Modeling microbial growth – which is better and why?

PokMan HO

Department of Life Sciences, Faculty of Natural Sciences,

Imperial College London

Approximate Word Count: 1094

Modeling microbial growth – which is better and why?

PokMan HO (CID: 01786076)

Abstract

Introduction

Phenological models are expected to fit data trends within its biological field. Yet due to different reasons, models developed and published from one sample may not fit the others. These reasons may be due to data variabilities, confounding factors, inaccurate assumptions or models being too-specific. This project is aimed at compare and contrast published phenological models on microbial population size data, highlighting which is a better model under what conditions. The hypotheses are:

- published phenological models are better than polynomials in describing microbial population size;
- appropriate phenological model(s) can be identified through distinguishable shapes of microbial population size; and
- parameters of data under each phenological model is clustered, similar with dataset best-described by the same model but different from those described by other models.

Methods

Experimental microbial population growth data library were divided into individual data subsets through six filters ("Temperature (in ^{o}C)", "Microbial clade", "growth substrate materials", "experimental replicate number", "population data recording unit" and "data source"). Records with data unit "OD_595" were scaled into optical density percentages (i.e. data*100) to facilitate general analyses workflow. Independent (or explanatory) variable was "Time (hr)" and dependent (or response) variable was "population size".

Some raw data were recorded in minutes (instead of hour). This record artifact was not corrected because of two reasons: 1. shape of curves were the main concern instead of independent variable's scale; and 2. the unit was consistent within each data subset.

Model assessment

Six candidate models were assessed, four phenological and two polynomial equations. They were "Verhulst (classical)"[1], "modified Gompertz"[2], "Baranyi"[3], "Buchanan"[4], "quadratic" and "cubic". NLLS was used only on the four phenological models and linear model-fitting was done on the two polynomials. Starting values selection (for phenological models only) was described below:

Initial (N0) and final (K) population sizes were selected to be the minimum and maximum values of each data subset respectively. Maximum growth rate (r.max) was selected by linear model through a recursive manner. For every iteration, population size data from the top 5% independent variable values were excluded from the linear model calculation. The data and slope would only be recorded if it was positive, higher adjusted R^2 value and larger slope than the recorded "best slope" value. After scanning from the maximum side, the best slope and its respective data were taken out and screened from the minimum side. Final best slope and x-intercept were regarded as the r.max and relative time lag (t.lag) of the population (in the source experiment) respectively. Time which this linear model intersected with K was regarded as the time achieving carrying capacity (t.K). Population data was then classified into three groups (gx) according to the time: g1 $\leq$ t.lag $<$ g2 $<$ t.K $\leq$ g3. 5% was chosen as the scanning threshold because I assumed this resolution was fine enough for achieving good starting values for NLLS fitting. Inputs for phenological modelswere listed below (popn & time were the dependent and independent variables respectively):

$$\begin{aligned}
\text{Verhulst (classical):} \quad & \text{popn} = f(\text{N0, K, r.max, time}) \\
\text{modified Gompertz:} \quad & \text{popn} = f(\text{N0, K, r.max, time, t.lag}) \\
\text{Baranyi:} \quad & \text{popn} = f(\text{N0, K, r.max, time, t.lag}) \\
\text{Buchanan:} \quad & \text{popn} = f(\text{N0, K, r.max, time, t.lag, gx})
\end{aligned}$$

All test starting values were than sampled from normal distribution with mean as the estimated value and standard deviation (sd) of 1. The sd value was chosen because of different reasons for each parameters. N0 and K were directly extracted from the raw experimental data, which could be assumed being an accurate estimate for that data subset (hence a small sd was logical). r.max was a guesstimated value from fitting linear models. This process could potentially be affected by extreme values in the data and hence a large sd should be preferred. 100 trials were done as a optimal value under a trade-off between efficiency and accuracy.

Only AIC[5–7] was used to select for optimal parameter values within each phenological model and best model between the six candidates for a data subset. Reasons would be listed in Discussion section. For models with more than one parameter sets as sharing the lowest AIC value, the first set of values from the random sampling trials were used for downstream analyses. AIC tolerance threshold was expanded to min(AIC)+2[8] to incorporate more accepted models for analyses.

Statistical analysis

Kruskal test was used for identify the best-fit model among all included model because the count was categorical and not assumed being normally-distributed. Pairwise Nemenyi comparisons would be carried out to identify the best test if p-value of the above test was significant.

Using principal component analysis (PCA), parameter weights could be observed across phenological models. Phenological models would be positively-correlating with a parameter if dataset observations were concentrated towards the positive side of the factor and vice versa. It was expected that datasets would cluster together (or being on similar positions) if parameter(s) were representing the observed data.

Main Assumptions

- there was no negative population growth (i.e. starting population was always lower than carrying capacity), so negative population growth data were set to zeros;
- estimated parameter estimates would always result in a global optimal status in parameter space through the non-linear least squares method (NLLS)

Computing tools

R (ver 3.6.0)[9] was used with “minpack.lm”[10] for computing non-linear least square statistics for model comparisons. “PMCMR”[11] was used for carrying out Kruskal test and “stats”[9] was used for PCA analysis.

Results

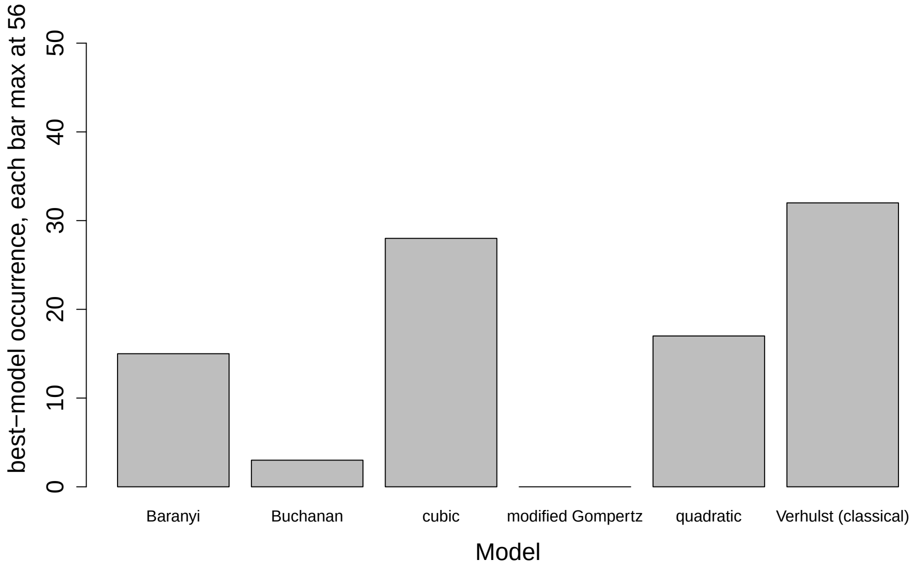

Figure 1: Barplot showing the number of "best model" identification under AIC model-selection methods with " Kruskal-Wallis rank sum test " statistic $X^2 = 5$, df = 5 , p = 0.42

From Fig.1, large fluctuations between each model to be described as "best-fit" were observed. However the occurrence difference was not statistical significant. Among the counts, there were 39 datasets with more than one "best-fit" models. Verhulst (classical) and cubic were the top two models selected as "best-fit" for the 56 datasets (32 for Verhulst (classical) and 28 for cubic). There are 10 datasets calling both "best-fit" at the same trial. Between Baranyi and quadratic, the counts were 15 and 17 respectively with 6 datasets calling both models "best-fit". The only outstanding performance was from modified Gompertz, which 0 datasets were called it as "best-fit".

In Fig.2, principal component 1 (PC1) was capturing 50 % variability. It was composed approximately by 0.58 N0, 0.6 K, -0.16 r.max and 0.53 t.lag. PC2 was capturing 24 % variability. It was composed approximately by -0.14 N0, -0.06 K, -0.99 r.max and -0.09 t.lag.

From Fig.2, Verhulst (classical) was having no observable patterns. Since modified Gompertz was not a "best-fit" candidate from any dataset, it could not be shown on the PCA analysis. Baranyi was generally observed being positively-correlated with r.max but negatively relating

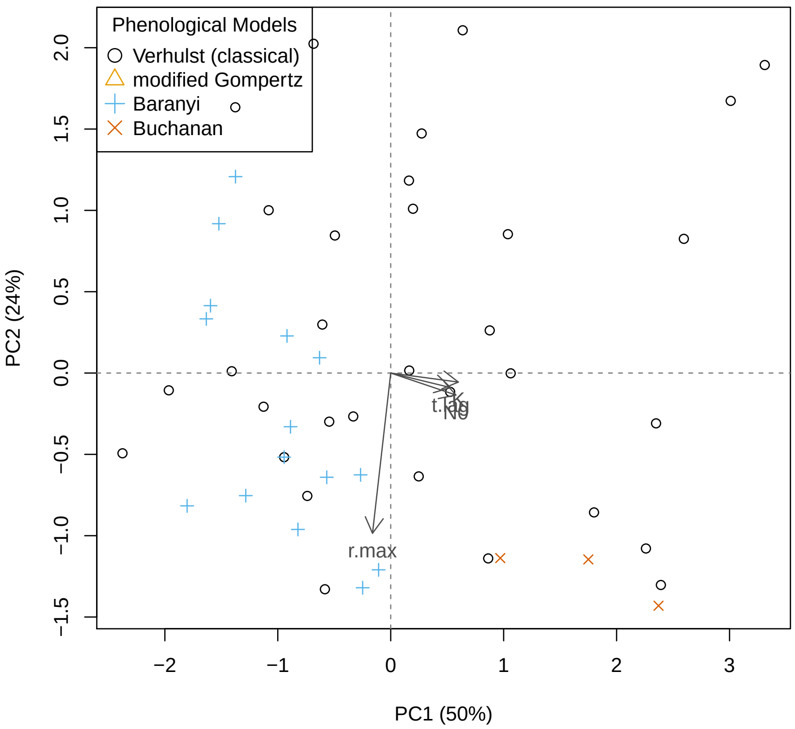

Figure 2: Biplot of Principal Component Analysis (PCA) comparing phenological models using estimated parameter values with "minimal AIC +2" evaluations.

to N0, K and t.lag. Buchanan was observed being positively correlated with all four parameters.

Discussion

AIC is considered the most suitable model-selection approach within model and between models in this project. Unlike BIC, AIC are more accurate with small sample sizes[12,13] and sparse data[13]. AIC did not assume a "true model" was under examination[14–16]. Since candidate models were not "nested model", BIC is not a better choice than AIC[17]. Hence the use of only AIC as model-selection criterion should be justified.

k

Conclusion

Published phenological models were data-specific, which none of them were found significantly performing better than the others in general. Parameters defined by these phenological models were appropriate, which none of them were having observable domination nor negligible weights on the function calculations.

Code and Data Availability

All scripts and data used for this report were publicity available at GitHub.

References

1. McKendrick, A. & Pai, M. K. XLV.—the rate of multiplication of micro-organisms: a mathematical study. *Proceedings of the Royal Society of Edinburgh* **31,** 649–653 (1912).

2. Gil, M. M., Brandão, T. R. & Silva, C. L. A modified Gompertz model to predict microbial inactivation under time-varying temperature conditions. *Journal of Food Engineering* **76.** Bugdeath, 89 –94. ISSN: 0260-8774. `http://www.sciencedirect.com/science/article/pii/S0260877405003389` (2006).

3. Baranyi, J, McClure, P., Sutherland, J. & Roberts, T. Modeling bacterial growth responses. *Journal of industrial microbiology* **12,** 190–194 (1993).

4. Buchanan, R., Golden, M. & Whiting, R. Differentiation of the effects of pH and lactic or acetic acid concentration on the kinetics of Listeria monocytogenes inactivation. *Journal of Food Protection* **56,** 474–478 (1993).

5. Johnson, J. B. & Omland, K. S. Model selection in ecology and evolution. *Trends in ecology & evolution* **19,** 101–108 (2004).

6. Akaike, H. in *Selected papers of hirotugu akaike* 199–213 (Springer, 1998).

7. Burnham, K. & Anderson, D. Model selection and multimodel inference: a practical information-theoretic approach. *Ecological Modelling.*

8. Burnham, K. P. & Anderson, D. R. Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research* **33,** 261–304 (2004).

9. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2019). `https://www.R-project.org/`.

10. Elzhov, T. V., Mullen, K. M., Spiess, A.-N. & Bolker, B. *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds* R package version 1.2-1 (2016). `https://CRAN.R-project.org/package=minpack.lm`.

11. Pohlert, T. *The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR)* R package (2014). `https://CRAN.R-project.org/package=PMCMR`.

12. Acquah, H. D.-G. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics* **2,** 001–006 (2010).

13. Kuha, J. AIC and BIC: Comparisons of assumptions and performance. *Sociological methods & research* **33,** 188–229 (2004).

14. Aho, K., Derryberry, D. & Peterson, T. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* **95,** 631–636 (2014).

15. Vrieze, S. I. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods* **17,** 228 (2012).

16. Yang, Y. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92,** 937–950 (2005).

17. Wang, Y. & Liu, Q. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock–recruitment relationships. *Fisheries Research* **77,** 220–225 (2006).

18. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6,** 461–464 (1978).

19. Kelley, C. T. *Iterative methods for optimization* (SIAM, 1999).

20. Turchin, P. *Complex population dynamics: a theoretical/empirical synthesis* (Princeton university press, 2003).