Delft University of Technology
Faculty Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Modeling the single seed oxygen consumption of germinating seeds

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree

BACHELOR OF SCIENCE
in
APPLIED MATHEMATICS

by

LINDA WIEGMAN

Delft, the Netherlands
August 2015

BSc thesis APPLIED MATHEMATICS

“Modeling the single seed oxygen consumption of germinating seeds”

LINDA WIEGMAN

Delft University of Technology

Responsible Professor

Dr. N.V. Budko

Other members of the thesis committee

Prof.dr.ir. C. Vuik

Dr. J.L.A. Dubbeldam

Dr.ir. M. Keijzer

August, 2015

Delft

Abstract

The rate of oxygen consumption by germinating seeds is considered to be one of the promising parameters for monitoring the seed status and a candidate for predicting germination and seed vigor. Nowadays, the single seed oxygen consumption patterns can be measured on a big scale and at a detailed time resolution using a Q_2 machine. However, interpretation of the data in terms of functioning of internal oxygen transport and overall seed properties is still hard, due to the lack of knowledge on and the complexity of these properties and processes. Modeling may be of great help in understanding the relation between germination and the oxygen consumption pattern of a seed. In this thesis a model for the single seed oxygen consumption inside a closed test tube is proposed, relating the measured oxygen concentration to the size of the active mitochondrial population. The analytical solution of this model is used for calibration of the experimental data and all calibrated results are analyzed. From this analysis an attempt is made to distinguish between germinating and non germinating seeds in order to determine which seed properties cause slow/fast germination. Based on this model a method introducing the seed and test tube volume is proposed in order to predict the observations when the sizes of these volumes change.

Contents

1 Introduction

1.1 Motivation

From January 28 to February 1, 2013 I participated in the 90th Study Group Mathematics with Industry (SWI) held at Leiden University. SWI is a combined industrial-academic week where mathematics is used to tackle industrial problems. About fifty to eighty mathematicians, working in industry or academia, come together during a full week to work together intensively on industrial problems. The format follows the original Oxford model, dating back to 1968, which is used worldwide in similar Study Groups.

One of the questions was formulated by Fytagoras. Fytagoras is a science orientated company with much expertise in the fields of sensor technology, seed technology and plant breeding. The question concerned the oxygen consumption of germinating seeds. Since I have always been very interested in the link between physics/biology and mathematics this subject drew my attention. After working on the problem for a week I decided to turn it into my bachelor thesis.

1.2 Problem statement

Seed germination refers to the physiological process of the growth of an embryo within a seed. It starts with water uptake by the seed (imbibition) and ends with the emergence of the embryo from its enclosing coverings. Although it is poorly defined, the event of germination is commonly associated with the moment the seed coat is being penetrated by the embryo's radicle [1]. Seed germination depends on both internal and external conditions. The most important external factors include temperature, water, oxygen and sometimes light. Various plants require different variables for successful seed germination. The variables depend on the individual seed variety. In agronomic and horticultural practice a fast and homogeneous germination is desirable for many crops. Since the germination time of seeds in seed batches often varies significantly, much effort is devoted to the selection and treatments of seeds to achieve simultaneous germination.

This thesis focuses on the dependency of seed germination on and the consumption of oxygen. Living seeds start respiration upon imbibitions and respiration accelerates at the moment the germination really commences. The availability of oxygen to the embryo in the seed depends on the oxygen concentration around the seed, the respiration rate of the cells (in the embryo) and oxygen transport through the seed towards the cells of the embryo. Especially the transport through the seed is still a black box.

Fytagoras has developed a so-called Q_2 machine by which single seed oxygen consumption can be measured at a detailed time resolution. The measured curves can be used in seed quality evaluation since the rate of oxygen consumption by germinating seeds is considered to be one of the promising parameters for the monitoring of the seed's status and a candidate for predicting germination and seed vigor [2][3]. However, the interpretation of the data in terms of physical, morphological and physiological properties and processes within the seed is still hard, due to lack of knowledge of basic aspects of gas exchange in seeds and its role in the germination process. It also greatly hampers advances in seed treatment possibilities, such as respiration controlled seed priming, advanced seed coating and pellets, as well as improvements in seed storage and seed longevity.

Currently oxygen consumption profiles are usually simply fitted using low-order polynomials or piecewise-linear curves and the germination behavior is predicted on the basis of certain features of these curves. The disadvantage of this approach lies with the parameters of these polynomials. They have no clear biological meaning and therefore these models are unsuitable for proper interpretation of the data.

1.3 Research objectives

The main goal is to develop a mathematical model for the oxygen transport and consumption within the seed that is in line with the current biological knowledge on these processes and the highly detailed time-resolved oxygen consumption measurements for single seeds. The model should allow interpretation of the observed data and characteristics in terms of the functioning of internal oxygen transport processes and overall seed properties or provide hypothesis on these that may be validated in future research.

In order to develop such a model we first take a closer look at the materials and measurements in Section 2 and we dive into the biological properties of seeds and the processes associated with their germination in Section 3. Secondly, we combine this with our mathematical knowledge and develop a model for the growth and proliferation of cells and their oxygen consumption in Section 4. The model contains several (unknown) parameters, is based on ordinary differential equations (ODE) popular in biological growth and proliferation studies and describes the two major stages of the germination process: the repair stage and the growth stage. By fitting the obtained equation(s) to our experimental data, statistics on these parameters are obtained and after analyzing these statistics and their corresponding fitted curves, an attempt is made to split the seed batches into different classes. The fitting and analysis of the model is performed on one seed batch first, Section 5, and the resulting hypothesis and observations are tested and verified on the complete data batch in Section 6. The seed and test tube volume are not taken into account even though they are likely to influence the germination process. Therefore in Section 7 a method is proposed to introduce the volumes into our model.

The eventual goal in this study is to find possible (cor)relations between parameters and link the parameter values to curve- and therefore seed properties in order to determine which seed properties cause slow/fast germination. This information will help seed technology and plant breeding companies in the selective breeding process and mutation of plant seeds.

2 Materials and measurements

2.1 Seeds

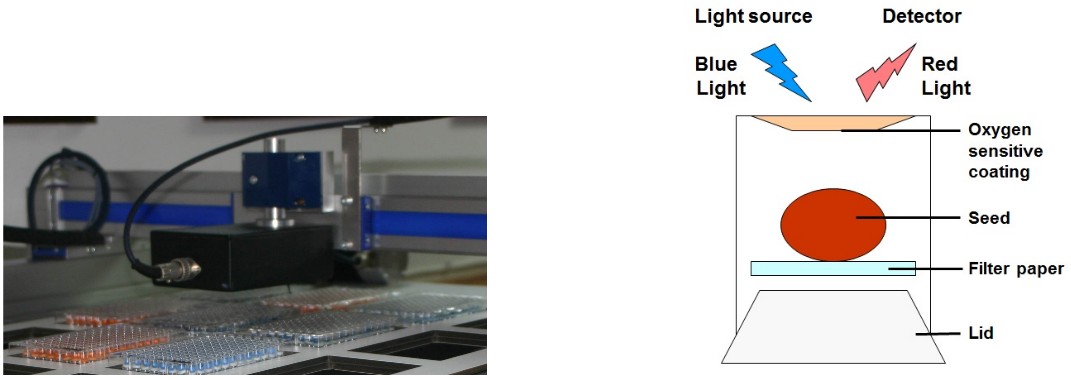
Fytagoras performed most of the single seed oxygen consumption measurements on barley seeds (*Hordeum vulgare* Var. Esterel 2003). Seeds from different batches with different germination properties were available and therefore this seed was very suitable for our research. Before measurements the seeds were treated with a powder to prevent fungal growth.

2.2 Respiration measurements

Single seed oxygen consumption measurements were performed by a Q_2 machine developed by Fytagoras. Q_2 stands for Quality and Quick but also for O_2, the element on which the whole instrument is based upon [2]. In a typical setup of such a machine a significant number of randomly selected seeds is placed in small sealed, cylindrical shaped containers. These containers, i.e. test tubes, are almost air-tight and contain enough water and air to suffice the germination conditions of the seed. When a single test tube is too large, the decrease in oxygen concentration is very small and therefore hard to measure. A test tube that is too small will cause the oxygen to deplete so fast, that we are left with only a few measurement points. It turned out that the best results in both time and oxygen level were obtained by using test tubes with a volume of 1.0 ml [14]. Therefore a 96 well plate (Eppendorf) was chosen for the machine. In each of these wells a single seed was placed on top of a circular paper filter soaked in water.

The oxygen level in each well is carefully measured using the Q_2 measuring technology. In this fluorescent-dye technique [2] [7], the top of each test tube is by a lid with an oxygen sensitive fluorescent coating which reacts with the oxygen present in the container. Plates are placed in the Q_2 machine and with the help of LED-technology blue light is sent through the coating and transformed into red fluorescent light. The quantity of fluorescent light is in proportion with the partial evaporation tension of oxygen in the container. By measuring the amount of light the level of oxygen present in the container can be determined very precisely.

A schematic view of the experimental setup and a picture of the complete Q_2 machine are shown in Figure 2.1.



(a) Q_2 machine [2]

(b) Experimental setup

Figure 2.1: Overview of the operation of the Q_2 machine

Each well was scanned automatically by the Q_2 machine at 30 minute time intervals. The oxygen level measurements were collected and stored in Excel data format and can easily be converted to data curves showing the oxygen consumption pattern of individual seeds. Based on this pattern we can categorize the seeds into different types [2]. The patterns of three types of seeds are shown in Figure 2.2.

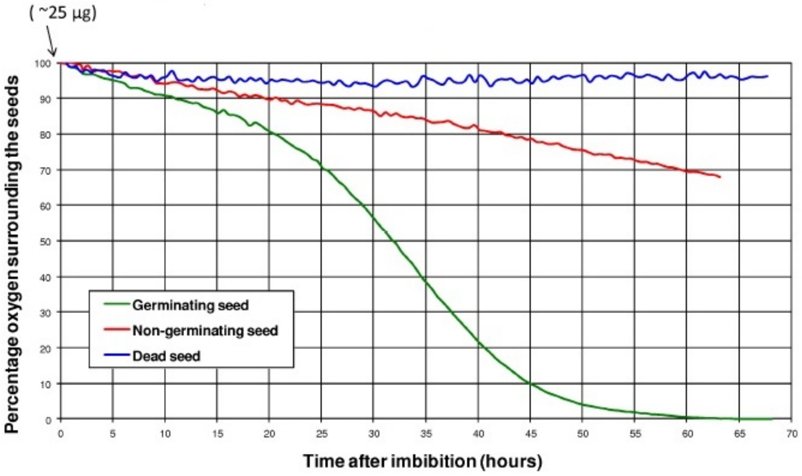


Figure 2.2: Oxygen consumption measurement curves of three types of seeds. [2]

The upper line in this figure displays no oxygen consumption and is therefore categorized by Van Asbrouck and Taridno [2] as a dead seed. The line showing a small and constant amount of oxygen consumption is categorized as a dormant seed. The concept of dormancy will be explained in Section 3. The lower line displays a huge increasing oxygen consumption and is therefore categorized as a germinating seed. Germinating seeds can be split into two groups: strong germinators, the seeds that show a sigmoid type of oxygen consumption pattern, and weak germinators, the seeds that show a linear or incomplete oxygen consumption pattern. The definition of a dormant seed and a weak germinator slightly overlap, which makes it hard to distinguish between the two. Also note that for a complete picture of the seeds, approximately two days were needed.

2.3 Data

Fytagoras has provided us with 112 sample batches, which amounts to data on circa 11,000 barley seeds. The batches are bundled in 7 files each containing 16 batches. Each file has its own time frame, varying from about 21 to 72 hours. Data curves of one of the batches from the first file are shown in Figure 2.3.

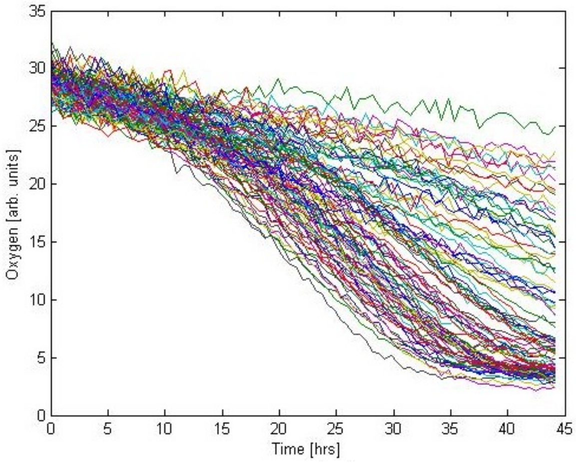


Figure 2.3: Oxygen consumption curves of one data batch.

`Matlab` was used for implementation and data fitting so the data originally stored in Excel files was converted to `.mat` files using the `xlsread` function. For details of the model and data fitting the reader is referred to Sections 4 and 5.

3 Biological processes and modeling assumptions

This study considers the germination process of (barley) seeds. By definition, germination commences when the dry seed, shed from its parent plant, takes up water (imbibition), and is completed when the embryonic root visibly emerges through the seed coat. Thereafter, there is seedling establishment, utilizing reserves stored within the seed, followed by vegetative and reproductive growth of the plant, supported by photosynthesis. These successive phases will not be discussed in this thesis.

In the mature, dry state the seeds are metabolically inactive (quiescent) and can withstand extremes of drought and cold. For example, dry seeds can be stored at -150 degrees Celsius for many years without harm. In this state the cells contain less than 10% water and the seed coat is (almost) fully impermeable to oxygen. Upon imbibitions, the dried cells are being repaired to healthy cells, which contain approximately 70% water. Only healthy cells are able to divide and grow so this repair is necessary for the germination process. Besides that, the healthier the cells the easier they absorb oxygen. During the repair process, which is shown in Figure 3.1, a little bit of oxygen is entering the cell with the water. This oxygen is used for energy production to facilitate the repair process. Within minutes of the water uptake the growth commences and the respiration increases. Respiration refers to aerobic respiration, the chemical process that convert nutrients from the starch into energy in the form of Adenosine Triphosphate (ATP) using oxygen. As far as the transport physics are concerned the material inside the seed is mostly water after this quick imbibition.

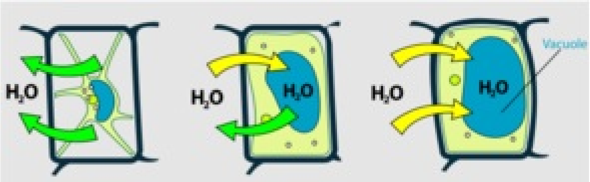


Figure 3.1: Cell repair [4]

In the transport of oxygen from the seedcoat to the embryo, several aspects have to be considered that influence the permeability and therefore the respiration rate. As was mentioned before, the seed coat gets more permeable to oxygen as it gets more saturated. But also the specific structure of the seed coat affects the permeability. It can be a paper-thin layer (peanut) or something way more substantial (coconut). After passively diffusing through the seed coat, the oxygen encounters a layer of starch cells. These cells are completely impermeable to oxygen so the oxygen molecules have to pass through the intercellular spaces. The size of these spaces depends on the shape and organization of the starch cells and this differs per seed type. The larger the spaces and the more organized the cells, the higher the permeability. In Figure 3.2 the structure of the barley seed coat and starch cells is shown.

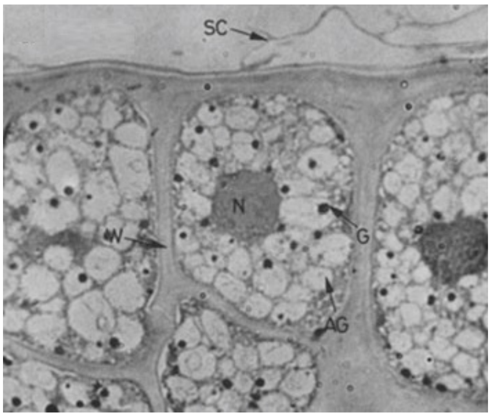


Figure 3.2: Microscopic view of the seed coat and part of the starch layer of a barley grain.
This image shows the outer seed coat (SC), the cell wall of a starch cell (W), its nucleus (N) and the protein storage vacuoles (AG) with globoids (G) (spherical crystalline inclusions that contain nutrients for plant growth)[1].

Now the oxygen has made its way through the starch and reaches the embryo cells where it passively diffuses through the cell walls and plasma membranes. The main assumption of our model is that most of the oxygen is being consumed by the mitochondria, which are small membrane-enclosed organelles located in the cytoplasm of the cells. In these organelles the aerobic respiration takes place to produce energy that is as a source of chemical energy for other processes inside the cell. This is why mitochondria are sometimes described as 'cellular power plants'. The number of mitochondria per cell can vary from a single one up to thousands of them. According to the Endosymbiotic Theory [5], mitochondria evolved from bacteria and still largely behave as such. The other organelles inside the seed use oxygen on such a small scale that they are considered relatively insignificant an thus the oxygen data provide a unique insight into the dynamics of the mitochondrial population.

Given that the material inside the seed is considered mostly water, we can apply Henry's law [6], which states that the measured oxygen concentration outside the seed (in the test tube) is proportional to the oxygen concentration inside the seed:

$$c_{\mathrm{aq}} = k_{\mathrm{H,cc}} \cdot c_{\mathrm{gas}}$$

where c_{aq} is the concentration of gas in a solution, water in this case, c_{gas} is the concentration of gas above that solution and $k_{\mathrm{H,cc}}$ is the Henry constant. Numerical simulations have shown that this Henry equilibrium is established within the current measurement interval of 30 minutes [14], which means that we can assume that the oxygen concentration inside our seed is uniform at all times and we can thus neglect any spatial variations. A resulting assumption is that the complex transport- and diffusion process of oxygen entering the embryo cells, as discussed earlier, can be neglected as well. The more sophisticated models based on partial differential equations (reaction-diffusion, transport etc.) introduce many more tuning parameters without providing any better fit for the data [8] [9] [10]. This means that one of the research objectives, namely interpreting the observed data and characteristics in terms of the functioning of internal oxygen transport processes, becomes more and more irrelevant. Next to this, our initially much more sophisticated simulations and models, have shown that we are also able to neglect the changes in seed shape and volume. Because of these assumptions it is not necessary to elaborate on the detailed dynamics of the growth and proliferation of the embryo cells during germination.

One thing left to consider is that not all seeds are able to germinate. Amongst these are the dead and dormant seeds discussed in Section 2. Dormant seeds are seeds that do not have the capacity to germinate in a specific period of time under any combination of normal, physical and environmental factors that otherwise is favorable for its germination [1]. The breaking of dormancy is possible and therefore dormancy is seen as a temporary block on germination.

This concludes the biological background. The next section will introduce a model based on the assumptions made in this section.

4 Model for oxygen consumption and growth in closed test tubes

In this section the model for the oxygen consumption of a germinating seed in a closed test tube is developed. This model relates the measured oxygen concentration $c(t)$ to the number of active mitochondria $n(t)$ and splits the germination process into two stages, a repair stage and a growth stage. An analytical solution is obtained and will be used for fitting the experimental data in subsequent sections.

4.1 Oxygen consumption and growth in closed test tubes

From the previous section we know that oxygen consumption happens mainly in the mitochondria. Experiments with isolated mitochondria show that the rate of this consumption is proportional to the concentration of oxygen [11]. Now assuming that this consumption rate is the same for all mitochondria and that both $c(t)$ and $n(t)$ are continuous in t we arrive at the following equation for the decrease in oxygen concentration:

$$\frac{dc}{dt} = -\alpha n c, \quad c(0) = c_0. \tag{4.1}$$

Mitochondria behave as a colony of bacteria, i.e. multiplying at a rate proportional to their number. Assuming that this rate is also proportional to the oxygen concentration we obtain the equation for the growth of our mitochondrial population:

$$\frac{dn}{dt} = \beta n c, \quad n(0) = n_0. \tag{4.2}$$

Both the oxygen consumption rate α and mitochondrial reproduction rate β are non-negative. It is natural to assume that α is constant while one may expect that β depends on the physiological state of the seed and therefore on time. As mentioned in Section 3, the common understanding is that embryo cells need to repair first before they can function properly. This means that the mitochondria inside these cells also need a certain repair phase before they are fully functional and able to replicate. The length of this repair phase is from now on denoted by t_r. We assume that all mitochondria are fully functional at the end of this repair phase and their physiological state remains the same for the rest of the germination process. These facts are taken into account by setting β equal to zero for $0 \leq t \leq t_r$ and take $\beta > 0$ constant for $t > t_r$.

The solution of the ODE system (4.1)-(4.2) is obtained by firstly dividing (4.2) by (4.1):

$$\frac{dn}{dc} = -\frac{\beta}{\alpha}. \tag{4.3}$$

During the repair stage Equation (4.2) turns into

$$\frac{dn}{dt} = 0, \quad n(0) = n_0$$

with the trivial solution

$$n(t) = n_0, \quad 0 \leq t \leq t_r.$$

For $t > t_r$ we use Equation (4.3) in combination with the initial conditions $n(t_r) = n_0$ and $c(t_r) = c_r$ to obtain:

$$n(t) = n_0 - \frac{\beta}{\alpha}[c(t) - c_r], \quad t > t_r,$$

resulting in the following solution for the amount of mitochondria at time t:

$$n(t) = \begin{cases} n_0 & 0 \leq t \leq t_r \\ n_0 - \frac{\beta}{\alpha}[c(t) - c_r] & t > t_r. \end{cases} \tag{4.4}$$

This shows that the number of mitochondria does not grow during the repair stage and is proportional to the oxygen concentration for $t > t_r$.

Substituting this result into Equation (4.1) we obtain the following equation for the oxygen consumption during the repair stage

$$\frac{dc}{dt} = -\alpha n_0 c, \quad c(0) = c_0$$

with solution

$$c(t) = c_0 e^{-\alpha n_0 t} \qquad \text{for } 0 \leq t \leq t_r,$$

attaining the value $c_r = c_0 e^{-\alpha n_0 t_r}$ at the end of this stage. For $t > t_r$ substitution results in the standard logistic equation:

$$\begin{aligned}
\frac{dc}{dt} &= -\alpha \left[n_0 - \frac{\beta}{\alpha}(c - c_r) \right] c \\
&= -\alpha n_0 c + \beta c^2 - \beta c_r c \\
&= \beta c \left(c - c_r - \frac{\alpha n_0}{\beta} \right) \\
\frac{dc}{dt} &= \beta c (c - \tilde{c}), \quad c(t_r) = c_r, \qquad (4.5)
\end{aligned}$$

where $\tilde{c} = c_r + \alpha n_0 / \beta$. The solution of this equation can be derived using separation of variables again and equals

$$c(t) = \frac{c_r(\beta c_r + \alpha n_0)}{\beta c_r + \alpha n_0 \; e^{(\beta c_r + \alpha n_0)(t - t_r)}}, \quad t > t_r. \qquad (4.6)$$

See Appendix A for the derivation. So for the oxygen concentration at time t we now have:

$$c(t) = \begin{cases} c_0 e^{-\alpha n_0 t} & 0 \leq t \leq t_r \\ \dfrac{c_r(\beta c_r + \alpha n_0)}{\beta c_r + \alpha n_0 \; e^{(\beta c_r + \alpha n_0)(t - t_r)}} & t > t_r. \end{cases} \qquad (4.7)$$

Seeds with high concentrations of mitochondria are able to deplete oxygen to a zero level, but seeds with a lower concentration of mitochondria only deplete oxygen to a certain minimal level $c_{\min}$ [11]. As the oxygen pressure approaches this level the oxygen consumption and multiplication of mitochondria slow down and eventually stop. The lowest level at which the seeds start to reduce respiration, and thus metabolism, as a response to the lack of oxygen is generally referred to as the COP (critical oxygen pressure) level, see Figure 4.1.

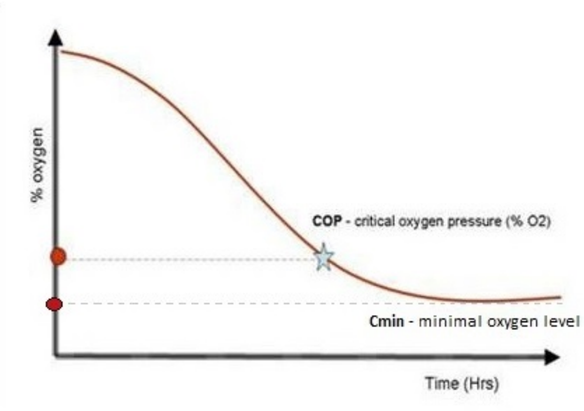


Figure 4.1: Graph showing $c_{\min}$ and the COP level.[2]

Also, the measurements may contain a systematic error due to poor calibrations. By substituting $c - c_{\min}$ for c in the above equations, these effects are taken care of. The final formula that can be used to approximate the measured data becomes:

$$c(t) = \begin{cases} c_{\min} + (c_0 - c_{\min})e^{-\alpha n_0 t} & 0 \le t \le t_r \\ c_{\min} + \dfrac{(c_r - c_{\min})[\beta(c_r - c_{\min}) + \alpha n_0]}{\beta(c_r - c_{\min}) + \alpha n_0 \; e^{(\beta(c_r - c_{\min}) + \alpha n_0)(t - t_r)}} & t > t_r. \end{cases} \tag{4.8}$$

with $c_r = c_{\min} + (c_0 - c_{\min})e^{-\alpha n_0 t_r}$.

Using data fitting we are able to determine the unknown parameters in Equation (4.8). The curve of relative mitochondrial population growth in a closed test tube is then simply found by simply evaluating

$$\frac{n(t)}{n_0} = \begin{cases} 1 & 0 \le t \le t_r \\ 1 - \dfrac{\beta}{\alpha n_0}(c(t) - c_r) & t > t_r, \end{cases} \tag{4.9}$$

which turns into

$$\frac{n(t)}{n_0} = \begin{cases} 1 & 0 \le t \le t_r \\ 1 + \dfrac{\beta}{\alpha n_0}(c_r - c_{\min}) & \\ -\dfrac{\beta}{\alpha n_0}\dfrac{(c_r - c_{\min})[\beta(c_r - c_{\min}) + \alpha n_0]}{\beta(c_r - c_{\min}) + \alpha n_0 \; e^{(\beta(c_r - c_{\min}) + \alpha n_0)(t - t_r)}} & t > t_r \end{cases} \tag{4.10}$$

after substituting our expression for $c(t)$.

We have now obtained the equation that can be used to fit our data but before moving on to the fitting, we take a closer look at the definition of one of the most important seed quality parameters, the germination time.

4.2 Germination time

Currently the Relative Germination Time (RGT) is used as a parameter to make inferences about the actual germination time of a seed. The value of this RGT is determined by finding the point of fastest decline rate in the oxygen concentration curve and extrapolating the tangential line at that point until it crosses the time axis [2], as is shown in Figure 4.2.

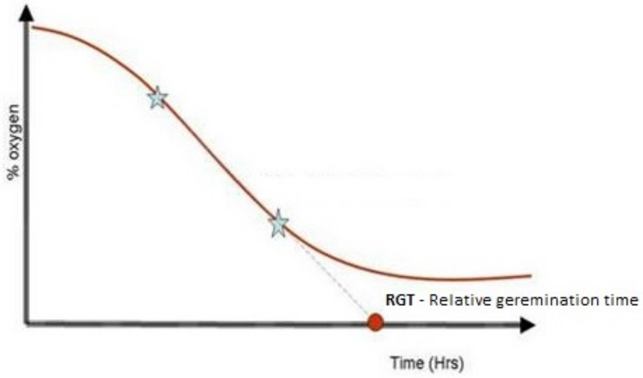


Figure 4.2: Graph showing the RGT value.[2]

From a biological view the point of fastest decline must lie in the growth stage of the graph and can be computed by setting the second derivative of $c(t)$ equal to zero:

$$\begin{aligned}\frac{dc}{dt} &= \beta(c - c_{\min})[(c - c_{\min}) - (\tilde{c} - c_{\min})] \\ \frac{d^2c}{dt^2} &= 2\beta(c - c_{\min}) - \beta(\tilde{c} - c_{\min}) = 0,\end{aligned}$$

so

$$2\beta(c(t) - c_{\min}) = \beta(\tilde{c} - c_{\min}).$$

Substituting the expressions for $c(t)$ and $\tilde{c}$ we find:

$$\begin{aligned}\frac{2\beta(c_r - c_{\min})[\beta(c_r - c_{\min}) + \alpha n_0]}{\beta(c_r - c_{\min}) + \alpha n_0\ e^{[(\beta(c_r - c_{\min}) + \alpha n_0)(t - t_r)]}} &= \beta(c_r - c_{\min}) + \alpha n_0 \\ \frac{2\beta(c_r - c_{\min})}{\beta(c_r - c_{\min}) + \alpha n_0\ e^{[(\beta(c_r - c_{\min}) + \alpha n_0)(t - t_r)]}} &= 1 \\ \alpha n_0\ e^{[(\beta(c_r - c_{\min}) + \alpha n_0)(t - t_r)]} &= \beta(c_r - c_{\min}) \\ e^{[(\beta(c_r - c_{\min}) + \alpha n_0)(t - t_r)]} &= \frac{\beta}{\alpha n_0}(c_r - c_{\min}) \\ [(\beta(c_r - c_{\min}) + \alpha n_0)(t - t_r)] &= \log\left(\frac{\beta}{\alpha n_0}(c_r - c_{\min})\right) \\ t - t_r &= \frac{\log\left(\frac{\beta}{\alpha n_0}(c_r - c_{\min})\right)}{(\beta(c_r - c_{\min}) + \alpha n_0)}.\end{aligned}$$

So the time coordinate t_m of this point of maximum decline is:

$$t_m = t_r + \frac{\log\left(\frac{\beta}{\alpha n_0}(c_r - c_{\min})\right)}{\beta(c_r - c_{\min}) + \alpha n_0} \tag{4.11}$$

and the corresponding oxygen concentration $c(t_m)$ is

$$c(t_m) = c_{\min} + \frac{1}{2\beta}[\beta(c_r - c_{\min}) + \alpha n_0].$$

The tangential line drawn at this point has a slope of

$$\left.\frac{dc}{dt}\right|_{t=t_m} = \beta[c(t_m) - c_{\min}][c(t_m) - \tilde{c}]$$

and is therefore given by

$$y(t) = c(t_m) + \beta[c(t_m) - c_{\min}][c(t_m) - \tilde{c}](t - t_m).$$

This line will cross the time axis at t_g with $y(t_g) = 0$, or:

$$\begin{aligned}c(t_m) + \beta[c(t_m) - c_{\min}][c(t_m) - \tilde{c}](t_g - t_m) &= 0 \\ (t_g - t_m) &= \frac{c(t_m)}{\beta[c(t_m) - c_{\min}][\tilde{c} - c(t_m)]} \\ t_g &= t_m + \frac{c(t_m)}{\beta[c(t_m) - c_{\min}][\tilde{c} - c(t_m)]} \\ &= t_m + \frac{2}{\beta(c_r - c_{\min}) + \alpha n_0} + \frac{4\beta c_{\min}}{(\beta(c_r - c_{\min}) + \alpha n_0)^2}\end{aligned}$$

which can also be written as:

$$t_g = t_r + \frac{2 + \log\left[\frac{\beta}{\alpha n_0}(c_r - c_{\min})\right]}{\beta(c_r - c_{\min}) + \alpha n_0} + \frac{4\beta c_{\min}}{(\beta(c_r - c_{\min}) + \alpha n_0)^2}. \tag{4.12}$$

We can conclude that this expression associated to germination time depends on all parameters of the problem in a rather complicated way. One thing that's clear is the fact that the germination is largely dependent on the repair time t_r. After calibration we've obtained all the unknown parameters and are able to calculate this estimated germination time for each seed.

5 Fitting the experimental data

Equation 4.8 describes the oxygen consumption during the two-stage germination process. For analyzing the data and its fits it is more convenient to consider the *normalized* equation, obtained by substituting $\frac{c(t)}{c_0}$ for $c(t)$:

$$\tilde{c}(t) = \begin{cases} \tilde{c}_{\min} + (1 - \tilde{c}_{\min})e^{-\alpha n_0 t} & 0 \leq t \leq t_r \\ \tilde{c}_{\min} + \dfrac{(\tilde{c}_r - \tilde{c}_{\min})[\beta(\tilde{c}_r - \tilde{c}_{\min}) + \alpha n_0]}{\beta(\tilde{c}_r - \tilde{c}_{\min}) + \alpha n_0 \; e^{(\beta(\tilde{c}_r - \tilde{c}_{\min}) + \alpha n_0)(t - t_r)}} & t > t_r, \end{cases} \tag{5.1}$$

with normalized parameters $\tilde{c} = \frac{c}{c_0}$, $\tilde{c}_{\min} = \frac{c_{\min}}{c_0}$ and $\tilde{c}_r = \frac{c_r}{c_0} = \tilde{c}_{\min} + (1 - \tilde{c}_{\min})e^{\alpha n_0 t_r}$. To make sure c_0 remains a tuning parameter in our fitting algorithm, we implement the above equation as

$$c(t) = c_0 \cdot \tilde{c}(t). \tag{5.2}$$

Otherwise all curves are forced to start at the same point as the data curves, which hardly ever results in the best fit because the data fluctuates a lot. From now on the data set will be referred to as (x, y) with x the vector with time coordinates and y the vector with oxygen concentrations. After fitting the data to the model function $c(t)$, the normalized results are obtained by dividing both the data and the estimated result by y_0, the initially measured oxygen concentration.

The oxygen consumption is thus described by Equations (5.2) and (5.1) in terms of the following five parameters:

1. c_0 - the initial oxygen concentration
2. $\tilde{c}_{\min}$ - the normalized minimal oxygen concentration $c_{\min}/c_0$
3. αn_0 - the product of the initial amount of mitochondria n_0 and the (average) oxygen consumption rate α of a single mitochondrion
4. β - the (average) mitochondrial reproduction rate
5. t_r - the duration of the repair stage

Since we are dealing with a function which is a nonlinear combination of the model parameters, we use a nonlinear least squares optimization algorithm to fit our observational data.

5.1 Nonlinear Least Squares

The method of nonlinear least squares is used to fit a set of m data points to a presumed nonlinear model with n unknown parameters ($m > n$). Its goal is to find parameters such that the model function fits the data points best. The basic approach is to approximate the model function by a linear one and to iteratively update the parameters.

We consider a set of m data points, $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ and a model function $f(x, \beta)$ where β is the vector of $n < m$ parameters, $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$. The goal is to find the parameters β that minimize the sum of squared residual errors r_i, i.e. the errors between the data points and the corresponding values predicted by the model function:

$$\begin{aligned} \arg\min_{\beta} \|R\|_2^2 &= \arg\min_{\beta} \sum_{i=1}^{m} r_i^2 \\ &= \arg\min_{\beta} \sum_{i=1}^{m} [f(x_i, \beta) - y_i]^2 \end{aligned} \tag{5.3}$$

The necessary condition for obtaining a minimum is that the gradient must be equal to zero [12], hence:

$$\nabla \left(\sum_{i=1}^{m} r_i^2 \right) = \sum_{i=1}^{m} \nabla \left(r_i^2 \right) = 2 \sum_{i=1}^{m} r_i \frac{\partial r_i}{\partial \beta_j} = 0 \qquad \text{for } j = 1, \cdots, n, \tag{5.4}$$

which can also be written as

$$2R^T J_R = 0, \tag{5.5}$$

with J_R is the Jacobian of R, $(J_R)_{ij} = \partial r_i / \partial \beta_j$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. All critical points meet this condition so to be sure that we are in fact dealing with a minimum, the Hessian, $\nabla^2 \left(\sum_{i=1}^m r_i^2 \right)$, must be positive definite in that critical point.

In a nonlinear system, the Jacobian is still dependent on β and Equation (5.5) has no closed-form solution in general. We start with an initial guess for the parameters and iteratively refine them to obtain a global minimum. There exist many iteration methods all with their own pros and cons. In this thesis the Curve Fitting Toolbox in `Matlab` is used for implementation of the nonlinear least squares algorithm. This toolbox has two algorithms, a trust-region-reflective algorithm and the Levenberg-Marquardt algorithm. `Matlab` recommends to use the trust-region-reflective algorithm if there are no constraints or only bound constraints and the Levenberg-Marquardt algorithm if the problem is underdetermined ($m < n$). Hence we chose the first of the two.

5.1.1 Trust-Region-Reflective algorithm

The basic idea of the trust-region approach is to approximate the function you want to minimize by a simpler function q, which reasonably reflects the behavior of the original function in a certain neighborhood N around the point β^k [13]. This neighborhood is the trust region. A trial step $\Delta\beta$ is computed by (approximately) minimizing over N. This is called the trust-region subproblem:

$$\arg\min_{\Delta\beta} \{ q(\beta^k + \Delta\beta), \ \Delta\beta \in N \}$$

The current point is updated to be $\beta \approx \beta^{k+1} = \beta^k + \Delta\beta$ if $f(\beta^k + \Delta\beta) < f(\beta^k)$; otherwise the current point remains unchanged and N, the trust region, is shrunk and the trial step computation is repeated. The key steps in this approach are choosing and computing the approximation q, choosing and modifying the trust region N, and accurately solving the trust-region problem.

In the standard trust-region method, the minimization function is approximated by a second-order Taylor series expansion about β^k. Our minimization function is of the form $r_i = f(x_i, \beta) - y_i$, where f is our model function with the following Taylor series expansion:

$$\begin{aligned} f(x_i, \beta) &\approx f(x_i, \beta^k) + \sum_j \frac{\partial f(x_i, \beta^k)}{\partial \beta_j} \left(\beta_j - \beta_j^k \right) + \frac{1}{2} \sum_j \sum_k \frac{\partial^2 f(x_i, \beta^k)}{\partial \beta_j \partial \beta_k} \left(\beta_j - \beta_j^k \right) \left(\beta_k - \beta_k^k \right) \\ &\approx f(x_i, \beta^k) + \sum_j J_{ij} \Delta\beta_j + \frac{1}{2} \sum_j \sum_k \Delta\beta_j \Delta\beta_k H_{jk(i)} \\ &\approx f(x_i, \beta^k) + \Delta\beta^T \nabla f(x_i, \beta^k) + \frac{1}{2} \Delta\beta^T H_i \Delta\beta. \end{aligned}$$

Also the neighborhood N is usually spherical or ellipsoidal in shape, which means that our minimization problem (5.3) has the following corresponding trust-region subproblem:

$$\arg\min_{\beta, \, \|\Delta\beta\|_2 \leq \delta} \left\{ \left\| \Delta y + \Delta\beta^T J + \frac{1}{2} \Delta\beta^T H \Delta\beta \right\|_2^2 \right\} \tag{5.6}$$

with $\Delta y = f(x, \beta^k) - y$, J the Jacobian of f, H the Hessian of f and $\delta > 0$.

There are many different algorithms for solving this trust-region subproblem. Such algorithms typically involve computation of a full eigensystem, which requires $\mathcal{O}(n^3)$ operations and can take a lot of time. Therefore the approach followed in the Curve Fitting Toolbox is to restrict the subproblem to a two-dimensional subspace S. Once the subspace S has been computed, finding a solution to the minimization problem (5.6) is trivial since the problem is only two dimensional. The two-dimensional subspace S is determined with the aid of the preconditioned conjugate gradient process described below. The solver defines S as the linear space spanned by s_1 and s_2, where s_1 is in the direction of the gradient of $(f(x_i,\beta)-y_i)^2$ and s_2 is either an approximate Newton direction, i.e. a solution to $Hs_2=-g$ with g the gradient, or a direction of negative curvature, $s_2^THs_2<0$. The philosophy behind this choice of S is to force global convergence (via the steepest descent direction or negative curvature direction) and achieve fast local convergence (via the Newton step, when it exists).

Preconditioned conjugate gradients methods are popular for solving large positive definite systems of linear equations $Hp=-g$, where H is symmetric. This iterative approach requires the ability to calculate matrix-vector products of the form Hv where v is an arbitrary vector. The symmetric positive definite matrix $M=EE^T$ is a preconditioner for H, i.e. a matrix that approximates H but is easier to invert, with $M^{-1}H$ a well-conditioned matrix or a matrix with clustered eigenvalues. The system is indirectly solved by

$$M^{-1}Hp=-M^{-1}g$$

or

$$E^{-1}H(E^{-1})^T\hat{p}=-E^{-1}g, \qquad \hat{p}=E^Tp,$$

which is solved for $\hat{p}$ first, then for p. This preconditioning technique is necessary to ensure fast convergence of the conjugate gradient method. The method starts by using g, he gradient of the residual, as a search direction. Each new search direction is constructed (from the residual) to be A-orthogonal to all the previous residuals and search directions. The algorithm exits when a direction of negative curvature is encountered, i.e., $p^THp<0$. The output direction, s_2, is either a direction of negative curvature or an approximate solution to the newton system $Hs_2=-g$ as mentioned above.

The trust-region-reflective algorithm can thus be summarized by the following steps [13]:

1. Determine the two-dimensional subspace with the aid of the preconditioned conjugate gradient method and formulate the two-dimensional trust-region subproblem.
2. Solve the minimization problem (5.6) to determine the trial step $\Delta\beta$.
3. If $\|r(x,\beta+\Delta\beta)\|_2<\|r(x,\beta)\|_2$, set $\beta=\beta+\Delta\beta$.
4. Adjust δ to reduce the size of the trust region N.

These steps are repeated until convergence.

These are the basics of the nonlinear least squares optimization algorithm programmed in `Matlab`. The following subsection discusses the implementation and results of this fitting algorithm applied to one data batch.

5.2 Implementation

The initial implementation of this algorithm is done with the second batch of barley seeds. This batch of 1520 seeds has the largest observation time and therefore the most data points as well. It is beneficial to have as many data points available as possible in order to get better and more reliable fits. We discuss our findings and some basic results before we apply the algorithm on the entire batch of barley data.

Since the algorithm is iterative choosing the initial parameters β^0 could be of great importance to the outcome of the model function. Especially when there are lots of local minima on the error surface of the model function, the choice of initial parameters is crucial. To avoid getting trapped in a local minimum and to ensure fast convergence of the iterative procedure the initial values should be as close as possible to the global minimum. This can be achieved by a grid search in the parameter space or by making some assumptions leading to a rough guess.

Since the first is tricky for complex model functions with many parameters we apply the latter. In case of an unsatisfactory fit different initial parameter values are chosen and this process is repeated until the quality of the fit is sufficient. To be confident that the minimum found is in fact the global minimum and not some local minimum one should widely differ the initial parameter values. When the same minimum is found regardless of these initial values, it is likely to be the global minimum.

Furthermore the range of parameters should also be considered. Implementing limits of the parameters avoids futilely large or infinitesimally values and expedites the fitting process. Therefore we have introduced some bounds on the five fitting parameters mentioned at the start of this section. Their lower bound is zero since all parameters are defined non-negative and the upper bounds are determined after the initial implementation without upper bounds.

Despite the fact that α and n_0 only occur as a product and therefore aren't regarded as individual parameters, previous research showed that it is mathematically advantageous to see them as such [14]. The extended six-parameter minimization problem proved to be easier to solve than the original five-parameter problem. Therefore we chose to directly implement the six-parameter formulation. However, the code in the previous research was written in Python and as it turns out, the implementation in `Matlab` shows no advantage in using the six-parameter versus the five-parameter model. In fact the fitting results are exactly the same under both models, which means that either implementation could be used.

In the process of determining adequate initial parameter values, we've also observed the minimization algorithm is quite sensitive to the initial choice of the repair time t_r. As we discussed above, this could indicate there are (lots of) local minima for t_r. To avoid ending up in such a local minimum we simply run the algorithm with two initial guesses for t_r, one close to zero and the other around 20 hrs. Comparing the resulting standard deviations of the residuals, i.e. the square root of the mean squared difference between the data and the fitted model (Root Mean Squared Error (RMSE) in the goodness-of-fit statistics), we could choose the better of the two fits, always resulting in a robust (global) minimum.

Implementing a simpler single-stage model which neglects the repair process ($t_r = 0$) shows the repair process is an essential aspect of our model. In Figure 5.1, three typical results for barley seeds are shown together with both the two-stage and one-stage model approximation. The fact that the one-stage model provides poorer fits is abundantly clear, especially for the curves of seeds with a longer estimated repair time. Standard deviations of both models are computed for each seed to quantify and confirm the benefits of the two-stage model. The two-stage model shows both a smaller mean of the deviations and more concentration about this mean, as can be seen from the density plots of the standard deviations for a batch of 1520 barley seeds in Figure 5.2.

Discussing the implementation resulted in a six-parameter implementation with clear importance of the repair phase of duration t_r. The finally obtained adequate initial values and upper and lower bounds for our fitting parameters are:

Parameter	c_0	$\tilde{c}_{\min}$	α	n_0	β	t_r
initial value	29	0.1	0.06	0.6	0.02	$0.1 \vee 20$
lower bound	0	0	0	0	0	0
upper bound	$\max(y_0)$	1.1	1.5	1.5	20	160

The tolerance is set to 10^{-8} meaning that the solver stops if the change in the residual or parameter value from one iteration to the next is less than 10^{-8}. These settings result in high-quality fit. The quality of a fit is usually determined by the goodness-of-fit statistic R-squared which is defined as the ratio of the expected variance (model variance) to the total variance:

$$\text{R-square} = \frac{\sum_i (f_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the observed data. R-squared can take on any value between 0 and 1 with a value closer to one indicating a better approximation of the data by our model function. For example, an R-squared value of 0.8982 means that the fit explains 89.82% of the total variation in the data about the mean. In this case the mean R-squared value is larger than 0.99.

The `Matlab` code for this implementation of the nonlinear least suqares optimization method is shown in Appendix B. In the next subsection some results of this implementation are discussed.

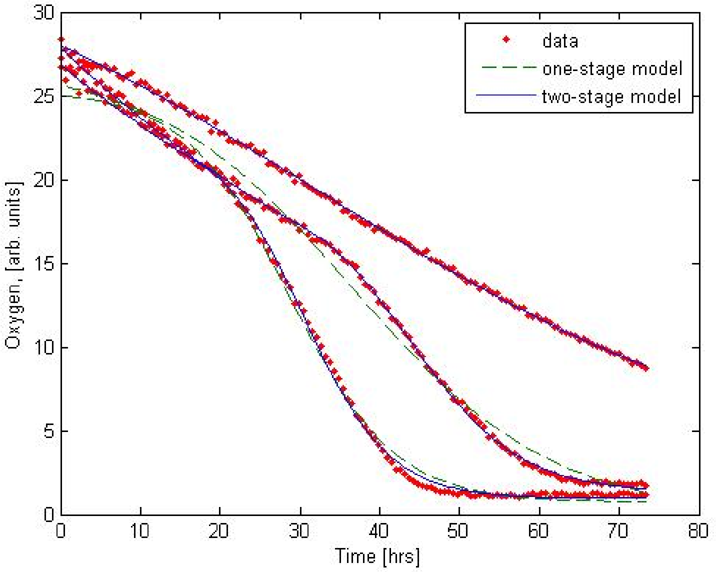


Figure 5.1: Oxygen consumption data for three barley seeds and the results fitting using a one-stage and two-stage model. The upper curve has a small repair time and the lower curves show larger repair times.

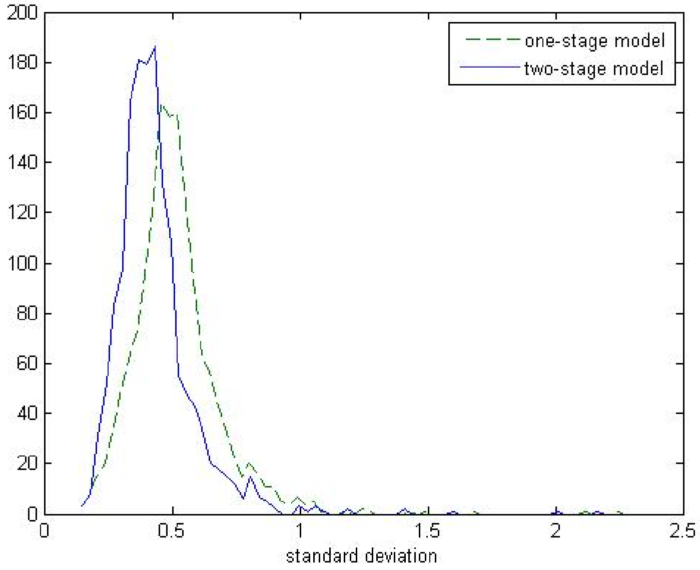


Figure 5.2: Histograms of the standard deviations of the two models showing the superiority of the two-stage model.

5.3 Initial results and discussion

Data fitting provides interesting statistics about the seeds and further insights into the model and the associated minimization algorithm. In this subsection we take a look at the resulting fits and the associated parameter values to find relations between the two that could help determine which seed properties cause slow/fast germination.

To show how inhomogeneous the seeds in a batch are, oxygen consumption curves and resulting fitted curves for the complete batch are presented in Figure 5.3, where the normalized data is shown next to the fitted curves. Compared to the non-normalized results it is much easier to compare the curves as they are no longer dependent on the measured initial oxygen concentration, which varies widely in our seed batch. From both the original and the normalized results one can conclude the shape of the oxygen consumption curves differ a lot. Parameter statistics could be helpful in determining the cause of these differences.

Some of the data curves in Figure 5.3 contain a lot of noise which makes the resulting fits less reliable. Therefore a method is proposed in the next section to filter these noised curves and all other curves that deliver a bad fit. Filtering these curves leads to more reliable parameter statistics and hence helps with better understanding the seed data.

Based on the clear gap between the curves in the graph it seems like the seed batch can be divided into two classes. The first class contains all seeds corresponding to the oxygen curves at the top of the graph, i.e. curves that do not reach a normalized oxygen value below 0.7 in the given observation time. Seeds in this class clearly do not germinate during the observation time, so, based on the definitions given in Section 2 and 3, we expect this class to consist of both dead and dormant seeds. From a biological view we expect this class to show large repair times small growth parameters. The second class contains the seeds corresponding to all the other curves, which are both germinating and non-germinating seeds. Here the non-germinating seeds can be subdivided into dormant seeds and weak germinators, causing the dormant seeds to be spread over two classes. As mentioned in Section 2, we are able to distinguish between germinating and non-germinating seeds by the shape of their oxygen consumption curve, but since the curve shape depends on multiple parameters, this separation is hard to accomplish based on these parameter values. Perhaps the value of c_{end}, the normalized oxygen concentration measured at the end of the observation time, will also come in handy to separate these two types of seeds. Whether or not the value of c_{end} is a valid parameter to separate the potential classes will be discussed after treating the parameter statistics and other observations.

Histograms of the three main parameters - $\alpha n_0, \beta$ and t_r - are shown in Figure 5.4. The parameter c_0 is almost completely determined by the highest measured oxygen level and therefore its meaning is quite clear. So is the meaning of c_{min}, the minimum oxygen level, but this value is not completely determined by the measurements as it is possible that seeds do not reach this level in the given observation time. Therefore we do consider this as a meaningful parameter, but since the results are highly related to the lowest measured oxygen level we do not show its histogram.

The distribution of αn_0 seems to resemble a normal distribution, which could suggest the variation in this parameter is caused by a large number of statistically independent factors. However, statistical tests (discussed in Section 6.1.1) reject this hypothesis, meaning that αn_0 is not normally distributed. Both β and t_r contain several outliers which means that their distributions are both skewed to the right. Looking at the values around the mode of the parameter distributions, Figure 5.5 we see that β indeed shows more observations on the left side of the histogram and t_r shows a relatively large number of seeds with extremely short repair times.

From a biological point of view a small repair time means that the seeds start growing almost immediately which should result in oxygen consumption curves that start descending right away. But taken separately these curves turn out to be quite ordinary compared to the total batch. This could indicate the extremely small values arise from a local minimum and are not the parameter values consistent with a global minimum. Hence we followed the procedure explained in the previous subsection and widely differed the initial values for t_r. Despite the fact that the resulting histograms for t_r are slightly different we keep ending up with a spike for extremely small values and thereby we can conclude that the small t_r values are no artifact of minimization. There are some remarks to be made about the curves corresponding to these small repair times. Firstly, none of the curves are part of the suspected class of dead/dormant seeds discussed earlier, which means that our fit still makes some sense biologically speaking. Secondly, the parameter β only takes on values from the left hand side of the histogram shown in Figure 5.5 which could indicate some kind of relation between β and t_r. Looking at the model equation, Equation (4.8), we see that β and $(c_r - c_{\min})$ always appear as a product. So reduction in t_r, which leads to the growth of the factor c_r, may indeed be compensated by a smaller factor β. Unfortunately neither these two nor the remaining three parameters show any significant mutual correlation. Lastly, during the analysis we could distinguish different shapes of curves much easier than when we considered all the curves from one batch. Looking at the parameter values it turned out that the parameters $c_{\min}$ and especially β play a significant role in the eventual shape of the curve. From a biological view one expects a faster descending curve for a larger β and the lower the oxygen level this curve is descending to is, the faster this minimum value is reached. These expectations were shown to be true for these small values of t_r but have to be verified for all values of t_r before we draw any conclusions. For the complete batch the same kind of results were obtained with respect to the influence of β. The expected influence of $c_{\min}$ however was not reflected by the results as seeds with approximately the same values for β but smaller values for $c_{\min}$ did not reach this minimum faster in all cases. However the shape of the curves if affected by $c_{\min}$ since $c_{\min}$ is highly related to the lowest measured oxygen level which on its turn determines where the curve will end. Results of the different graph shapes are shown in Figure 5.6.

As can be further observed from the rightmost histogram of Figure 5.5, neglecting the extremely small values, the main peak value of the repair time t_r is around 12 hours. This result is in agreement with the average time it takes mitochondria to develop *cristae* - folds of the inner membrane characteristic of mature mitochondria (Carrie et al, 2012). This indicates the repair process of a seed ends with the presence of full grown cristae and that these cristae are crucial in the oxygen consumption, which makes sense given the fact that cristae increase the surface of the inner mitochondrial membrane and thus the efficiency of oxygen uptake.

So far we haven't discussed the outlying parameter values, but understanding where these outliers originate from is an important aspect of understanding our model and the seed data. Firstly, we look at the outliers of t_r, or, specifically, all curves with $t_r > 50$. Next to the expected large amount of dormant/dead seeds we encounter several seeds with differently shaped curves. These seeds consume much more oxygen during their long repair phase and seem to start growing rapidly afterwards which is reflected by their large values for the growth parameter β. The transition from the repair stage to the growth stage is indicated by a 'bump' in the oxygen consumption curve. Is it likely that these seeds are possible of germination but mostly remain in the repair phase too long to complete germination in the given time. Therefore they are not actually dormant, but they do not seem to germinate in the given time either. How we classify these and other non-dead and non-dormant seeds depends on our interpretation of the definition of germination in terms of oxygen consumption curves. Is a seed for instance germinated if it has depleted oxygen to almost its own minimal level, or does germination require a depletion of oxygen below a certain fixed level? These classification issues will be discussed at the end of this section.

Secondly, we consider the outliers of β, i.e. all curves with $\beta > 0.5$. These curves are shaped just like the ones with large β we just encountered which means that this specific shape is more determined by the value of β than by the value of t_r. Once more the prominent influence of β on the curve shape is made clear. Data and fitted curves corresponding to a selection of the outliers are shown in Figure 5.7. Some of the data curves corresponding to large β values show a lot of noise or weird jumps and therefore will be filtered from our seed batch which indicates the importance of filtering to the reliability of our parameter statistics

So far no clear correlation between parameters has been found. Dividing the batch into two or more classes based on seed/curve properties could shed more light on which parameter values and combinations of them lead to a slow/fast germination and whether there is any correlation between these parameters within classes. In the beginning of this subsection c_{end} was proposed as a classification criterion with a clear mark at 0.7 for this seed batch, separating the dead and part of the dormant seeds from the rest of the seeds. Our expectation that this class contains large values for t_r and small values for β turns out to be true with a few exceptions, but since there are only 35 seeds in this class we can't draw any general conclusions. If such a clear mark for a large value of c_{end} also shows in the rest of the data batches and the associating parameters show the same results, this classification is solid. Note however that the other data batches have smaller observation times, meaning that their c_{end} is defined as the normalized oxygen concentration at a different, smaller time t_{end}. This results in overall higher values for c_{end}, causing the suspected mark separating the dead (and dormant) seeds from the rest of the seeds to be higher as well. Whether such a mark exist and if so at which level, will be discussed in the next section. That leaves us to the rest of the seeds in this batch. Although they all deplete oxygen to a normalized level below 0.7, they will not all germinate in the giving time and we are left with a class that contains both germinating as non-germinating seeds. In addition, the latter consists of both dormant and weak germinating seeds. Hence further classification is desired. Our hypothesis is that c_{end} could play a role in this classification as well. Or, more specifically, we believe that there is a second mark, at some value of c_{end}, separating the germinating seeds, i.e. both the strong ones and the weak ones that do germinate in time, from the non germinating seeds, i.e. both the dormant and weak germinating seeds that don't finish germination in time. Based on this hypothesis our interpretation of the definition of germinating seeds becomes "germinating seeds are seeds that deplete oxygen below a certain level in the given time". Whether or not this hypothesis turns out to be true and what the concerning mark and corresponding time frame are will be discussed in the next section where we regard results on all barley data.

This section discussed the implementation and results of our model applied to a batch of 1520 barley seeds. In the previous paragraphs some observations with corresponding hypothesis or needed actions were posed. Firstly, it's necessary to filter the results from bad fits to obtain more reliable parameter statistics. Secondly we did not find any correlations between the fitting parameters but we believe classification might help resolve this. Our hypothesis is that c_{end} is a good selection criterion for dividing our seed batch into (at least) the three following classes:

1. Dead/highly dormant seeds
2. Non-germinating seeds
3. Germinating seeds

The parameter statistics of these classes have to be analyzed to detect possible (cor)relations and see if we can determine which properties cause slow/fast germination. The next section addresses both the filtering and the classification.

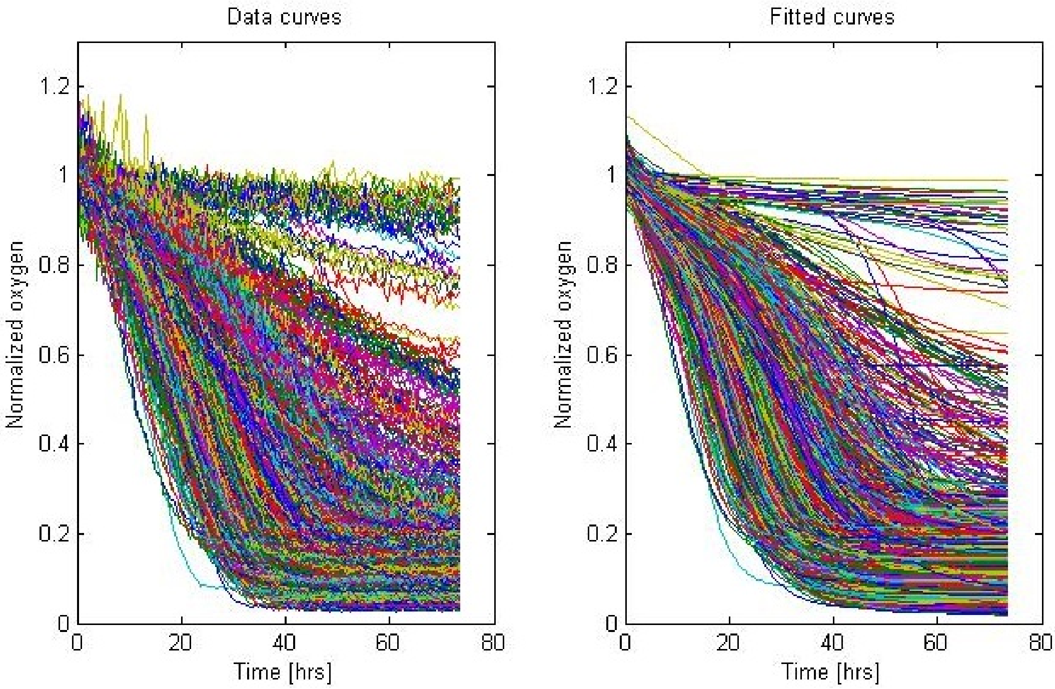


Figure 5.3: Normalized oxygen consumption data and fitting results for a batch of 1520 barley seeds.

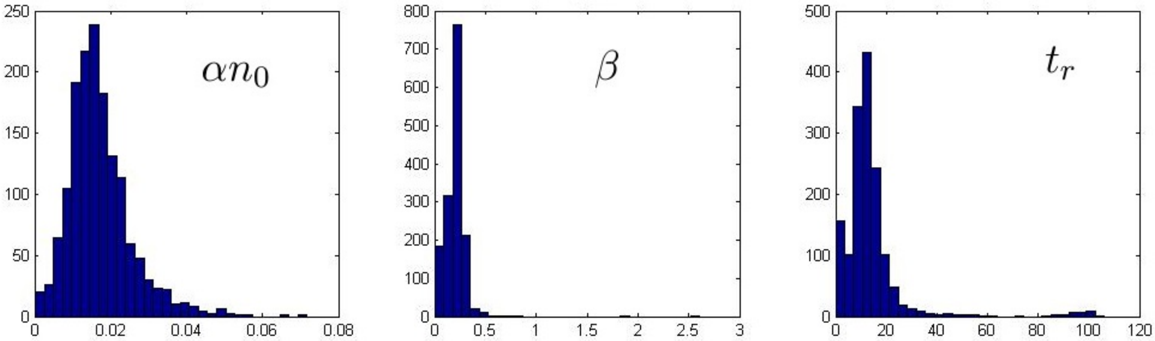


Figure 5.4: Histograms of the three main parameters αn_0, β and t_r obtained by fitting the model on a batch of 1520 barely seeds.

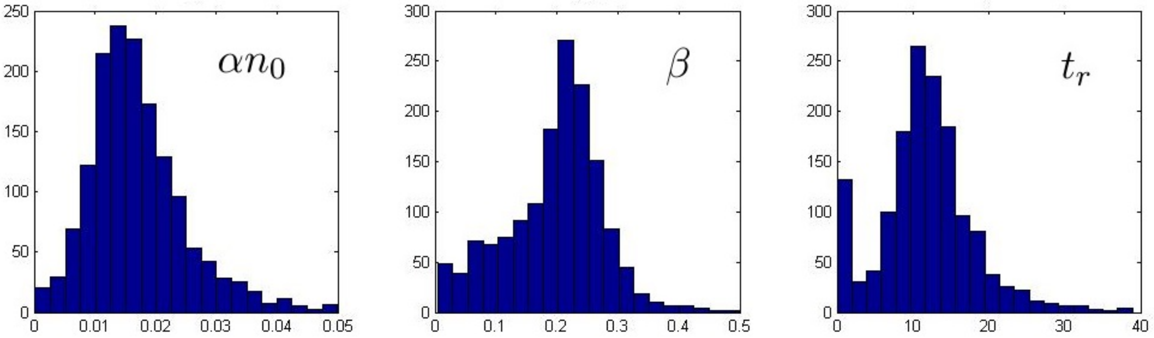


Figure 5.5: Zoomed-in histograms of the three main parameters αn_0, β and t_r.

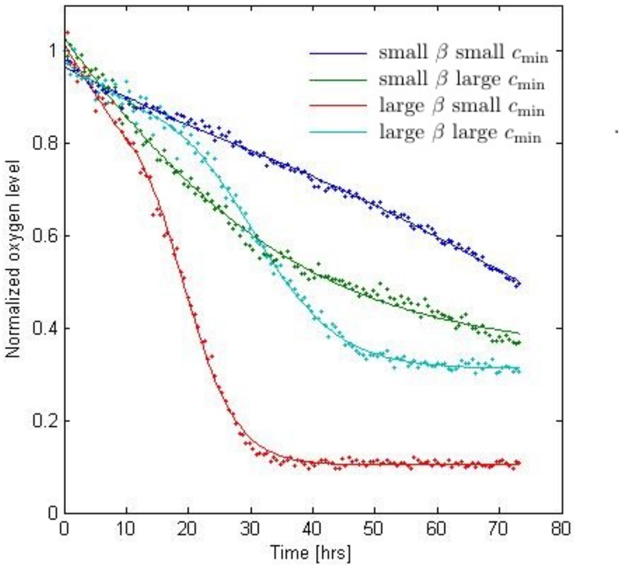


Figure 5.6: Different graph shapes showing the influence of β and $c_{\min}$.

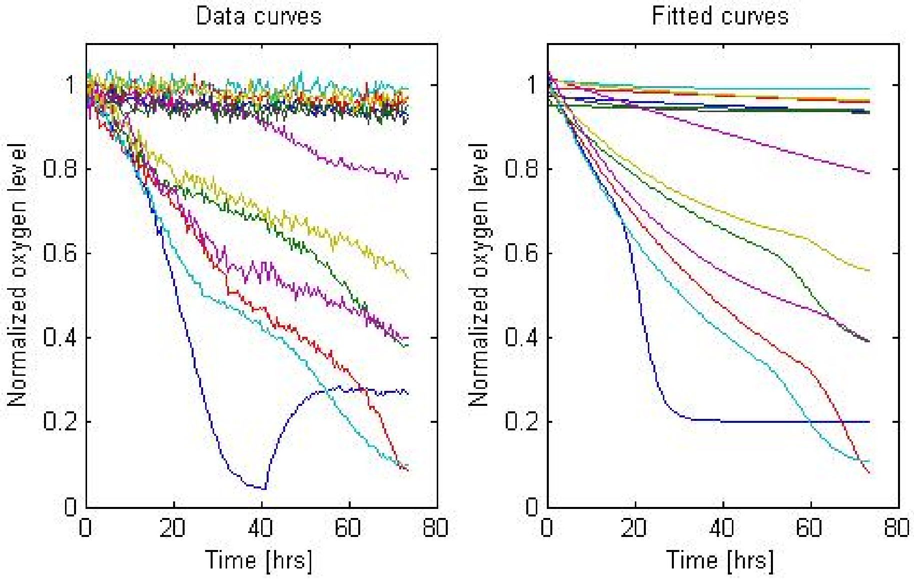


Figure 5.7: Oxygen consumption data and fitting curves corresponding to some of the extremely large values for t_r (at the top of the graph) and β (rest of the graph).

6 Results on all barley data

The fitting algorithm described in the previous chapter is applied to all available barley data, about 11.000 curves, resulting in the following parameter distributions:

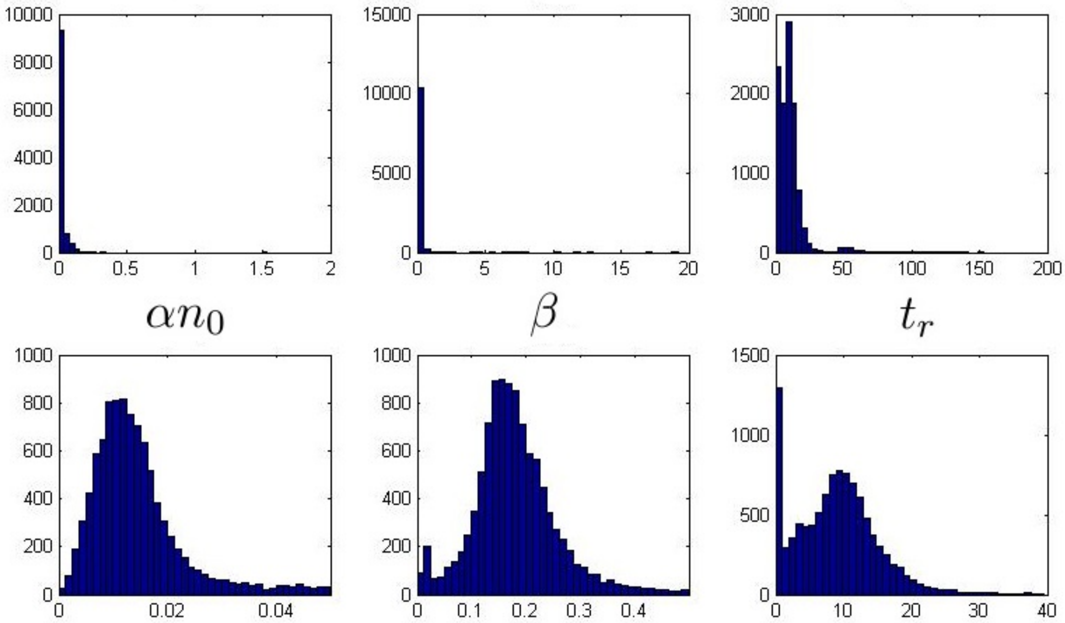


Figure 6.1: Histograms and zoomed-in histograms of the three main parameters αn_0, β and t_r obtained by fitting the model to the complete batch of 10.640 barley seeds.

Compared to the results on 1520 barley seeds all histograms are more right-skewed, especially those of αn_0 and β. The skewness of αn_0 even changed from 1.36 to 14.67. Around the mode all histograms have smoothed, as one would expect when increasing the sample size. This part of the histogram of β is more bell shaped than before and has an obvious spike for small values. For t_r the spike for small values already noted in the previous section, is even more significant and the histogram shows more observations to the left of the mode than before. Neglecting the extremely small values, the main peak is around 10 hours, which is a little smaller than the 12 hours discovered in the previous section. This could indicate that the cristae don't have to be full grown but just highly developed for the repair process to end.

Further observations about the parameter distributions will be made after filtering the bad fits and unreliable predictions from our obtained results. Subsequently we will treat the classification method as posed in the previous section.

6.1 Filtering

6.1.1 Filtering bad fits

By filtering the bad fits from our results we hope to obtain more reliable parameter statistics and perhaps get rid of some of the extreme values on the way. Bad fits are fits where the estimated values deviate too much from the original data. Therefore, in order to detect these bad fits, we have to look at the residuals. In case of a good fit, where the estimation is very close tot the original data, the residuals will be concentrated around zero. The worse the fit gets, the more large residuals will emerge and the larger the variance will become. Residuals concentrated around a point far away from zero, indicating that the errors tend to one side of the data curve, could also indicate a bad fit. The latter happens rarely as we will see in this section.

Detecting bad fits thus comes down to finding the fits with residuals that have a large variance, which is the same as a large RMSE value (see Section 5.2), or a large absolute mean. The variances of our complete batch are concentrated around 0.45 and vary up to a value of 79. A histogram of the biggest part of the values, up to the observations equal to two, is shown in Figure 6.2.

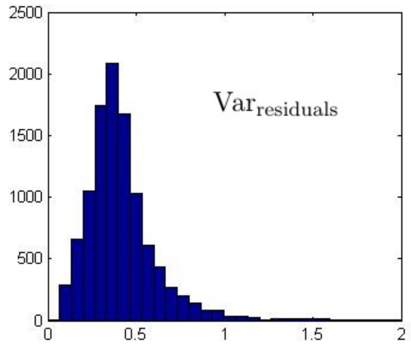


Figure 6.2: Histogram of the variance of the residuals showing a around the mode of 0.45.

Looking at the data curves compared to their fitted curve corresponding to large variances, we've concluded that twice the mean of all variances (≈ 0.9) is a good threshold for distinguishing between the good and the bad fits. This results in 384 bad fits, which is about 3.1% of the batch. The (absolute) means of the residuals are concentrated around zero and vary up to 0.18. All the fits belonging to large means that also visually give bad fits are already classified as bad fits based on the large variance of their residuals. Therefore we discard the mean of the residuals as an indication and only use the variance to detect bad fits.

In detecting the bad fits it became clear that there are three possible reasons for the fitting algorithm to deliver a bad fit:

1. The seed data contains a lot of *noise*.
2. The seed data contains weird *jumps*.
3. The seed data does not contain a lot of noise or weird jumps but the model delivers a bad fit anyway.

The latter means that the model equation is inadequate, i.e. the fitting algorithm is simply unable to estimate the observed data closely under the given model equation. Furthermore, the three reasons are not mutually exclusive.

Residuals have both a random component caused by for example measurement errors or noise, and a systematic component caused by for example inadequacy of the model equation. Assuming that the noise we're dealing with is in fact randomly distributed, the residuals of bad fits due to the first reason will deal with a large random component. The second and third reason on the other hand give rise to a larger systematic component in the residuals.

In cases like noise, a single measured value is often regarded as the weighted average of a large number of small effects. Using generalizations of the Central Limit Theorem, stating that the distribution of the sum of a large number of random variables will tend towards a normal distribution, we expect that this would often (though not always) produce a final distribution of residuals that is approximately normal. Therefore we expect it to be very likely that the distribution of the residuals of a good fit, with only a small random noise component, resembles a normal distribution with a mean close to zero and a very small variance. If a systematic error is introduced then we expect it to become less likely that the distribution will resemble a normal distribution. These expectations are shown to be true in the subsequent part of this section.

By looking at the curves and corresponding histograms of the residuals, shown in Figure 6.3, we can clearly see the difference between good and bad fits. (Note that the scale of the x-axis is smaller for the histogram of the good fit.) Additionally, the histogram of the good fit shows a resemblance to the normal distribution, as expected. To quantify this resemblance we use the Jarque-Bera (JB) test. This is a goodness-of-fit test of whether the sample data have the *skewness* and *kurtosis* matching a normal distribution. Skewness is a measure of the asymmetry of the probability distribution of a real valued random variable about its mean and kurtosis is a measure of the "peakedness" of the probability distribution of a real-valued random variable. Samples from a normal distribution have an expected skewness of zero and an expected kurtosis of 3. The test assumes that the data originates from a normal distribution with unknown mean and variance (the null hypothesis) and calculates the probability of obtaining the given data or even "worse" data, under this assumption. "Worse" in this case means even further away from the null hypothesis and closer to the alternative hypothesis that the data does not originate from a normal distribution. If this probability, also called p-value, is smaller than the significance level of 0.05 then the null hypothesis is rejected and we can assume that the underlying distribution is not a normal distribution. If the p-value is not less than 0.05 then the test has no result and we have insufficient evidence to conclude that the underlying distribution is indeed a normal distribution. However, the larger this p-value is, the more likely it becomes that our distribution resembles a normal distribution. The result for the good fit is a p-value of approximately 0.82 and corresponding skewness and kurtosis values of 0.0022 and 3.26 and therefore it is indeed likely that the residuals of a good fit resemble a normal distribution.

Comparing the histogram of the good fit to the histogram of the bad fit caused by a lot of noise, we observe a similar resemblance to the normal distribution. This is also reflected in the JB test statistics showing a p-value of 0.83 and a skewness and kurtosis of -0.0606 and 3.21 respectively. The fact that the distributions of this good and bad fit show resemblance is not strange considering the residuals of both fits arise mostly from noise in the data. The only difference is the amount of noise, which is reflected by a larger variance and doesn't change the shape of the distribution.

In the third bad fit a small overall systematic error is introduced caused by the inadequacy of our model. This systematic error increases (respectively decreases) the residuals that are already present due to noise. In the histogram this is reflected by a gap for residuals around zero, two spikes for values close to zero and a larger spread. While the skewness of the distribution remains close to zero, attaining a value of 0.0077, the kurtosis becomes much smaller than 3 due to the spikes, attaining a value of 2.19. Therefore it is less likely that this distribution resembles a normal distribution, as anticipated. This claim is supported by the JB test, resulting in a p-value of 0.11. The larger the overall systematic error is, the lower the kurtosis will become, eventually leading to rejection of the null-hypothesis. If the (large) systematic error tends to one side of the data, then the skewness will increase which will also lead to rejection of the null-hypothesis. This is the case with the second bad fit, showing a locally large systematic error caused by a weird 'jump' in the data. Though the resulting residuals have larger positive values, the amount of negative values is much higher, causing the distribution to be right-skewed. Skewness and kurtosis attain values of 0.8939 and 2.78 and the JB test rejects the hypothesis that this data originates from a normal distribution with a very small p-value of 0.0036.

We can conclude that histograms corresponding to either a good fit or a bad fit solely due to a large amount of noise, will likely resemble a normal distribution. This shows the beauty of data fitting: that a random measurement error, in this case noise, has no influence on the quality of the resulting fit. Therefore these fits can not in fact be qualified as bad fits and they will be returned to the batch of good fits. The histograms corresponding to bad fits caused by model errors show less resemblance to the normal distribution due to a higher skewness and/or lower kurtosis. In case of a jump in the data, the resemblance to a normal distribution is completely gone and we're dealing with a very high skewness and/or very low kurtosis. Therefore we can easily separate the jumpy data from the other bad fits. Distinguishing between the noised data and the inadequate model is much harder, especially when last two reasons are slightly combined. However, comparing the corresponding p-values gives the desired result in most cases.

In this subsection we've shown that bad fits can be filtered from our data by looking at both the distribution and corresponding variance of their residuals. This resulted in the following two criteria for detecting a bad fit:

1. The variance is higher than twice the mean of the variances of all residuals.
2. The result of the JB test is equal to one, rejecting the hypothesis that the distribution is normal, or equal to zero with a corresponding p-value lower than 0.55, making it not likely enough that the distribution is normal.

These two criteria result in 308 bad fits, which is about 2.9% of the batch. Filtering the bad fits results in more reliable parameter values. The histograms of the parameter values after filtering are similar to the ones for the complete batch, showed in 6.1, and therefore the new histograms are not yet displayed and we will continue the filtering first.

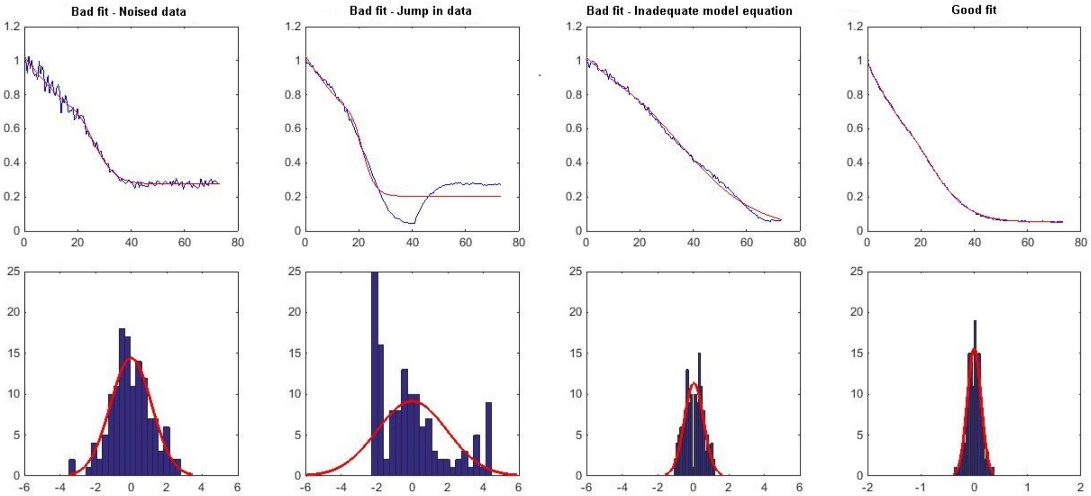


Figure 6.3: Curves and corresponding histograms of the three types of bad fits compared to those of a good fit.

6.1.2 Filtering unreliable predictions

The hypothesis from the previous section is that our seed batch can be divided into (at least) three different classes based on the normalized oxygen concentration at the end of the observation time, c_{end}. Since all batches have different observation times, we cannot compare their c_{end} directly but have to extrapolate the fitted results to a general observation time first. In Section 2 it was noted that in order to obtain a complete picture of the seed germination, approximately two days were needed. Therefore we extrapolate, if necessary, to 48 hours. This extrapolation is carried out by evaluating the fitted result at a vector of values running from t_{end} to 48 hours, where t_{end} is the original observation time of the measurements. Extrapolation is subject to greater uncertainty and a higher risk of producing meaningless results, therefore prediction intervals of the extrapolated values have to be taken into account before drawing any conclusions. A prediction interval is an estimate of an interval in which future values will fall, with a certain probability, given what has already been observed. Hence an extremely wide 95% prediction interval around the estimated concentration c_{end}, means that the value of this concentration is very unreliable. Looking at the original observation times of all seven batches, a time of only 21 hours is observed in case of the third batch. Due to this low observation time, there are very little data points available. Although the resulting fit might be very close to the original data, both the confidence intervals of the estimated parameters and the prediction intervals of the estimated function values are very wide due to this small amount of data points (in some cases even tenth orders of magnitude wider than for the other batches). This causes the estimated parameters and function values to be already very unreliable. Extrapolation will only introduce more uncertainty en therefore we remove these seeds are removed from our batch.

But besides these seeds, extrapolation can also cause a huge uncertainty for more reliable fits with observation times closer to 48 hours. Therefore we also discard the seeds with a very wide prediction interval for the value of c_{end} and seeds for which the the extrapolation drastically enlarges this interval. An example is shown in Figure 6.4 where the data and extrapolated fit are shown together with the prediction bounds of the fitted curve. The prediction bounds were constructed using the `predint` function in `Matlab`.

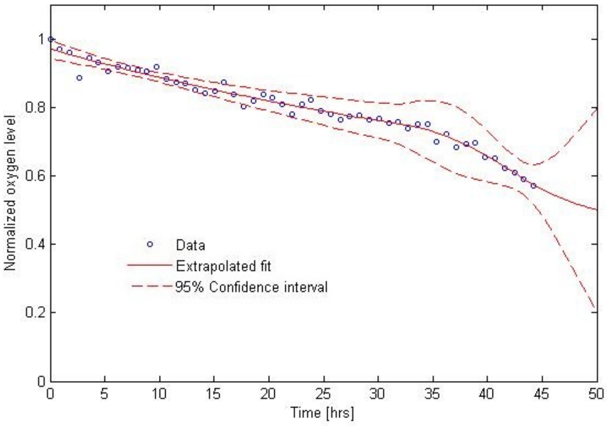


Figure 6.4: Data and corresponding fitted result extrapolated to 48 hours with a 95% prediction interval, showing the unreliability of the extrapolated value for c_{end}.

After filtering the bad fits and unreliable predictions we are left with a batch of 8551 barley seeds with parameter distributions shown in Figure 6.5. Compared to unfiltered results we see that αn_0 shows less observations for large values and also β and t_r have lost some of their extremely large values. The peak for small values of t_r is a little less high, but besides that the overall shapes of the distributions remained the same.

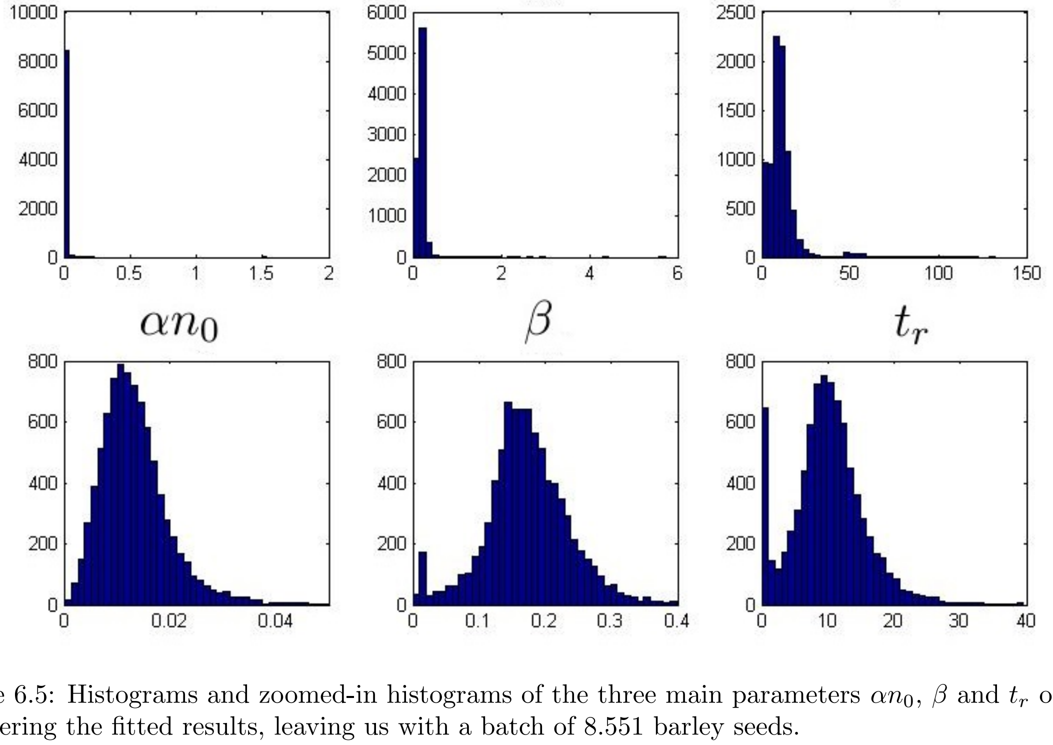


Figure 6.5: Histograms and zoomed-in histograms of the three main parameters αn_0, β and t_r obtained by filtering the fitted results, leaving us with a batch of 8.551 barley seeds.

Filtering the bad fits and unreliable parameters showed us that some of the observed extreme parameter values were caused by our fitting algorithm or deviations in the seed data. Some of the extreme values are however still present, meaning that they could in fact have a biological meaning. In the next section an attempt is made to divide our seed batch into different classes with the goal to show (cor)relations between parameters and indicate which combination of parameter values causes fast/slow germination.

6.2 Classification

The total batch of barley seeds has now been reduced to a smaller batch with more reliable parameters and values for c_{end}, the normalized concentration at $t = 48$ hours. On this batch the classification proposed in Section 5 is performed. The hypothesis was that c_{end} could be used as a classification parameter, splitting our batch into (at least) the following three classes:

1. Dead and highly dormant seeds.
2. Non-germinating seeds
   The rest of the dormant seeds and weak germinators that don't germinate within the given time of 48 hours.
3. Germinating steeds
   Strong germinators and weak germinators that do germinate within the given time of 48 hours.

This separation, and the separation between the first and second class in particular, originated from the clear gap between the normalized end concentrations in batch 2. Looking at the histogram of c_{end} for the complete batch shown in Figure 6.6, we do not observe a similar gap around a large value but only a slight decrease in the amount of observations around 0.87. This means that the reason to split the dormant seeds over two classes disappears and we will start by focusing on the classification between germinating and non-germinating seeds. Afterwards we discuss the possibilities of achieving the following sub-classification:

1. Non-germinating seeds
   (a) Dead seeds
   (b) Dormant seeds
   (c) Weak germinators that don't germinate within the given time
2. Germinating seeds
   (a) Weak germinators that do germinate within the given time
   (b) Strong germinators

based on the different seed types mentioned in Section 2. In the ideal case, each subclass has its own specific parameter distribution or a distinct characteristic separating its seeds from the seeds in other classes. We will thus take a close look at the parameter distributions and possible (cor)relations between parameters in each classification attempt.

6.2.1 Germinating versus non-germinating seeds

In Section 5 the following interpretation of the definition of germination was given: "germinating seeds are seeds that deplete oxygen below a certain level in the given time". The given time has already been set at 48 hours, so that leaves the oxygen level. In the literature, the only relation between the oxygen consumption curve and the time of germination is given by the relative germination time (RGT) described in Section 4.2. However, the RGT value can be the same for seeds with very different oxygen consumption patterns that certainly do not germinate at the same time. The relative germiniation time is therefore of no use in our classification problem. As an example two of those curves with the same RGT value are shown in Figure 6.7.

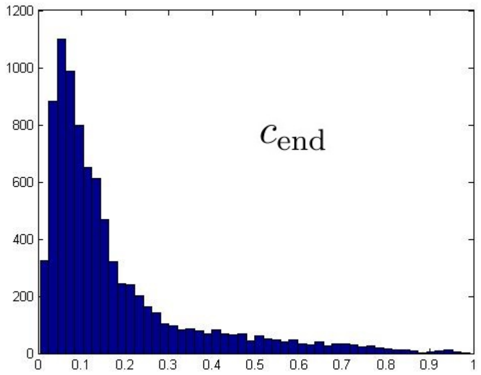


Figure 6.6: Histogram of c_{end}, the estimated normalized oxygen level at $t = 48$ hours, for the complete batch of 8551 barley seeds.

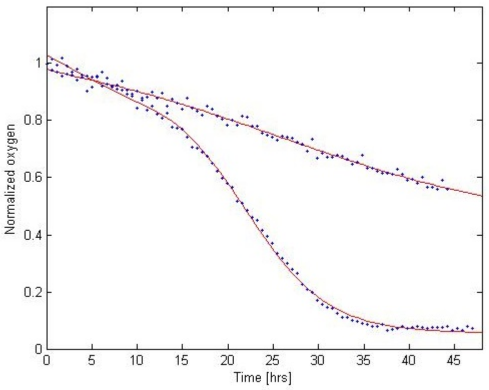


Figure 6.7: Data and corresponding fitted result of two seeds with the same RGT value but a completely different oxygen consumption pattern.

Since this was the only lead to finding a link between the time of germination and the normalized oxygen level, we will now have to determine the threshold separating the germinating from the non germinating seeds by considering several values for c_{end} and comparing the results. Intuitively the threshold must be chosen closer to zero than to one and we must make sure that the amount of germinating versus non germinating seeds remains realistic. For instance, a 95% germination in the given time is not really realistic given the wide spread of oxygen consumption patterns in the batch. Therefore the maximum threshold is set at $c_{\text{end}} = 0.25$, meaning that a seed has to consume 75% of the available oxygen in order to germinate. In this batch 80% of the seeds suffices that condition. On the other hand the threshold can't be too close to zero either, meaning that almost no seeds germinate in the given time. The minimum threshold is thus set at $c_{\text{end}} = 0.05$ resulting in a 18% germination. The goal of classification is to gain more insight in the influence of parameter values on germination and finding possible (cor)relations between these parameters. Therefore we regard both the distributions as the scatter plots of all parameters, while we vary the level of c_{end}.

Overall we can conclude that there is more correlation between parameters in the germinating class compared to the complete batch showing no correlation at all, see Figure 6.9. Changing the threshold defining this germinating class however, does not make much of a difference to either the shape of the parameter distributions or the correlation between them. Therefore we set the threshold separating the germinating from the non germinating seeds at 0.173, which is equal to the mean of c_{end}, resulting in a 70% germination. Corresponding scatter plots and parameter distributions are shown in see Figure 6.8. The correlations between the parameters seem to be all slightly positive, but no clear relationship can be derived from the scatter plots.

The non germinating seeds show practically the same plots as the total batch of seeds, meaning that there is no correlation at all between parameters in the non germinating class.

Looking at the parameter values of the germinating and non germinating seeds, which are more clearly shown in Figure 6.10, we see that the non germinating class contains most large parameter values, especially for t_r. The distributions of the germinating class slightly resemble the distributions of the parameters of the complete batch around the mode, Figure 6.5. The clear difference is the large amount of observations for small β values in the complete batch. This makes sense given the fact that a small β indicates a slow growth and therefore makes it less likely that the seed is able to deplete oxygen below the level of 0.173 in the given time. Furthermore it seems like both αn_0 and β could be normally or log-normally distributed. Unfortunately, the JB test rejects the null-hypothesis in all cases, meaning that none of these assumptions turn out to be true. The non germinating parameter distributions resemble those of the complete batch, which calls for further investigation. In the following two subsections both the class of germinating seeds as the class of non germinating seeds will be further investigated.

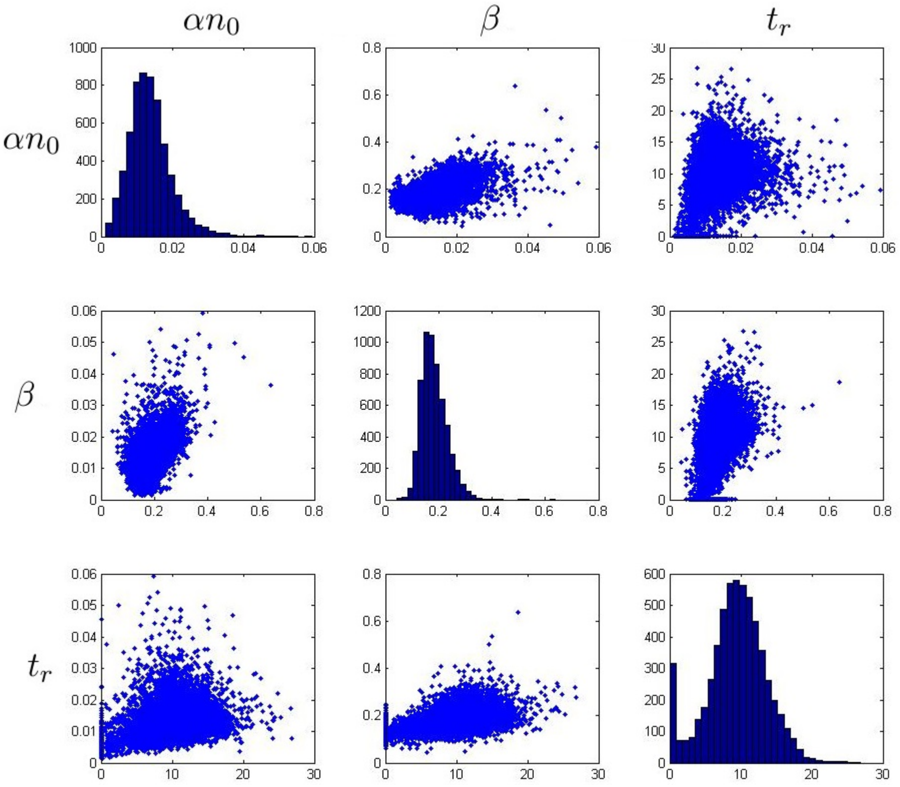


Figure 6.8: Scatter plots and histograms of the parameters from the germinating class.

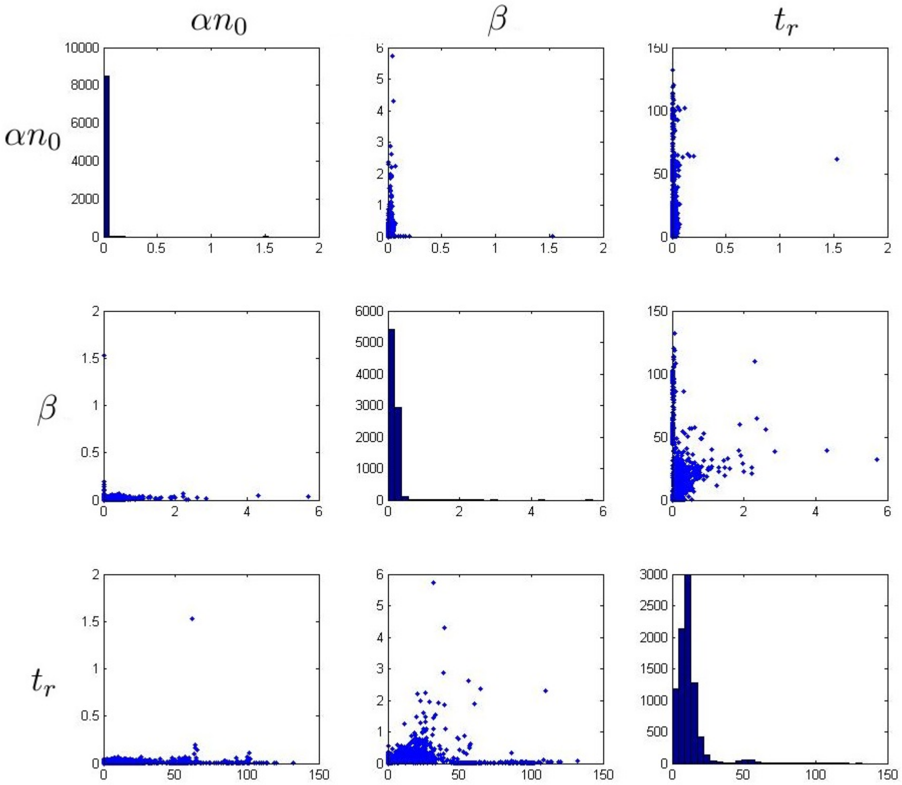


Figure 6.9: Scatter plots and histograms of the parameters from all the seeds in the batch.

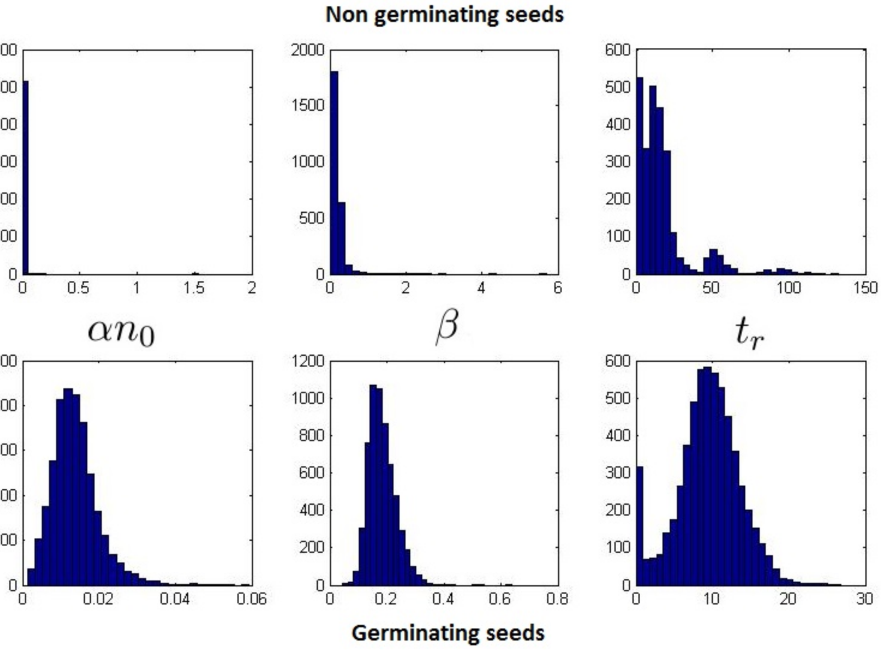


Figure 6.10: Histograms of the three main parameters for both the germinating and the non germinating seeds.

6.2.2 Germinating seeds

As suggested in Section 2, there are two types of germinating seeds, strong germinators and weak germinators. The difference between them is that strong germinators show a sigmoid type of oxygen consumption pattern while weak germinators show a linear or incomplete pattern. Distinguishing between these two thus comes down to finding a parameter or parameter combinations that divide(s) our class into two groups, based on the shape of the oxygen curve. As mentioned in Section 5.3 this is very hard to accomplish and a split based on c_{end} is proposed. However this does not give the desired results as the curve shapes are spread trough-out both groups. Therefore we return to the three main parameters to see if one these can be used as a classification parameter.

In Section 5 the significant role of β in determining the shape of the oxygen curve became apparent. And indeed, distinguishing between large and small β, the group of small β shows more 'flat' curves while the group of large β consists of mostly sigmoid type curves. Two examples are shown in Figure 6.11.

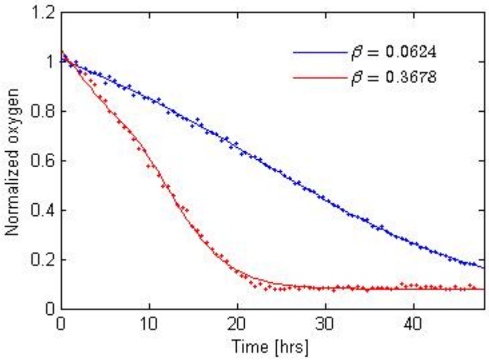


Figure 6.11: Oxygen consumption patterns of two seeds with different values for β, showing a clear difference in shape.

However, this clear separation only holds for extremely large and extremely small values for β. For all the values around the mean of β, the curves can have either of the two shapes or a shape somewhere in between. Here lies the problem in classifying oxygen consumption curves based on shape that was already mentioned in Section 5.3. A clear split between shapes in a batch of thousands of seeds is highly unlikely. Furthermore the extreme small values of β occur in conjunction with small values of t_r and the same holds for the large values. This shows the slight correlation between the parameters that was already showed in the previous sections.

Although a clear separation is not possible, we showed that the value of β gives a good indication in whether a germinating seed is a strong or a weak germinator.

6.2.3 Non germinating seeds

The class of non germinating seeds by definition consists of the following three types of seeds: dead seeds, dormant seeds and weak germinating seeds. The difference between dead seeds and the other two is that dead seeds do not deplete oxygen at all, while the other two do. Dormant seeds only consume a small constant amount of oxygen while weak germinators have a wide range of possible oxygen consumption patterns, varying from a linear one up to an unfinished sigmoid one.

In Section 5.3 the value of c_{end} was chosen to separate the dead and part of the dormant seeds from the rest of the batch. In this case there was no clear mark for c_{end} and since we are more interested in keeping all the dormant seeds together, we will look for other separation criteria first. Considering the parameter distributions of the non-germinating seeds shown at the top of Figure 6.10, t_r seems like an interesting parameter with regard to splitting up the batch into several groups. Since t_r has the most influence on the repair stage and not so much on the growth stage, and given that t_r is not even slightly correlated to one of the other parameters, we have to take another parameter into account that regulates the growth stage of the germination process. Therefore we will look for combinations of t_r and β.

First we use $t_r = t_{\text{end}} = 48$ as a threshold. We expect all seeds with a repair time larger than t_{end} to be either dead or dormant. However, this turns out to be false because not all of the seeds show a linear oxygen consumption pattern. The seeds that don't do not belong to the class of dead or dormant seeds but to the class of weak germinators. Further classification is needed to isolate the dead and dormant seeds. Here the growth parameter β steps in. As it turns out, all dead and dormant seeds are consistent with a small value for β, as was expected in Section 5.3. This makes sense since both types of seeds show no growth at all during the given observation time. Now that we've isolated the dead and dormant curves, we can easily distinguish between them by using c_{end}. Seeds that hardly deplete any oxygen and therefore have a high value for c_{end} (> 0.9) are dead and seeds that deplete more oxygen are dormant.

The hereby selected curves of both the dead and dormant seeds are shown in Figure 6.12. One of these curves shows an exponential depletion of oxygen. This is the only curve in the entire batch having a very large value for αn_0, meaning that the seed rapidly depletes oxygen during the repair stage but practically stops depleting oxygen afterwards. This is a unique case that can not be put in any class based on shape, but comes closest to a dormant seed based on the definition.

Since we have now isolated all dead and dormant seeds, the remaining seeds must be weak germinators. These weak germinators have a wide variety of oxygen consumption patterns and therefore also take on a wide spread of variables. Their mutual correlation however did slightly improve compared to the correlation of the parameters of all non germinating seeds as can be seen in Figure 6.13.

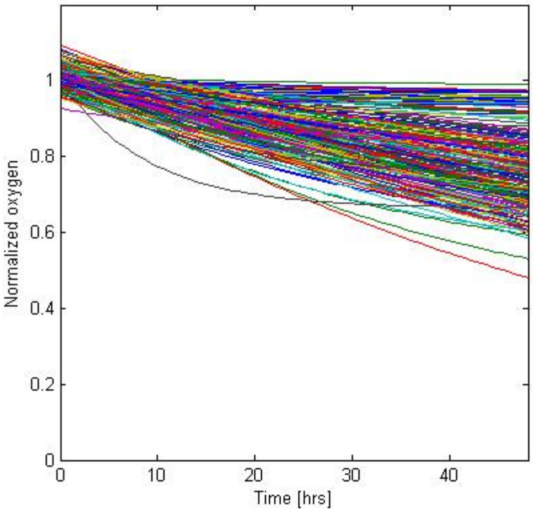


Figure 6.12: Fitted curves of both dead and dormant seeds.

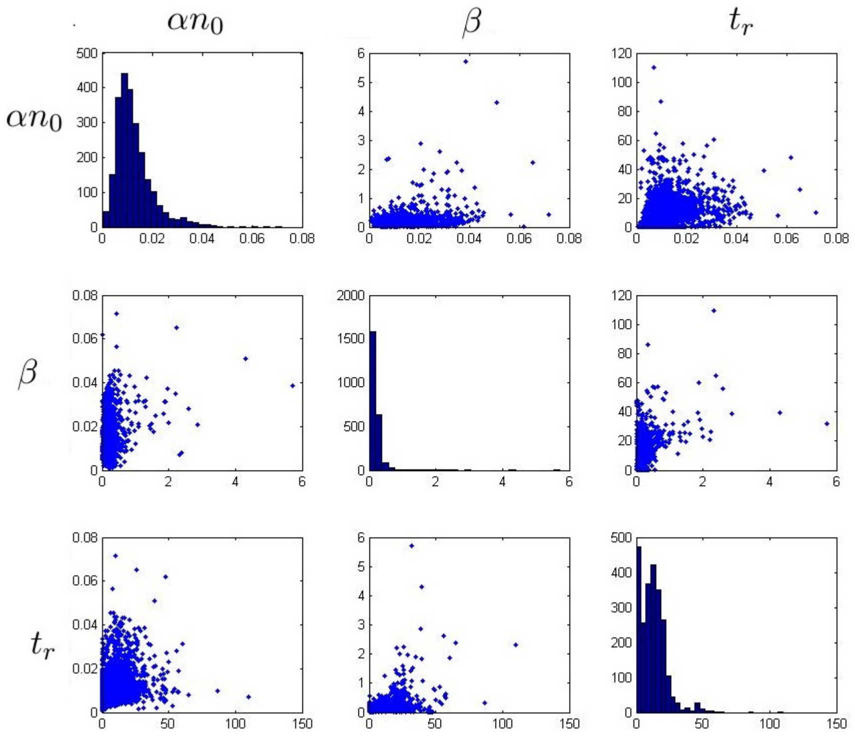


Figure 6.13: Scatter plots and histograms of the parameters from the weak germinators in the non germinating class.

6.2.4 Conclusion Classification

The classification attempts in the previous subsections have lead to the following three discussed classes

1. Germinating seeds
2. Non germinating seeds (These are the weak germinators that are able to germinate, but not in the given observation time)
3. Dead and Dormant seeds

Histograms of the parameters of all three classes are shown in Figure 6.14. The histograms of αn_0 of both the germinating and the non germinating class seem similarly shaped, which could indicate that variation in αn_0 is mostly caused by a large amount of statistical factors that are practically the same in both classes. Furthermore the germinating seeds have less skewed distributions compared to the non germinating seeds, which shows that our threshold for c_{end} also "bounds" our parameters. Compared to the total batch, correlations between parameters in the classes have improved, but still no really clear correlations were found.

In discussing the germinating batch, we concluded that distinguishing between slow and fast germinators is extremely hard, but that the value of β gives a good indication. In the non germinating class β was also used but this time in combination with t_r to separate the dead and dormant seeds from the rest of the batch. We can thus conclude that the course of the germination process is highly determined by values of the parameters t_r and β. For example, a seed with a large β and an average t_r is more likely to germinate in the given time than a seed with a small β and an average t_r.

All observations have pointed out that a clear classification based on c_{end} is very hard to achieve. However we can conclude that some combinations of β, t_r and c_{min} are more likely to result in germinating seeds than other and therefore these three values have to be watched very closely.

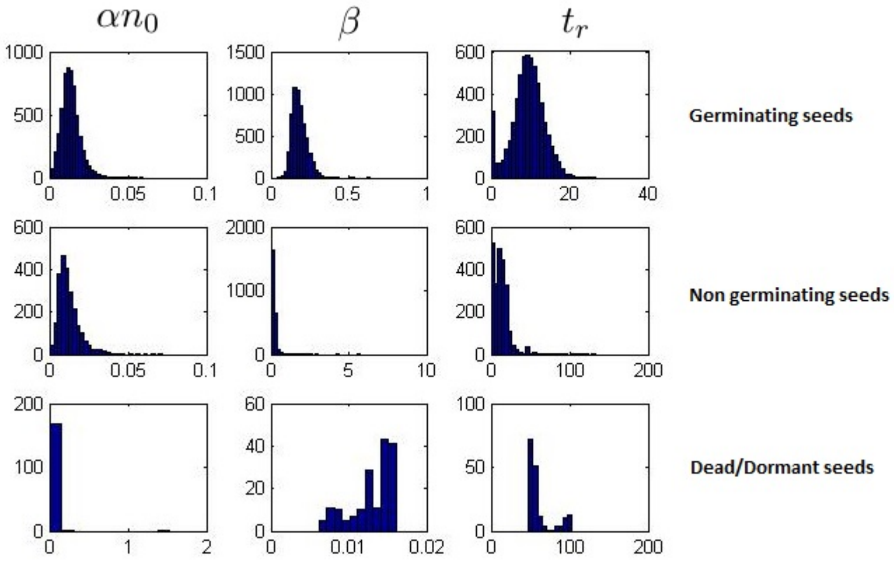


Figure 6.14: Histograms of the parameters for the three different classes.

7 Volume predictions

In the previous sections, a model for oxygen consumption and growth in closed test tubes was proposed and fitted to a large amount of experimental data. All experiments were performed with equally sized test tubes and (almost) equally sized seeds. Therefore the influence of the seed volume and test tube volume were not taken into account. However, both volumes do influence the germination process as will be showed in this section.

Define the test tube volume as V_1 with oxygen concentration $c_1(t)$ and the seed volume as V_2 with oxygen concentration $c_2(t)$. The oxygen measured concentration is $c_1(t)$ but we are interested in $c_2(t)$, the oxygen used by the seed. According to the Henry law, mentioned in Section 3, the concentration of oxygen inside the test tube is proportional to the oxygen concentration inside the seed:

$$c_1(t) = k_{H,cc} \cdot c_2(t),$$

where $k_{H,cc}$ is the Henry constant, from now on referred to as k. Differentiating the above equation with respect to t results in

$$\frac{dc_1}{dt} = k\,\frac{dc_2}{dt}. \tag{7.1}$$

Furthermore we have conservation of mass, meaning that all oxygen consumed in V_1 has to come from V_2:

$$\begin{aligned} \frac{d}{dt} m_1 \bigg|_{V_1} &= \frac{d}{dt} m_2 \bigg|_{V_2} \\ \frac{d}{dt}\,(V_1\, c_1) &= \frac{d}{dt}(V_2\, c_2), \end{aligned} \tag{7.2}$$

where m_1 and m_2 are the total masses of oxygen in respectively V_1 and V_2.

Assuming the volume changes caused by growth of the seed are very small and very slow, we can say that both volumes remain constant over a small period of time and Equation (7.2) can be written as:

$$V_1 \frac{dc_1}{dt} = V_2 \frac{dc_2}{dt}. \tag{7.3}$$

The equation for oxygen consumption inside the seed, $\frac{dc_2}{dt}$, discussed in Section 4 is given by:

$$\frac{dc_2}{dt} = -\alpha n(c_2 - c_{\min}).$$

Substituting this results in

$$V_1 \frac{dc_1}{dt} = -V_2 \alpha n(c_2 - c_{\min}).$$

After applying the Henry from Equation (7.1), we end up with an equation only containing c_2:

$$V_1 k \frac{dc_2}{dt} = -V_2 \alpha n(c_2 - c_{\min}),$$

resulting in the following equation for the oxygen consumption of the seed

$$\frac{dc_2}{dt} = -\alpha \frac{V_2}{kV_1} n(c_2 - c_{\min}). \tag{7.4}$$

We see that both volumes are taken into account by the change of the oxygen consumption rate from α to $\alpha \frac{V_2}{kV_1}$. Increasing the size of the test tube causes an increase in V_1 resulting in a lower oxygen consumption rate.

Since $\frac{V_2}{kV_1}$ is a constant, we can easily find the analytical solution by simply substituting $\alpha\frac{V_2}{kV_1}$ for α in Equation (4.8). Therefore the predicted oxygen consumption curves, when changing the test tube volume, can be obtained by adjusting the fitted parameter results for α for all seeds. Of course, the prediction intervals, with respect to α, of these estimated results have to be taken into account to check the reliability of these predictions.

Similarly we can we can predict the possible oxygen consumption patterns for larger seeds, i.e. larger V_1, in the same test tubes. Since Fytagoras has carried out the single seed oxygen consumption measurements, under the same conditions, on other seeds besides barley, there is some data available to investigate the validity of these predictions. Unfortunately, we were unable to include this in this thesis, so this will be regarded as further research.

8 Conclusion and discussion

The main goal in this thesis was to develop a mathematical model describing the oxygen transport and consumption within germinating seeds. It had to be in line with the current biological knowledge on these processes and the highly detailed time-resolved oxygen consumption measurements for single seeds, obtained with Q_2 measuring technology. This should allow interpretation of the observed data and characteristic in terms of internal oxygen transport processes and overall seed properties.

In developing this model, the basic premise is that most of the oxygen is being consumed by the mitochondria and that they largely behave as a colony of bacteria. Also, by the Henry law, we can assume that the oxygen concentration inside the seed is uniform at all times resulting in the assumption that the complex processes of transport and diffusion of oxygen can be neglected. Based on these assumptions a simple ODE model is developed, relating the measured oxygen concentration to the number of active mitochondria. The model splits the germination process into two stages, an initial repair stage and a subsequent growth stage.

An analytical solution is obtained and fitted to the experimental data using a standard nonlinear least squares optimization algorithm in `Matlab`. It turns out that the simple two-stage model provides excellent fits for the experimental data. Furthermore, the fit results in several biologically interesting parameters of which the main three are the oxygen consumption rate by mitochondria times the initial amount of mitochondria, the growth rate of the mitochondrial population and the duration of the repair stage. Especially the last two parameters have a large influence on the oxygen consumption pattern of a seed. The longer the duration of the repair stage, the less likely it becomes that the seed will germinate during the given observation time. Similarly, the higher the growth parameter is, the more likely it becomes for the seed to germinate in the given time. Furthermore our experiments show a large peak for small repair times and, discarding this peak, a mean of about 10 hours, which is close to the previously discussed time of 12 hours when mitochondria contain fully developed cristae.

Initial results also show that filtering is necessary in order to obtain more reliable parameter statistics. This is done in two different ways, firstly by removing the bad fits and secondly by removing the unreliable predictions for the normalized concentration at $t = 48$ hours, c_{end}. The reliable predictions for c_{end} are used in the classification, which attempts to divide the seed batch into different classes, each with their own characteristic properties. Results show that a clear separation between these classes is hard to accomplish. However, the parameters in the class of germinating seeds, i.e. seeds that deplete oxygen to a level below the mean of the normalized end concentration at $t = 48$ hours, show more correlation compared to the parameters of the total batch and are therefore biologically interesting to investigate. Next to this the parameters attain values from a more restricted range, meaning the all the outlying parameter values are found in the non germinating class.

The definition for germinating seeds used here, is our interpretation of the general definition of germination in terms of oxygen consumption. Whether this definition coincides with the actual occurrence of germination is unknown. Therefore determining a correct definition of germination in terms of oxygen consumption curves, with the use of observations, would greatly contribute to the classification issue and help us in determining what exact combinations of parameters cause a fast/slow germination. For now all we can say is that seeds with certain combinations of β, t_r and probably c_{end} are more likely to germinate than seeds with other combinations. Here a large β, small t_r and small c_{end} are in favour of germination.

For research into the influence of the volumes of both the seed and test tube, an adjusted model is proposed with a new oxygen consumption rate containing α, both volumes and the Henry constant k. This model can be used to predict the effect of changing one of these volumes, to the oxygen consumption patterns. These predictions should afterwards be validated by comparing it to experimental data.

References

[1] J.D. Bewley, K. J. Bradford, H.W.M. Hilhorst, and H. Nonogaki. *Seeds: Physiology of development, germination and dormancy.* 3rd Edition, Springer, New York, Heidelberg, Dordrecht, London (2013).

[2] J. Van Asbrouck and P. Taridno. *Using the single seed oxygen consumption measurements as a method for determination of different seed quality parameters for commercial tomato seed samples*, Asian Journal of Food and Agro-Industry, (2009), pp. S88-S95.

[3] K. J. Bradford, P. Bello, J.-C. Fu, and M. Barros *Single-seed respiration: a new method to assess seed quality*, Seed Science and Technology, 41 (2013), pp. 45-54.

[4] http://opencurriculum.org/5358/cell-transport-and-homeostasis/

[5] L. Margulis *Symbiosis in cell evolution.* 2nd Edition, W.H. Freeman, San Francisco, (1993), pp. 452

[6] W. Henry *Experiments on the Quantity of Gases Absorbed by Water, at Different Temperatures, and under Different Pressures*, Philosophical Transactions of the Royal Society (London), No. 93, pp.29–274

[7] H.C. Gerritsen, R. Sanders, A. Draaijer, C. Ince, and Y.K. Levine *Fluorescence lifetime imaging of oxygen in living cells*, Journal of Fluorescence, 7 (1997), no. 1, pp. 11-15.

[8] N. Collis-George, and M.D. Melville *Models of oxygen diffusion in respiring seed*, J. Exp. Botany, 25 (1974), No. 89, pp. 1053-1069.

[9] K.E. Dionne *Oxygen transport to respiring myocytes*, J. Biol. Chem., 265 (1990), pp. 15400-15402.

[10] N. Budko, A. Corbetta, B. van Duijn, S. Hille, O. Krehel, V. Rottschëafer, L. Wiegman, and D. Zhelyazov *Oxygen transport and comsumption in gerinating seeds* Proceedings of the 90th European Study Group Mathematics with Industry, pp. 5-30

[11] E. Gnaiger, R. Steinlechner-Maran, G. Méndez, Th. Eberl, and R. Margreiter *Control of mitochondrial and cellular respiration by oxygen*, J. Bioenergetics and Biomembranes 27 (1995), No. 6, pp. 583-596.

[12] C.T. Kelly *Iterative Methods for Optimization*, SIAM, Philadelphia 1999

[13] http://nl.mathworks.com/help/optim/ug/least-squares-model-fitting-algorithms.html

[14] N.V. Budko, F.J. Vermolen, S. Hille, and B. van Duijn *Predicting germination time from single seed oxygen consumption measurements* [Unpublished]

A Derivation of Equation (4.6)

We ended up with

$$\frac{dc}{dt} = \beta c(c - \tilde{c}), \quad c(t_r) = c_r,$$

where $\tilde{c} = c_r + \alpha n_0/\beta$.
Now separation of variables results in:

$$\frac{1}{c(c-\tilde{c})}\ dc = \beta\ dt$$

which is the same as

$$-\frac{1}{\tilde{c}}\left(\frac{1}{c} - \frac{1}{c-\tilde{c}}\right)\ dc = \beta\ dt.$$

Integrate both parts to obtain

$$\begin{aligned}
-\frac{1}{\tilde{c}}(\log(c) - \log(c-\tilde{c})) &= \beta t + C_0 \\
\log(c) - \log(c-\tilde{c}) &= -\tilde{c}\beta t + C_0 \\
e^{\log(c)-\log(c-\tilde{c})} &= C_0 \cdot e^{-\tilde{c}\beta t} \\
\frac{c}{c-\tilde{c}} &= C_0 \cdot e^{-\tilde{c}\beta t} \\
c &= C_0 \cdot (c-\tilde{c})e^{-\tilde{c}\beta t}
\end{aligned}$$

where C_0 is an arbitrary constant. We write this in the form $c(t)$ equals some function of t:

$$\begin{aligned}
(1 - C_0\ e^{-\tilde{c}\beta t})\ c(t) &= -C_0\ \tilde{c}e^{-\tilde{c}\beta t} \\
c(t) &= \frac{-C_0\ \tilde{c}e^{-\tilde{c}\beta t}}{1 - C_0\ e^{-\tilde{c}\beta t}} \\
c(t) &= \frac{\tilde{c}}{1 - \frac{1}{C_0}e^{\tilde{c}\beta t}}.
\end{aligned}$$

To obtain an expression for the constant in the equation above we substitute $c(t_r) = c_r$:

$$\begin{aligned}
c_r &= \frac{\tilde{c}}{1 - \frac{1}{C_0}e^{\tilde{c}\beta t_r}} \\
1 - \frac{1}{C_0}e^{\tilde{c}\beta t_r} &= \frac{\tilde{c}}{c_r} = 1 + \frac{\alpha n_0}{c_r\beta} \\
-\frac{1}{C_0} &= \frac{\alpha n_0}{c_r\beta}e^{-\tilde{c}\beta t_r} \\
C_0 &= -\frac{c_r\beta}{\alpha n_0}e^{\tilde{c}\beta t_r}
\end{aligned}$$

Substituting our constant and the expression for $\tilde{c}$, we arrive at our solution:

$$\begin{aligned}
c(t) &= \frac{c_r + \frac{\alpha n_0}{\beta}}{1 + \frac{\alpha n_0}{\beta c_r}e^{(c_r\beta + \alpha n_0)(t-t_r)}} \\
&= \frac{c_r(\beta c_r + \alpha n_0)}{\beta c_r + \alpha n_0\ e^{(\beta c_r + \alpha n_0)(t-t_r)}}, \quad t > t_r \quad \square
\end{aligned}$$

B Matlab code

B.1 FitAllData.m

```
clear all;
clc;

load('AllBarleyData.mat','Data','Time');

[Nsamp,Nexp]=size(Data);

% Initial parameter values
x0=[0.06 0.02 29 0.1 0.6 0.1];
x1=x0;
x1(end)=20;

aFit=zeros(Nexp,1);
bFit=zeros(Nexp,1);
c_0Fit=zeros(Nexp,1);
c_mFit=zeros(Nexp,1);
n_0Fit=zeros(Nexp,1);
t_rFit=zeros(Nexp,1);
dataFit=NaN(Nsamp,Nexp);
RMSE=NaN(Nexp,1);
Rsquare=NaN(Nexp,1);
residuals=NaN(Nsamp,Nexp);

%% Fit
for i=1:Nexp
    t=Time(:,ceil(i./95));
    t(isnan(t))=[];
    ydata = Data(:,i);
    ydata(isnan(ydata))=[];
    t(size(ydata)+1:end)=[];

    if isempty(ydata)==0
        % Comparing the quality of the fits under both initial conditions
        [fitresult1,gof1,output1] = createFit(t, ydata, x0);
        [fitresult2,gof2,output2] = createFit(t, ydata, x1);
        if gof1.rmse <= gof2.rmse
            fitresult=fitresult1;
            gof=gof1;
            output=output1;
        elseif gof1.rmse>gof2.rmse
            fitresult=fitresult2;
            gof=gof2;
            output=output2;
        end
        % Save the fitting results
        aFit(i)=fitresult.a; % fitted value alpha
        bFit(i)=fitresult.b; % fitted value beta
        c_0Fit(i)=fitresult.c_0;
        c_mFit(i)=fitresult.c_m; % fitted value c_m
        n_0Fit(i)=fitresult.n_0; % fitted value n_0
        t_rFit(i)=fitresult.t_r; % fitted value t_r
        dataFit(1:length(ydata),i)=fitresult(t);
        RMSE(i)=gof.rmse;
        Rsquare(i)=gof.rsquare;
        residuals(1:length(ydata),i)=output.residuals;
    end
end
alpha=aFit.*n_0Fit;
```

```
%% Normalize the data and corresponding fits
Data2=bsxfun(@rdivide, Data(1:end,:),Data(1,:));
dataFit2=bsxfun(@rdivide, dataFit(1:end,:),Data(1,:));
```

B.2 createFit.m

```
function [fitresult, gof, output] = createFit(time, expdata, x0)
%CREATEFITS(TIME,EXPDATA)
%  Data for 'untitled fit 1' fit:
%      X Input : time
%      Y Output: expdata

%  Output:
%      fitresult : a cell-array of fit objects representing the fits.
%      gof : structure array with goodness-of fit info.
%
%  See also FIT, CFIT, SFIT.

%% Fit: 'Create fit'.
[xData, yData] = prepareCurveData(time, expdata);

% Set up fittype and options.
ft = fittype('ModelEquation(time,a,b,c_0,c_m,n_0,t_r)',...
    'independent', 'time', 'dependent', 'y' );
opts = fitoptions( ft );
opts.Display = 'Off';
opts.StartPoint = x0;                       % Initial parameter values
opts.Lower = [0 0 0 0 0 0];                 % Lower Bound
opts.Upper = [1.5 20 500 1.1 1.5 160];      % Upper Bound
opts.TolFun = 1e-08;                        % Tolerance model function
opts.TolX = 1e-08;                          % Tolerance paramaters

% Fit model to data.
[fitresult, gof, output] = fit(xData, yData, ft, opts);
```

B.3 ModelEquation.m

```
function f = ModelEquation(x,a,b,c_0,c_m,n_0,t_r)
f=zeros(size(x));
c_r=c_m+(1-c_m)*exp(-a*n_0*t_r); % Concentration at the end of repair stage

for i = 1:length(x)
    if x(i)<=t_r; % Repair stage
        f(i)=c_0*(c_m+(1-c_m)*exp(-a*n_0*x(i)));
    else  % Growth stage
        f(i)=c_0*(c_m+...
            ((c_r-c_m)*(b*(c_r-c_m)+a*n_0))/...
            (b*(c_r-c_m)+a*n_0*exp((b*(c_r-c_m)+a*n_0)*(x(i)-t_r))));
    end
end
```