Model Comparison of Population Growth Curves

Xiaoya Zhang/CID:01830605

Department of Life Science

Imperial College London

Word Count: 2026

Abstract

A population grows exponentially when its abundance is low and its resources unconstrained (Malthusian principle). However, this growth then slows and eventually stops as resources become limiting. There may also be a time lag before the population growth really takes off at the start. When resources become limited, this growth slows and eventually stops. The contribution of bacterial-feeding nematodes to litter decomposition and nutrient mineralization depends, in part, on the abundance of particular nematode species. Population dynamics will be constrained by edaphic factors, food availability and food quality (Vsenette, RC and Ferris, H,1998). In this project, one mathematical model best fits an empirical dataset should be chosen from three mathematical models. Several models (logistic, Gompertz, Baranyi, Buchunana) were compared by assessing the fits of data. After comparison, the third dataset population growth was selected to finish the project analysis. After data processing and model fitting, a proper mathematical model has been selected to best fits the empirical dataset.

1 Keywords

Data preparation, Model selection, population growth, model Comparison, modified Gompertz model, Baranyi model, Buchanan model, classical model

2 Introduction

Fluctuations in abundance (density) of individual populations play a key role in ecosystem dynamics and emergent functional characteristics such as carbon fixation or disease transmission rates. If the surface-area-to-volume ratio of such an 'expanding system' remains unchanged logarithmic growth can continue indefinitely, and in a constant environment the system enters a time-dependent 'exponential state'. Autocatalysis is not involved in the logarithmic growth of an expanding system; but when an autocatalytic stage is included the growth curve can exhibit a typical log-phase during which growth rate is virtually independent of concentrations of source material above a threshold level (Perret, CJ,1960). When resources become limited, this growth slows and eventually stops. The contribution of bacterial-feeding nematodes to litter decomposition and nutrient mineralization depends, in part, on the abundance of particular nematode species. Population dynamics will be constrained by edaphic factors, food availability and food quality. (Venette, RC and Ferris, H,1998). Data contains measurements of change in biomass or number of cells of microbes over time have been given.

The two main fields, Pop Bio and Time, should be focused on. The project includes four parts: Data preparation, NLLS fitting including obtaining starting values, plotting and analysis.

3 Models

There popular models are the modified Gompertz model (Zwietering et. al., 1990), the Baranyi model (Baranyi, 1993), and the Buchanan model (or three-phase logistic model; Buchanan, 1997).

1. The Buchanan model in particular is capable of capturing the lag phase before the population starts rowing exponentially, often seen in microbial population growth.

2. The modified Gompertz model has been used frequently in the literature to model bacterial growth. The Gompertz equation is capable of fitting survival curves which are linear, those which display an initial lag region followed by a linear region, and those which are sigmoidal (Linton, RH and Carter, 1995). The Baranyi model was used to fit the four commonly observed survival curves: linear curves, those with a lag phase, those with a tailing phase and sigmoidal curves. It was validated by using published experimental data for thermal inactivation of Listeria monocytogenes Scott A heated in infant formula and compared with the modified Gompertz equation. For the prediction performance, the Baranyi model was better and more robust than the modified Gompertz equation (Xiong, R and Xie, 1999).

$$N_t = Ae^{-e^{\frac{r_{max}e(t_{lag}-t)}{A}+1}} A = ln(\frac{N_{max}}{N_0}) \tag{1}$$

This equation was first proposed by Zwietering et. al.in 1990

3. The Baranyi model introduces a new dimensionless parameter that represents the initial physiological state of the cell. The length of the lag phase is determined by the value at the time of inoculation and the environment after inoculation. Thus, the definition of lag is independent of the shape of the growth curve, and the effects of the previous environment are separate from the effects of the present environment. This makes it possible to simulate growth without a lag period after inoculating from growth-friendly media to growth-friendly new media. One of the attractive points of the Baranyi model, besides its good predictive capabilities, is the fact that

it is a truly dynamic model in the sense that it can deal with time varying environmental conditions (Grijspeerdt, Koen and Vanrolleghem, Peter, 1999).

$$N_t = N_0 + r_{max} A_t - ln(1 + \frac{e^{r_{max}} A_t - 1}{e^{N_{max} - N_0}}) A_t = t + \frac{1}{r_{max}} \cdot ln(\frac{e^{-r_{max} t} + h_0}{1 + h_0}) t_{lag} = \frac{ln(1 + \frac{1}{h_0})}{r_{max}} \quad (2)$$

4 Methods

4.1 Data

1. The data preparation script processes the data for later use. Data was prepared for fitting using python and it is easier to view data. Firstly, one column was inserted named 'ID', which gives each row of data a name, making it more convenient to distinguish and subset data. Secondly, the data was subset by some essential columns 'data. Species + data. Temp. map(str) + data. Medium + data. Citation'. All records with value of NA were excluded. Getting and Outputting all unique IDs using unique method is next step. Then, all unique IDs are transferred into dataFrame form and size of ID was achieved which is 284. After that, the subset run through a for loop to check if every single ID has more than 8 records. Each ID's records size is calculated and 8 is the proper number I chose to fit with the moat parameters. Finally, each ID was plotted and kept to a .csv file for later use called 'data.csv'.

2. The NLLS fitting script obtained starting values. First, csv file was read. Data and ID list then be transferred to dataframe form for later use. Second, get data of the first ID and output data size of the first ID. Then all time parameters of the first ID were achieved as x, and put in order from small to large. All popbio parameters of the first ID were achieved as y, and sorted in order from small to large. Sort index methods were used for sorting time and popbio. After that, two lists were built for storing new slope and point, which are stapeNew and pointList. Then the first slop in ID1 was calculated by two points, the first one and the second one. The slope was calculated by using the difference between the vertical coordinates of two points divided by the difference between the horizontal coordinates of two points. N0 was achieved by taking the vertical coordinate of first point. Nx was achieved by taking the vertical coordinate of the last point which is the largest one. A for loop is used to calculate the slope at each point to

get the maximum one. By assuming that the maximum slope named stape, compare the slope of each point one by one and replace stape with the larger one. In this way, the largest slope of ID 1 was got. Then, in the same way using a for loop on outermost layer of the code, all largest slopes of all IDs were achieved. After that, Tlag, N0, Nmax, Rmax, A values of all IDs can be obtained easily. Finally, all data were put in a list and stored as a csv file in my data folder.

4.2 Computing tools

1. Python 3.6.7 was used for data processing and preparation. In the data exploration and preparation part, pandas package was used for all data frame manipulations. The data were transferred to data frame form once they were got, which means all data processing is finished in the form of dataframe, which is more familiar to me. Matplotlib.pylab package was used for plotting.

2. Python 3.6.7 was also used for the NLLS fitting script part. Starting values were obtained in this script. Pandas package was used to read .csv files as well as for data frame manipulations. Math package was applied for mathematic calculating. A1 was obtained by calculating log(Nmax) divided by log(N0).

3. R .RStudio is used in final plotting and model fitting script. The reason is that R can finish plotting. Data list was created to a .csv file in NLLS fitting script. The. csv file was read and change into dataframe form. Four model functions were built, which are gompertz model, baranyi model, buchanan model and classic model.

4. Bash is used in MiniProject.sh script for running Latex, preparation, NLLS fitting and R script. MiniProject.sh is A single script which runs the whole project, right down to compilation of the LaTeX document.

5 Results

Data preparation: The data preparation script was completely finished. Data was prepared for fitting. Data subsetting by ID was finished which is more convenient for data processing. All records with value of NA were excluded. The valid data generates a .csv file classified by their IDs named 'data.csv', which was stored in DATA directory.

NLLS fitting: Starting values of each ID were successfully obtained in this script. Starting values including ID, Tlag, N0, Nmax, Rmax and A have been put in one csv file named 'LIST.csv'. This file was also stored in DATA directory.

Model fitting: Model fitting script was not completely finished. Four model functions were specified. Dataset was fitted to four model functions. They are gompertz model, baranyi model, buchanan model and classic model.

6 Discussion

As expected, three models have been analyzed. Although there is no final result in this project, there are still some areas that can be developed and expanded in the future.

In terms of model selection, in addition to the above three models, there are other mathematical models that can be used to analyze data. The first order kinetic model, the Buchanan model and Cerf's model, can model a linear survival curve, a survival curve with a shoulder and a survival curve with a tailing, respectively. However, they are not suitable for fitting a sigmoidal survival curve. The three models were integrated into a new model that was capable of fitting the four most commonly observed survival curves: linear curves, curves with a shoulder, curves with a tailing (biphasic curves) and sigmoidal curves (Xiong, R and Xie, 1999). This gives us a new idea, the integration model, which means that several different models can be integrated together.

As to the model selection, there are some methods to choose a good one. Given a data set, you can fit thousands of models at the push of a button, but how do you choose the best? With so many candidate models, overfitting is a real danger. Is the monkey who typed Hamlet actually a good writer? Choosing a model is central to all statistical work with data. Real-data examples are complemented by derivations providing deeper insight into the methodology, and instructive exercises build familiarity with the methods (Claeskens, Gerda and Hjort, 2008).

With regard to the data, the results should not be accurate due to there is not enough data has been collected. If more data were collected and an-

alyzed, the results would be more accurate.

In the future works, this project should be completed successfully. All the points mentioned above should be improved. This will increase the accuracy of data model analyze.

References