# A probabilistic model of cell size reduction in Pseudo-nitzschia delicatissima (Bacillariophyta)

Roland Schwarz, Matthias Wolf, Tobias Müller


www.elsevier.com/locate/yjtbi

# A probabilistic model of cell size reduction in Pseudo-nitzschia delicatissima (Bacillariophyta)

Roland Schwarz, Matthias Wolf, Tobias Müller

PII: S0022-5193(09)00054-X
DOI: doi:10.1016/j.jtbi.2009.02.002
Reference: YJTBI 5459

To appear in: *Journal of Theoretical Biology*

Received date: 27 October 2008
Revised date: 19 December 2008
Accepted date: 4 February 2009

Cite this article as: Roland Schwarz, Matthias Wolf and Tobias Müller, A probabilistic model of cell size reduction in Pseudo-nitzschia delicatissima (Bacillariophyta), *Journal of Theoretical Biology* (2009), doi:10.1016/j.jtbi.2009.02.002

# A probabilistic model of cell size reduction in *Pseudo-nitzschia delicatissima* (Bacillariophyta)

Roland Schwarz[a,b,*], Matthias Wolf[a], Tobias Müller[a]

[a] *Department of Bioinformatics, University of Würzburg, Am Hubland, 97074 Würzburg*
[b] *Institute of Hygiene and Microbiology, University of Würzburg, Josef-Schneider-Straße 2 / E1, 97080 Würzburg*

**Abstract**

The pennate planktonic diatom *Pseudo-nitzschia delicatissima* is very common in temperate marine waters and often responsible for blooms. Due to its surrounding rigid silicate frustrule the diatom undergoes successive size reduction as its vegetative reproduction cycle proceeds. Since a long time the life cycle of diatoms has raised scientific interest and some years ago extensive samples of *Pseudo-nitzschia* have been taken from coastal waters. Mating and cell size reduction experiments were carried out and served us as a data basis for a probabilistic model of cell size reduction.

We applied a homogenous non-stationary continuous-time Markov chain to model the development of individual diatoms from an initial size of about 80 $\mu$m until cell death which occured when the size reached its low at about 18$\mu$m. In contrast to conventional curve fitting models we are capable of calculating confidence intervals for estimates of the population ages as well as integrate the process of auxospore formation into the model. We thus propose a unique way to describe the stationary size distribution in a diatom population in terms of cell division and auxospore formation probabilities of its individuals.

*Corresponding author

## Introduction

*Pseudo-nitzschia delicatissima* (Cleve) Heiden is a chain-forming pennate planktonic diatom which is often found in abundance in temperate marine waters. Surrounded by a rigid silicate frustrule which provides mechanical protection against predators [1], the diatom undergoes progressive size reduction as its vegetative reproduction cycle proceeds. The original cell size is usually restored by entering a phase of sexual reproduction upon formation of an auxospore, which is not surrounded by a silicate shell and thus capable of expansion [2, 3]. Besides the intriguing nature of the process of diatom reproduction in itself, it has been shown that the diatom cell size plays an important part in community analyses [4]. The silicate shell itself is another focus of research with potential applications in modern technology and nanosciences [5]. Until now, most mathematical analyses of the processes involved have focused on linear models incorporating covariates [6] or were content with estimating single characteristics like average division rates or maximum size reduction rates [7, 8, 9]. Without doubt all these studies have given significant insights into the cryptic life cycle of diatoms. Still, to our knowledge no model has yet been proposed which is capable of modelling the size reduction of individual cells stochastically and thus might reveal the immanent mechanics of size distributions within a population.

In 2005, carried out mating experiments on three *Pseudo-nitzschia delicatissima* samples taken from the Gulf of Naples (Mediterranean Sea, Italy). Progenies of the original samples were grown in monoclonal cultures to prevent any sexual reproduction, and their continuous reduction in cell size was measured over a time period of 265 days using light microscopy [7]. The observerd cell sizes ranged from 80$\mu$m apical axis length down to 18$\mu$m after 265 days, until the cells finally died (for an illustration of the original dataset, see Fig. 1).

In order to achieve this desired individual modeling of size reduction, we combined the data collected by Amato and co-workers [7] with a modeling approach based on Markov chains (MCs). Markov chains are a family of memory-

less stochastic processes in which individuals assume certain states, and can move from one state to another with a given probability. Any object can only be in one state at any given point in time, and the transition probability to reach the next state only depends on the current state of the object (Markov property). In the homogenous case, transition probabilities additionally do not change over time [10].

Random processes comparable to these have been successfully used earlier and in similar settings. One is the stochastic description of the Polymerase Chain Reaction (PCR) which was modelled by Weiss and von Haeseler [11] using a randomly bifurcating tree or branching process to describe the step-wise doubling of the amplified DNA molecules under a certain error rate. To also model the mutations which can occur during the error prone replication process, the authors superimposed a Poisson process with an estimated mutation rate. This approach of modeling the PCR was successfully applied and extended by the authors [12] and other researchers [13, 14] as well.

We show here how an individual stochastic model of the cell division and sexual reproduction cycle of a diatom can lead to interesting insights about the size distribution in the population under study. The model is trained and verified on distributions of three independently sampled populations. We describe and validate the reliability of our results by back-estimation of the population ages and formalize the integration of additional covariates like changing climatic conditions or spore stages for spore forming species.



## Material and Methods

*Markov chains*

A Markov chain $X$ is a random process, i.e. a family $X_t : t \in T$ of random variables indexed by some set $T$. If $T = \mathbb{R}$ we call $X$ a continuous-time random process. $X$ takes values in a usually discrete *state space* $S$ such that every $X_t$ is a discrete random variable that takes one of $|S|$ possible values [10].

The transition rates (in the continuous-time case) from one state to another

are given in the $|S| \times |S|$ rate matrix $Q$, such that

$$\pi_t = \pi_0 \ e^{tQ},$$

where $\pi_t$ is the state distribution at time point $t$ and $\pi_0$ is the starting distribution of the chain. [10].

*Fisher Information and the variance of the MLE*

The definition of the *Fisher Information* from a sample of $n$ observations and any PDF $f$ is based upon the *score*

$$s(X;\theta) = \frac{\partial}{\partial\theta} \log f(X;\theta) \tag{1}$$

of an observation. The information is then given by

$$\begin{aligned} I(\theta) &= -\mathbb{E}\left(\frac{\partial\ s(X;\theta)}{\partial\theta}\right) \\ &= -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right) \\ &= -\int\left(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right) f(x;\theta)dx \end{aligned} \tag{2}$$

From the additivity of the information for independent samples it follows, that

$$I_n(\theta) = nI_1(\theta) = nI(\theta). \tag{3}$$

(for proofs see e.g. [15, 16]).

Computation of the variance of the maximum likelihood estimator (MLE) is based on its property of asymptotic normality, i.e.

$$\frac{(\hat{\theta}_n - \theta)}{se} \rightsquigarrow N(0,1),$$

where $se$ is given by the inverse Fisher Information [16]

$$se \approx \sqrt{1/I_n(\theta)}.$$

*Gaussian Mixture Models*

Gaussian Mixture Models are a method for unsupervised clustering of datasets. Several normal density distributions with different means and variances are mixed and fitted to the original dataset until the approximated density resembles the original as closely as possible. The memberships of individual data points to one of the gaussian distributions form the cluster predictions. Optimization of the fit is usually achieved bei evaluation of the Bayesian Information Criterion (BIC) for an Expectation Maximization (EM) fitted model. Mixture Model estimation of the diatom datasets was performed using the "mclust" R library [17, 18].

## Results

From the original work of we received three independently sampled *Pseudo-nitzschia* populations, named F1-5 (Fig. 1), F1-13 and F1-14. Population F1-5 served as a training dataset throughout the analyses, whereas populations F1-13 and F1-14 were used for validation purposes only.

The Markov chain applied for the modelling of size reduction in *P. delicatissima* populations is a state-discrete continuous-time homogenous Markov chain. It consists of a fixed number of states or cell size classes. The population develops continuously over time and as such mitosis — and therefore a size reduction step — can occur at any given point in time for any of the individuals. The probability of a state transition, i.e. one or a certain number of cell divisions, to occur does not change over time and only depends on the size of the cell.

In our initial model we assumed that cells die once they have reached the lowest size class and no covariates were included. With natural size reduction per generation being typically very low ($< 1.5$ $\mu$m gen$^{-1}$, [7]) and cell sizes measured with a high variation from 18 to 80 $\mu$m this would have led to more than 40 different states. As the measurements in the experimental datasets were only taken about once every week with longer breaks between days 70 and 139 and after day 153, the samples were not fine-grained enough to identify those 40

state transitions. We therefore chose to reduce the number of states to an still informative number by maximum likelihood estimation of a gaussian mixture model.

*Data discretization*

In order to find an appropriate number of states which still reflect the gradual stepwise reduction of cell size and thus allow for an accurate modelling but are easy to handle, we applied a gaussian mixture model (see Fig. 2) to find the most likely number of size categories and their respective limits (Table 1). See material and methods for details.

To prevent overfitting, we evaluated the Bayesian Information Criterion (BIC) for ML estimates with different numbers of clusters (Fig. 3). We found that the original size distributions are fitted best by seven categories having different variances and assigned each measurement to one of the seven size classes, which themselves each covered a size range between 5.5 and 17.5 $\mu$m. For the discretized data, see table 2.

*Estimation of an initial rate matrix*

In order to find the optimal rate matrix for our training population, we chose to first setup an initial rate matrix by estimation of the holding times of the states using the linear regression model (Fig. 1) applied earlier. Afterwards this initial rate matrix was further optimized by numerical optimization of the corresponding MLE for the Markov chain given the training data. The first approximation step was merely included to improve the condition of the optimization problem and to increase convergence speed.

Holding times can serve as estimators for the initial rate matrix $Q$. The holding time for state $i$ is exponentially distributed with a mean equal to $1/q_{ii}$, where $q_{ii}$ is the i'th main diagonal entry of the rate matrix $Q$. We put further constraints on the rate matrix by modelling the system as a pure birth process with a $n \times n$ rate matrix

$$Q = \begin{pmatrix} -q_{12} & q_{12} & & 0 \\ & \ddots & \ddots & \\ & & -q_{n-1\ n} & q_{n-1\ n} \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{4}$$

($q_{ij} > 0$ for all $i, j$), thus allowing only transitions from one state to the next "smaller" state with no increase in size or skipping of states allowed, corresponding to the biological intuition that cell division occurs successively. The inverse regression function of the first model (marked in red in figure 1)

$$x = y^{\frac{1}{b}}\ e^{-\frac{a}{b}} \tag{5}$$

was used to estimate the holding times using the previously determined state size bounds. The resulting holding times reached from 2.99 days for the second state to 168.7 days for the sixth state (see table 3). The same rate with opposite sign was then used as the transition rate into the successive state thereby keeping the generator matrix consistent.

*Finding the optimal Q*

After initialization of the rate matrix with sensible starting values the transition rates of the MC were improved by numerical optimization of the likelihood, to find the set of parameters that most likely could generate the measured experimental data.

Therefore, let $K$ be the number of states of the Markov chain, $L$ be the number of datasets (i.e. the number of distinct points in time at which measurements occured) and $N$ be the (constant) number of observations per measurement. Aditionally, let the vector $t$ of length $L$ hold the actual times at which the measurement occured (in days).

Then

$$\pi_{t_j} = \pi_0\ e^{t_j Q}$$

describes the state distribution after $t_j$ days. $x_{ij} \in \{1 \dots K\}, i \in \{1 \dots N\}, j \in \{1 \dots L\}$ is the observed state of the individual $i$ at time point $j$, and the like-

lihood of the parameters $t$ and $Q$ given the data is

$$\begin{aligned} L(Q) &= \prod_{i=1}^{N}\prod_{j=1}^{L}\left[\pi_{t_j}\right]_{x_{ij}} \\ &= \prod_{i=1}^{N}\prod_{j=1}^{L}\left[\pi_0\ e^{t_j Q}\right]_{x_{ij}}. \end{aligned}$$

For simplicity we define $n_{ij}$ as the number of individuals in state $i$ at time point $j$. Substituting $n$ for $x$ in the likelihood and taking the logarithm yields

$$\mathcal{L}(Q) = \sum_{j=1}^{L}\sum_{i=1}^{K} n_{ij}\ log\left(\pi_0\ e^{t_j Q}\right)_i. \tag{6}$$

The desired $\text{argmax}_Q\,\mathcal{L}(Q)$ was then found numerically by a modified version of the quasi-Newton BFGS algorithm (L-BFGS-B, [19]) using the R software package [20].

*Model evaluation*

First evaluations of the predicted size distributions against the measured showed, that the initial rate matrix using the holding times as approximations for the division rates was sensible. With a MSE of 0.3913 (log-likelihood -395.47) between the relative state frequencies of the experimental and predicted data this could be improved by the numerical optimization procedure which reduced the MSE to 0.3144 (log-likelihood -352.33). To further test our model, we estimated the population age $t$ from each time slice of the original dataset using the optimized rate matrix $Q$ with the MLE (Eq. 6) and compared it to the exact times of the experiments (Figure 4). Not surprisingly, the training dataset could be estimated very precisely for the training dataset (blue). The two other strains seemed to be only roughly comparable in terms of their size development (red and green solid line). This could be for two reasons, either the F1-13 and F1-14 populations indeed developed a unique size reduction dynamic which differs significantly from our training population or they simply evolve faster but not different in principle than the F1-5 model. In order to remove any such linear growth effect we estimated rate calibration factors (RCF) for the

F1-13 and F1-14 populations to normalize the time line of the three different Markov chains. This was achieved by a linear model on the estimated times for each of the two additional populations against the estimated times of the training set. The slope of the resulting two regression lines was then used as a linear factor for the time estimates (dotted green and red lines in figure 4).

The RCF compensated for the different overall growth rates but time estimations after normalization still showed some major differences. Between days 0 and 35 growth rates were generally overestimated for both F1-13 and F1-14 populations. Day 35 yielded again a very precise estimate of the population age for all three populations whereas for later measurements the age was again overestimated for the F1-14 and underestimated for the F1-13 population (Figure 4).

*Confidence intervals*

After computation of the optimal rate matrix it was now possible to determine the variance of the ML estimator using the *Fisher Information* in order to compute confidence bounds for the time estimation.

In the concrete case of our Markov chain we have a family of distributions parametrized over $t$

$$\pi_t(i) = \left(\pi_0 \; e^{tQ}\right)_i ,$$

which give the probability to be in state $i$ at time point $t$. The log-likelihood of *one* observation and therefore the score function (compare eq. 1) is given as

$$\begin{aligned} s(i;t) &= \frac{\partial}{\partial t} \log f(i;t) \\ &= \frac{\partial}{\partial t} \log \left(\pi_0 \; e^{tQ}\right)_i \\ &= \frac{\left(\pi_0 \; Q \; e^{tQ}\right)_i}{\left(\pi_0 \; e^{tQ}\right)_i}. \end{aligned}$$

Accordingly, by eq. (2) the information of *one observation* is given through

$$\begin{aligned} I_1(t) &= I(t) \\ &= -\mathbb{E}\left[\frac{\partial}{\partial t}\,\frac{\left(\pi_0\; Q\; e^{tQ}\right)_i}{\left(\pi_0\; e^{tQ}\right)_i}\right] \\ &= -\mathbb{E}\left[\left(\pi_0\; Q^2\; e^{tQ}\right)_i \left(\pi_0\; e^{tQ}\right)_i^{-1} + \left(\pi_0\; Q\; e^{tQ}\right)_i \left(-\pi_0\; e^{tQ}\right)_i^{-2} \left(\pi_0\; Q\; e^{tQ}\right)_i\right] \\ &= -\mathbb{E}\left[\frac{\left(\pi_0\; Q^2\; e^{tQ}\right)_i}{\left(\pi_0\; e^{tQ}\right)_i} - \frac{\left(\pi_0\; Q\; e^{tQ}\right)_i^2}{\left(\pi_0\; e^{tQ}\right)_i^2}\right] \\ &= -\sum_{i=1}^{K}\left[\left(\frac{\left(\pi_0\; Q^2\; e^{tQ}\right)_i}{\left(\pi_0\; e^{tQ}\right)_i} - \frac{\left(\pi_0\; Q\; e^{tQ}\right)_i^2}{\left(\pi_0\; e^{tQ}\right)_i^2}\right)\left(\pi_0\; e^{tQ}\right)_i\right] \\ &= -\sum_{i=1}^{K}\left(\left(\pi_0\; Q^2\; e^{tQ}\right)_i - \frac{\left(\pi_0\; Q\; e^{tQ}\right)_i^2}{\left(\pi_0\; e^{tQ}\right)_i}\right). \end{aligned} \tag{7}$$

Therefore by the additivity of the score (see eq. 3) the overall information is

$$I_n(t) = \sum_{i=1}^{K} N_{ij}\left(\left(\pi_0\; Q^2\; e^{tQ}\right)_i - \frac{\left(\pi_0\; Q\; e^{tQ}\right)_i^2}{\left(\pi_0\; e^{tQ}\right)_i}\right), \tag{8}$$

where $N_{ij}$ is the number of observations in state $i$ at time point $j$.

Since the MLE is asymptotically normal, the approximate 95% confidence bound for the time estimates is given by

$$\hat{t} \pm 1.96\sqrt{1/I_n(t)}. \tag{9}$$

For the confidence bounds of the time estimates of the diatom populations see table 4.

*Model extension*

Besides the probabilistic assessment of the significance of estimates, the major advantage of individual stochastic modeling of diatom state transitions is the possibility to also include the state of sexual reproduction into the model. Even though no direct experimental data on the probability of entering the size-recreating state of sexual reproduction was available, at least some rough estimates could be taken from the literature. These data were merely used to illustrate the principle approach of integrating this factor: It is generally accepted that smaller cells have a higher probability of entering the sexual phase,

typically when they reach about 30%-40% of their maximum cell size [3]. Yet, more recent reports have stressed, that auxospore formation can occur over almost the whole range of cell sizes, somewhere between 20%-90% of the initial cell size [21, 7]. In general, sex is rare in diatoms and can reach frequencies of about 4% in blooms [22] or be as low as 0.1 % as reported in *A. subarctica* and *F. kerguelensis* [23, 24].

To illustrate the possibility of integrating sex phases into our model we proceeded as follows: The number of states was extended to 8, the 8th stage representing auxospore formation. For simplicity, a transition into the sex phase was only possible from stages 6 and 7 (apical axis length $< 24\mu m$, i.e. within the preferable 30%-40% range), with a higher rate from stage 7 and both rates adjusted such that the auxospore stage probability in the stationary distribution of the chain never exceeded 4%.

The stationary distribution introduced by the sexual phase had its modus at state 5, with an expected mean state of 3.8 (not considering the sex state). For a detailed view of the distribution and the convergence of the chain into this distribution over time, see figure 5.

Even though we lacked the necessary experimental data to verify our results, the chosen rate matrix yielded a holding rate of $-0.5$ for the sex state 8. This means that according to our model, the length of the sexual phase should be exponentially distributed with a mean length of 2 days, a result which is in line with mating experiments in *Pseudo-nitzschia* [21].

**Discussion and conclusions**

The intruiging life cycle of diatoms has always drawn the attention of many researchers. But even though various experimental projects have been carried out, most of the published articles have been solely descriptional. The number of mathematically oriented works on the subject is still low, with most of the available papers focusing on estimating single characteristics of diatoms, typically division rates or size reduction per generation under differing photoperiods, nutrient status, salinities or water temperatures. This is typically achieved

by applying straightforward statistical methods, like linear regression. In this manner, the project presented here has several advantages: (i) Modelling diatom size reduction as an individual-based stochastic process is certainly closer to biological reality than a regression line on the e.g. median of cell sizes of a population. (ii) We are able to integrate the auxospore formation process into the model without changing the model paradigm. This has the immediate effect that cell state transitions can be modeled, even when the size distribution has reached stationarity and the mean cell size is constant, i.e. where standard linear models of size reduction fail. (iii) The stochastic nature of the process gives us the opportunity to compute statistical properties, like confidence bounds, for all model parameters, i.e. the rate matrix, time estimates, mean size reduction, etc.

As a drawback we must note, that the fitting of the Markov chain, even with a reduced number of states, where one state transition accounts for more than one cell division, requires several precise measurements of the size reduction process. Preferably, measurements should be taken on an individual basis, i.e. in our case the division rates of individual diatoms in a population should be measured, something which is certainly difficult to achieve. Fitting the Markov chain only by means of its marginals as in our example is suboptimal, but feasible. To increase model accuracy, more measurements per time interval would be desirable, especially if we aim at a more fine-grained model, where one state transition accounts for one cell devision.

But even with our relatively reduced sets of experimental data we successfully fitted the model and were able to estimate population ages for the F1-13 test population. In order to comment on the differences between measured and estimated ages for the F1-14 population, additional measurements would be essential. Fitting the Markov chain on a variety of populations taken under comparable experimental or environmental conditions would give us the opportunity to see whether our rate matrix is general.

Nevertheless, our proposed principle of modeling diatom size development is advantageous. An accurate model is not only capable of modeling typical

size distributions, but also able to detect divergence from the stationary distribution of the chain, even when other characteristics like mean cell size remain constant. This might be important for the detection of abnormal population developments, like blooms, or for their modeling. In summary, the description of the size distribution of a population by means of cell divisions of individual diatoms is novel. The advantages more than make up for the increased demand for experimental data and might give valuable insights into the development of the single cell as well as population wide effects like blooms or oscillary size distributions due to seasonal effects. It is the first integrated model of the complete life cycle of a diatom and as such a valuable tool for both bio-mathematicians and diatom researchers alike and a basis for further development of diatom size reduction models.

**Acknowledgements**

We wish to thank Alberto Amato, Domenico D'Alelio and Marina Montresor for providing us with the necessary experimental data and for fruitful discussions. We thank an anonymous reviewer for improving the manuscript.

## References


[1] C. E. Hamm, R. Merkel, O. Springer, P. Jurkojc, C. Maier, K. Prechtel, V. Smetacek, Architecture and material properties of diatom shells provide effective mechanical protection., Nature 421 (6925) (2003) 841–843. doi:10.1038/nature01416.
URL http://dx.doi.org/10.1038/nature01416

[2] D. G. Mann, Patterns of sexual reproduction in diatoms, Hydrobiologia 269-270 (1) (1993) 11–20.
URL http://dx.doi.org/10.1007/BF00027999

[3] F. E. Round, R. M. Crawford, D. G. Mann, Diatoms: Biology and Morphology of the Genera, Cambridge University Press, 1990.

[4] P. Snoeijs, S. Busse, M. Potapova, The importance of diatom cell size in community analysis, Journal Of Phycology 38 (2) (2002) 265–272.

[5] C. E. Hamm, The evolution of advanced mechanical defenses and potential technological applications of diatom shells, Journal Of Nanoscience And Nanotechnology 5 (1) (2005) 108–119.

[6] M. Mizuno, Influence Of Cell-Volume On The Growth And Size-Reduction Of Marine And Estuarine Diatoms, Journal Of Phycology 27 (4) (1991) 473–478.

[7] A. Amato, L. Orsini, D. D'Alelio, M. Montresor, Life cycle, size reduction patterns, and ultrastructure of the pennate planktonic diatom Pseudonitzschia delicatissima (Bacillariophyceae), Journal Of Phycology 41 (3) (2005) 542–556.

[8] J. Fehling, K. Davidson, S. S. Bates, Growth dynamics of non-toxic Pseudonitzschia delicatissima and toxic P. seriata (Bacillariophyceae) under simulated spring and summer photoperiods, Harmful Algae 4 (4) (2005) 763–769.

[9] D. H. Jewson, Life-Cycle Of A Stephanodiscus Sp (Bacillariophyta), Journal Of Phycology 28 (6) (1992) 856–866.

[10] G. Grimmett, D. Stirzaker, Probability and Random Processes, Oxford University Press, 2001.

[11] G. Weiss, A. von Haeseler, Modeling the polymerase chain reaction., J Comput Biol 2 (1) (1995) 49–61.

[12] G. Weiss, A. von Haeseler, A coalescent approach to the polymerase chain reaction., Nucleic Acids Res 25 (15) (1997) 3082–3087.

[13] N. Saha, L. T. Watson, K. Kafadar, A. Onufriev, N. Ramakrishnan, C. Vasquez-Robinet, J. Watkinson, A general probabilistic model of the PCR process., Conf Proc IEEE Eng Med Biol Soc 4 (2004) 2813–2816. doi:10.1109/IEMBS.2004.1403803.
URL http://dx.doi.org/10.1109/IEMBS.2004.1403803

[14] N. Saha, L. T. Watson, K. Kafadar, N. Ramakrishnan, A. Onufriev, S. Mane, C. Vasquez-Robinet, Validation and estimation of parameters for a general probabilistic model of the PCR process., J Comput Biol 14 (1) (2007) 97–112. doi:10.1089/cmb.2006.0123.
URL http://dx.doi.org/10.1089/cmb.2006.0123

[15] B. W. Lindgren, Statistical Theory, Chapman & Hall/CRC, 1993.

[16] L. Wasserman, All of Statistics, Springer, 2005.

[17] C. Fraley, A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, Journal Of The American Statistical Association 97 (458) (2002) 611–631.

[18] C. Fraley, A. Raftery, mclust: Model-Based Clustering / Normal Mixture Modeling, r package version 3.1-5 (2008).
URL http://www.stat.washington.edu/fraley/mclust

[19] R. H. Byrd, P. Lu, J. Nocedal, C. Y. Zhu, A Limited Memory Algorithm for Bound Constrained Optimization, SIAM Journal on Scientific Computing 16 (6) (1995) 1190–1208.
URL citeseer.ist.psu.edu/byrd94limited.html

[20] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2006).
URL http://www.R-project.org

[21] N. A. Davidovich, S. S. Bates, Sexual reproduction in the pennate diatoms Pseudo-nitzschia multiseries and P-pseudodelicatissima (Bacillariophyceae), Journal Of Phycology 34 (1) (1998) 126–137.

[22] R. M. Crawford, The Role Of Sex In The Sedimentation Of A Marine Diatom Bloom, Limnology And Oceanography 40 (1) (1995) 200–204.

[23] P. Assmy, J. Henjes, V. Smetacek, M. Montresor, Auxospore formation by the silica-sinking, oceanic diatom Fragilariopsis kerguelensis (Bacillariophyceae), Journal Of Phycology 42 (5) (2006) 1002–1006.

[24] D. H. Jewson, Size-Reduction, Reproductive Strategy And The Life-Cycle Of A Centric Diatom, Philosophical Transactions Of The Royal Society Of London Series B-Biological Sciences 336 (1277) (1992) 191–213.

## Figure legends

Figure 1: Original data showing the decrease in cell size over time. The blue and red lines are fitted linear regression models with only the predictor (blue) or both predictor and outcome log-transformed (red). The inverted regression function of model 1 (red) was later used for approximation of the holding times. Both models showed a high goodness-of-fit with a $R^2$ of 0.92 and an overall p-value $<$ 2.2e-16.

Figure 2: Combined density distribution of the seven scaled normal densities with means given in the figure legend. These seven clusters were used to divide the original size range into seven discrete categories, each represented by a state in the Markov chain.

Figure 3: Results of the mixture model fitting process. The number of clusters in the data is evaluated against the Bayesian Information Criterion (BIC) for both a variable variance (dashed red) and constant variance (solid black) model. Note, that in both cases seven clusters provide the best fit to the original data density.

Figure 4: Plot of estimated population ages against the residuals (estimated - real) for the original dataset (blue, F1-15) and the two test datasets (F113, red and F114, green). The graph shows time estimates both with (dotted) and without (solid) timeline normalization through rate calibration.

Figure 5: Convergence of the MC into the stationary distribution over time from an initial state distribution of $\pi_0 = (1, 0, 0, 0, 0, 0, 0)$.

## Tables

| State 1 | State 2 | State 3 | State 4 | State 5 | State 6 | State 7 |
|---|---|---|---|---|---|---|
| 80.0 | 74.5 | 57.0 | 48.0 | 37.5 | 21.0 | 14.0 |

Table 1: Lower bounds of the states found by MLE mixture model fitting (in $\mu$m).

| days | 16 | 29 | 35 | 42 | 49 | 60 | 70 | 139 |
|---|---|---|---|---|---|---|---|---|
| **State 1** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **State 2** | 0 | 0.487 | 0.128 | 0 | 0 | 0 | 0 | 0 |
| **State 3** | 0 | 0.436 | 0.667 | 0.641 | 0.231 | 0.077 | 0.026 | 0 |
| **State 4** | 0 | 0.051 | 0.154 | 0.256 | 0.564 | 0.436 | 0.103 | 0 |
| **State 5** | 0 | 0.026 | 0.051 | 0.077 | 0.179 | 0.41 | 0.769 | 0 |
| **State 6** | 0 | 0 | 0 | 0.026 | 0.026 | 0.077 | 0.103 | 1 |
| **State 7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| days | 146 | 153 | 188 | 265 |
|---|---|---|---|---|
| **State 1** | 0 | 0 | 0 | 0 |
| **State 2** | 0 | 0 | 0 | 0 |
| **State 3** | 0 | 0 | 0 | 0 |
| **State 4** | 0 | 0 | 0 | 0 |
| **State 5** | 0 | 0 | 0 | 0 |
| **State 6** | 1 | 1 | 0.949 | 0 |
| **State 7** | 0 | 0 | 0.051 | 1 |

Table 2: Relative state frequencies in the original dataset, after discretization.

| State 1 | State 2 | State 3 | State 4 | State 5 | State 6 |
|---|---|---|---|---|---|
| 5.077451 | 2.994306 | 15.589484 | 14.984200 | 31.949934 | 168.739609 |

Table 3: Holding times (in days) of the states found by MLE mixture model fitting. Note that the last state (7) was also the death state and thus has an infinite holding time.

| **real** | 0 | 13 | 19 | 26 | 33 | 44 |
|---|---|---|---|---|---|---|
| **estimate** | 0.500 | 10.522 | 16.852 | 22.507 | 31.255 | 41.400 |
| **lower** | 0.042 | 7.783 | 12.918 | 17.544 | 24.781 | 33.267 |
| **upper** | 0.958 | 13.260 | 20.785 | 27.470 | 37.728 | 49.533 |

| **real** | 54 | 123 | 130 | 137 | 172 | 249 |
|---|---|---|---|---|---|---|
| **estimate** | 53.082 | 136.284 | 136.284 | 136.284 | 144.350 | 300.000 |
| **lower** | 43.058 | 105.290 | 105.289 | 105.289 | 110.204 | 197.434 |
| **upper** | 63.107 | 167.279 | 167.279 | 167.279 | 178.495 | 402.566 |

Table 4: Lower and upper 95% confidence bounds of the time estimates of the F15 Population.

Figure 1



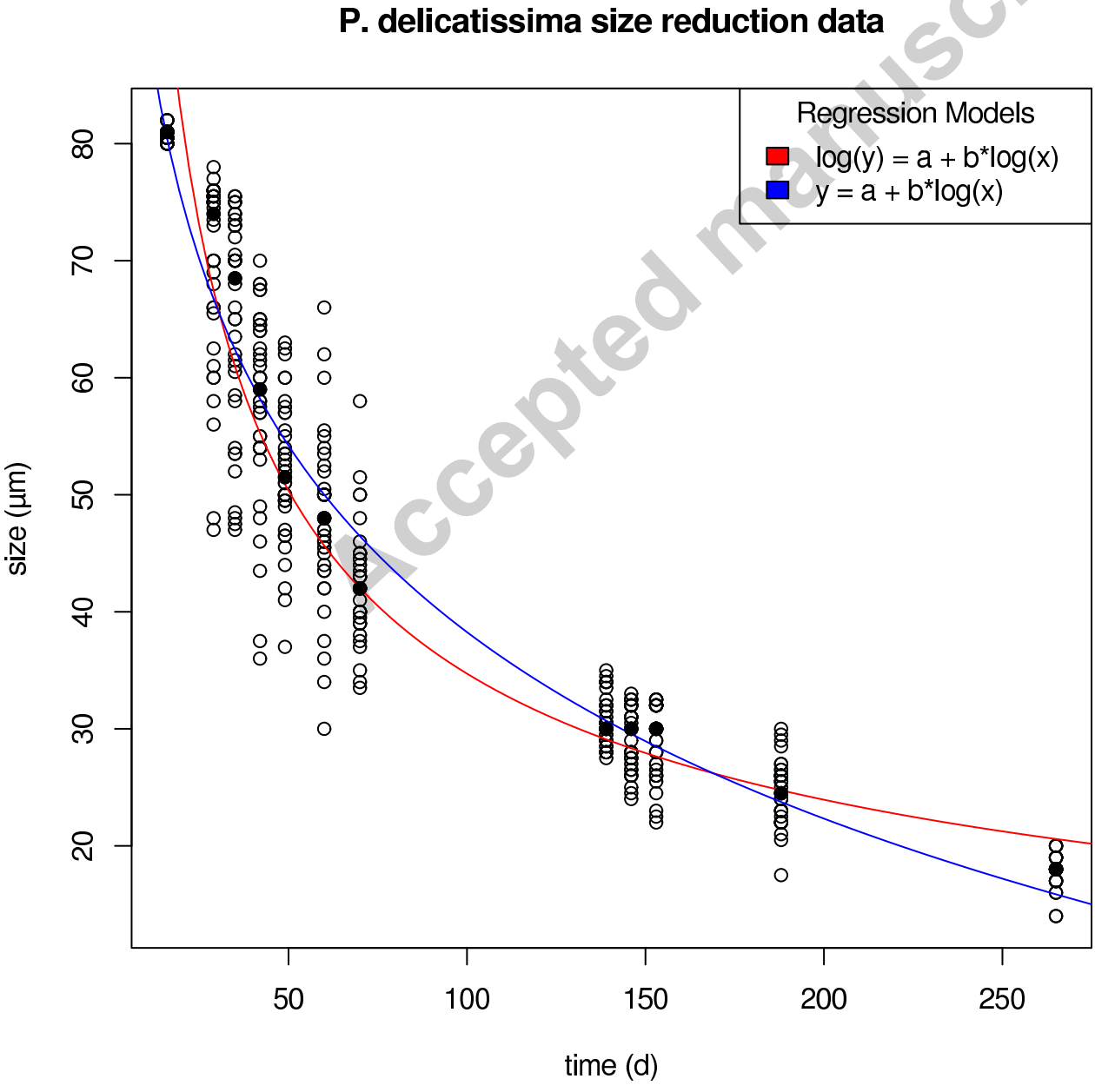

Figure 2



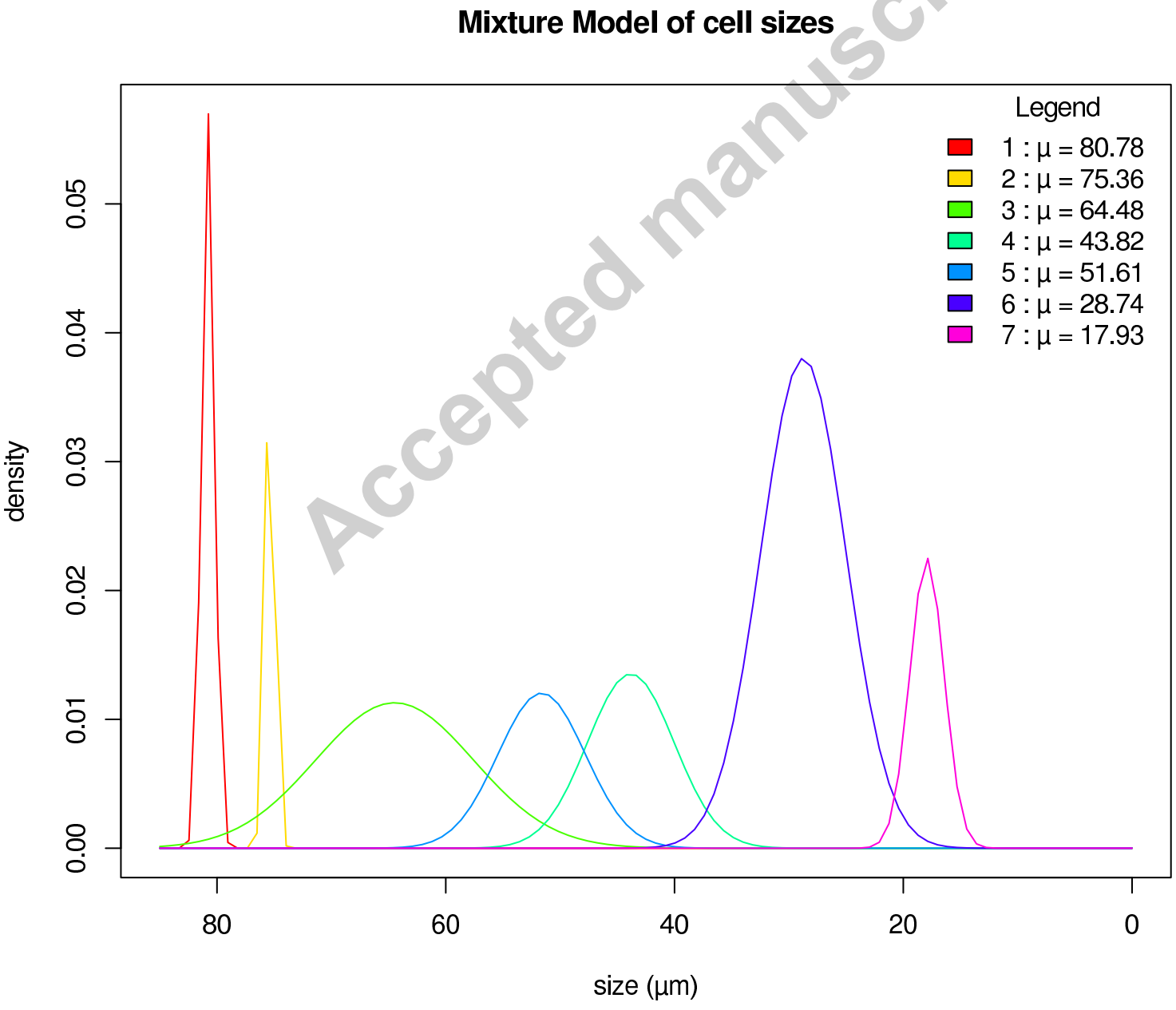

Figure 3



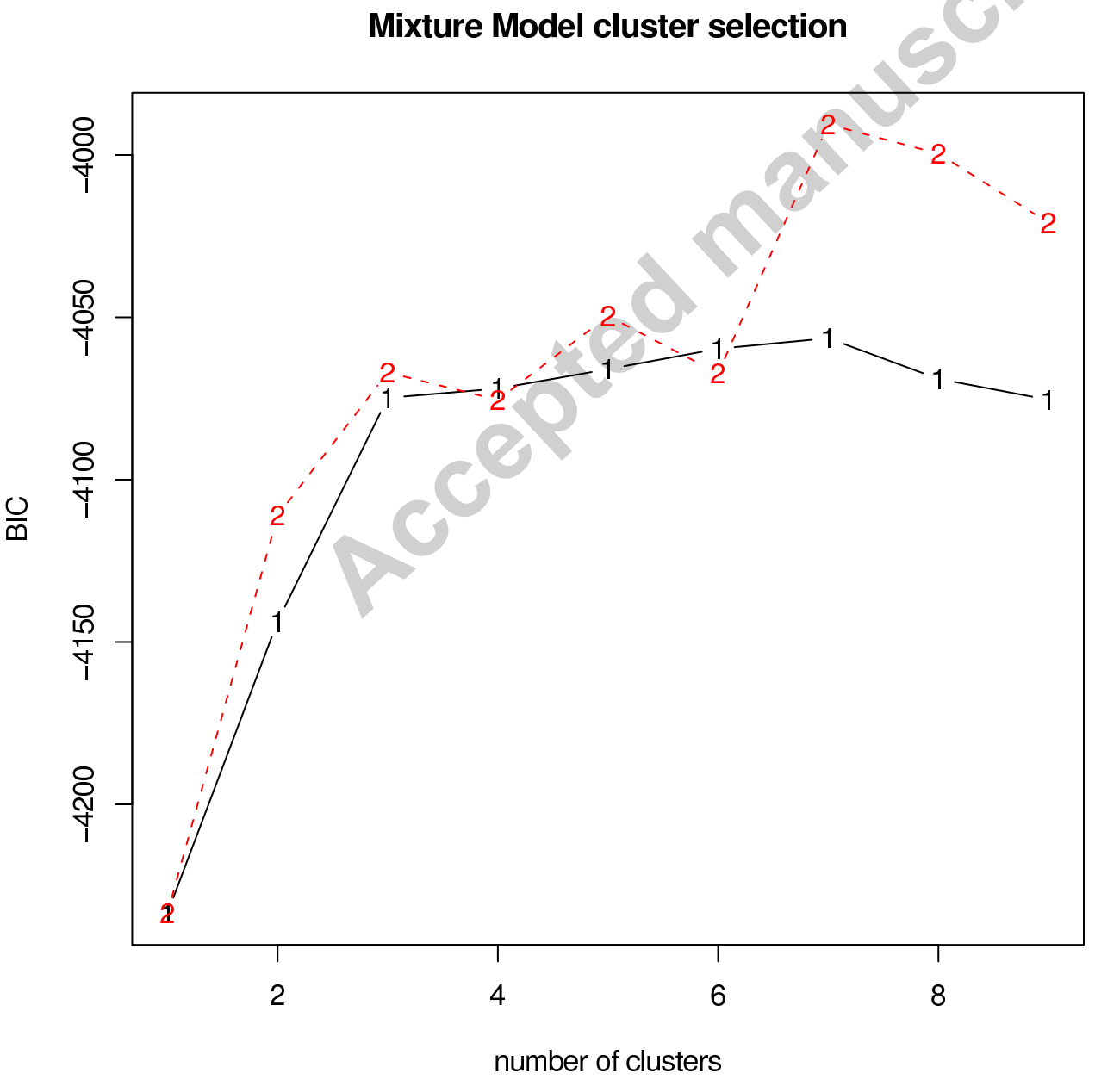

Figure 4



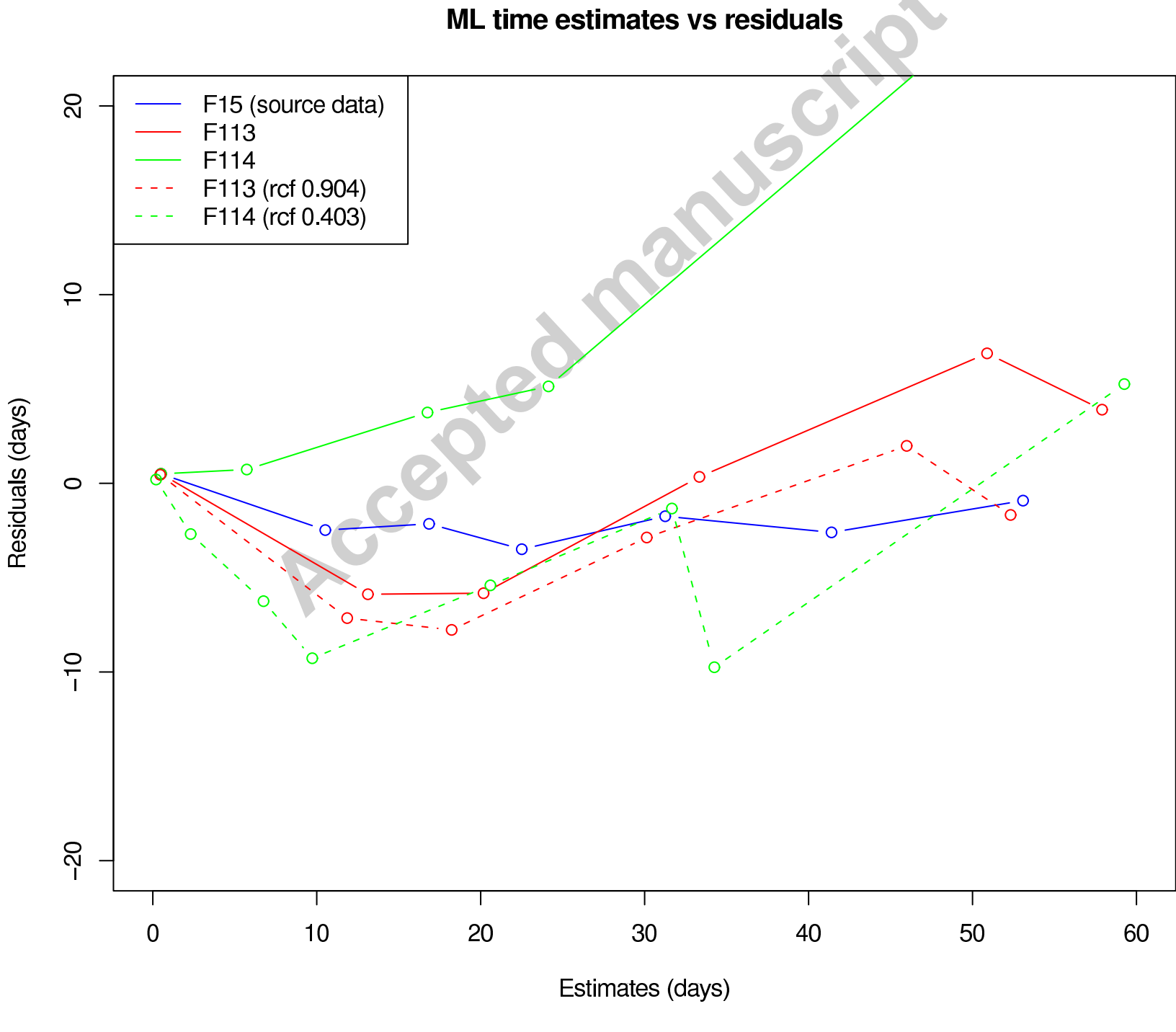

Figure 5