

Logistic model best fits  
population growth data  
using ordinary Linear and  
Non-linear model fitting  
methods

Junyue Zhang

This report is presented for the  
Computing Miniproject

Life Sciences  
Imperial College London  
03 December 2021  
Word count: 3480

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Methods</b>	<b>4</b>
2.1	Data . . . . .	4
2.2	Models . . . . .	4
2.3	Model fitting . . . . .	6
2.4	Plotting and analysis . . . . .	6
2.5	Computing tools . . . . .	6
<b>3</b>	<b>Results</b>	<b>7</b>
<b>4</b>	<b>Discussion</b>	<b>14</b>
<b>5</b>	<b>Conclusion</b>	<b>16</b>

1

## Abstract

2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20

The most prevalent cause of food spoilage and food poisoning during storage, when containing pathogen, is microbial growth. Therefore, five different mathematical models are applied to fit the population growth data using ordinary linear and non-linear least squares methods to find the ones which best fit the dataset. The linear quadratic and cubic polynomial models can be easily applied to fit the data using the linear regression in R. Then a non-linear least squares model fitting function called nlsLM in R can be used to fit each mechanistic model after defining the starting values of the primary parameters. The count of logistic model with the minimum AIC and BIC values is the largest, which indicates the logistic model best fits the population growth data on the whole. While the quadratic model is the most inaccurate one to fit the data, but it can capture the “mortality phase”. The cubic polynomial model and modified Gompertz model generally fit the population growth data well at similar levels. Meanwhile, the Baranyi model poorly fits the whole dataset since the model fitting fails to converge in more than half of the subsets. This study shows that on the whole, the logistic model is best suited for the population growth data.

## 21 1 Introduction

22 The most prevalent cause of food spoilage and food poisoning during stor-  
23 age, when containing pathogen, is microbial growth (Gram et al., 2002) [6].  
24 A large number of mathematical models have been adopted to describe the  
25 properties of microbial growth and fitting precision between the actual val-  
26 ues and fitted values (Baty et al., 2004) [4]. Baranyi and Roberts (1994)  
27 [3] reported that costs related to laboratory challenge testing of foods can  
28 be reduced significantly if methods for making realistic predictions can be  
29 established. A growth curve can be used to delineate the time-dependent  
30 increase in the microbial population in a closed system and record the count-  
31 able cells at certain time intervals during the evolution of population (Winsor  
32 and Charles P, 1932) [16]. And under a variety of environmental conditions,  
33 growth curves for specific micro-organisms are essential to most of the pre-  
34 dictive methods and mathematical models either partially or completely  
35 (Perni et al., 2005) [13].

36 Extensive research has been carried out to investigate microbial growth pat-  
37 terns and the factors that influence them. Zwietering et al. (1990) [17]  
38 compared numerous sigmoidal functions statistically to describe a bacte-  
39 rial growth curve by using the t test and F test and found that the modified  
40 Gompertz equation was satisfactory enough to delineate the bacterial growth  
41 and easy to use. In addition, a series of members of the family of growth  
42 models have been proposed and developed by Baranyi and other researchers  
43 successively. For example, a non-autonomous differential equation was put  
44 forward by Baranyi et al. (1993) [2] to describe the dynamics of growing  
45 bacterial cultures. In particular, the dynamic model proposed by Baranyi  
46 and Roberts (1994) [3] is widely used, the new approach can depict bac-  
47 terial growth in an environment where the factors vary with time, such as  
48 temperature and pH.

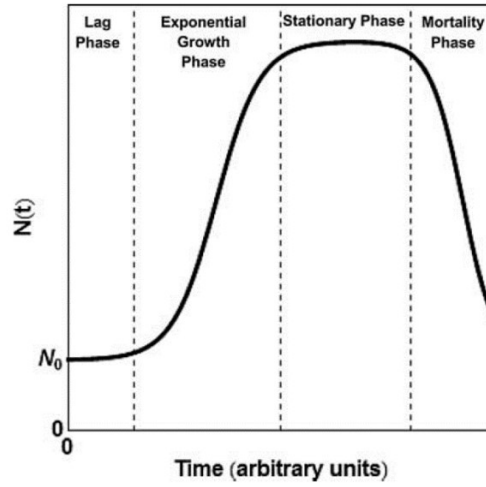


Figure 1: Four distinguishable phases of bacterial growth in a closed habitat (Micha and Corradini, 2011) [12]

Typically, there are four distinguishable phases of bacterial growth in a closed habitat as shown in Figure 1 above. They are “lag phase”, “exponential growth phase”, “stationary phase”, and “mortality phase” (McKellar and Lu, 2003) [9]. But the “mortality phase” is ignored in most cases and only the first three phases are investigated in the food microbiology since foods will become harmful and inedible long before the “mortality phase” starts, and sometimes even before the “stationary phase” (Micha and Corradini, 2011) [12].

My objective of this study is to investigate how well different mathematical models, such as those based on population growth (mechanistic) theory vs. those based on phenomenological ones, fit the population growth data across the unique ID. The population growth data covers measurements of change in biomass or number of microbes cells over time and is collected through lab experiments throughout the world. Five different mathematical models can be applied to fit the population growth data using ordinary linear and non-linear least squares methods to find the ones which best fit the dataset. Additionally, the five mathematical models involve not only phenomenological quadratic and cubic polynomial models but also non-linear mechanistic models of population growth. After that, Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) can be used to

69 compare the models (Aho et al., 2014) [1].

## 70 **2 Methods**

### 71 **2.1 Data**

72 The population growth data was acquired through lab experiments around  
73 the world and includes measurements of change in biomass or the number  
74 of microbial cells over time. There are two main variables in the population  
75 growth data, PopBio (abundance) is the population or biomass measure-  
76 ment regarded as the response variable and Time is the measurement time  
77 regarded as the independent variable. Furthermore, a unique population  
78 growth curve can be identified by combining Species, Medium, Tempera-  
79 ture, and Citation these four variables to create a new independent variable  
80 called ID. As a result, numerous subsets of the population growth data can  
81 be constructed according to each unique ID. After the data wrangling, no  
82 missing value was found in each column. There are some negative values  
83 in the Time column, and these values are very close to zero in general, so  
84 they are set to be zeros to make the data more realistic and meaningful. To  
85 obtain the log-transformed values, only the positive values of PopBio are  
86 saved since the independent variable of the logarithmic functional has to  
87 be greater than 0, which is followed by deriving a log-transformed PopBio  
88 column.

### 89 **2.2 Models**

90 To start with, two linear mathematical models were applied to fit each subset  
91 of the population growth data, namely the phenomenological quadratic and  
92 cubic polynomial models (Johnson et al., 2004) [7]. The equations (1) (2)  
93 of these two models are shown below respectively.

$$94 \qquad y = ax^2 + bx + c \qquad (1)$$

$$95 \qquad y = ax^3 + bx^2 + cx + d \qquad (2)$$

97 The quadratic and cubic curve models can capture the curvature in data  
98 (Motulsky et al., 2004) [10]. For non-linear mechanistic models, three dif-  
99 ferent growth rate models were adopted to fit each subset in a similar way

(Bolker et al., 2013) [5] . These mechanistic models are the logistic model, modified Gompertz model (Zwietering et. al., 1990) [17], and the Baranyi model (Baranyi, 1993) [2]. A classical logistic equation (3) is displayed below.

$$N_t = \frac{N_0 K e^{rt}}{K + N_0(e^{rt} - 1)} \quad (3)$$

Here  $N_0$  is the initial population size,  $N_t$  is population size at time  $t$ ,  $r$  is the maximum growth rate, and  $K$  is the maximum possible population abundance and is also called carrying capacity (Samraat Pawar, 2021) [11].

However, there is a time lag in the microbial population growth. Before starting the exponential growth, bacteria need to spend some time adapting to the new environment or growth media and activating genes relating to intaking nutrition and metabolic processes (Rolfe et al., 2012) [15]. And the modified Gompertz model (Zwietering et. al., 1990) [17] can be applied to capture the time lag when modeling the bacterial growth. The Gompertz equation (4) is a bit more complicated as shown below.

$$\log(N_t) = N_0 + (N_{max} - N_0)e^{-e^{\frac{r_{max} \exp(1)}{(N_{max} - N_0) \log(10)} + 1} (t_{lag} - t)} \quad (4)$$

Here  $t_{lag}$  is the lag time before the exponential growth,  $r_{max}$  is the maximum growth rate tangent to the inflection point, and  $\log(\frac{N_{max}}{N_0})$  is the log ratio of the carrying capacity and the initial population size. In addition, the Gompertz model is in the log scale and designed to fit the log-transformed population growth data (Samraat Pawar, 2021) [11].

Besides the Gompertz model, the Baranyi model (Baranyi, 1993) [2] can also be used to describe the lag phase. The Baranyi equation (5) is displayed below.

$$y(t) = y_0 + \mu A(t) - \ln\left(1 + \frac{e^{\mu A(t)} - 1}{e^C}\right) \quad (5)$$

$$A(t) = t + \frac{1}{\mu} \ln(e^{-\mu t} + e^{-\mu \lambda} - e^{-\mu(t+\lambda)}) \quad (6)$$

Here  $y$  is the population size in log scale at time  $t$ ,  $y_0$  is the initial population size in log scale,  $y_{max}$  is the final population size in log scale,  $\mu$  is the maximum growth rate,  $C$  is the difference between  $y_0$  and  $y_{max}$  in log scale, and  $\lambda$  is the lag time before the exponential growth (Pla et al., 2015) [14].

## 131 2.3 Model fitting

132 The linear quadratic and cubic polynomial models can be easily applied to  
133 fit the data using the `lm` function in R. The goodness-of-fit is evaluated  
134 by the summary function and the coefficients of models can be acquired  
135 subsequently.

136 For the non-linear mechanistic models, several starting parameters need to  
137 be defined previously. To start with, a linear model can be used to fit the  
138 data. And the slope of the Time variable can be adopted to be the starting  
139 value of the maximum growth rate. For the logistic model, the starting  
140 values of  $N_0$  and  $K$  are assigned to be the lowest population size and highest  
141 population size respectively.

142 For the modified Gompertz model and Baranyi model, the starting values  
143 of initial population size and final population size are set to be the lowest  
144 population size and highest one both in the log scale respectively. To find the  
145 last time point of the lag phase, the whole dataset is sliced into the first third  
146 of it, and the lag time is the time point where the second-order derivative of  
147 the differentials is maximal. Then a non-linear least squares (NLLS) model  
148 fitting function called `nlsLM` in R can be used to fit each model using the  
149 starting values derived above, followed by obtaining the model summary.  
150 Moreover, the `try` function is applied in the non-linear least squares model  
151 fitting to capture errors when the fitting does not converge.

## 152 2.4 Plotting and analysis

153 To observe the goodness-of-fit of each model, every unique population growth  
154 curve is plotted with these five models overlaid. In addition, the main coef-  
155 ficients of each model as well as AIC, BIC, and R squared values of model  
156 fitting are separately saved in csv files for subsequent review and analysis.  
157 For the subsets of the population growth data, the number of times that  
158 the AIC or BIC value of each model is the lowest in the same subset can be  
159 acquired to find the models which best fit the dataset.

## 160 2.5 Computing tools

161 For this mini-project, R is the main scripting language for data preparation,  
162 model fitting, and final plotting and analysis, since it is more convenient

163 and coherent for me to use the same language. And I also use bash to  
164 compile the LaTeX report and run the whole project. In R, nls.lm package  
165 is necessary because the nlsLM function is used, and ggplot2 is loaded to  
166 generate beautiful figures for plotting.

### 167 **3 Results**

168 As outlined in the introduction, in order to find the models which best fit the  
169 dataset, five different mathematical models are applied to fit the population  
170 growth data across the unique ID. After running the model fitting script,  
171 several csv files are generated to display the coefficients of each model as well  
172 as the corresponding AIC, BIC, and R squared values. As the phenomeno-  
173 logical models, quadratic and cubic polynomial models fit the data linearly.  
174 For the mechanistic ones, the modified Gompertz model fitting succeeds 208  
175 times and fails 77 times and the Baranyi model succeeds 139 times and fails  
176 146 times. However, the logistic model fits each data subset successfully.



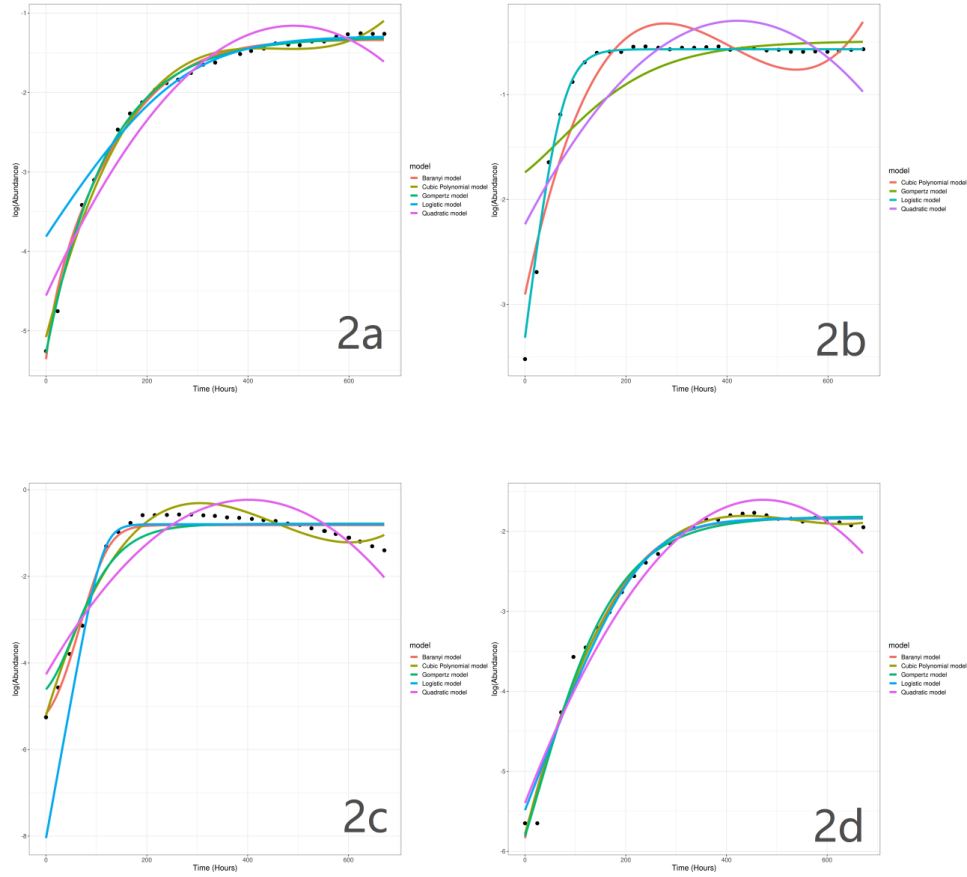


Figure 2: Four distinct plots of different data subsets

177 After running the plotting and analysis script, 285 population growth scatter  
 178 plots are created with the mathematical model curves overlaid. Four distinct  
 179 plots of different data subsets are displayed in Figure 2 above. As shown in  
 180 figure 2a, the quadratic model fits the data imprecisely and only a few data  
 181 points fall on the curve. This model roughly describes the change in biomass  
 182 over time, but it captures the “mortality phase” of population growth after  
 183 the carrying capacity has been reached. The cubic polynomial model fits the  
 184 data more accurately than the quadratic one and the cubic curve even fits  
 185 the data points well at the lower end. However, in contrast to the quadratic  
 186 model, the cubic curve goes up unexpectedly after the “stationary phase”  
 187 of population growth is reached. For the non-linear mechanistic models, the

188 logistic model visibly deviates from the data at the initial stage, but it highly  
189 fits the remaining part of the data. And it can be seen that the Gompertz  
190 model and Baranyi model both fit the data perfectly and the curves of the  
191 three mechanistic models overlap nicely at the “stationary phase”.

192 As shown in figure 2b, similarly, the quadratic model fits the data roughly.  
193 It seems that the quadratic curve only passes through two data points, and  
194 it has little to do with the variation in the data. As expected, the quadratic  
195 curve also captures the “mortality phase” of population growth in this case.  
196 However, the cubic polynomial model fits the data poorly at this time and  
197 the cubic curve also rises after the “stationary phase”. For the mechanistic  
198 models, the Baranyi curve fails to show up since there might have some  
199 errors in the model fitting process. It can be observed that the Gompertz  
200 model also fits the data poorly and it diverges from the data on the whole.  
201 Fortunately, the logistic model fits this data subset excellently and nearly  
202 all the data points fall on the logistic curve.

203 As indicated in figure 2c, both the quadratic model and cubic model can  
204 catch the decrease in the population size after the maximum value is reached  
205 and they can delineate the trends in the data sketchily. For the mechanistic  
206 ones, the logistic model obviously deviates from the data in the beginning.  
207 While the Baranyi model slightly diverges from the data after finishing the  
208 “exponential growth phase” and the Gompertz model fails to fit the data  
209 well in general. Additionally, these three non-linear models overlap together  
210 in the final stage, deviating from the declining data points.

211 As indicated in figure 2d, similarly, the quadratic model roughly depicts the  
212 changes in data and captures the decrease in the data points after the car-  
213 rying capacity has been attained. In this case, the cubic model fits the data  
214 nicely and even approaches the data at the lower end. For the mechanistic  
215 models, the logistic model again diverges from the data at the early stage  
216 but fits the remaining section well. The Gompertz model and Baranyi model  
217 have very similar effects on data fitting for this subset. These three non-  
218 linear models are highly overlapped after the “exponential growth phase”  
219 has finished and unable to describe the downward trajectory.

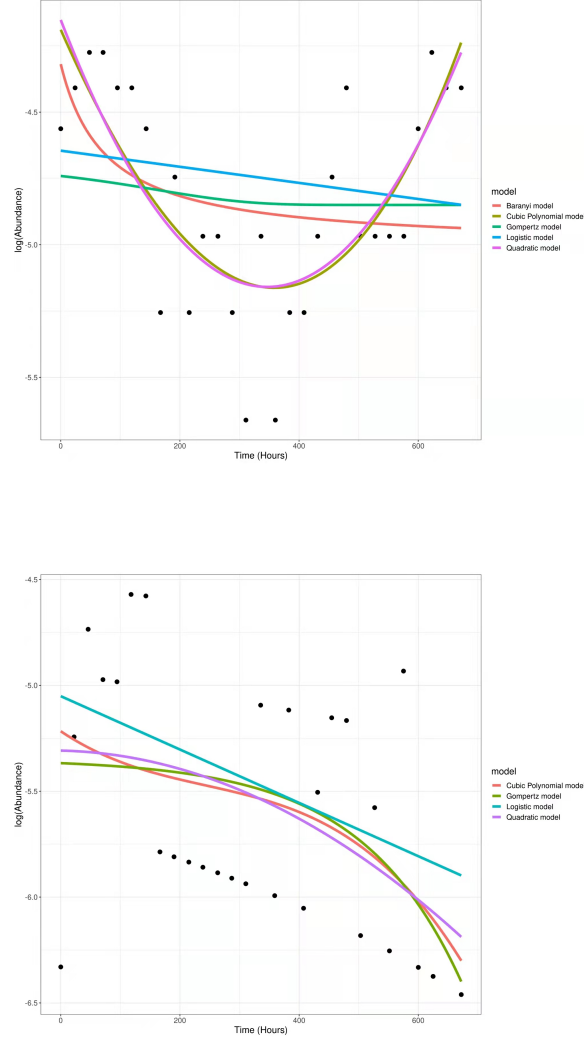


Figure 3: Subsets with scattered and irregular data points

220 However, as displayed in figure 3, data points of some subsets are extremely  
 221 scattered and irregular, and no model can fit these data well. And it is  
 222 apparent in Figure 4 that the logistic model completely fails to fit the data  
 223 at the lag phase, while the rest of the models can delineate the data well  
 224 before the “exponential growth phase” is reached.

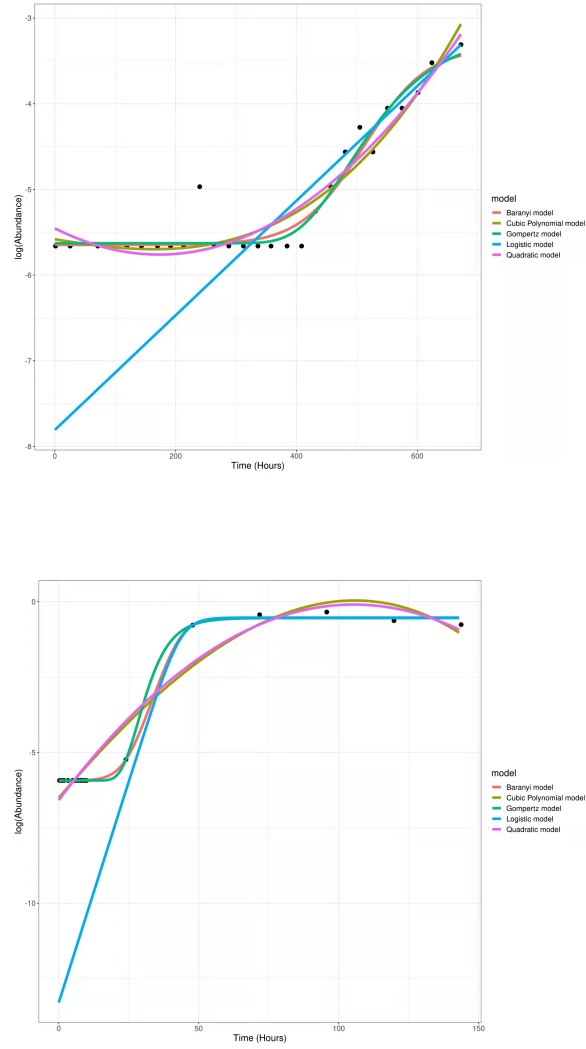


Figure 4: Subsets with distinctive lag phases

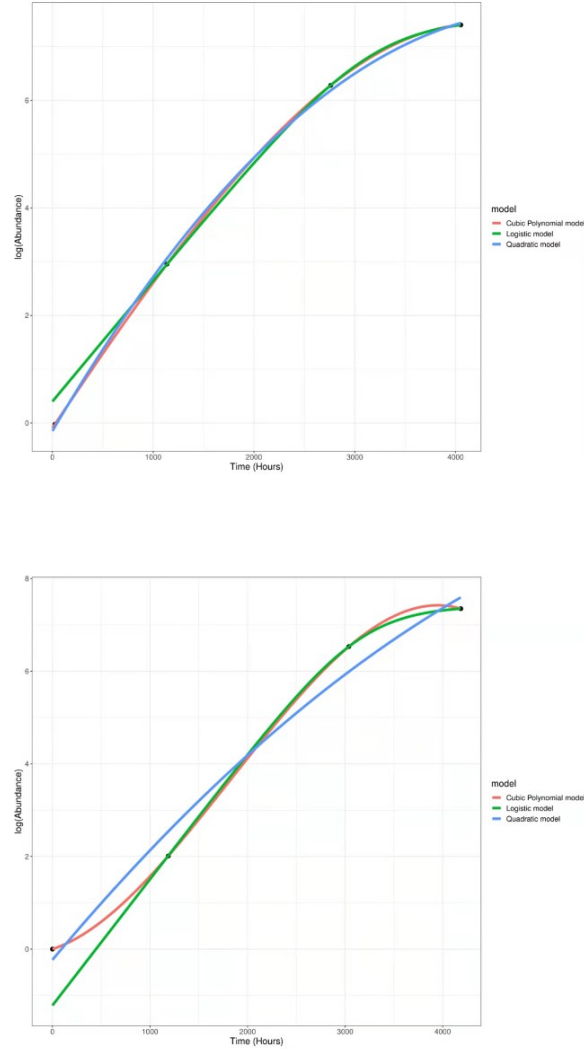


Figure 5: Subsets with only a few data points

Especially, what stands out in Figure 5 is that these subsets have only a few data points, and the number of microbial cells increases nearly linearly over time. Therefore, the quadratic and cubic polynomial models fit the data really well, the AIC and BIC values of the cubic models are calculated as negative infinity. While the Gompertz and Baranyi curves do not appear because the model fitting doesn't converge.



Figure 6: Number of cases with the smallest AIC/BIC value in each model

231 As for comparing and selecting models, the AIC and BIC values of these five  
 232 mathematical models are calculated for all the 285 subsets. Interestingly, it  
 233 can be observed that the number of cases when the AIC and BIC values of  
 234 the logistic model are the lowest among the five models is 88. And the count  
 235 of logistic model with the minimum AIC and BIC values is the largest. As  
 236 demonstrated in Figure 6, the quadratic model has the lowest AIC and BIC  
 237 values 26 times, the cubic polynomial model has the lowest AIC and BIC

values 71 times, the Gompertz model has the lowest AIC and BIC values 71 times, and the Baranyi model has the lowest AIC and BIC values 29 times.

## 4 Discussion

As stated above, five different mathematical models are adopted to fit the population growth data throughout the unique ID using ordinary linear and non-linear least squares methods to determine the models that best fit the dataset. One interesting finding is that although the quadratic model fits the data roughly, it can capture the “mortality phase” of population growth after the carrying capacity has been reached in most subsets. There are several cases of cubic polynomial model fitting for the population growth data. This model fits the data nicely in some subsets but poorly in others. In some cases, its curve goes up unexpectedly after the “stationary phase” is reached but in other cases, it captures the decrease in the population size after the maximum value is achieved.

For the mechanistic ones, the logistic model usually obviously diverges from the data at the “lag phase” but fits the remaining section pretty well in most subsets. The Gompertz model fitting can successfully converge in the majority of subsets. It fits the data perfectly in some cases but poorly in others, deviating from the data on the whole. And the Baranyi model fitting fails to converge in more than half of the subsets. But its curve fits the data excellently in general as long as the model fitting converges successfully. Furthermore, in most cases, the three non-linear models overlap together after the “stationary phase” has reached and are unable to describe the downward trajectory of the data. One noticeable finding is that if subsets have only a few data points, the cubic model fits the data remarkably.

Moreover, because the lower AIC or BIC, the better, the results of AIC and BIC values surprisingly suggest that the logistic model best fits the population growth data on the whole, while the quadratic model is the most inaccurate one to fit the data. The cubic polynomial model and modified Gompertz model generally fit the population growth data well at similar levels. The former is a phenomenological model and the latter, as a mechanistic model, fails to fit the data 77 times out of 285 subsets. Meanwhile, the Baranyi model poorly fits the whole dataset since the model fitting is

271 unable to converge 146 times. And Baranyi model emerges as a winner 29  
272 times out of the 139 successful convergences.

273 Unfortunately, the findings above are contrary to previous studies (Zwieter-  
274 ing et al., 1990) [17] which have suggested that the Gompertz model fitted all  
275 growth curves better than the linear, quadratic,  $t$ th-power, logistic, and ex-  
276 ponential models and was considered the best model to delineate the growth  
277 data in most cases. A possible explanation for this may be that the starting  
278 values of the primary parameters are not set to be quite appropriate by us-  
279 ing the non-linear least squares method, leading to the inaccurate Gompertz  
280 models. In addition, it has been suggested that the Baranyi model was more  
281 advantageous than the Gompertz model in terms of goodness-of-fit at higher  
282 growth rates (Baranyi et al., 1993) [2]. And López et al. (2004) [8] also re-  
283 ported that the Baranyi model showed a remarkably outstanding ability to  
284 fit the microbial growth data than the Gompertz model. The findings above  
285 again seem to be inconsistent with the previous research. This discrepancy  
286 may be explained by the facts that the starting values of the Baranyi model  
287 parameters deviate from the proper ones using the non-linear least squares  
288 method and the population growth data itself contains a large number of  
289 problematic and disorganized subsets. Although the findings above do not  
290 support the previous studies, they can still provide new insights into the  
291 phenomenological and mechanistic model fitting for the population growth  
292 data.

293 The findings are not very encouraging since they disagree with the previous  
294 research. The inconsistency may be due to the inaccurate starting values  
295 of the important parameters in the mechanistic models, problematic and  
296 scattered values in the population growth dataset, and the single criterion  
297 (AIC/BIC) for model comparison and selection.

298 Therefore, further research can be undertaken to set the starting values  
299 in the mechanistic models more precisely such as using the rolling regres-  
300 sion, remove the low-quality values in the dataset, and adopt more compli-  
301 cated methods to compare and select models such as Akaike Weights and  
302 Likelihood-Ratio test.



## 303 5 Conclusion

304 To sum up, five different mathematical models are applied to fit the popula-  
305 tion growth data using ordinary linear and non-linear least squares methods  
306 to find the ones which best fit the dataset. The count of logistic model with  
307 the minimum AIC and BIC values is the largest, which indicates that the  
308 logistic model best fits the population growth data on the whole. And the  
309 logistic curve usually visibly diverges from the data at the “lag phase” but  
310 fits the remaining section pretty well in most subsets. The quadratic model  
311 is the most inaccurate one to fit the data, but it can capture the “mortality  
312 phase” of population growth after the carrying capacity has been reached  
313 in most cases. The cubic polynomial model and modified Gompertz model  
314 generally fit the population growth data well at similar levels. Meanwhile,  
315 the Baranyi model poorly fits the whole dataset since the model fitting fails  
316 to converge in more than half of the subsets. But its curve fits the data  
317 nicely in general as long as the model fitting converges successfully. Al-  
318 though the findings are contrary to the previous research, they can still  
319 provide new insights into the phenomenological and mechanistic model fit-  
320 ting for the population growth dataset using ordinary linear and non-linear  
321 least squares methods.

## 322 References

- 323 [1] Ken Aho, DeWayne Derryberry, and Teri Peterson. Model selection for  
324 ecologists: the worldviews of aic and bic. *Ecology*, 95(3):631–636, 2014.
- 325 [2] J Baranyi, TA Roberts, and P McClure. A non-autonomous differential  
326 equation to model bacterial growth. *Food microbiology*, 10(1):43–59,  
327 1993.
- 328 [3] József Baranyi and Terry A Roberts. A dynamic approach to predicting  
329 bacterial growth in food. *International journal of food microbiology*,  
330 23(3-4):277–294, 1994.
- 331 [4] Florent Baty and Marie-Laure Delignette-Muller. Estimating the bac-  
332 terial lag time: which model, which precision? *International journal of*  
333 *food microbiology*, 91(3):261–277, 2004.

- 334 [5] Benjamin M Bolker, Beth Gardner, Mark Maunder, Casper W Berg,  
335 Mollie Brooks, Liza Comita, Elizabeth Crone, Sarah Cubaynes, Trevor  
336 Davies, Perry de Valpine, et al. Strategies for fitting nonlinear ecological  
337 models in r, ad m odel b uilder, and bugs. *Methods in Ecology and*  
338 *Evolution*, 4(6):501–512, 2013.
- 339 [6] Lone Gram, Lars Ravn, Maria Rasch, Jesper Bartholin Bruhn, Al-  
340 lan B Christensen, and Michael Givskov. Food spoilage—interactions  
341 between food spoilage bacteria. *International journal of food microbi-*  
342 *ology*, 78(1-2):79–97, 2002.
- 343 [7] Jerald B Johnson and Kristian S Omland. Model selection in ecology  
344 and evolution. *Trends in ecology & evolution*, 19(2):101–108, 2004.
- 345 [8] Sophie López, M Prieto, Jan Dijkstra, Mewa Singh Dhanoa, and Jim  
346 France. Statistical evaluation of mathematical models for microbial  
347 growth. *International journal of food microbiology*, 96(3):289–300, 2004.
- 348 [9] Robin C McKellar and Xuewen Lu. *Modeling microbial responses in*  
349 *food*. CRC press, 2003.
- 350 [10] Harvey Motulsky and Arthur Christopoulos. *Fitting models to biological*  
351 *data using linear and nonlinear regression: a practical guide to curve*  
352 *fitting*. Oxford University Press, 2004.
- 353 [11] Samraat Pawar. Themulquabio.
- 354 [12] Micha Peleg and Maria G Corradini. Microbial growth curves: what the  
355 models tell us and what they cannot. *Critical reviews in food science*  
356 *and nutrition*, 51(10):917–945, 2011.
- 357 [13] Stefano Perni, Peter W Andrew, and Gilbert Shama. Estimating the  
358 maximum growth rate from microbial growth curves: definition is ev-  
359 erything. *Food microbiology*, 22(6):491–495, 2005.
- 360 [14] María-Leonor Pla, Sandra Oltra, María-Dolores Esteban, Santiago An-  
361 dreu, and Alfredo Palop. Comparison of primary models to predict  
362 microbial growth by the plate count and absorbance methods. *BioMed*  
363 *research international*, 2015, 2015.
- 364 [15] Matthew D Rolfe, Christopher J Rice, Sacha Lucchini, Carmen Pin,  
365 Arthur Thompson, Andrew DS Cameron, Mark Alston, Michael F

- 366 Stringer, Roy P Betts, József Baranyi, et al. Lag phase is a distinct  
367 growth phase that prepares bacteria for exponential growth and involves  
368 transient metal accumulation. *Journal of bacteriology*, 194(3):686–701,  
369 2012.
- 370 [16] Charles P Winsor. The gompertz curve as a growth curve. *Proceedings*  
371 *of the National Academy of Sciences of the United States of America*,  
372 18(1):1, 1932.
- 373 [17] MH Zwietering, Il Jongenburger, FM Rombouts, and KJAEM  
374 Van’t Riet. Modeling of the bacterial growth curve. *Applied and envi-*  
375 *ronmental microbiology*, 56(6):1875–1881, 1990.