

Imperial College London

**An Information-Theoretic Approach to Model Selection**

Department of Life Sciences

Hovig Artinian

MSc CMEE

March 6, 2020

Word count: 1653

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Materials &amp; Methods</b>	<b>3</b>
2.1	Data Preparation . . . . .	3
2.2	Data Analysis . . . . .	4
2.2.1	Model Fitting . . . . .	4
2.2.2	Model selection . . . . .	4
2.3	Computing Tools . . . . .	6
<b>3</b>	<b>Results</b>	<b>7</b>
3.1	Model Fitting . . . . .	7
3.2	Model Selection . . . . .	7
3.2.1	Case 1 . . . . .	7
<b>4</b>	<b>Discussion</b>	<b>7</b>
<b>5</b>	<b>Conclusion &amp; Future Work</b>	<b>10</b>
<b>6</b>	<b>Acknowledgements</b>	<b>10</b>

# 1 Introduction

Modelling nature has been the general interest in the field of ecology for over a century [1]. A crucial part of this process involves model selection. The basic approach is the null hypothesis testing, where biological inferences are made based on whether or not a suggested hypothesis is rejected [2]. This would be based on arbitrary criteria such as p-values, confidence intervals, and t-tests, that are set by statisticians as general rules of thumb. Amidst this mayhem, however, some scientists have started to shift their workflow towards more robust approaches that rely on advanced mathematical theories.

In this project, the information-theoretic approach is used for model selection. This methodology is based on information theory [3].

Preceding the model selection phase, however, is finding parameter estimates of models by fitting them to data. 7 models are used in this study: linear, quadratic, cubic, logistic [4], gompertz [5], baranyi [6], and buchanan [7]. The first three are phenomenological, which means their parameters have no biological significance. For the mechanistic models, however, 4 parameters are involved and are described in Table 1 below.

Parameter	Description	Models included
$N_{max}$	carrying capacity	logistic, gompertz, baranyi, buchanan
$N_0$	initial abundance value	logistic, gompertz, baranyi, buchanan
$r_{max}$	growth rate	logistic, gompertz, baranyi, buchanan
$t_{lag}$	time taken for lag phase	gompertz, baranyi, buchanan

Table 1: Parameters involved in the mechanistic models used in this study with their corresponding descriptions.

19 The Non-Linear Least Squares (NLLS) technique was performed for  
20 model fitting, which allows models to detect non-linear patterns found in  
21 data. Once parameter estimates are found, model selection is performed by  
22 calculating several information-theoretic criteria such as AIC,  $AIC_c$ , BIC,  
23  $R^2$ , AIC differences, likelihood of models, Akaike weights, and evidence ra-  
24 tios.

25 The main objective of this project is to show how sometimes finding a  
26 "best" model is not always ideal; rather, a multimodel inference approach  
27 would be optimal instead.

## 28 **2 Materials & Methods**

### 29 **2.1 Data Preparation**

30 The starting dataset consisted of 4387 samples. No missing abundance  
31 values were detected. However, negative values were present. The smallest  
32 value was largely negative and, hence, removed. To deal with the rest,  
33 while still minimizing the amount of data points lost, the smallest value was  
34 added to the whole data, and then removed (to avoid having zero as an  
35 abundance value). The end result was a dataset with 4385 values. Next,  
36 each species/temperature/medium/citation/replicate was grouped together,  
37 resulting in 305 unique IDs. Finally, the new dataset was saved to be used  
38 for data analysis.

## 39 2.2 Data Analysis

### 40 2.2.1 Model Fitting

41 Non-linear least squares (NLLS) fitting was used to fit all 7 models to  
42 each unique group in the new dataset.

43 To work with this method, starting parameter values must first be pro-  
44 vided. The better the starting values, the more precise the estimated pa-  
45 rameter values will be.

46 For phenomenological models, finding the starting values was straight-  
47 forward (they were set to 1). In the case of mechanistic models, on the other  
48 hand, more computation was needed. The starting values of  $N_{max}$  and  $N_0$   
49 were set to be the highest and lowest abundance values in the dataset, re-  
50 spectively. That of  $r_{max}$  was less direct. A straight-line was fit to the first  
51 50% of the dataset, and its slope was assigned as the starting value of  $r_{max}$ .  
52 Lastly, the intersection point between the fitted tangent line and the hori-  
53 zontal line at  $y = N_0$  was set to be the starting value of  $t_{lag}$ .

54 Next, the actual fitting was performed, where residuals for each models to  
55 be fit were provided using the newly found starting values. For each model,  
56 if the fit converged, the estimated parameters were saved in a variable;  
57 otherwise, the estimated parameter values were set to 0.

### 58 2.2.2 Model selection

59 For model selection, first, Akaike's Information Criterion (AIC) [8] was  
60 calculated for each model. The AIC value gives the quality of each model  
61 relative to the other models in the set used for fitting the data. Hence,

the model with the lowest AIC score is preferred. For models where the sample size (n) to number of parameters (K) ratio was less than 40 (arbitrary suggestion),  $AIC_c$  was calculated instead. It is known to be the second-order variant of AIC derived by Sugiura in 1978 [9]. The difference is in the bias-correction term added to the AIC value of the model. This makes sure that when sample size and number of parameters are close, the model gets heavily penalized.

Next,  $AIC^1$  differences ( $\Delta_i$ ) were calculated to get a better idea of the empirical support provided by each model in the set; thus, obtaining a relative ranking of all the models in the set. The formula is given as follows:

$$\Delta_i = AIC_i - AIC_{min} \quad (1)$$

where  $AIC_i$  is the AIC value of the  $i$ th model and  $AIC_{min}$  is the AIC value of the "best" model in the set. The larger the difference, the less likely it is for that model to be the best. This likelihood was quantified and calculated by the following formula:

$$\mathcal{L}(g_i|x) \propto e^{-\frac{1}{2}\Delta_i} \quad (2)$$

Contrary to  $\Delta_i$ , the larger the relative likelihood of a model, the better chances it has to be the preferred model. For better interpretation, the likelihood of the models were normalized, adding to 1. This normalized

---

<sup>1</sup>To avoid repetition, the term "AIC" will be used to mean both AIC and  $AIC_c$ . There is, however, a distinction between them, as explained, and the reader should bear that in mind.

79 form is known as the Akaike weight,  $w_i$  of a model:

$$w_i = \frac{e^{-\frac{1}{2}\Delta_i}}{\sum_{r=1}^R e^{-\frac{1}{2}\Delta_r}} \quad (3)$$

80 where  $R$  is the number of models.

81 Lastly, evidence ratios were calculated to show how reliable it is to believe  
82 that the best model is actually a good fit for the data. To put it in another  
83 way, evidence ratios reveal whether or not the best model in a set of models  
84 works best alone, or within a group consisting of one or more of the other  
85 models in the set. It can be found by simply calculating the ratio of Akaike  
86 weights:

$$Evidence\ ratio = \frac{w_1}{w_j} \quad (4)$$

87 where  $w_1$  is the Akaike weight of the model with the lowest AIC value, and  
88  $w_j$  is the Akaike weight of the  $j$ th model.

## 89 2.3 Computing Tools

90 Several programming languages were used to create the different aspects  
91 of this project.

92 **R** was used for: (1) data exploration/preparation - playing around with  
93 data in R is fast and intuitive, (2) plotting - really good packages that  
94 produce nice-looking plots, (3) finding information-theoretic criteria - since  
95 calculating statistical measures is fast and straightforward in R. Packages  
96 used: *dplyr* [10] for data manipulation and *ggplot2* [11] for plotting. **Python**  
97 was used to perform heavy computation, especially model fitting since the

98 package for NLLS fitting in Python is much more robust than the one in R.  
99 Packages used: *Pandas* [12] for using dataframes, *NumPy* [13] for scientific  
100 and numeric computing, and *LMFIT* [14] for NLLS fitting. **L<sup>A</sup>T<sub>E</sub>X** was used  
101 for typesetting. **Bash** was used to glue the R and Python scripts together.  
102 **Git** was used for version control of all codes/scripts/workflow.

## 103 3 Results

### 104 3.1 Model Fitting

105 Model fitting convergence success rate was 100% for 5 out of the 7 models  
106 used for NLLS fitting. The baranyi model converged 80.98% of the time  
107 when fitted to all 305 groups, while the the logistic model did not converge  
108 with over half the total number of groups in the dataset (44.92%)

109 Figure below gives a general idea of what some of the data look like and  
110 how the models were able to be overlaid.

111 However, visualizing the plots is not a reliable way of determining good  
112 fits. Looking at the model selection criteria can be more helpful in this case,  
113 which is shown in the next section.

### 114 3.2 Model Selection

#### 115 3.2.1 Case 1

## 116 4 Discussion

117 Although Akaike's Information Criterion is recognized as a major mea-  
118 sure for selecting models, it has one major drawback: The AIC values lack



119 intuitivity (easy to use and understand) despite higher values meaning less  
120 goodness-of-fit. For this purpose, Akaike weights come to hand for calculat-  
121 ing the weights in a regime of several models. Additional measures can be  
122 derived, such as (AIC) and relative likelihoods that demonstrate the prob-  
123 ability of one model being in favor over the other.

124

125       **AIC:**

126       good thing: it accounts for overfitting - it has a penalty that increases as  
127       more parameters are added. Adding parameters will almost always improve  
128       the goodness of the fit.

129       An individual AIC value, by itself, is not interpretable due to the unknown  
130       constant (interval scale). AIC is only comparative, relative to other AIC  
131       values in the model set; thus such differences  $\Delta_i$  are very important and  
132       useful.

133       It is important to note here that AIC values do not give any information  
134       about the goodness-of-fit of a model to the data. Rather, they show how  
135       each model performs relative to the other ones in the set. Hence, a single  
136       AIC value of a model has no meaning.

137       **BOOK:**

138       **Ambivalence:**

139       The inability to ferret out a single best model is not a defect of AIC or  
140       any other selection criterion. Rather, it is an indication that the data are  
141       simply inadequate to reach such a strong inference. That is, the data are  
142       ambivalent concerning some effect or parametrization or structure.

143       In such cases, all the models in the set can be used to make robust in-

144 ferences: multimodel inference.

145

146 The AIC differences ( $\Delta_i$ ) and Akaike weights ( $w_i$ ) are important in rank-  
147 ing and scaling the hypotheses, represented by models. The evidence ratios  
148 (e.g.,  $w_i/w_j$ ) help sharpen the evidence for or against the various alternative  
149 hypotheses. All of these values are easy to compute and simple to under-  
150 stand and interpret.

151

152 The principle of parsimony provides a philosophical basis for model se-  
153 lection, K-L information provides an objective target based on deep theory,  
154 and AIC,  $AIC_c$ ,  $QAIC_c$ , and TIC provide estimators of relative, expected  
155 K-L information. Objective model selection is rigorously based on these  
156 principles. These methods are applicable across a very wide range of sci-  
157 entific hypotheses and statistical models. We recommend presentation of  
158  $\log(\mathcal{L}(\hat{\theta}))$ , K, the appropriate information criterion (AIC,  $AIC_c$ ,  $QAIC_c$ , or  
159 TIC),  $\Delta_i$ , and  $w_i$  for various models in research papers to provide full infor-  
160 mation concerning the evidence for each of the models.

161

162 **Do not mix null hypothesis testing with information-theoretic**  
163 **criteria:**

164 Some authors state that the best model (say  $g_1$ ) is *significantly* better than  
165 another model (say  $g_6$  based on a  $\Delta$  value of 4-7. Alternatively, sometimes  
166 one sees that model  $g_6$  is rejected relative to the best model. These state-  
167 ments are poor and misleading. It seems best not to associate the words  
168 significant or rejected with results under an information-theoretic paradigm.

169 Questions concerning the strength of evidence for the models in the set are  
170 best addressed using the evidence ratio (Section 2.10), as well as an anal-  
171 ysis of residuals, adjusted R<sup>2</sup>, and other model diagnostics or descriptive  
172 statistics.

## 173 5 Conclusion & Future Work

174 studying the death phase

## 175 6 Acknowledgements

176 I would like to extend my gratitude to my fellow colleagues Pablo Lechon  
177 and Sam Turner for engaging in insightful discussions with me about the  
178 miniproject, as well as pulling an all-nighter the night before the submission  
179 deadline.

## References

- [1] Sharon E. Kingsland. *Modeling Nature: Episodes in the History of Population Ecology*. University of Chicago Press, 1995.
- [2] Jerald B. Johnson and Kristian S. Omland. *Model selection in ecology and evolution*, 2004.
- [3] S. Guiasu. *Information theory with applications*. McGraw-Hill, New York, NY, 1977.

- [4] R. Pearl and L. J. Reed. On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation. *Proceedings of the National Academy of Sciences*, 1920.
- [5] M. H. Zwietering, I. Jongenburger, F. M. Rombouts, and K. Van't Riet. Modeling of the bacterial growth curve. *Applied and Environmental Microbiology*, 1990.
- [6] József Baranyi and Terry A. Roberts. A dynamic approach to predicting bacterial growth in food. *International Journal of Food Microbiology*, 1994.
- [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. *Food Microbiology*, 1997.
- [8] H Akaike. Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd international symposium on information theory. *Second International Symposium on Information Theory*, 1973.
- [9] Nariaki Sugiura. Further Analysis of the Data by Anaike' S Information Criterion and the Finite Corrections. *Communications in Statistics - Theory and Methods*, 1978.
- [10] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *Package 'dplyr'*. 2018.

- [11] Hadley Wickham. *ggplot2*. 2009.
- [12] Wes McKinney and PyData Development Team. Pandas - Powerful Python Data Analysis Toolkit. *Pandas - Powerful Python Data Analysis Toolkit*, 2015.
- [13] © Copyright 2017 NumPy developers. NumPy NumPy, 2017.
- [14] Matthew Newville, Antonino Ingargiola, Till Stensitzki, and Daniel B. Allen. LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python. *Zenodo*, 2014.