
Mini Project

Comparing Mathematical Models for Bacterial Growth Curves Fitting

Word Count: 2656

Yuchen Yang

MRes Computational Method for Ecology and Evolution

Department of Life Sciences

Imperial College London

yuchen.yang19@imperial.ac.uk

Abstract

This study compares the Baranyi, Logistic, Gompertz, and the basic Polynomial(Cubic) model's behaviour in a bacterial growth data set using AIC , ΔAIC , and $AICweight$. The result indicates that the **Gompertz** model is the the best model for fitting this general data set, (best model to fit a data set group for a total of **182** times, taking up **62.5%** of the total). The study also used a subset of **Tetraselmis tetrahele** in the data set to examine temperatures effects on model fitting's behavior. Using the Kruskal-Wallis Test, All models have shown a significant differences ($Baranyip-value = 0.001679$, $Logisticp-value = 0.01684$, $Gompertzp-value = 0.03801$, $Polynomialp-value = 0.01672$) in its $AICweight$ across different temperature groups. Finally, this study conducted analysis on models behaviour on a death phase of a growth curve.

1 Introduction

Mathematical models are used to understand and predict the growth of bacterial population for quite a long time. Bacterial growth in batch culture follows a distinct set of phases - lag phase, exponential phase and stationary phase. which would perfectly fit a sigmoid shape. The Gompertz model(Gompertz 1825) and the Logistic model (Verhulst 1838) are the most frequently used sigmoid models fitted to growth(Tjørve and Tjørve 2017). However most of the time nowadays, taking measurement errors and the expected growth kinetics into consideration, food microbiology researches would use the logarithm of cell concentrations as raw data. Some of the papers are using those classical sigmoid growth models as above, but since the population is in logarithm of the cell concentrations, the results are not as good(Baranyi 2011). Baranyi also pointed out that the reason those two models, while being very popular and useful, has their own limitations, especially in fitting bacterial data, is that the original model's curve does not remain sigmoid shape, but rather a monotone shape in log-space. Hence they could fail to capture the lag phases, and so that the use of such models are limited(Baranyi 2011).

It has been argued that the empirical models such as modified Gompertz and Logistic models were not

as preferred as the more mechanistic models such as the Baranyi model even with the modifications for it to work better in log-space(Juneja et al. 2009). in order to compare and evaluate similarities and differences between the models behaviour on a general bacterial database, this study has chosen the Baranyi (Baranyi, Roberts, and McClure 1993), the Logistic, the Gompertz(both using the log-transformed formulation of Zwietering)(Zwietering et al. 1990) and the basic Polynomial(Cubic) model, from the most mechanistic to the least, to better understand the behavioral differences. I have chosen the version of four parameters for all of the models.

the Polynomial model is introduced mainly to benchmark models' behavior on a fourth stage that has not been very often mentioned - death phase. since A Polynomial model may be quite good at capturing decline in population size after some maximum value (the carrying capacity) has been reached (the "death phase" of population growth)(Buchanan 1918), it is interesting to see how different models are behaving in that period of time too.

To recap on the objectives:

- This study would first compare the above mentioned models behavior for a general bacterial data set, and determine which one fits better.
- This study would help to understand different co-variates' (specifically focus on temperature) effect on model fitting performance.
- This study will touch base on the death phase performances.

2 Data

2.1 Basic Features

The data set consists of data from multiple researches, which contains measurements of change in biomass or number of cells of microbes over time and were collected through lab experiments across the world. the original data can be found in the `Data` directory with the name `LogisticGrowthData.csv`, detailed column meta data could be seen in `LogisticGrowthMetaData.csv`. The two main fields of interest are called `PopBio`(abundance), and `Time`.

The data set consists of 4387 rows of data from 10 researches, covering a total of 17 temperature groups from 0 to 37°C, 45 species, and 18 mediums. At the first glance, the *min* of the `Time` is -25.2632, the *min* of the `PopBio` is -668.284, suggesting further data wrangling is required, but there are no missing values in the data.

2.2 Pre-Processing

2.2.1 Re-Calibrate

The data set is grouped into 305 groups based on its `Temp`, `Species`, `Medium`, `Citation`, and `Rep` to better facilitate the study. The group sample size is rather small, with a *mean* of 14.36066 and a *median* of 12. Furthermore, each group's starting point $O(T_{\{0\}}, N_{\{0\}})$ is used as a reference to move the whole group's data to the origin in order to get rid of any negative values and make better sense.

2.2.2 Logarithm

Since all the model we have used is either designed for the log space, or being log-transformed, to better unify the results and deal with measurement errors and etc., a new column called `logPopBio`, which is the natural log of the original `PopBio`, is added and used for all future analysis.

2.2.3 Death Phase

To better understand the fitting performance of all models during the "death phase", which is a decline of population after reaching the maximum population, a column named `deathPhase` is added. By default the value is 0, and when $\frac{N_{max} - N_{tlast}}{N_{max} - N_{min}} > 20\%$, the value is set to 1, which indicate there might be a death phase in the graph.

3 Methods

3.1 Models

For this particular study, I have used transformed/modified models to make sure they all have four parameters to fit and should work well within a log-space:

3.1.1 Baranyi

$$y = N_{max} + \ln \left(\frac{-1 + e^{r_{max} \cdot t_{lag}} + e^{r_{max} \cdot t}}{e^{r_{max} \cdot t} - 1 + e^{r_{max} \cdot t_{lag}} \cdot e^{N_{max} - N_0}} \right) \quad (1)$$

3.1.2 Logistic

$$y = N_0 + \frac{N_{max} - N_0}{1 + e^{\frac{4 \cdot r_{max} \cdot (t_{lag} - t)}{N_{max} - N_0} + 2}} \quad (2)$$

3.1.3 Gompertz

$$y = (N_{max} - N_0) \cdot e^{-e^{\frac{r_{max} \cdot e \cdot (t_{lag} - t)}{N_{max} - N_0} + 1}} + N_0 \quad (3)$$

3.1.4 Polynomial(Cubic)

$$y = C_0 + C_1 \cdot t + C_2 \cdot t^2 + C_3 \cdot t^3 \quad (4)$$

In model 1, 2, and 3, the four parameters to fit are N_0 , N_{max} , r_{max} , and t_{lag} , where N_0 is initial cell culture (Population) density, N_{max} is maximum population density (aka "carrying capacity"), r_{max} is the maximum growth rate (the tangent to the inflection point), and t_{lag} is the x-axis intercept to this tangent. Parameters like A (in Gompertz) and h_0 (in Baranyi) are obtained and transformed using the four parameters. Meanwhile model 4 has four parameters with no biological meaning at all.

3.2 Starting Value

For this fitting practice, all starting values for model (4) have been set to 1, since the parameters are purely mathematical, and not biologically relevant, and it would fit regardless. For the rest of the models, methods for obtaining estimates for starting values are described in table 1:

Parameter	Method	Equation
N_0	Get the min of N(PopBio)	$N_0 = \min(PopBio)$
N_{max}	Get the max of N(PopBio)	$N_{max} = \max(PopBio)$
r_{max}	Get the max of slopes for all adjacent points(approximate tangent to the inflection point)	$r_{max} = \max(\frac{\Delta N}{\Delta t})$
t_{lag}	Get the $y = N_0$ intercept to the line with r_{max}	$t_{lag} = t_{rmax} - \frac{N_{rmax}}{r_{max}} + N_0$

Table 1: Overview on getting starting values

3.3 Evaluation

Coefficient of determination R^2 is used to determine converged but bad fit, and will be used to exclude data set groups that do not make sense. Since the package used in python did not return R^2 , it is calculated by using the already available SS_{res} and calculated SS_{tot} using values of the data set group and model fit estimate.

AIC and ΔAIC are used to populate the grading system for models. ΔAIC is calculated as $\Delta AIC_i = AIC_i - AIC_{min}$. Most of the time, it is recommend not to use AIC without the bias correction term, which is $AICc = AIC + \frac{2k^2+2k}{n-k-1}$, unless $\frac{n}{k} < 40$, where k is the total number of parameters of a model(degree of freedom for a model), and n is the sample size of the working data set group(Burnham and Anderson 2002). While in this study, all of the model has the same parameter count of 4, making $AICc$ and AIC essentially the same.

BIC is calculated as $\ln(n)k - 2\ln(\hat{L})$ whereas $AIC = 2k - 2\ln(\hat{L})$, \hat{L} represents the maximized value of the likelihood function of the model. Making the differences between AIC and BIC purely based on the penalty part. In order for AIC and BIC to have a difference, $\ln n$ needs to be than 2, meaning the sample size should at least be larger than 8. Since the average sample size of data set groups is $n_{mean} = 14.36066$ as mentioned, there should be a difference between the two standards. Essentially

speaking, *BIC* is different because it prefers less complicated models when sample size is larger, however all models used in this study have 4 parameters($n = 4$), which means they have the same degrees of freedom, and that renders the *AIC* and *BIC* less different than each other. δBIC would still be used in the grading process, but it is likely to give the same result.

P are given based on each group's fitting results. In this particular study, 2 system would be used, one using *AIC*, the other using *BIC*. 1 point will be given to a model using $\Delta AIC/BIC < 2$, where 2 is set to be the threshold for a substantial difference in performance.(Burnham and Anderson 2004; Kass and Raftery 1995).

Since this study is going to compare model performance across different co-variate groups, and the size of different co-variate groups S_{group} various. In order to be able to compare, a relative indicator is needed other than the P_{model} for the comparison, here we use Akaike weights $w_i(AIC)$ (Wagenmakers and Farrell 2004), calculated as in equation 5

$$w_i(AIC) = \frac{e^{-0.5 \cdot \Delta_i AIC}}{\sum_{k=1}^K e^{-0.5 \cdot \Delta_k AIC}} \quad (5)$$

3.4 Computing Tools

This project used *Python* for data preparation and fitting, *R* for plotting all graphs and doing analysis, *LaTeX* for report writing, and *Bash* to glue scripts together and make it automated and reproducible. Packages used in *Python* are *pandas*, *math*, *numpy*, and *lmfit*. *ggplot2* and *reshape2* is used in *R* for plotting.

4 Result

4.1 Post-Processing

4.1.1 Bad Fit

Using values obtained according to section 3.2 Starting Value, 2 of the all fit returns a *result.success == False* which would indicate a unsuccessful fit due to not being able to converge. Upon further in-

vestigation, this is most certainly caused by having inappropriate initial value(larger than it should) for r_{max} . In order to get better initial values for the fit, a *while* loop is introduced for a fit failing to converge, where it randomly samples a uniform distribution of $(\frac{r_{max}}{3}, r_{max})$ to be the new r_{max} . There are other cases where there is an exception for *valueError* and the fit is aborted. Based on the error message, it is caused by having negative values in the $\log()$ function for the Banranyi model or overflow for $\exp()$ for all the model. This requires further investigation on either manipulate the raw data more, or set boundaries to parameters, and will not be done in this particular study.

4.1.2 Bad Data

A first look at R^2 overview from figure 1 indicates quite some out-liners in R^2 for all models. These indicate that although there is a fit for the corresponding data using that model, the goodness of fit is relatively bad(a low value in R^2). Here a threshold is set to $R^2 = 0.8$, since this value, as shown in the graph would help to exclude maximum of out-liners without losing too much information.

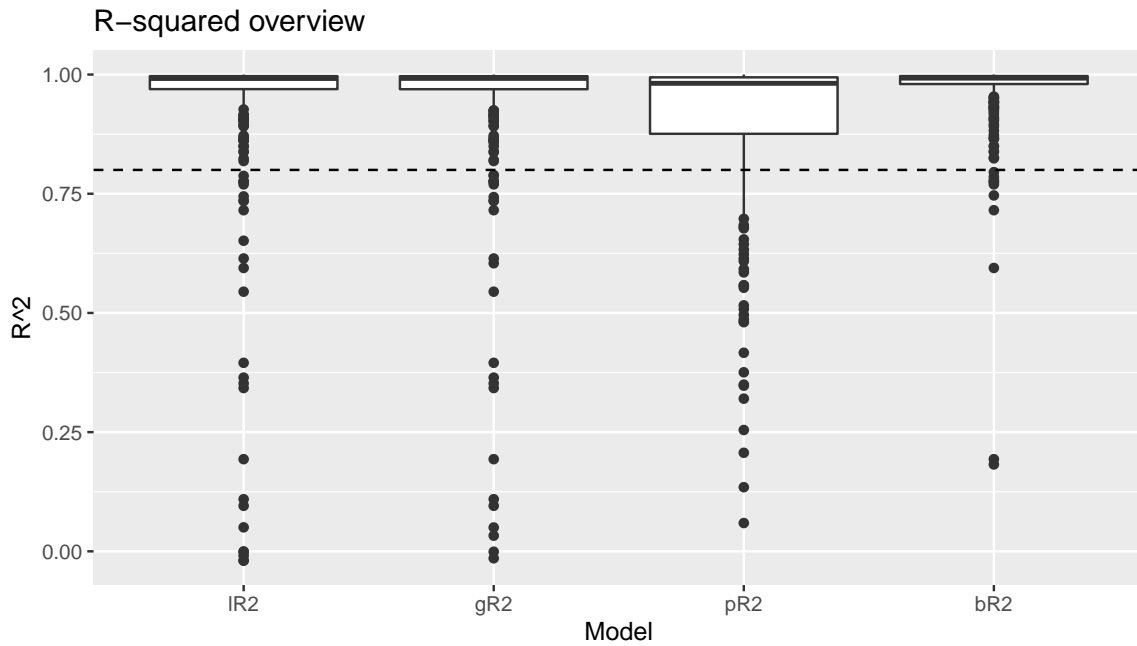
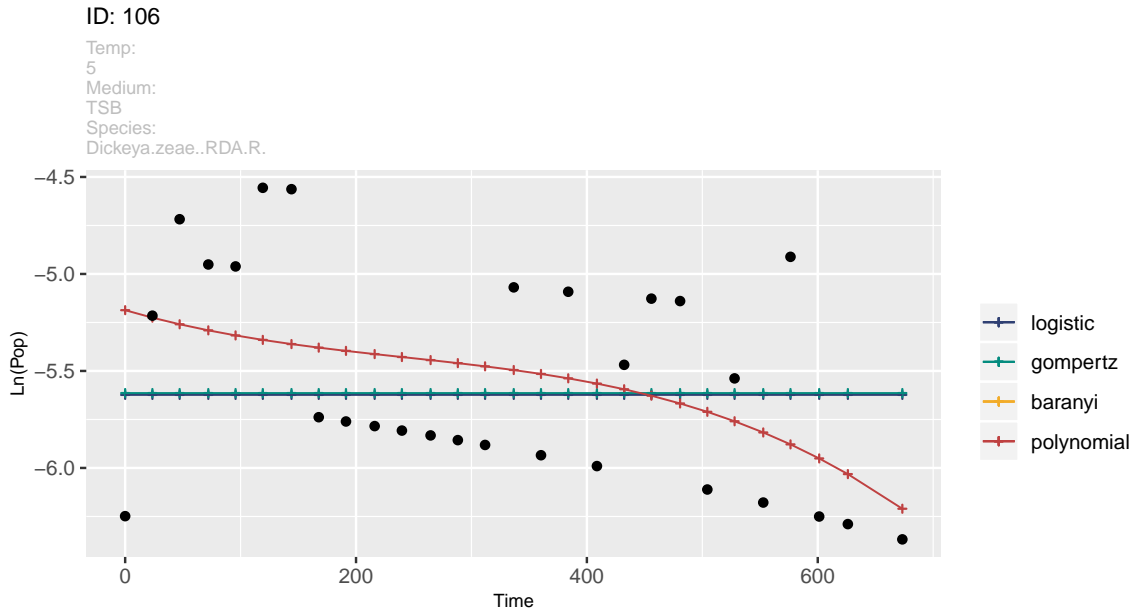


Figure 1: R^2 overview for each model, where $lR2 = Logistic$, $gR2 = Gompertz$, $pR2 = Polynomial$, $bR2 = Baranyi$

Figure 2 is an example of a data set where either all models' fit is bad($R^2 < 0.75$), or some of the model did not fit, and rest of the model indicated a bad fit. Before grading the model, data set groups

that are similar to figure 2, are excluded from the final evaluation data. Upon examination, all of these cases happened because of the fact that the data set group itself is showing a bad shape, as could be seen in figure 2. A total of 14 sets are excluded in this step, leaving the final count of data set groups to be 290. The majority of the data set groups are showing a good value in R^2 , Which also indicates a pretty decent starting value choice.



106 .pdf

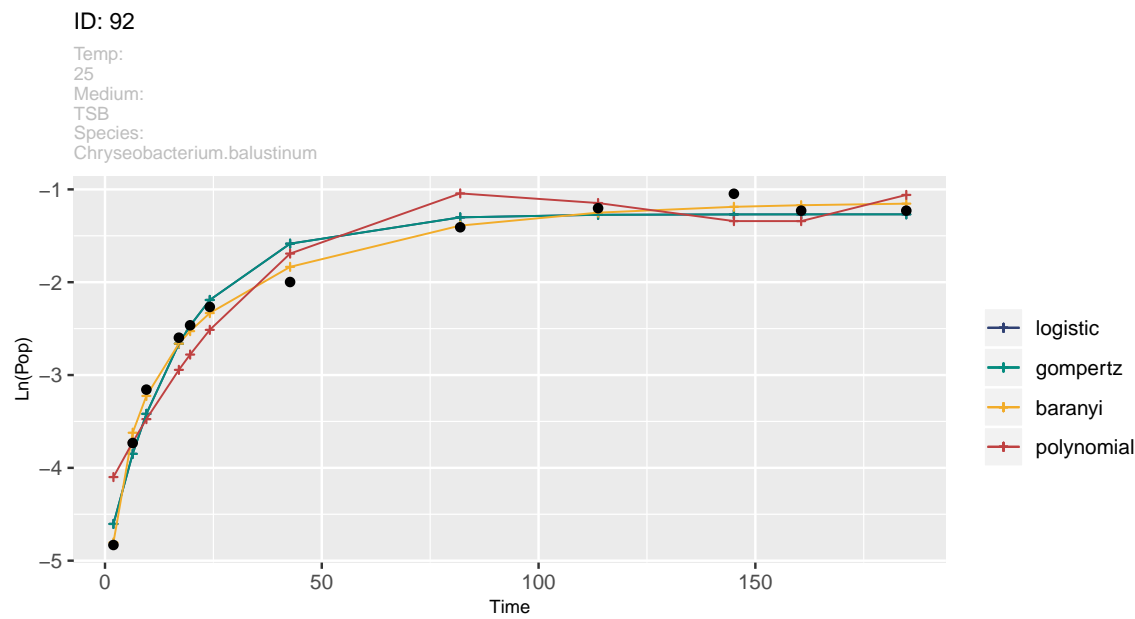
Figure 2: Example of deleted data set, where the Baranyi model didn't fit, and the rest three models showed bad fitting.

4.1.3 Grading

This part will take data set group $ID = 92$ as an example to better illustrate the grading process. As can be seen in the evaluation matrix shown in table 2 and graph shown in figure 3 for this specific data group, all four model fit well(all with a $R^2 > 0.91$). And according to the rule for the grading system illustrated in section 5, only the Baranyi model gets 1 point for having a $\Delta AIC < 2$, and all other model would not get a point for all of their ΔAIC is greater than 2. As can be seen, ΔBIC in this particular study is giving exactly the same results, therefore BIC would not be considered in the following analysis.

	AIC_{g2}	$\Delta_{g2}AIC$	$w_{g2}AIC$	BIC_{g2}	$\Delta_{g2}BIC$	R^2	Point
<i>Baranyi</i>	-50.92886	0	9.996976e-	-48.98923	0	0.9942026	1
			01				
<i>Gompertz</i>	-33.33531	17.59355	1.511741e-	-31.39568	17.59355	0.9748831	0
			04				
<i>Logistic</i>	-33.33458	17.59428	1.511190e-	-31.39495	17.59428	0.9748816	0
			04				
<i>Polynomial</i>	-18.89898	32.02988	1.108328e-	-16.95935	32.02988	0.9163562	0
			07				

Table 2: Evaluation matrix, ID=92



92 .pdf

Figure 3: Example of a data set, where all models fit.

4.2 Fit Result

4.2.1 Overall

The final result for there models' fitting performances in the general data set can be seen in table 3. The **Gompertz** model has the highest P amongst all. It is the the best model for fitting this general data set, regarded as the best model to fit a data set group for a total of **182** times, taking up **62.5%** of the total group size. The **Logistic** and **Baranyi** performed quite similarly with a percentage of **54.6** and **52.6** respectively.

Baranyi	Gompertz	Logistic	Polynomial	Total Groups
153(52.58%)	182(62.5%)	159(54.6%)	77(24.6%)	291

Table 3: Overall Score, with percentage of Score over total size in bracket.

4.2.2 Co-variate Group Fit

Upon further examination of the data set, **Tetraselmis tetrahele** is used to conduct the analysis of model's performance differences across temperature groups in the next section, as it is the only specie in the data set that allows such analysis(5 repetitions under same medium in different temperature groups). However, this part will only illustrate the overall fit results in different temperature groups. The data set used for this study contains 17 temperature groups from 0 to 37°C. The final fit result for each group using each model can be seen in table 4. **10 out of 17 (58.82%)** groups, the **Gompertz** model still kept the lead amongst others. The **Logistic** and **Baranyi** performed quite similarly again, taking the lead for **5 out of 17(29.41%)** and **6 out of 17(35.29%)** groups respectively.

Temperature($^{\circ}C$)	Baranyi	Gompertz	Logistic	Polynomial
0	0	5	1	0
2	5	8	7	5
4	8	6	5	6
5	13	11	12	7
6	4	3	5	4
7	4	14	6	2
8	5	9	5	3
10	10	18	11	4
12	10	7	7	5
15	24	29	24	14
16	6	4	7	0
20	19	21	20	15
25	19	15	17	5
30	7	9	9	4
32	3	3	3	0
35	14	16	17	3
37	2	4	3	0

Table 4: Detailed point for temperature groups

Fit results in data set groups with and without a death phase are illustrated in table 5. The **Gompertz** model performed the best in data set groups without a death phase with, and was out performed by **Baranyi** when there is a death phase.

Death Phase	Baranyi	Gompertz	Logistic	Polynomial
Yes	19(73.1%)	14(53.8%)	16(61.5%)	8(30.8%)
No	134(52.3%)	168(63.4%)	143(54.0%)	69(26.0%)

Table 5: Detailed point for death phase groups, with percentage of the total group size following in brackets.

4.3 Comparisons In Performances

In order to have a more accurate and reasonable analysis on model's performance across different temperature groups, **Tetraselmis tetrahele** is chosen for this part of the analysis. A total of 5 temperature groups are presented ($T = 5$, $T = 8$, $T = 16$, $T = 25$, $T = 32$). Differences in the models performance which was indicated in $wAIC$ can be seen in figure 4. Using the Kruskal-Wallis Test, All models have shown a significant differences ($Baranyip - value = 0.001679$, $Logisticp - value = 0.01684$, $Gompertzp - value = 0.03801$, $Polynomialp - value = 0.01672$) in its $wAIC$ across different temperature groups at .05 significance level, indicating temperature may have effected these models' performance for this specific subset of data. The **Baranyi** have the most significant differences across all groups.

Further comparing the pairwise difference shows that the **Baranyi** has the most significant differences in pair $8^{\circ}C - 16^{\circ}C$ ($p - value = 0.007937$), the **Gompertz** has the most significant differences in pair $8^{\circ}C - 25^{\circ}C$ ($p - value = 0.01587$), the **Logistic** has the most significant differences in pair $8^{\circ}C - 16^{\circ}C$ and pair $5^{\circ}C - 16^{\circ}C$ ($p - value = 0.007937$), while the **Polynomial** has significant differences in pair $5^{\circ}C - 16^{\circ}C$, pair $8^{\circ}C - 16^{\circ}C$, and pair $16^{\circ}C - 25^{\circ}C$ ($p - value = 0.007937$).

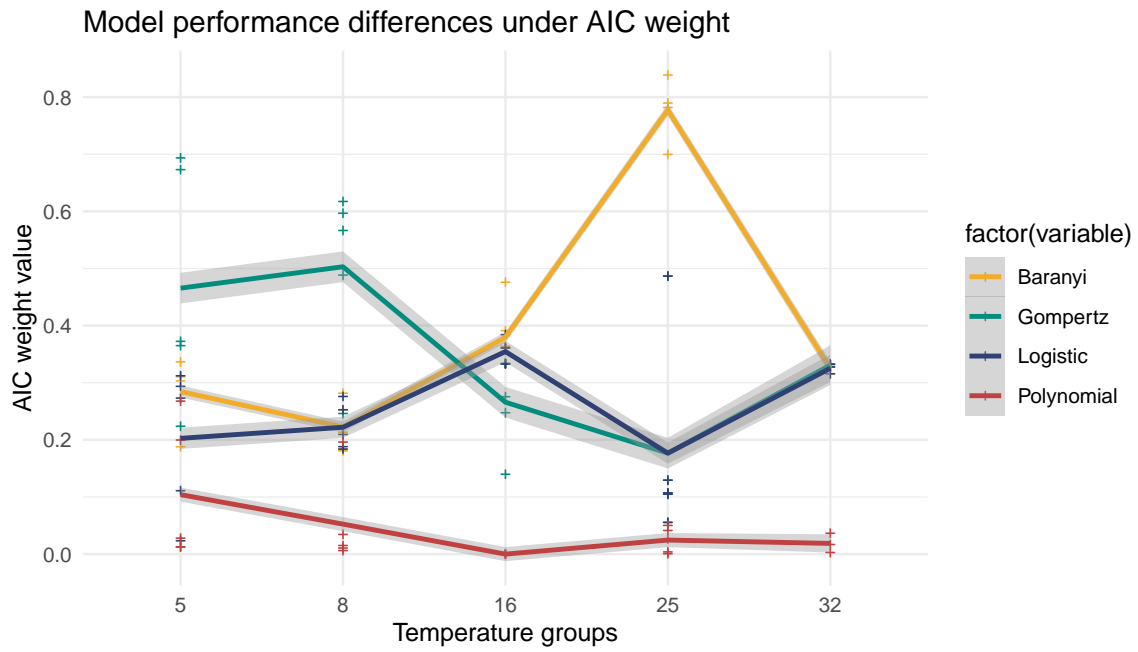


Figure 4: wAIC of different model across temperature groups

190 Similarly for the death phase groups(seen in figure 5), **Baranyi** and **Gompertz** model have shown
 191 a significant differences ($p - value = 0.04993$, $p - value = 0.02158$) in date set with and without a
 192 death phase at .05 significance level, suggesting that having a death phase in the data set group would
 193 probably effect these models' performances. The **Polynomial** and **Logistic** model have not shown a
 194 significant differences ($p - value = 0.6555$, $p - value = 0.8133$) in their performance across the two
 195 groups.

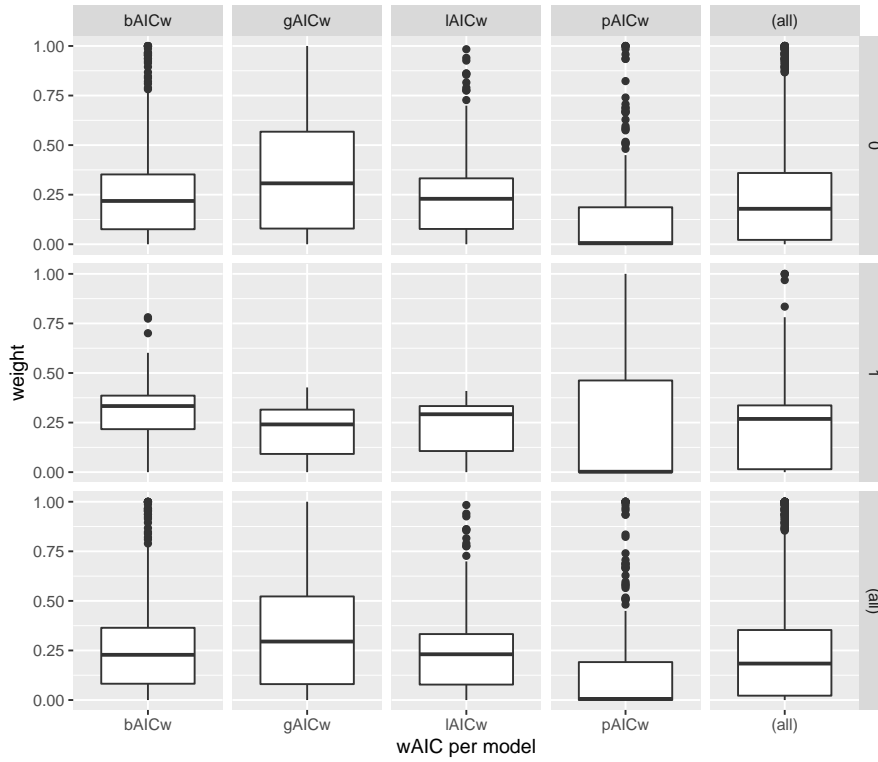


Figure 5: wAIC of different model for dataset with and without a death phase

5 Discussion

5.1 Reasons For Performance Difference

Although the previous analysis has concluded there are performance differences for all four models, especially for the **Baranyi**, and some vague trends can be seen as in figure 4, the results are only based on a rather small data subset ($size = 92$) for **Tetraselmis tetrahele**. It is hardly an indication for anything solid.

Also the difference could be caused by different sample sizes, measurement errors, and so on. However, these errors are very hard to quantify.

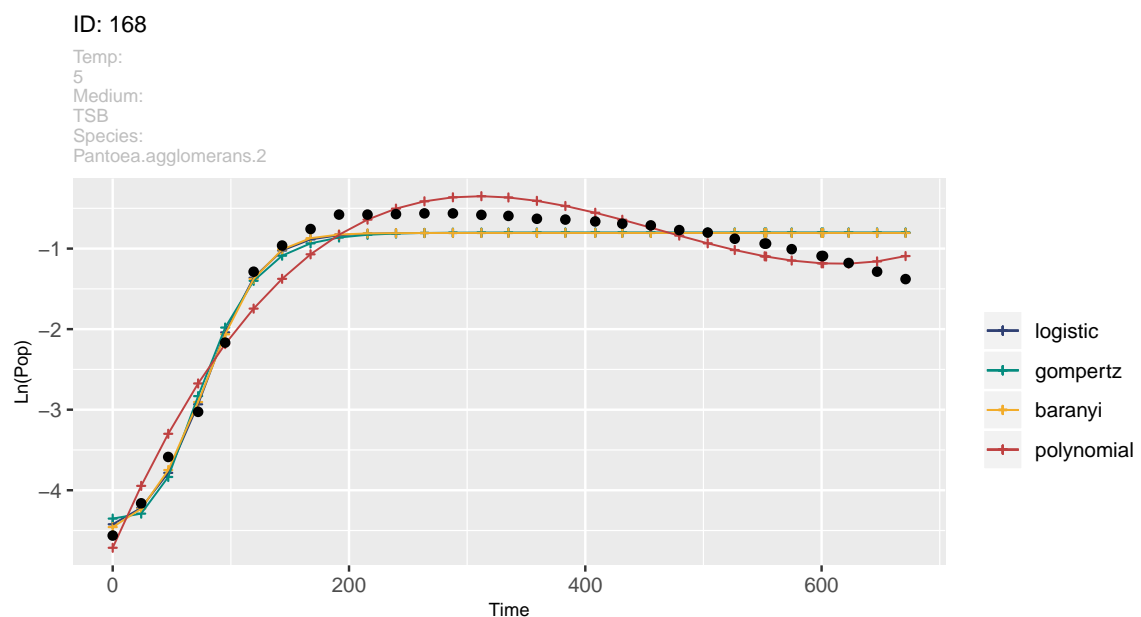
The fundamental question for model performance difference is probably much aligned with the fact on how temperature will effect the shape of population growth curve for bacteria, since models(excluding the **Polynomial**) used in this study indicate a very similar shape regardless.

A proper investigation in the relationship between temperature and models' performances should require much more experimental data, and the analysis should be conducted in a larger scale in order to limit the error brought by data.

5.2 Death Phase And The Polynomial Model

The **Polynomial**(Cubic) model is introduced to mainly illustrate the disadvantages other models have regarding the fourth stage of bacterial growth curves - death phase, since it allows the curve to have a decreasing phases after hitting the maximum point.

In reality, as shown in table 5, the behavior of the Polynomial model is not better than any of the other models. A closer look at a data set group that has a death phase (figure 6) reveals some of the reasons why it is the case.



168 .pdf

Figure 6: Example of a data set, where all models fit.

As can be seen, the data set group has shown a decline since $Time = 200$, and models other than Polynomial are fitting the decline using a straight line. Although the Polynomial model is somehow catching the decline, the shape of the cubic model is relatively different from the 4-phased growth curve. Catching the decline in the data set group does not necessarily help with the overall fit result.

However the comparison does show the disadvantages of the other three models used in this study at catching the decline after the peak. It also shows that the Polynomial model, due to its natural of having an increase after the decline, is not the best at fitting the death phase.

This could be a very rough first step of working on finding or building models that could work well

²²⁶ for the 4-phased growth curve.

References

- Baranyi, J, TA Roberts, and P McClure (1993). “A non-autonomous differential equation to model bacterial growth”. In: *Food microbiology* 10.1, pp. 43–59.
- Baranyi, József (2011). “Modelling and parameter estimation of bacterial growth with distributed lag time”. PhD thesis. szte.
- Buchanan, RE (1918). “Life phases in a bacterial culture”. In: *The Journal of Infectious Diseases*, pp. 109–125.
- Burnham, Kenneth P and David R Anderson (2002). “A practical information-theoretic approach”. In: *Model selection and multimodel inference, 2nd ed. Springer, New York*.
- (2004). “Multimodel inference: understanding AIC and BIC in model selection”. In: *Sociological methods & research* 33.2, pp. 261–304.
- Gompertz, Benjamin (1825). “XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c”. In: *Philosophical transactions of the Royal Society of London* 115, pp. 513–583.
- Juneja, Vijay K et al. (2009). “Mathematical modeling of growth of Salmonella in raw ground beef under isothermal conditions from 10 to 45 C”. In: *International journal of food microbiology* 131.2-3, pp. 106–111.
- Kass, Robert E. and Adrian E. Raftery (1995). “Bayes Factors”. In: *Journal of the American Statistical Association* 90.430, pp. 773–795. ISSN: 01621459. URL: <http://www.jstor.org/stable/2291091>.
- Tjørve, Kathleen MC and Even Tjørve (2017). “The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family”. In: *PloS one* 12.6.
- Verhulst, Pierre-François (1838). “Notice sur la loi que la population suit dans son accroissement”. In: *Corresp. Math. Phys.* 10, pp. 113–126.
- Wagenmakers, Eric-Jan and Simon Farrell (2004). “AIC model selection using Akaike weights”. In: *Psychonomic bulletin & review* 11.1, pp. 192–196.
- Zwietering, MH et al. (1990). “Modeling of the bacterial growth curve”. In: *Appl. Environ. Microbiol.* 56.6, pp. 1875–1881.