Comparisons of Models Applied in Microbial Population Growth

Congjia.chen (Congjia.Chen21@imperial.ac.uk)

Dec,2021

[1]Word Count: 3058

[2]CMEECoursework,Department of Life Sciences, Imperial College of Science, Technology and Medicine

Abstract

Microbial growths is highly correlated with human society. Therefore, having knowledge of the microbial growth is essential so that human can anticipate or control their growth under particular conditions. Mathematical models have been proved to be functional in microbial growth anticipation. However, lack of empirical model comparison with universal data will lead some bias to the model selection. In this report, based on the model fitting and model selection on 285 empirical data sets, non-linear model performs better than linear model.Among the non-linear model, logistic model is sufficient for simple growth situation,while Gompertz and Baranyi can handle more complicated situation like lag phase. However, all of the non-linear models such as logistic model and Baranyi model have the defect that they can not be applied to fit the death phase properly. Segmented model and external factors calibration might be a potential strategy to optimize the result.

1 Introduction

Microbial growths is highly correlated with human society, for example, the yeast growing in wort to make beer, the pathogenic bacteria to make human sick and the microbe which leads to food spoilage. Therefore, having knowledge of the microbial growth is essential so that human can anticipate or control their growth under particular conditions[1]. In contrast to multi-cellular organisms, microbial growth is measured by population growth, either by counting the number of cells or by increasing the overall mass. However, current methods of population measurement are relatively complicated[2].

In a closed system, the growth curve of the microbial population can be divided into four phases: (1) Lag Phase, (2) Log(exponential) phase, (3)Stationary phase, (4) Death phase[3]. In particular, after adapting the new environment to the lag phase, the microbial population increases exponentially while its abundance is low and resources are not limited. This growth slows down and ends up when resources become scarce.

Figure 1 shows that the application of mathematical model is dramatically increasing on microbial researches.

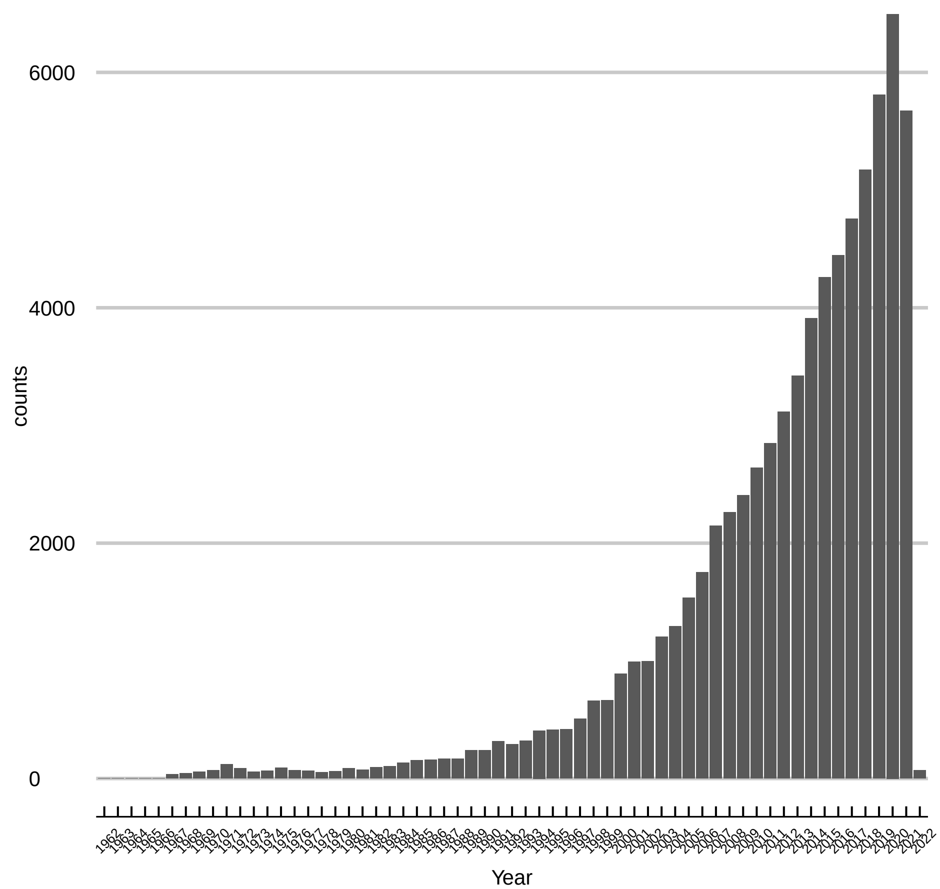


Figure 1: The searching result from PudMed.gov with key words : Microbial Modeling

There are currently many mathematical models with different benefits and drawbacks to illustrate the population growth curve[4]. The most widely used mathematical models are the Logistic model[5],modified Gompertz model[5] and Baranyi model[4]. Theoretically, if the per-capita growth rate of a population is held constant (with limitless resource), the microbe will lead to exponential unbound growth. However, it is not biologically realistic. Therefore, one common way to address this deficiency is to use the logistics model. In which,after the exponential growth, the rate will decrease to zero as the population approaches a fixed value, also know as carrying capacity. Although the logistic model is apparently simple and suitable for some situations, it is still not generic enough in capturing other phenomena. Therefore, the modified Gompertz model (Gompertz)[5] and Baranyi model [4] which considered the influence of lag phase duration were created. Although some literature had examined the goodness of fit among several models[6, 7], most of them focus on the growth curve in one microbial species which might be not generic for other species or under other conditions.

Therefore, Which model's performance is better to describe the general microbial population growth is the main concern of this study. Based on 285 different empirical experiments data collected around the world with different traits like species,growing medium,growing temperature etc., this study focused on fitting and compare 6 different models (including linear models and Non-linear models) to all the lab experiments data. Both Akaike information criterion (AIC) and Bayesian information criterion (BIC) [8] were used as criterion to assess the model fitting and to evaluate the performance of model. However, given that the AIC and BIC performed 98% similar selection, for the concern of efficiency and universality, AIC will be used as the main criterion in this study. Moreover, Whether the environment (temperature and medium), measurements (Units) will influence the performance of the model will be assessed as an extra support to the model selection strategy.

2 Data and Methods

2.1 Data Set and Preparation

The data set used in the study is called LogisticGrowthData.csv.The field names are defined in a meta-file called LogisticGrowthMetaData.csv. Both files are accessible in :

https://github.com/nedchen2/CMEECourseWork/tree/master/MiniProject/data.

The two main fields of interest are PopBio (abundance with different units) and Time (Hours). Based on the combination identifier (unique temperature-species-medium-citation combinations), 285 independent empirical experiments were labelled with unique id from 1 to 285. Given the log(exponential) phase during the growth, we did ln transform to the abundance (PopBio) and named the result as LogPopBio. Among the data sets, there were some negative abundance which were replaced with a small number 0.0000000000000000001.

2.2 Mathematical Models

This study covers 6 models. In following equations, T is the time points variables. $LogPopBio(T)$ is defined as the ln(abundance) at the time point T

Simple Linear Regression model(named as Straightline in the report)

$$LogPopBio(T) = A_0 + A_1 T + A_2 T^2 \tag{1}$$

Quadratic polynomial model

$$LogPopBio(T) = A_0 + A_1 T + A_2 T^2 \tag{2}$$

Cubic polynomial model

$$LogPopBio(T) = A_0 + A_1T + A_2T^2 + A_3T^3 \tag{3}$$

Logistic model

$$LogPopBio(T) = \frac{N_0Ke^{r_{max}t}}{K + N_0(e^{r_{max}}t - 1)} \tag{4}$$

Gompertz model

$$LogPopBio(T) = N_O + (K - N_0)e^{-e^{\frac{r_{max}e^1(t_{lag}-T)}{(K-N_0)log(10)}+1}} \tag{5}$$

Baranyi model[4]:

$$LogPopBio(T) = N_0 + r_{max}(T + \frac{1}{r_{max}}log(e^{-r_{max}T}) + e^{-r_{max}t_{lag}} - e^{-r_{max}(T+t_{lag})} - \frac{log(1 + ((e^{r_{max}(T+\frac{1}{r_{max}}lo}}{} \tag{6}$$

The parameters used in non-linear least squares model have biological meanings. N_0 is the initial population size, r_max is the maximum growth rate , and K is carrying capacity (maximum possible abundance of the population),t_lag is the duration of the delay before the population starts growing exponentially. The algorithm I used to estimate the parameters above will be introduced in subsequent part.

2.3 Model Fitting

With the development of computer and model fitting, multiple software can be applied least square algorithm for calculating the parameters of different models including linear (OLS) and non-linear model (NLLS). Levenberg-Marquardt algorithm [9] was applied to search and optimal parameter estimates (or minimize the residual sum of squares,i.e. RSS). Maximum iteration number was set to be 200.

The start value estimate algorithm required in the non-linear model will be listed in table 2:

Table 1: The algorithms to estimate the start value

Parameters	Algorithm
N_0	minimum population of certain experiment
K	maximum population of certain experiment
r_{max}	the slope of straight line model fitting of certain experiment
t_{lag}	the time where the maximum differentials of population takes place

2.4 Model Selection

For the sake of the high accuracy, adjusted R-squared would be applied to compare the Linear Models[10]. However, when compares the non-linear and linear models, AIC and AIC_c and BIC of all models were calculated using following equations[8, 11]:

Akaike information criterion (AIC)

$$AIC = -2ln(Likelihood) + 2k \tag{7}$$

Bayesian information criterion (BIC)

$$BIC = -2ln(Likelihood) + kln(n) \tag{8}$$

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \tag{9}$$

Where n is sample size, k is number of parameters in the model, and $Likelihood$ is maximized likelihood function value of the model. AIC and BIC take both goodness of fit and model complexity into consideration to evaluate the performance of the model[8].Although the formula of AIC and BIC are quite similar, they are distinguishing in theoretical bases[8]. The criterion was compared based on the empirical data considering the universality and efficiency to select the most appropriate one.

Model selection strategy.

Assuming that we are comparing a set of models with AIC values, we need a new data called ΔAIC which equals to the difference between a given model and the model with the lowest AIC.

Rough AIC: When one model's AIC equals to the minimum of the whole sets of AIC in different models, these models are identified as the best models.

Strict AIC: When one model's ΔAIC is less than 2[12], these models are identified as the plausible best models.

2.5 Extra Factors to Support The Model Selection

Measurements: In the data set,there are 4 different units which means the experiments were recorded by different measurements. Therefore, we want to know that if the measurement is relevant to the model performance.

Temperature: In original data set, the temperatures ranged from 0 to 37. In order to better categorize the temperature[13], we roughly classify the temperature between 0 - 5 into ”cold lover”, the temperature between 5-20 into ”middle cold lover”, and the temperature between 20-37 into ”middle hot lover”.

Medium: Finally, we classify the medium which is more routine such as milk and chicken as ”Nature”, and the medium which is much more lab-based like TSB and MRS as ”Artificial”.

After classification, the best fitted number of six types of models in every category was counted. Then the frequency would be transformed into percent and plotted to bar plot.

2.6 Computing Software

For the reason of powerful and convenient application, R 4.1.1 were used as the main computing tools for data wrangling,model fitting and visualization. Several packages such as ggthemes, ggpubr, tidyverse (for data wrangling and visualization), and minpack.lm (for model fitting) were imported in the study for different purpose. A bash script was used to compile LaTex. Python3, particularly with the subprocess package, was used as a work flow control tool to run all the scripts.

3 Results

3.1 Linear Least Squares Model Fitting

Simple linear regression model (straight line),quadratic polynomial,cubic polynomial were applied to fit the data. Adjusted R-squared are used as the criterion to evaluate the model fitting.

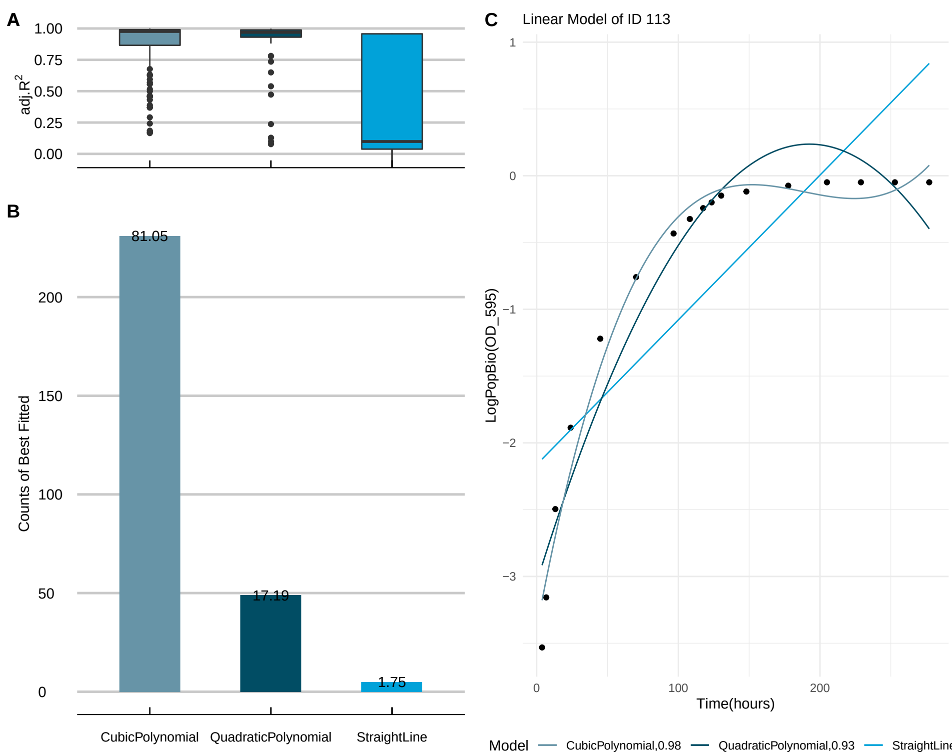


Figure 2: Linear model fitting. A.The box plot to show the range of adjusted R-squared in different models. B.The bar plot to show the best fitted counts in different models. C.An example of linear fitting

In Figure 2.A, the diagram shows that the cubic polynomial and quadratic polynomial model had a higher fitted R-squared value (close to 1). In the Figure 2.B, the bar plot shows the distribution of 3 different models frequency which best fits the 285 experiments. As demonstrated in the figure, the counts of best fitted cubic polynomial is much larger than other two models, which accounts for 81.05% percent of the 285 experiments. In the Figure 2.C, the figure shows the fitting of three linear models of experiment ID 113. The adjusted R-squared value are displayed after the model name. The cubic polynomial fitting shows the best fitting, while the Straightline Model shows the worst fitting. However, Figure 2.C also illustrates that the cubic polynomial model can not fit the stationary phase properly,despite that the R-squared are 0.98.

3.2 Comparison between AIC, AIC_C, BIC

In order to compare the OLS with NLLS model, we will use AIC,AIC_C or BIC as criterion in the comparison and selection. To simplify the analysis procedure, I decided to compare the three criterion first. A criterion which could best represented the model selection result would be chosen as the criterion for subsequent research.

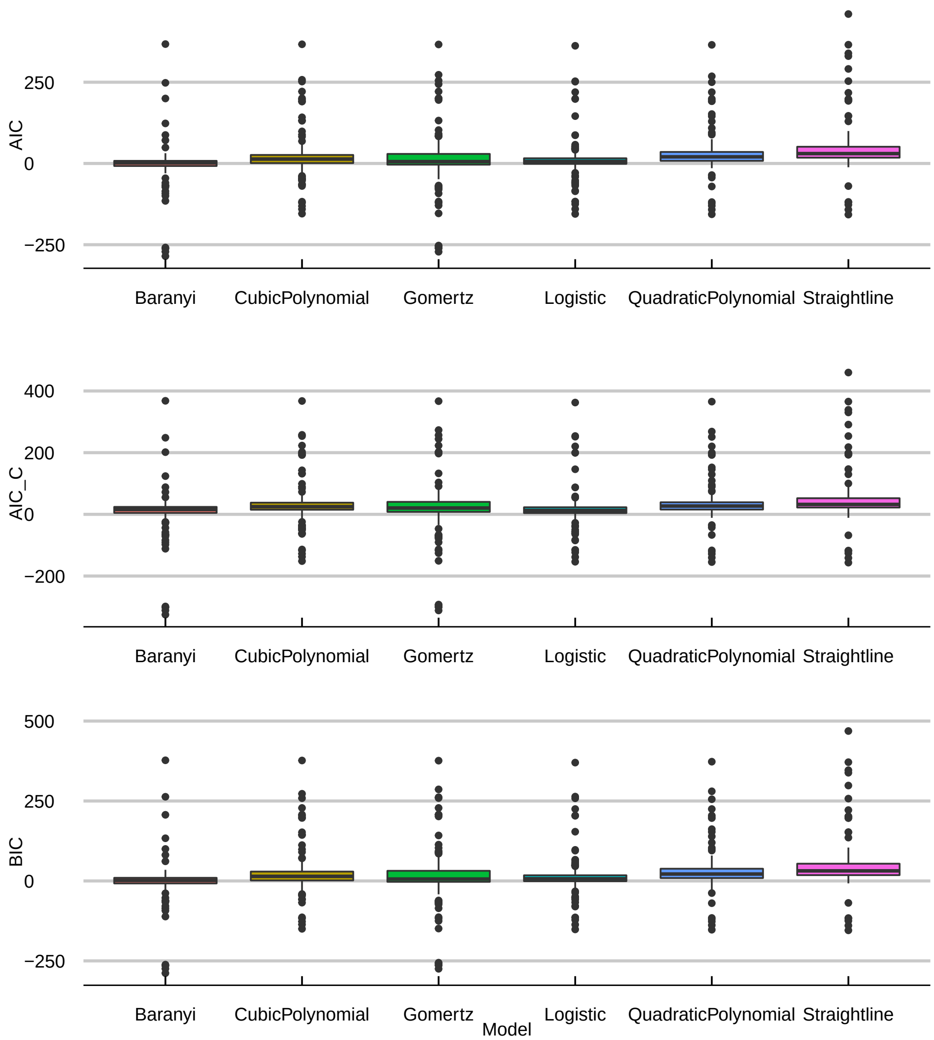


Figure 3: Box plot of different models show the range of AIC,BIC,AIC_C. The colours of Box represent different plausible models

Generally, in all criteria, Straightline model shows the largest average value. NLLS models including logistics,Gompertz,Baranyi models shows lower average value. Consequently, the figure in different criteria shows the similar pattern of value distribution. After Comparing the model selection result based on rough AIC algorithm, the result shows that 97.54% of the best fitted result are supported both by AIC and BIC. However, only about 50% of the best fitted result are supported both by AIC and AIC_C. For the universality and efficiency of the study, the decision was made to use AIC in the subsequent report.

3.3 Linear and Non-linear Model Fitting

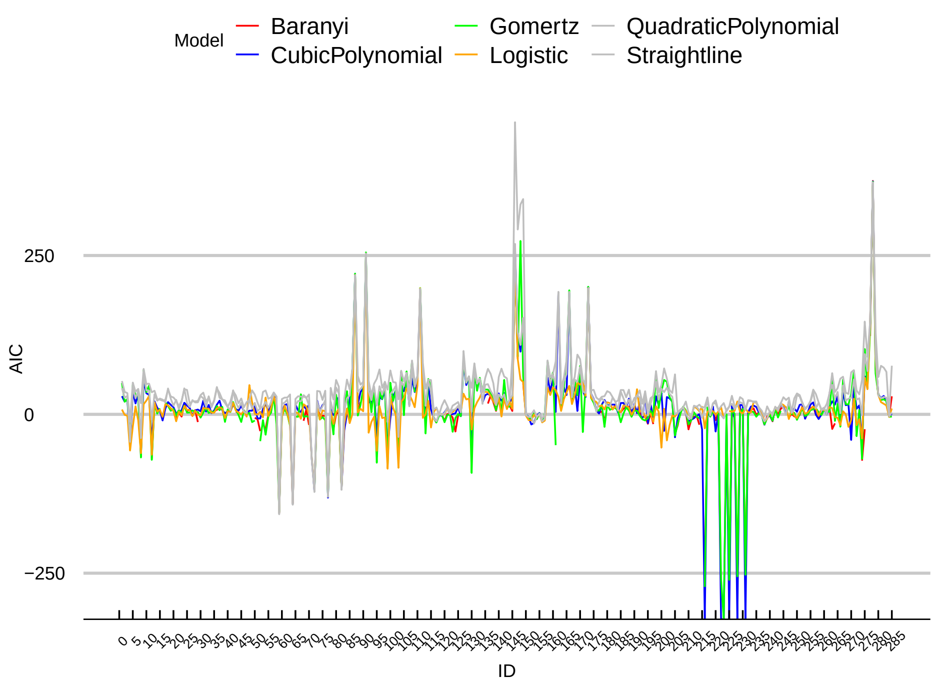
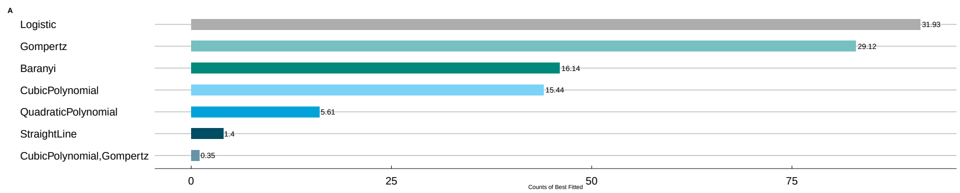


Figure 4: AIC distribution of different models across the experiment ids. The colours of lines represent different plausible models

Figure 4 shows the AIC value distribution of different models across the experiment ids. Based on the result of the preceding linear model fitting, the misfitted quadratic polynomial model and the straight line model will be defined as grey lines in Figure 4. As illustrated in the figure, a large portion of the yellow curve shows the lowest AIC value across the IDs axis. However,implicit difference in distribution of the AIC value in the various models can also be observed.



B table 1

	Best_Model_Type	Freq	Criterion	Percent(%)	SuccessCounts	RelativePercent(%)
1	Logistic	91	AIC	31.9298246	284	32.0422535
2	Gompertz	83	AIC	29.1228070	277	29.9638989
3	Baranyi	46	AIC	16.1403509	185	24.8648649
4	CubicPolynomial	44	AIC	15.4385965	285	15.4385965
5	QuadraticPolynomial	16	AIC	5.6140351	285	5.6140351
6	StraightLine	4	AIC	1.4035088	285	1.4035088
7	CubicPolynomial,Gompertz	1	AIC	0.3508772	285	0.3508772

table 2

	Model_type	Freq	Percent(%)
1	Baranyi	88	30.877193
2	CubicPolynomial	85	29.824561
3	Gompertz	120	42.105263
4	Logistic	127	44.561404
5	QuadraticPolynomial	31	10.877193
6	StraightLine	11	3.859649

Figure 5: Best Fitted Models. A.The bar plot of best fitted models based on rough AIC criteria. The number illustrated beside the bar is percentage (counts of best fitted/total experiments number); B. table 1 is the table based on rough AIC. Freq column shows the counts of best fitted model.SuccessCounts column shows the successfully modelling counts for each models. RelativePercent column shows the relative best fitted percentage (Freq/SuccessCounts);B. table 2 is the table based on strict AIC

During the modelling, some data set could not be fitted by NLLS model.The reason might be that the start value is too far from the correct answer. Therefore, relative best fitted percent which indicates the best fitted counts over the successfully modelled counts will be used to generalize this issue, as illustrated in Figure 5.B(table1). Figure 5(A) shows that based on rough AIC selection algorithm, 3 non-linear models (Logistic model, Gompertz model, Baranyi model) are the top 3 best fitted models. In one experiments, both cubic polynomial and Gompertz are considered as the best model. 31.93% of the experiments fits the Logistic model best. 29.12% of the experiments fits the Gompertz model best. 16.14 % of the experiments fits the Baranyi models best. 15.44% of the experiments fits the cubic polynomial best. Although the difference between the performance of cubic polynomial and Baranyi model is small, considering the relative best fitted percentage (demonstrated in Figure 5.B(table 1), the performance of Baranyi model is much better than cubic polynomial. The Figure 5.B(table 2) shows the result based on strict selection algorithm. The result shows that the best fitted percent of Logistic and Gompertz model raises dramatically, indicating that the performance of these two models are much better than other models.

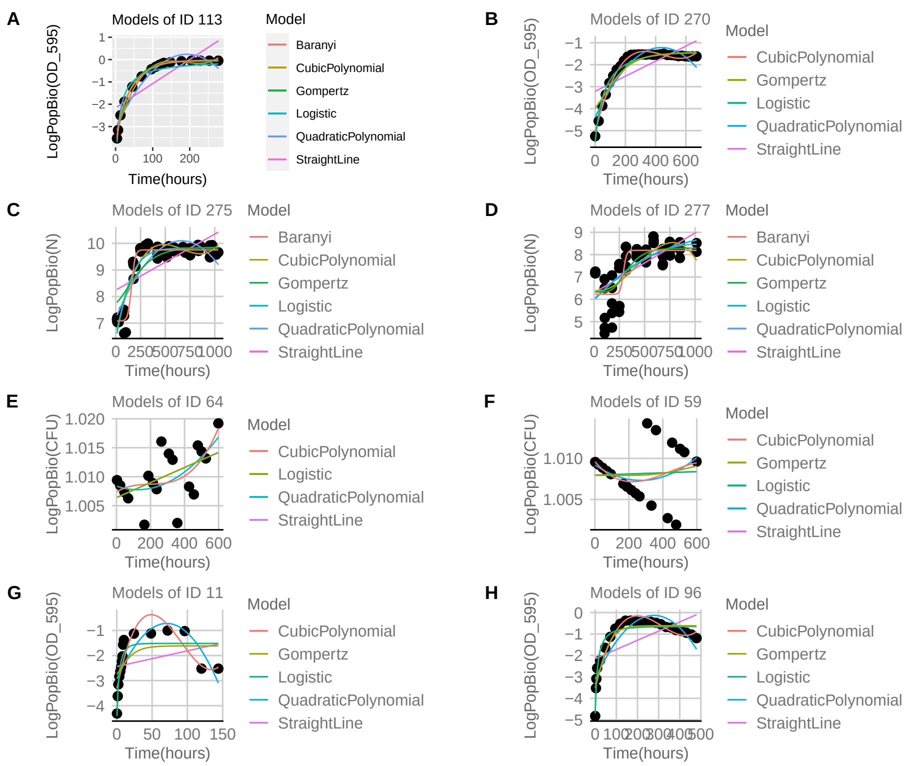


Figure 6: Examples of Model fitting in 285 experiments. Different models are represented in different colours.

The Figure 6 A,B show logistic growth phase curve and stationary phase curve.The difference of various NLLS models is not explicit in these cases. In the Figure6 C,D, sigmoid growth curve is displayed. We can also find that the non-linear model includig Baranyi, Gompertz models performs much better when the lag phase exists. However, although the Baranyi and Gompertz model can fit the lag phase, exponential phase and stationary phase, they can not fit the death phase properly,as illustrated in Figure 6 G,H. In the Figure6 E,F, the scatter shows unexpected pattern. In both cases, the straightline model shows the lowest AIC.

3.4 Models Performance of Different Measurements, Temperature, Medium

After classifying the best fitted result into different categories, the percentage of each models in different category was displayed.

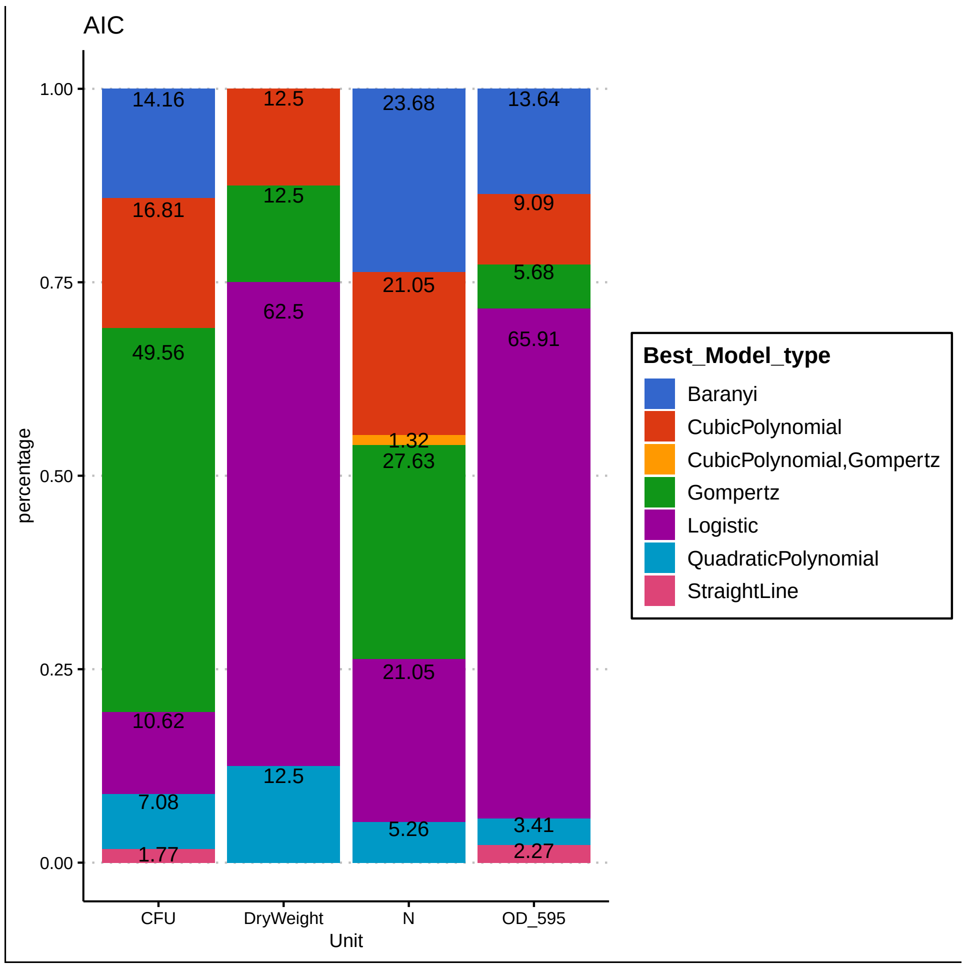


Figure 7: Best Fitted Model in different measurements. Different models are represented in different colours

Figure 7 shows that logistic models perform well (accounting for over 60% of best fitted model) for OD595 and DryWeight, while for CFU, Gompertz performs well.

According to previous work[**?**], we roughly classify the temperature between 0 - 5 into "cold lover", the temperature between 5-20 into "middle cold lover", and the temperature between 20-37 into "middle hot lover".

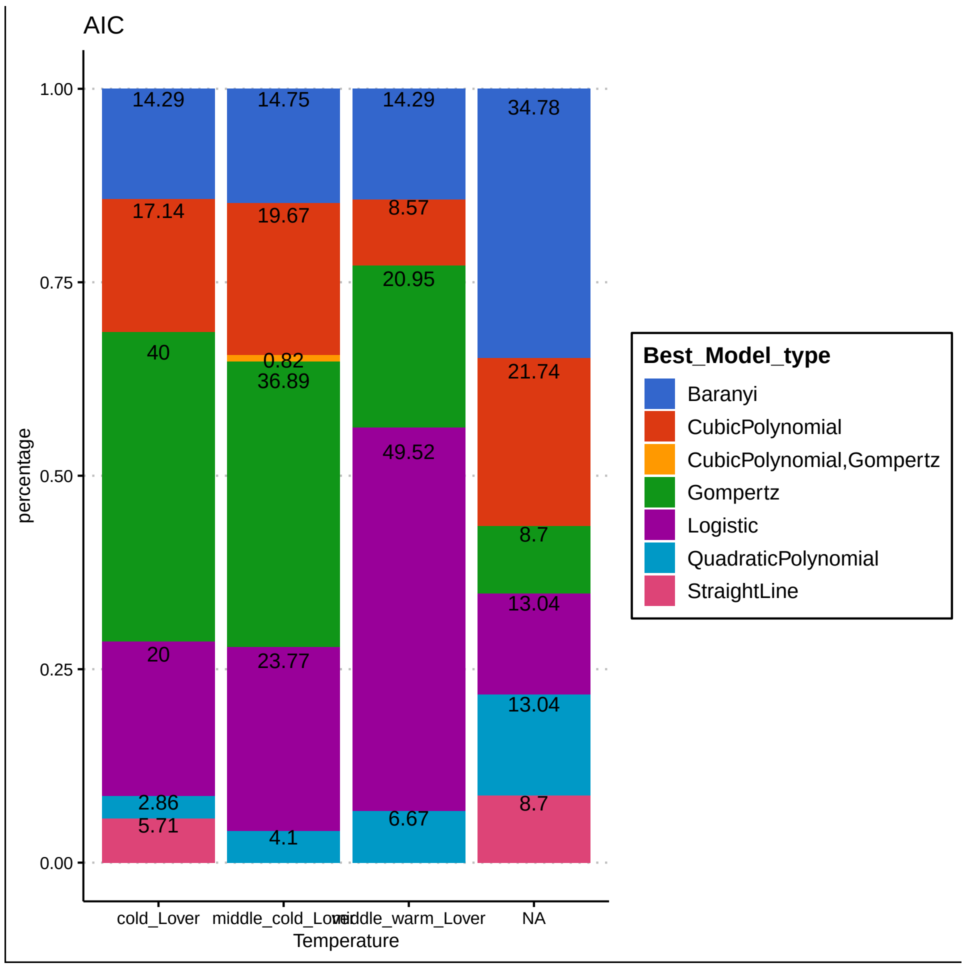


Figure 8: Best Fitted Model in different Temperatures. Different models are represented in different colours

Generally, Logistic model performs better in middle warm. Moreover, as the temperature grows, the best fitted percentage accounted by logistic model tends to grow. Additionally, the straightline model only displays in cold temperature.

Finally, we classify the medium which is more natural such as milk and chicken as "Nature", and the medium which is much more artificial like

TSB and MRS as "Artificial". After that, we count the best fitted models in "Nature" and "Artificial".

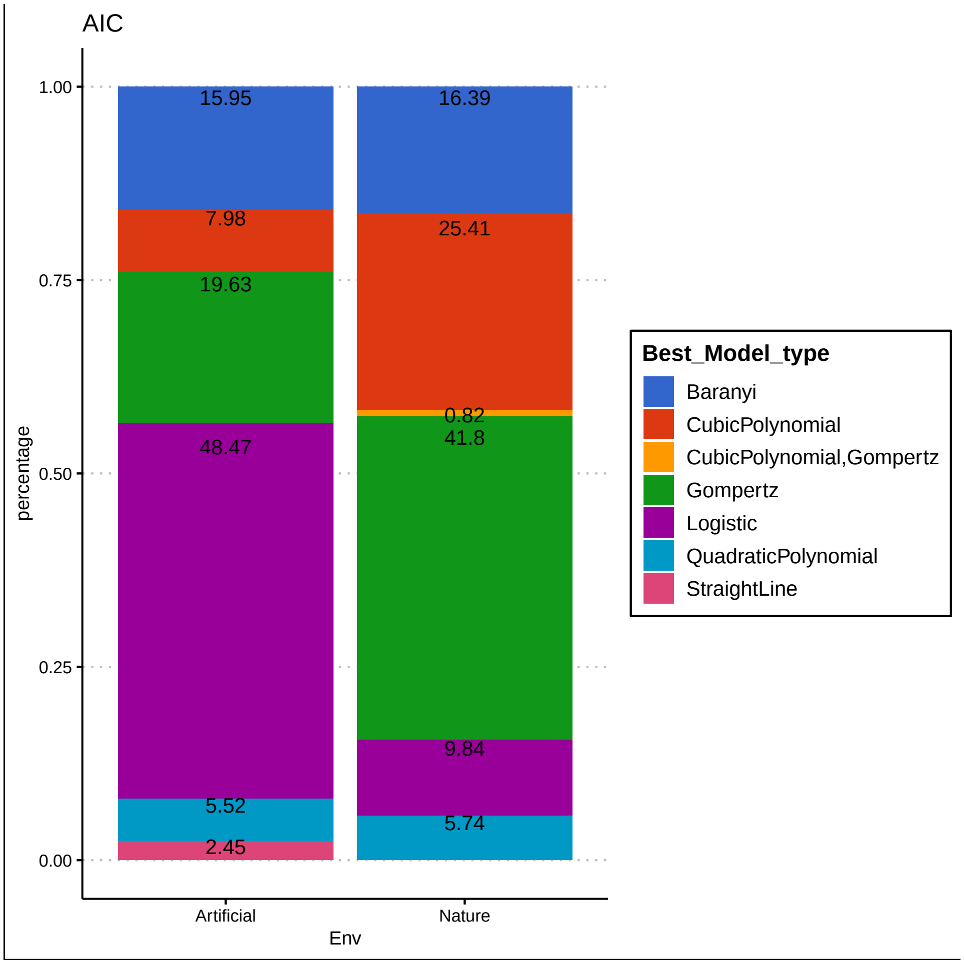


Figure 9: Best Fitted Model in different Medium.Different models are represented in different colours

The results shows that the Logistics model will be preferred in "Artificial" medium, while the Gompertz models will be preferred in "Nature" medium.

4 Discussion

Understanding how microbial organisms grow under particular conditions is essential for human to study population dynamics and microbe ecosystem[3]. In terms of microbial growth prediction and evaluation, choosing an universal and appropriate mathematical model which is supported by empirical experiments and valid data is an inevitable issue[14]. Currently, many models are pointed out to fit the microbial growth model. Therefore, to test the universality and application of certain model on empirical data, I did model fitting and selection on 285 worldwide experiments data set by 6 typical models.

NLLS Model fitting is challenging because of the start value choice. For example, the r_{max} , I have tried 3 methods including using the maximum derivatives of cubic polynomial model, using the A2 parameter of quadratic polynomial model and the and using the slope of the simple linear regression model. I choose the last method which have the highest successful fitting rate through the 285 experiments data set. However, the Baranyi model's successful fitting rate is still low (65%). Additionally, I did not randomly sample the start value to optimize the result of fitting. Therefore, efforts should be taken in the future to optimize the fitting by either improving the algorithm of the start value, or random sampling the start value.

Based on the comparison of various model criterion, in our study, almost 98% of the best fitted result is both supported by AIC and BIC as the best model. Therefore, for the concern of universality and efficiency, I decided to use AIC as the main criterion in further research.

During model selection, I define two selection rules : rough AIC and strict AIC to evaluate the performance of models. Firstly, regardless of the rules, non-linear least squares models (include Logistc,Gompertz,Baranyi models) perform better than linear models. In terms of the linear model, cubic polynomial model performs the best.

The results also showed that under strict AIC selection rules, logistic model is sufficient in fitting microbial growth for about 44% of experiments in the data set. Gompertz model is sufficient in fitting microbial growth for about 40% of experiments in the data set. Given the low successfully modelled rate, Baranyi Models can also performs better if we could improve the model fitting by some methods such as changing the start value algorithms.

Investigating the result with the ideal microbial growth curve in a "closed habitat", we can find some drawbacks among all models, regardless of linear or non-linear model.For instance, although cubic polynomial is good to fit the log(exponential) phase in empirical data, the cubic polynomial curve can not fit the lag phase and stationary phase properly, as illustrated in Figure 6 C,D and Figure 2 C. Similarly, although the Baranyi and Gompertz model can fit the lag phase, exponential phase and stationary phase, they can not fit the death phase,as illustrated in Figure 6 G,H. Furthermore, Straightline model might be a good indicator of abnormal data set, as illustrated in Figure 6 E,F. Consequently, although no models are univeral for all conditions, given the simple and efficient formula, logistic models can be considered as the sufficient model for most cases. However, logistic models can not handle the lag phase and death phase. Therefore, combine the model selection strategy with segmented growth phase might be a solution for handling different situations.

Apart from the segmented model strategy, we may use other factors to assist the model selection. For example, as illustrated in Figure 7, the logistic models are preferred (65%) when using OD595 as the microbial population measurements. However, the reason is unknown. Another example is that, the logistic models are preferred when the temperature getting warmer, which may be partly explained by that when the temperature increasing, the growth pattern may be more similar to logistic growth[13]. Finally, we classify the medium which is more natural such as milk and chicken as "Nature", and the medium which is much more artificial like TSB and MRS as "Artificial". The results shows that the Logistic model will be preferred in "Artificial" medium, while the Gompertz models will be preferred in "Nature" medium. The phenomenon might be explained by that in the "Nature" medium,the lag phase would be more apparent which is preferred by the Gompertz models.

In conclusion, based on the model fitting and model selection on 285 empirical data sets, non-linear model performs better than linear model. Among the non-linear model, logistic model is sufficient for simple growth situation,while Gompertz and Baranyi can handle more complicated situation like lag phase. However, all of the non-linear models such as logistic model and Baranyi model have the defect that they can not be applied to fit the death phase properly. Segmented model and external factors calibration might be a potential strategy to optimize the result.

5 Supplementary Materials

Please find the complete model fitting plot of every experiments in
https://github.com/nedchen2/CMEECourseWork/tree/master/MiniProject/results/
Filename Format: ID + ALL Model Plot.png
Please find the model fitting result data for each models in
https://github.com/nedchen2/CMEECourseWork/tree/master/MiniProject/results

Filename Format: Model type + result.csv

Please find the complete best model result depending on AIC,BIC,AIC_C in https://github.com/nedchen2/CMEECourseWork/tree/master/MiniProject/results

Filename Format: Best Model + criteria.csv

Please find the code in https://github.com/nedchen2/CMEECourseWork/tree/master/MiniProject/code

References

[1] William R. Shoemaker, Stuart E. Jones, Mario E. Muscarella, Megan G. Behringer, Brent K. Lehmkuhl, and Jay T. Lennon. Microbial population dynamics and evolutionary outcomes under extreme energy limitation. 118(33). Publisher: National Academy of Sciences Section: Biological Sciences.

[2] Thomas Egli. Microbial growth and physiology: a call for better craftsmanship. 6:287.

[3] Micha Peleg and Maria G. Corradini. Microbial growth curves: What the models tell us and what they cannot. 51(10):917–945. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10408398.2011.570463.

[4] J. Baranyi, P. J. McClure, J. P. Sutherland, and T. A. Roberts. Modeling bacterial growth responses. 12(3):190–194.

[5] M. H. Zwietering, I. Jongenburger, F. M. Rombouts, and K. van 't Riet. Modeling of the bacterial growth curve. 56(6):1875–1881.

[6] R Xiong, G Xie, A. S Edmondson, R. H Linton, and M. A Sheard. Comparison of the baranyi model with the modified gompertz equation for modelling thermal inactivation of listeria monocytogenes scott a. 16(3):269–279.

[7] R. L Buchanan, R. C Whiting, and W. C Damert. When is simple good enough: a comparison of the gompertz, baranyi, and three-phase linear models for fitting bacterial growth curves. 14(4):313–326.

[8] Jouni Kuha. AIC and BIC: Comparisons of assumptions and performance. 33(2):188–229. Publisher: SAGE Publications Inc.

[9] Henri P Gavin. The levenberg-marquardt algorithm for nonlinear least squares curve-fitting problems. page 19.

[10] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. 7:e623.

[11] Eric-Jan Wagenmakers and Simon Farrell. AIC model selection using akaike weights. 11(1):192–196.

[12] Adrian E. Raftery. Bayesian model selection in social research. 25:111–163. Publisher: [American Sociological Association, Wiley, Sage Publications, Inc.].

[13] Andrea De Silvestri, Enrico Ferrari, Simone Gozzi, Francesca Marchi, and Roberto Foschino. Determination of temperature dependent growth parameters in psychrotrophic pathogen bacteria and tentative use of mean kinetic temperature for the microbiological control of food. 9:3023.

[14] Kathleen M. C. Tjørve and Even Tjørve. The use of gompertz models in growth analyses, and new gompertz-model approach: An addition to the unified-richards family. 12(6):e0178691. Publisher: Public Library of Science.