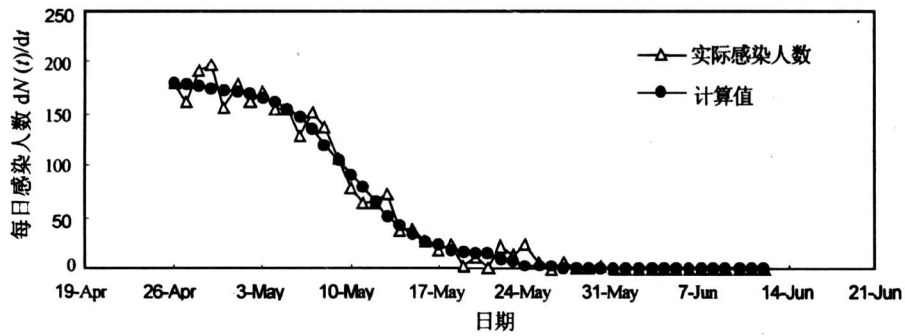
(a) 每日 SARS 感染人数实际值与计算值比较（北京）



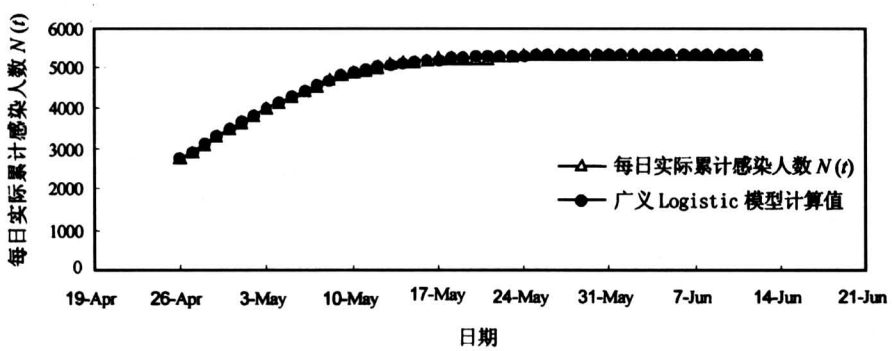
(b) SARS 感染累计人数 $N(t)$ 实际值与计算值比较（北京）

图 1　SARS 广义的 Logistic 生长模型（北京）

Fig.1　A realization of the generalized Logistic growth model of Beijing 2003 SARS



(a) 每日 SARS 感染人数实际值与计算值比较（全国）



(b) SARS 感染累计人数 $N(t)$ 实际值与计算值比较（全国）

图 2　SARS 广义的 Logistic 生长模型（全国）

Fig.2　A realization of the generalized Logistic growth model of Mainland China 2003 SARS

经比较，采用 Gompertz 函数[4]进行过程预测效果较好。Gompertz 函数是典型的 S 型曲线，对应于生命历程同样有 4 个阶段：初始缓慢增长阶段、加速阶段、减速阶段、稳定阶段。超 Gompertz 函数或广义 Gompertz 函数为

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \lim_{\beta\to 0}\frac{rN}{K^{\beta\gamma}}\left(\frac{K^{\beta}-N^{\beta}}{\beta}\right)^{\gamma} = rN\left[\ln\left(\frac{K}{N}\right)\right]^{\gamma}, \tag{5}$$

当 $\gamma=1$ 时，上式即是经典的 Gompertz 函数，其解为

$$N(t) = K\left(\frac{N_0}{K}\right)^{\mathrm{e}^{-rt}}。 \tag{6}$$

经典的 Gompertz 函数也可以由如下的 Richards 增长方程，通过右边除以 β，接着取 $\beta\to 0$ 的极限值得到

$$\frac{\mathrm{d}N}{\mathrm{d}t} = rN\left[1-\left(\frac{N}{K}\right)^{\beta}\right]。 \tag{7}$$

式（6）拐点处的 $N(t)_{\inf}$ 和 $t_{\inf}$ 值分别为

$$N_{\inf} = K\mathrm{e}^{-1}, \tag{8}$$

$$t_{\inf} = \frac{\ln\left[\ln\left(\frac{K}{N_0}\right)\right]}{r}。 \tag{9}$$

有时，经典的 Gompertz 函数也可用下列代数式直接给出：

$$N(t) = K\exp(1-a\exp(-bt)), \tag{10}$$

式中 b 为群集系数。

此外，笔者也分别运用了如下 Pearl 模型、Von Bertalanffy 模型、修正指数模型：

$$N(t) = \frac{K}{1+a\exp(-bt)}, \tag{11}$$

$$N(t) = K\left\{1-\frac{1}{3}\exp[-a(t-b)]\right\}^{3}, \tag{12}$$

$$N(t) = K + ab^{t}。 \tag{13}$$

上述各式特征参量中饱和值 K 的确定可采用两种方法：

1）首先由广义的 Logistic 生长模型，经优化处理得到（约束条件：$K>N(t)$，$N(t)$ 为最近一次累计感染人数），再代入上述各式，再次用无约束优化算法，求得参量 a，b，c 等。

2）直接运用各模型函数，视所有模型参量 K，a，b，c 等为变数，经约束优化求解得到，其中约束条件为 $K>N(t)$，$N(t)$ 为最近一次累计感染人数。

注意，所有优化算法中，目标函数均为残差平方和最小。这两种方法所得过程模拟结果，将因参量不同而不同，但一般差别不会太大，可以接受。k 值也可以因为采取了不同防治策略而为时变[5]。亦即采用不同的外部干预形式，事件演化的道路将极为不同。

上述各模型预测结果见表 2、图 3（K 不固定），结果显示经典 Gompertz 模型预测效果最好，并且预测曲线趋于稳定由快到慢的顺序依次为：Pearl 模型＞Von Bertalanffy 模型＞Gompertz 模型＞修正指数模型。

表 2 广东、北京、全国 SARS 几个生长模型模拟结果

Table 2 The predicted results of Guangdong, Beijing and Mainland China 2003 SARS using growth models

生长模型		广东 (2003.4.26—6.12)	北京 (2003.4.26—6.12)	全国 (2003.4.26—6.12)
Pearl 模型	a	0.1371	8.0982	3.4952
	b	0.1074	0.1917	0.1544
	K	1 529.1039	2 523.0000	5 404.1002
Von Bertalanffy 模型	a	0.1247	0.1338	0.1276
	b	−16.7865	6.7905	−2.3612
	K	1 516.3183	2 523.0000	5 385.6542
Gompertz 模型	a	0.1277	2.9422	0.82524
	b	0.1099	0.1388	0.12768
	K	1 514.9714	2 537.0379	5 442.85
修正指数模型*	a	−180.1327	−1844.3256	−3192.1477
	b	0.8518	0.8978	0.8946
	K	1 516.4740	2 560.3934	5 404.8864

下面以经过预处理的北京 SARS 全部观察数据（2003－04－19～2003－06－12）为例，介绍经典 Gompertz 模型过程预测结果。

首先运用广义的 Logistic 生长模型调优得到的饱和值 $K=2\,527$，直接辨识经典 Gompertz 函数式（6）中的参量得 $r=0.1458$，其中 $N_0=105$，此时残差平方和为 36 404.1042，并且如以四舍五入计，6 月 11 日事件将结束。另外的特征参量为：$N(t)_{\inf}=929.61$，$t_{\inf}=7.94$，后者对应于爆发峰值时间 4 月 26 日。

*修正指数模型不属于生长模型，而是增长模型，在不具备前期观测数据时，可以用来确定饱和值和达到饱和值的时间点。这里，尤其适合于广东 SARS 数据列的预测；对于北京 SARS，仅使用峰值爆发时间后数据列，即 2003－04－26～2003－06－12

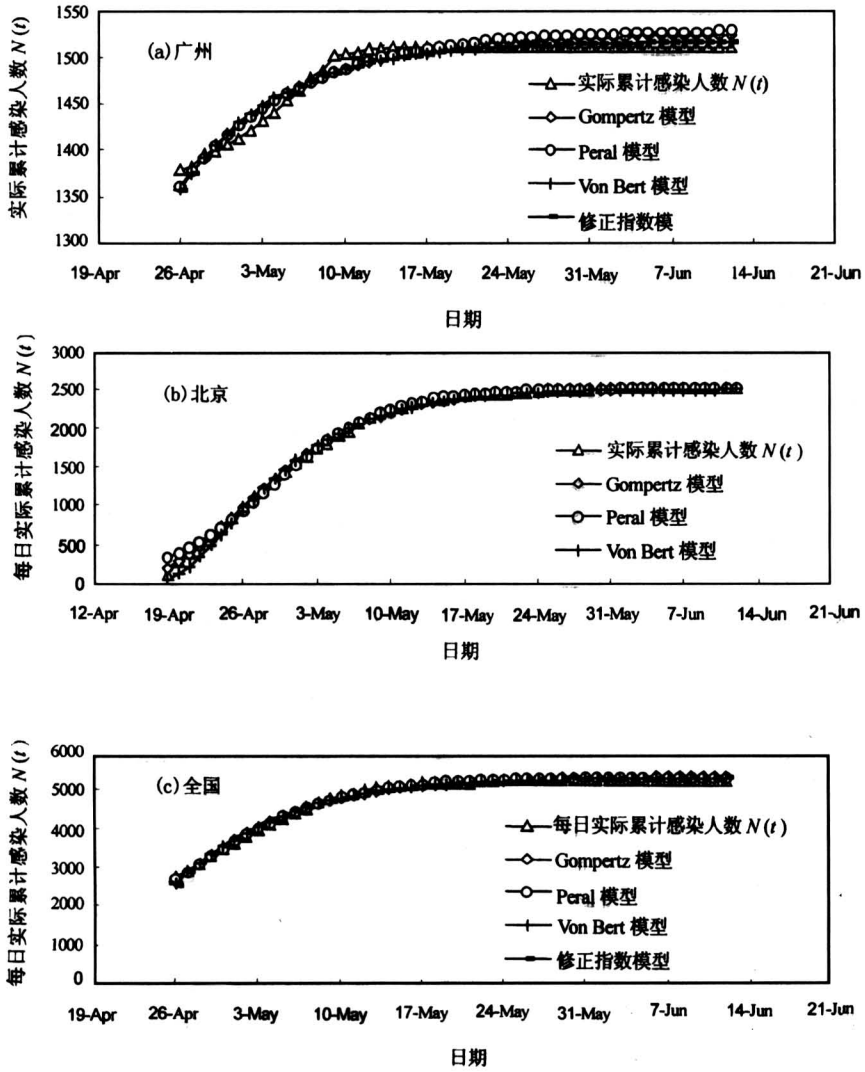


图 3　SARS 感染累计人数 $N(t)$ 实际值与各模型预测值比较

Fig.3　Predicted population size of Beijing 2003 SARS using different models and optimization projects

如果对式（10）中的所有参量进行辨识，如表 2 所示，则得 $K=2\,537$，$a=2.9422$，$b=0.1388$，此时残差平方和为 22 529.6951，并且如以四舍五入计，6 月 13 日事件将结束。

经典 Gompertz 函数模拟结果见表 3，可以看出与实际数据较吻合。

4　结论

1）SARS 感染累计的人数演化过程总体上可以划分为初始缓慢增长、加速、减速和稳定四个阶段，遵从广义生长模型。

2）就实际效果而言，为了能运用具有准确性、一致性的数据进行较高精度预测，卫生突发事件的统计数字及时、全面、准确地公布极其重要。

3）根据最新发展情况，适时更新预报非常重要（如可以运用 Bayesian Method 以及耦合时变参数[5]）。预测结果的准确性有赖于决不放松警惕，不出现“反弹”。对于复杂的长期发展过程，需要运用分段函数进行预测。

4）生长模型预测中，原始数据的预处理很重要；预测的峰值爆发点、饱和值、趋稳时间与实际事件过程的吻合程度，可作为判断生长模型预测结果满意与否的标志。

5）对于具有四阶段的生长过程，为了做出较好的预测，所需要的数据列的长度及所要求数据列覆盖哪些生长阶段仍需要做进一步研究。深入理解

事件发展过程的物理机制，是发展合适预测模式的关键，如广义的流行病模型（SIR）是有前景的[5]。进一步工作应以疑似、确诊、排除（包括排除、治愈、死亡）三者协同演化为主线建立合适的模型。

致谢：感谢王仁铎教授、张钧锋博士对本文的评阅、指正和鼓励，也感谢审稿人极好的建议和修改意见。

表3 经典 Gompertz 函数模拟结果

Table 3 The predicted results of Beijing 2003 SARS using Gompertz function

No.	日期	每日感染人数			累计感染人数		No.	日期	每日感染人数			累计感染人数	
		实际	$K=2\,526$	K 不限	实际	预测（K 不限）			实际	$K=2\,526$	K 不限	实际	预测（K 不限）
1	19－Apr	105	162	105	105	196	29	17－May	15	17	19	2420	2407
2	20－Apr	96	73	77	279	273	30	18－May	14	15	16	2434	2424
3	21－Apr	102	89	91	381	364	31	19－May	3	13	14	2437	2438
4	22－Apr	109	104	104	490	469	32	20－May	7	11	13	2444	2451
5	23－Apr	113	117	115	603	583	33	21－May	0	10	11	2444	2462
6	24－Apr	117	126	122	720	706	34	22－May	12	9	10	2456	2471
7	25－Apr	124	132	127	844	833	35	23－May	9	8	8	2456	2480
8	26－Apr	144	135	129	988	963	36	24－May	25	7	7	2490	2487
9	27－Apr	126	135	129	1114	1091	37	25－May	9	6	6	2499	2493
10	28－Apr	85	132	126	1199	1218	38	26－May	5	5	6	2504	2499
11	29－Apr	148	127	122	1347	1339	39	27－May	8	4	5	2512	2504
12	30－Apr	93	121	116	1440	1455	40	28－May	2	4	4	2514	2508
13	1－May	113	113	109	1553	1564	41	29－May	3	3	4	2517	2512
14	2－May	83	105	101	1636	1665	42	30－May	3	3	3	2520	2515
15	3－May	105	96	93	1741	1758	43	31－May	1	2	3	2521	2518
16	4－May	62	88	86	1803	1844	44	1－Jun	1	2	2	2522	2520
17	5－May	94	79	78	1897	1922	45	2－Jun	0	2	2	2522	2523
18	6－May	63	71	70	1960	1992	46	3－Jun	0	2	2	2522	2524
19	7－May	89	64	63	2049	2056	47	4－Jun	0	1	2	2522	2526
20	8－May	87	57	57	2136	2112	48	5－Jun	0	1	1	2522	2528
21	9－May	41	50	51	2177	2163	49	6－Jun	0	1	1	2522	2529
22	10－May	50	44	45	2227	2208	50	7－Jun	1	1	1	2523	2530
23	11－May	38	39	40	2265	2248	51	8－Jun	0	1	1	2523	2531
24	12－May	39	34	35	2304	2284	52	9－Jun	0	1	1	2523	2532
25	13－May	43	30	31	2347	2315	53	10－Jun	0	1	1	2523	2532
26	14－May	23	26	28	2370	2343	54	11－Jun	0	0	1	2523	2533
27	15－May	18	23	24	2388	2367	55	12－Jun	0	0	1	2523	2533
28	16－May	17	20	21	2405	2388	56	13－Jun	0	0	0	2523	2533

参考文献

[1] Tsoularis A, Wallace J. Analysis of logistic growth models[J]. Mathematical Biosciences, 2002, 179: 21～55

[2] McCann T L, Eifert J D, Gennings C, et al. A predictive model with repeated measures analysis of Staphylococcus aureus growth data [J]. Food Microbiology, 2003, 20: 139～147

[3] Narushin V G, Takma C. Sigmoid model for the evaluation of growth and production curves in laying hens[J]. Biosystems Engineering, 2003, 84(3): 343～348

[4] Impagliazzo J. Determinstic aspects of mathematical demography[M]. Springer-Verlag, Belin, Heidelberg, 1985

[5] Banks R B. Growth and diffusion phenomena: mathematical frameworks and applications [M]. Springer-verlag, Berlin, Heidelberg. 1994. 19～27; 126～147

Predicting SARS for Guangdong, Beijing and Mainland China 2003 Cases

Wang Jianfeng

(*Institute of Mechanics, Chinese Academy of Sciences, Beijing* 100080, *China*)

[**Abstract**] The study is aimed at choosing a better predictive model for the accurate description of SARS in Guangdong, Beijing and Mainland China in 2003. Observation and general experience have shown a sigmoid type of curve consisted of four phases comparable to the phases of the SARS growth in 2003: an initial lagging period, a period of accelerating change, a period of decelerating change, and a stationary period. In order to model the SARS system, a generalized Logistic growth function has been adopted in the paper. With the officially published data, the main features of evolution of the SARS population size have been obtained using the generalized Logistic growth model by optimizing technique. Then, for getting evolutionary process prediction, several classical S-models such as the Pearl, the Gompertz, Von Bertalanffy, and Richards are tested. The practice of calculations has found that the Gompertz model gives the most accurate results where fitting criteria are estimated as residual sum of squares (RSS).

[**Key words**] SARS; generalized Logistic growth model; Gompertz function; prediction; optimization