The relative success of Gompertz models in fitting population growth data

**Abstract: (**background, objectives, methods, main results, conclusions, discussions)

Assessing the relationship between population size and time is of the essence in ecology and evolution filed. In this project, I attempt to study this relationship by operating model fitting, the powerful and versatile technique. In this project, three widely used models: polynomial(cubic) model, logistic model [Buchanan, 1997?]and Gompertz model [cite] l (Zwietering et al. 1990] was chosen to be fitted using lm(), and nlsLM() function of the R language which is based on the Levenberg-Marquardt algorithm. All three models are fitted successfully on several data[cite] obtained from published papers studying population growth of bacteria and phytoplankton under various experimental conditions in this project without any serious problem. The goodness-of-fit of each model were compared by making inference through comparing the AICc, AIC and BIC values. After comparison, the Gompertz model gave the overall best fit for the data. The cubic and logistic model gave similar performance. The general outperformance of the Gompertz model may because XXX.

**///////////////////**

the result shows that the Gompertz model is the best among the three models chosen in this project.

I chose the non-linear model …

Many of the problems we are facing in nature are nonlinear. For addressing more unsolved questions, In ecology, specifically, the study about the relationship between population size is a typical example and has a wide variety of applications in …. Bacteria live almost everywhere on earth. The sigmoidal curve has extensive application in the real world (example and citation).

/////////////////////

**Introduction (“expectations” for the readers /** why growth rates are important to study in biology/[abstract](https://en.wikipedia.org/wiki/Abstract_(summary)) or summary, [preface](https://en.wikipedia.org/wiki/Preface), acknowledgements, and [foreword](https://en.wikipedia.org/wiki/Foreword).**)**

Population density arranges along the time playing an essential role in ecology functions and evolutionary processes. Model fitting, as a powerful and versatile approach applied in a wild variety of research analysis, gives us a way of studying the relationship between population size and time. By fitting models, estimating the relationship between variables, we can make further inference to explore the mechanism behind it. Whereas there are loads of models to choose, it can be confusing which one is the most appropriate to use. So, in this project, I will represent you a general idea of the model fitting and model selection process.

Base on theory and subject knowledge, the polynomial(cubic) model, logistic model [Buchanan, 1997?]and Gompertz model [cites (Zwietering et al. 1990] are chosen for this project to fit.

– polynomial, logistic and Gompertz model —

Model fitting as has several reasons to be used: *the*

ofThe model-fitting has wildly application in agriculture, ecology, medicine and food safety.

**Methods**

{states briefly how each of the scripting languages (bash, git, R, Python) was used and what packages within them were used and a justification of why.}

///////////////////

**2.1 Computing tools**

\*\*\*Data processing, model fitting and results in visualization are performed with R due to its excellent graphing capability and numerous packages for statistical analysis and data handling, including reshape2, minipack. lm and ggplot2 packages used in this program. Python is an ideal tool to build an automated workflow to analyse data with the subprocess module, especially useful when the program involves many different languages. Therefore, I use shell commands in Python script to run R scripts and compile LATEX to create the report, which guarantees this program is fully reproducible.

R: non-linear model fitting,

Python: subprocess the Rscript

Git: upload the scripts

///////////////////

2.2 **data set**

285 data sets were collected from published works, study microbes and phytoplankton, contain populational growth data.

///////////////////////

**2.2 Models and Equations**

**2.2.1 models and the biological meaning of the parameters**

The linear cubic model, mechanistic logistic model and Gompertz model [Zwietering et. al., 1990] are chosen in this project to evaluate the data set respectively. The models split the curves of bacterial population growth into three phases: the lag phase at the beginning and the stationary phase at the end, those two phases have the specific growth rate of zero, and the exponential phase where the bacterial population increases nearly linearly over time, which corresponded to the parameter: r\_max in model equations. In reality, however, the population doesn’t growth promptly after getting into the new environment. So, to capture the lag phase, the more complicated growth model: Gompertz model[1] was introduced in this project which is asymmetrical compared with the logistic model and has one more parameter t\_lag to represent the last time point before the population exponentially growth. The other two parameters used in this model fitting is N\_0 and N\_max, respectively represents the logarithm of minimum and maximum population size in data.

**2.2.2 equations**

1. polynomial cubic model
2. logistic model
3. Gompertz model

//////////////////////////////

**2.3 Model fitting**

**2.3.1 data standardization (data.R)**

1. delete the population size of negative numbers which do not have any biological meaning.

2. The population size in each data set was log-transformed for better analysing when the population growth is still in the lag phase with too small size and comparing the Gompertz model, which models the log-transformed variables, with other two models when calculating the AICc, AIC and BIC value of each model in the identical standard.

2.3.2 **models fitting on experimental data sets** (Gompertz.R and Logistic.R)

Two parallel R scripts, Gompertz.R and Logistic.R. were written containing the functions to operate the model fitting, and through criteria comparison return needed data.

1. define the model\_logistic and model\_gompertz functions to return the equations of logistic and Gompertz models.
2. parameter estimation: the methods of estimating the preliminary parameters as starting value to fit the models will be represented in table1.
3. Sample the starting value 1000 times around the preliminary starting values with the factor of 0.2 follow the normal distribution.
4. In GuessStart function: fitting the models using the nlsLM() function from minpack.lm package in R. It is possible to fail when fitting the non-linear model with too far-reach starting value, so within the function, using the tryCatch function in base R to return the error and avoid stop the whole process.
5. Repeat the GuessStart function fitting 1000 different starting values sampled in step 3 for each model.
6. Comparing the 1000 results and return the best successful-fitting result with lowest AICc, which contains the evaluation data: AICc, AIC, BIC and R^2, starting values of parameters used and the plot point predicted by the predict() function in base R. The AIC and BIC values are calculated by the in-built function in R. To calculate the AICc, the equation: XXX was used. Besides, the R square calculation equation is XXX.

**2.3.3 compare models and visualization (Compare\_Models\_and\_plot.R)**

1. read the evaluation data and predict plot points got from step 2.3.2.

2. using the lm() function to fit the cubic linear model to the data sets and get the same evaluation data as logistic and Gompertz model.

3. write the evaluation data of three models in the data frame and using them to compare the goodness-of-fit of each model.

4. plot the actual data points overlap with predicted lines to visualize the effects of the fitting.

////////////////////

2.4 model selection

To compare the goodness-of-fit of the models, rather than the conventional null hypothesis testing approach, The AICc, AIC and BIC were adopted as model selection criteria. In which several competing hypotheses are simultaneously confronted and among model selection criteria, AIC and AICc are generally favoured because it has its foundation in Kullback–Leibler information theory [AIC site(3)].

**Results**

**Generally,** all the models were fitted without major problems and all the data sets were fitted successfully by cubic and logistic models, 277 by Gompertz model.

The typical fitted model visualizations look like (figure.1). The plot of the fitted curves shows that all three models fitted well, the big difference between the Gompertz model and the other 2 models are the ability of catch the lag phrase, which follows the …. No big difference from visualization comparison, so that more precise inference needs to be made by comparing the quantitative results, which in this project are comparison criteria: AICc, AIC and BIC.

By comparing the AICc, AIC and BIC value of each model in all fitted 277 data sets, the winning frequencies of Gomeperts model are: 144, 203 and 199, of the logistic model are: 114, 37 and 42, of the cubic model are: 19, 37, 36. By all means, the Gompertz model gave the best fit to the data (figure.2).

By analysing and evaluating the goodness-of-fit on population growth data of the chosen models

Exam the residuals?

Parameters’ confidential interval?

(plot histogram)

**Discussion** (*reminds the reader about what the original goals of the study were, states out key findings succinctly, and then discuss their implications in the wider context and then finished off with some caveats and a conclusion that delivers the final takeaway messages.*)

The logistic model has one fewer parameter than do the cubic and Gompertz equation and can be fitted to most of the data sets used in this study.

Generally, although the Gompertz model is the best among the three models chosen int this project, the cubic model is good enough to roughly predict the population growth kinetics, which can be used as a quick way of representing the good enough curve of the population growth without much effort.

Cause after logarithm the population data, at the lag phase, the logistic model always deviate from the beginning of the curve which is the period the bacteria is getting used to the environment and preparing for the exponential growth, I chose the Gompertz mod[cite]

The data used in this project is from various published experiments with different cultures, as we know the environmental conditions such as temperature(cite), pH(cite),… (cite) is key factors could influence the growth of the bacteria, which was not considered in this article. This project is mainly analysing the goodness of the models by fitting in randomly data chosen from bacteria area publishes.

Visually looking at the line representing the predicted values based on the fitted models and the actual points,

As we’ve already known the population growth follows the sigmoidal curve regular[site], some of the badly fitted models may have non-reasonable data. Doesn’t have any stable relationship or follow any reasonable distributions. After fitting all the model and check the data of those bad results of choosing comparison criteria, I discard 8 data set which does not show any regular among the points. By comparing the reasonable results we can see that…

identifiability properties[estimating parameters Baranyi model 1999]

the Gompertz model is consistently favoured.

The Gompertz model gave the overall best fit to the data

A measure of ‘goodness-of-fit’

The logistic model has one fewer parameter than do the other two models used in this study and can be fitted to all of the data sets in this study. If we do not concern too much about the lag time in analysis, the logistic model is a comparative precise and easier to fit model.

Another aspect of consideration is

From the equation of calculating the the AICc based on AIC, we can see that if there are more parameters in model equations, theoretically you will get higher AICc which means the fitting effects are worse. While the Gompertz model has one more parameter, some data sets give back higher AICc than Logistic model, Gomperts model is still the win more times model.

\*\*\* Successful nonlinear least-squares fitting depends on finding good initial estimates of parameters, which may occasionally be difficult.

\*\*\*The extensively used Gompertz model may be satisfactory to fit growth curves, but in contrast with previous results (Gibson et al., 1987; Zwietering et al., 1990), other models are shown to be more accurate for fitting microbial growth curves (Baranyi et al., 1993; Dalgaard, 1995; Dalgaard et al., 1994; Membre´ et al., 1999; Schepers et al., 2000).

# Reference\

[1] M. H. Zwietering, I. Jongenburger, F. M. Rombouts, and K. van ’t Riet, ‘Modeling of the Bacterial Growth Curve’, *Appl. Environ. Microbiol.*, vol. 56, no. 6, pp. 1875–1881, 1990, DOI: 10.1128/AEM.56.6.1875-1881.1990.

[2] M. Peleg and M. G. Corradini, ‘Microbial Growth Curves: What the Models Tell Us and What They Cannot’, *Crit. Rev. Food Sci. Nutr.*, vol. 51, no. 10, pp. 917–945, Dec. 2011, DOI: 10.1080/10408398.2011.570463.

[3] M. N. Gibbs and D. MacKay, ‘Variational Gaussian process classifiers’, *IEEE Trans. Neural Netw.*, vol. 11 6, pp. 1458–64, 2000.

\* checklist

1. check spell

2. use vector graphic

3. take-home messages: tables, figures, legends, text caption.

4. Avoid sub-sectioning (with headers) the Introduction and Discussion sections as it breaks the flow of your “narrative”. But subset the methods and results

6. README files

7. abstract just linear and non linear model

8. no cite in abstract

9. compile latex shell script

10. Figure ledgend species which data set

https://www.newgenapps.com/blog/6-reasons-why-choose-r-programming-for-data-science-projects/s