CHALMERS | UNIVERSITY OF GOTHENBURG

MASTER'S THESIS

A Mathematical Investigation of Breast Cancer and Tumor Growth

Statistical Analysis and Stochastic Modeling

HANNA M. MATTSSON

PIERRE K.U. NYQUIST

Department of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2009

Thesis for the Degree of Master of Science (30 credits)

A Mathematical Investigation of Breast Cancer and Tumor Growth

Statistical Analysis and Stochastic Modeling

Hanna M. Mattsson Pierre K.U. Nyquist

CHALMERS | UNIVERSITY OF GOTHENBURG

Department of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE – 412 96 Gothenburg, Sweden
Gothenburg, December 2009

Abstract

This thesis deals with aspects of probability and statistics applied to breast cancer research. The first part of the thesis concerns data from *in vitro* experiments with breast cancer cells. The effect on aggregate counts and morphological parameters of the cells by surrounding (simulated) tissue's stiffness is analyzed using methods from linear mixed models theory. The analysis indicates that certain parameters are significantly different for different tissue stiffness.

The second part of the thesis deals with stochastic modeling related to initial tumor growth. A mathematical model for tumor growth is studied and certain types of randomness are introduced in it. Using statistical methods we give a general characterization of numerical solutions to the random system. Large deviation techniques are used to obtain results for probabilities related to the random system and sample paths from the random and the deterministic systems are compared.

Keywords: Breast cancer cells; Tissue stiffness; Regression analysis; Repeated measurements; Linear mixed models; Oncogenic mutations; Stochastic modeling; Random perturbations; Large deviations

Acknowledgements

We want to thank our supervisor Patrik Albin for being the truly inspiring teacher that he is and for providing us with the opportunity to travel to UNC Chapel Hill for our master's thesis research. We also want to thank him for his valuable comments and input towards the end of our stay.

Our deepest thanks to Professor M Ross Leadbetter and Professor Amarjit Budhiraja, our supervisors at UNC Chapel Hill. Without any obligation to do so, they took us on and has provided us with some really interesting problems during our stay. They are not only brilliant mathematicians but also incredibly kind and inspiring people and we have learned an immense amount during our time with them.

Thanks to Dr Karen Burg for the most pleasant visit to Clemson University. We will remember our encounter with the Dala horse for a long time.

We again want to thank Professor Leadbetter and his wife Mrs Winsome Leadbetter for their unmatched hospitality and generosity. Not only towards us, but to our visiting families as well. They have made sure that we have never had to worry about anything outside of mathematics and for this we are truly grateful.

Finally, we want to thank all our new friends from the department and from our track club. You have all made this time very special and given us a glimpse of what America is.

Contents

1 Introduction **1**

I Statistical analysis of breast cancer cell data **3**

2 Introduction: Statistical data analysis **5**
2.1 Experimental setup 5
2.2 Experimental data 6
2.3 Objectives 8

3 Exploratory data analysis **9**
3.1 Number of aggregates 9
3.2 Perimeter of aggregates 13
3.3 Area of aggregates 16
3.4 Circularity of aggregates 20
3.5 Comments 22

4 Repeated measurements for longitudinal data **25**
4.1 The linear mixed model 26
4.1.1 Covariance structures 27
4.2 Application of LMM to repeated measurements data 28

5 Modeling of count and perimeter data **31**
5.1 Analysis of aggregate count 31
5.1.1 Modeling the covariance structure 31
5.1.2 Regression analysis 34
5.2 Analysis of the perimeter of the aggregates 38
5.2.1 Modeling the covariance structure 38
5.2.2 Regression analysis 40
5.3 Comments 43

II Stochastic modeling of tumor growth **45**

6 A mathematical model for initial tumor growth **47**

6.1 Model for initial tumor growth 47

6.2 Results for deterministic initial tumor growth 51

6.3 Comments 53

7 Stochastic behavior in the model for tumor growth **55**

7.1 Random perturbations 55

7.2 Random mutation parameters 58

7.2.1 Increased response to a mitotic activator 60

7.2.2 Escape from chemical control 68

7.3 Comments 73

8 Comparison of random and deterministic system **75**

8.1 Introduction to large deviations theory 75

8.2 Random perturbations 78

8.3 Random mutation parameters 82

8.4 Comments 84

9 Conclusions **87**

Chapter 1

Introduction

The National Cancer Institute (NCI) estimates that during 2009, 192,370 women in the U.S. will be diagnosed with breast cancer and 40,170 will die from the disease [27]. The NCI further goes on to state that based on the rates from 2004-2006, approximately 12% of the women born today will develop breast cancer at some point in their lives – it is the third most common type of cancer.

Scientists of all trades have joined in to try and fight cancer disease in general, and in particular this is so for breast cancer. Perhaps a bit surprisingly, mathematicians and statisticians have joined this fight in a rather successful way. The relatively young branch of the mathematical sciences that is referred to as *mathematical biology* is flourishing, with mathematicians providing models for all kinds of biological processes, cancer included. Although not an exhaustive list, Professor Crooke's (Vanderbilt University) reference list [10] shows the increase in mathematical publications related to cancer over the last 50 years.

One important aspect of mathematicians contribution to cancer research is the topic to define models for tumor growth. Examples of such models, based on diffusion equations, are given by Adams ([1], [2]). DeLisi and Rescigno ([12], [13]) provide a model for the interaction between tumor cells and a certain type of immune response, an important aspect for tumor growth. Based on predator-prey models, they characterize the growth of a tumor and conclude on how different initial conditions (tumor size, aggressiveness of immune response etc.) affect the outcome. Stepping away somewhat from cancer cells, certain aspects of cell movement have been modeled using stochastic differential equations (SDEs) ([15], [26]).

Aside from defining new models for tumor growth or similar characteristics of cancer cells, mathematicians (and perhaps especially statisticians) usually contribute in the following aspects:

- Statistical design and analysis.
- Stochastic modeling.

Statistical methodology lends itself naturally to the analysis of data from experiments regarding the behavior of cancer cells. Everything from basic statistics to advanced modeling methods comes to use when trying to understand the properties of such cells. In stochastic modeling, one

makes use of exact mathematical descriptions of biological phenomena and introduce randomness in different ways. It is important to understand how randomness will affect a certain model since cell behavior seldom is deterministic, "noise" might therefore be necessary to include in order for a model to adequately represent reality.

This thesis has been written at the University of North Carolina at Chapel Hill as a part of the NSF funded project *Emerging frontiers in 3-D breast cancer tissue test system.* Principal investigator is Dr Karen Burg of the bioengineering department at Clemson University, South Carolina. The aim of the project is to [8]

> *"enhance knowledge of breast cancer cellular and biomolecular behavior as an interactive function of a combination of oxygen level and tissue stiffness, by developing experimental and analytical tools to develop tissues of hierarchial structure and to assess normal and cancer cells within this framework."*

Our work contribute to the project in both ways mentioned above. Experiments that relate to the dependency of breast cancer cells behavior on the surrounding tissue's stiffness have been carried out at Clemson University. The obtained data is analyzed using various statistical methods. Furthermore, stochastic modeling is considered for a model of initial tumor growth. Different types of randomness are introduced in the model and the outcome is evaluated and compared to the deterministic case.

The thesis is divided into two parts - Part I describes the methods and results for analysis of experimental data and Part II deals with aspects of stochastic modeling related to initial tumor growth. In Part I, Chapter 2 describes the experimental setup used at Clemson University and corresponding obtained data. An exploratory analysis of the data is presented in Chapter 3, in which different subsets of the data are considered in accordance with test procedures. Some methods from linear mixed models are introduced in Chapter 4 and aspects of the experimental data is then further investigated, using such methods, in Chapter 5. In Part II, Chapter 6 introduces a model for tumor growth developed by Sherratt and Nowak [28]. In Chapter 7 the model for tumor growth is studied when containing random components of different types. Chapter 8 then compares the "new" stochastic models to the original deterministic one. Section 8.1 gives a brief introduction to large deviations theory, methods from which are then used to conclude on the general probabilistic behavior of the stochastic models. Finally, a short summary of conclusions from both parts is presented in Chapter 9.

Part I

Statistical analysis of breast cancer cell data

Chapter 2

Introduction: Statistical data analysis

As mentioned in Chapter 1, experiments conducted at Clemson University have provided data related to the behavior and properties of breast cancer cells. Section 2.1 describes the experimental setup and Section 2.2 the data obtained with that setup. The aims and objectives for analysis of the data as stated by grant proposals and experimental summaries (see e.g. [5] and [8]) are given in Section 2.3.

2.1 Experimental setup

The experiments aim to investigate how human breast cancer cells, MCF-7 cells, behave in different types of substrates. More specifically, it is how the stiffness of the different substrates affect the cells that is of most interest. The long-term goal associated with these experiments is to be able to use observations of morphological parameters from microscopic imaging to predict surface stiffness. This is clearly in line with the overall objectives of the project mentioned in the introductory Chapter 1.

MCF-7 cells were seeded onto the top of an agarose cellular mixture in 24 different cell culture plate wells. To obtain substrates with different stiffness, the cellular mixtures contained different amounts of agarose. The cellular mixture consisted of agarose, gelatin and phosphate-buffered saline (PBS). The percentage of agarose in the solution was between 0.75 and 3.0. The use of an agarose boundary was an attempt to minimize surface tension between the cellular mixture and the surrounding container. However, some substrates still showed signs of surface tension and therefore agarose content does not necessarily give a prediction of the gel stiffness. Henceforth, substrates with different agarose contents will therefore be referred to as different *populations*.

The 24 wells were distributed over eight populations with three samples each. An image was taken of each sample once a day for a total of fourteen days. During these days there were three media additions (days 1, 2 and 3) and three media replacements (days 5, 8 and 12). The replacement/addition of media was always done after the imaging on the corresponding day. Figure 2.1 shows examples of the type of images taken of the wells. In wells containing the substrates with the lowest percentages of agarose (0.75% and 1.0%), cancer cells sunk through

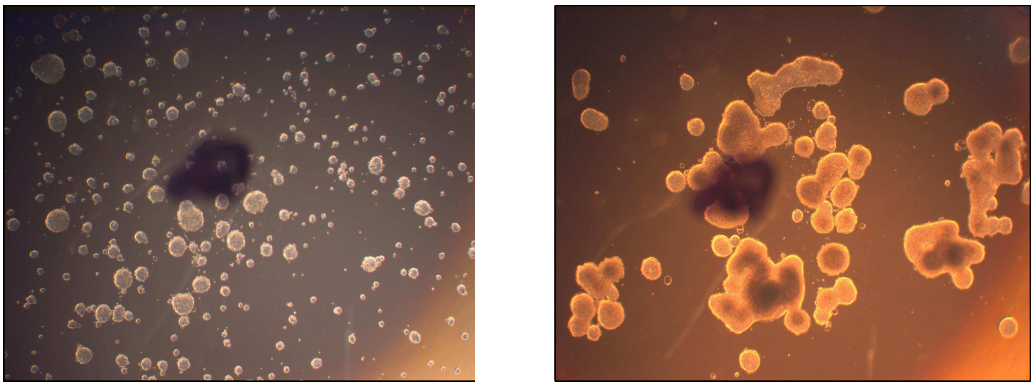

Figure 2.1: Images taken from one of the wells at day 2 (left) and at day 7 (right).

the gel and attached to the bottom of the well. Also, in some of these wells the gel was weak and fell apart. Moreover, the replacement/addition of media was observed to displace aggregates in the samples.

2.2 Experimental data

Data from the experiments was obtained by processing the daily images of the samples. The image processing was done in ImageJ under the protocol described in [5]. Figure 2.2 shows an original image together with the corresponding processed image.

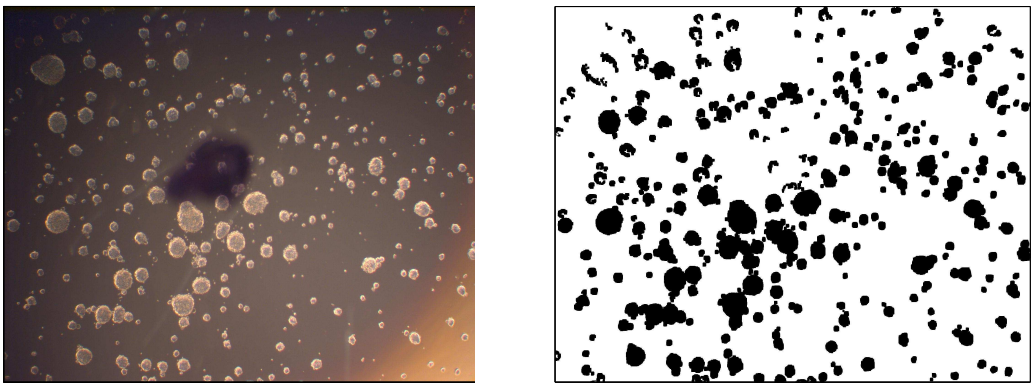

Figure 2.2: Image taken at day 2 from Figure 2.1 and the corresponding processed image.

The protocol for image processing basically separates large aggregates from the background in the original images, as seen in Figure 2.2. From the processed images the following data was obtained (k refers to a specific aggregate, j to a sample and i to the population):

- For a sample, the total aggregate count ($N_{i,j}$), average aggregate size and the average fraction of the image that was covered by aggregates (in pixels).

- For a specific aggregate, the (x,y) coordinates of the center of the aggregate, aggregate size (area, $A^k_{i,j}$), circularity ($C^k_{i,j}$) and perimeter ($P^k_{i,j}$).

Images from samples in which cancer cells sunk through the cellular mixture or in which the gel fell apart (mentioned in the previous section) could not be processed. Thus the wells with a cellular mixture consisting of either 0.75% or 1.0% agarose had to be ignored and no data was obtained from them. Moreover, images from late time points were sometimes processed in an erroneous way, e.g. spaces that were "locked within" aggregates were included in the aggregate cluster in the processed image. Figure 2.3 shows one such example.

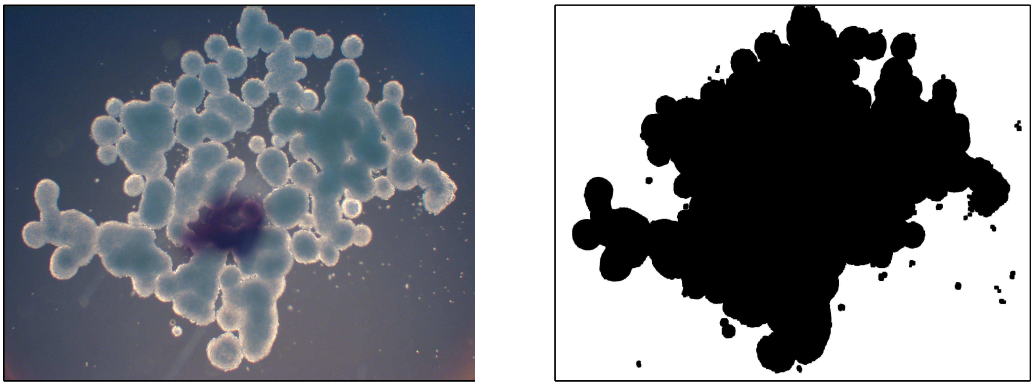

Figure 2.3: Original image (left) and processed image (right) for the same sample as in Figures 2.1 and 2.2 for day 12.

Due to the above described errors, we have chosen to exclude all data from late time points (after day 8). Experimental data is thus available for days 1-8 for the populations in Table 2.1. The notation in the table is henceforth used to refer to the different substrates. Note that for the subscript i in the variables, $i = 1$ corresponds to population A, $i = 2$ to population B and so forth. In [31] days 9 and 10 are also used for analysis. Due to the replacement of media that

Table 2.1: Agarose content in the cellular mixture for the different populations.

Population	**Agarose content**
A	1.25%
B	1.5%
C	2.0%
D	2.25%
E	2.5%
F	3%

took place on day 8 the samples were again disturbed, possibly changing the characteristics of the samples as compared to days 6-8. The choice to either include or exclude days 9 and 10 is not obvious and both cases may be argued. Here we have chosen the latter option. Furthermore,

due to the media replacements and additions we have chosen to analyze the data divided into different subsets. More specifically, the time periods days 1-8, days 1-6 and days 6-8 have been analyzed separately.

2.3 Objectives

As stated above, the hope is that the experimental setup described in Section 2.1 will enable the use of microscopic observations of morphological parameters to conclude on substrate stiffness. To achieve this, an investigation in the opposite direction is needed, i.e. it must be concluded which morphological parameters (if any) that are affected by the substrate stiffness.

Due to stiffness and percentage of agarose not being equivalent measures (see Section 2.1), objectives are here stated in terms of relations between population, rather than substrate stiffness, and morphological parameters. The main questions to be asked of the data, stated in [5], then translates to

- Do aggregates have different size or shape for different populations?
- Is the tendency of cells to cluster together and the rate at which this happens different for different populations?

Chapter 3

Exploratory data analysis

In the previous chapter the data obtained from experiments at Clemson University was introduced. As was stated there, the desire is to conclude whether or not different populations have different characteristics regarding cell clustering. In this chapter, exploratory data analysis is performed. The measures **number of aggregates**, **perimeter**, **area** and **circularity** given in the data are briefly investigated separately in Sections 3.1-3.4.

In the notations for the different measures, the time index t is usually dropped whenever the whole time period ($t \in \{1,2,...,8\}$) is consider or if it is explicitly stated which time period that is considered.

Due to the aforementioned replacement and addition of media during the experiments the analysis is generally performed for different subsets of the data, days 1-8 and days 6-8. Recall from Chapter 2 that there was no outside interference during days 6-8 of the experiments. Thus we consider this specific time period to see if the data show any signs of different behavior compared to days 1-8. In one instance the time period days 1-6 is also analyzed separately.

Due to the type of data (repeated measurements data) the aim is to, following this exploratory analysis, use linear mixed model (LMM) methods for modeling. Therefore, as will be further explained in the next chapter, the normality of the data is of particular interest and is investigated here.

It should be noted that the analysis presented in this chapter is only meant as an introductory one, used to give a sense of the data and to conclude what measures to be further concerned with.

3.1 Number of aggregates

Days 1-8

Figure 3.1 shows the number of aggregates for each sample during days 1-8. Clearly there is a decrease in the number of aggregates over time for all samples. However, the rate of decrease of aggregates in the samples is higher in the initial state of the experiment than later on. Noticeable is that the range of number of aggregates at day 1 for different samples is substantially larger

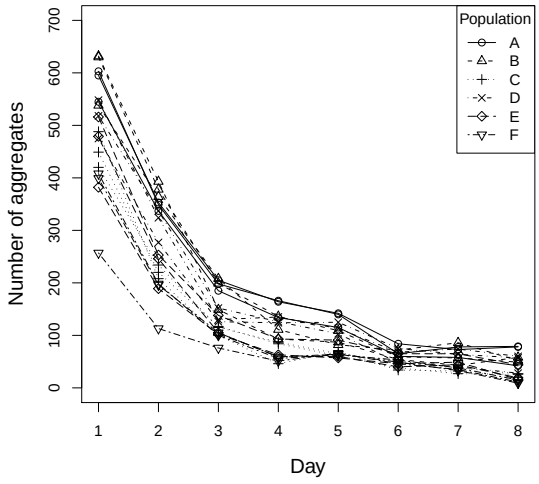

Figure 3.1: The number of aggregates, $N_{i,j}$, for each sample during days 1-8.

than the corresponding range at day 8. Furthermore, Figure 3.2 shows the total and relative difference respectively in number of aggregates between day 8 and day 1 for all samples. The difference is plotted against the amount of agarose in the cellular mixture and the letters A-F indicates the corresponding population. For convenience, a smoother is also included in the figure. From Figure 3.2 we observe that populations with less amount agarose in their cellular

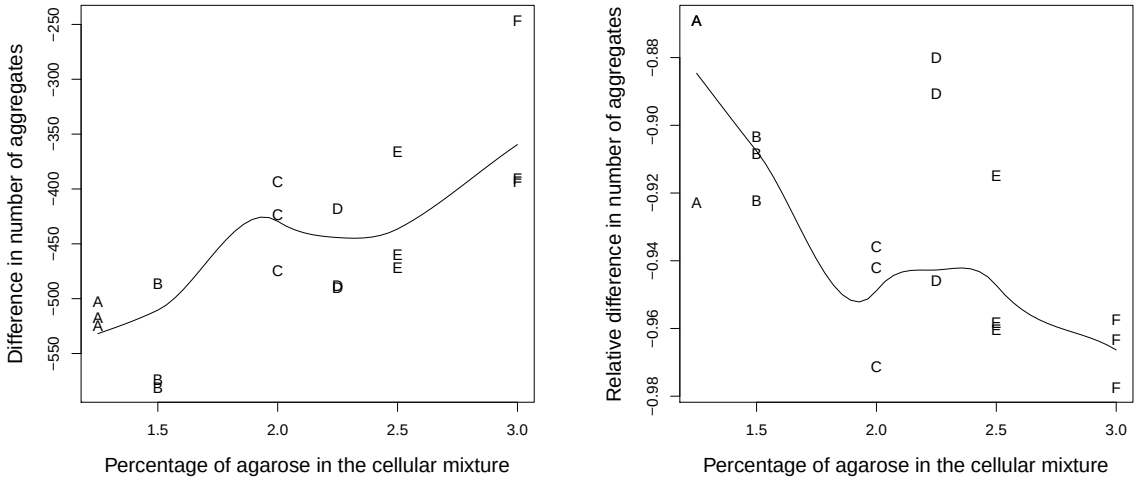

Figure 3.2: Total (left) and relative (right) difference in number of aggregates, $N_{i,j}$, between day 8 and day 1.

mixture have a larger total decrease of number of aggregates from day 1 to day 8. However, the relative difference

$$\frac{N_{i,j}(8) - N_{i,j}(1)}{N_{i,j}(1)}$$

is larger (in magnitude) for populations with a higher percentage of agarose. As seen in Figure 3.1, the samples from these populations contain fewer aggregates on day 1, which explains the difference between total and and relative difference when comparing populations. It should be noted that the relative differences are still quite close for all populations, ranging from -0.88 to -0.98.

Figure 3.3 shows a normal Q-Q plot for the number of aggregates during the days 1-8, indicating a poor fit between the distribution of the response ($N_{i,j}$) and a theoretical normal distribution. $\log(N_{i,j})$ yields a better fit between the sample distribution and a theoretical normal distribution, which also can be seen in Figure 3.3. However a Shapiro-Wilk test (used to

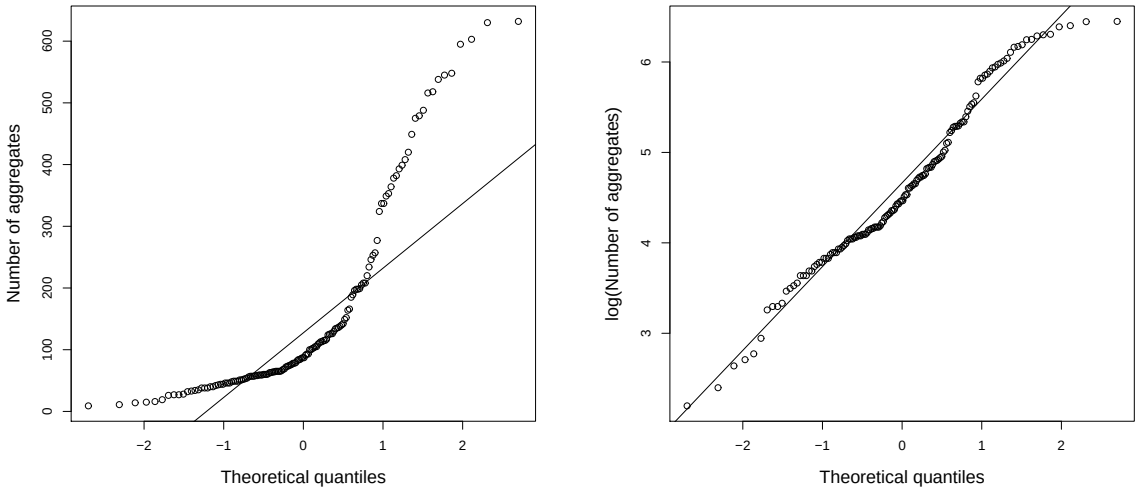

Figure 3.3: Normal Q-Q plots for $N_{i,j}$ (left) and $\log(N_{i,j})$ (right) during the days 1-8.

test the null hypothesis of normally distributed observations) indicates, with a p-value of 0.0064, that although transformed the data still cannot be considered to be normally distributed.

Days 1-6

Due to the different rates of decrease in number of aggregates over days 1-8 an exploratory analysis of the time period days 1-6 is performed (in addition to days 1-8 and 6-8). Figure 3.4 shows the number of aggregates for all samples during days 1-6. As was observed in Section 3.1, there is a larger decrease of aggregates during the first few days of the time period for all samples. Also, the range of aggregate counts for the different samples at day 1 is wide compared to the range of counts at day 6. Figure 3.5 shows the total and relative difference respectively in $N_{i,j}$ between day 6 and day 1 for all samples. The total difference follows the same pattern as for the time period days 1-8 while the relative difference of aggregates now seem to be quite equal for all populations.

A normal Q-Q plot for days 1-6 is shown in Figure 3.6. Not surprising, since the data of days 1-8 showed similar tendencies, the number of aggregates is not normally distributed. Figure 3.6 also shows the normal Q-Q plot for $\log(N_{i,j})$ and although the plot indicates heavy tails, the fit

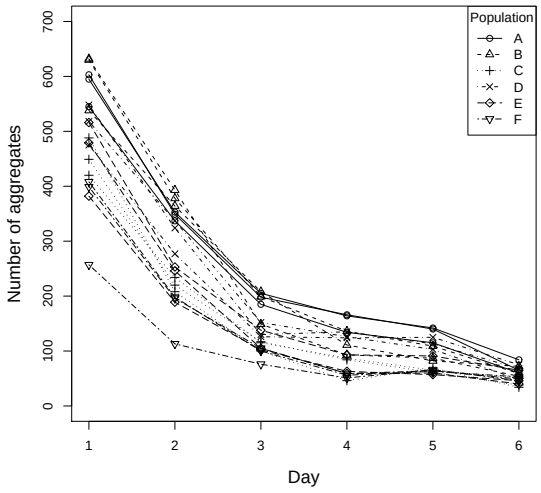

Figure 3.4: The number of aggregates for each sample during days 1-6.

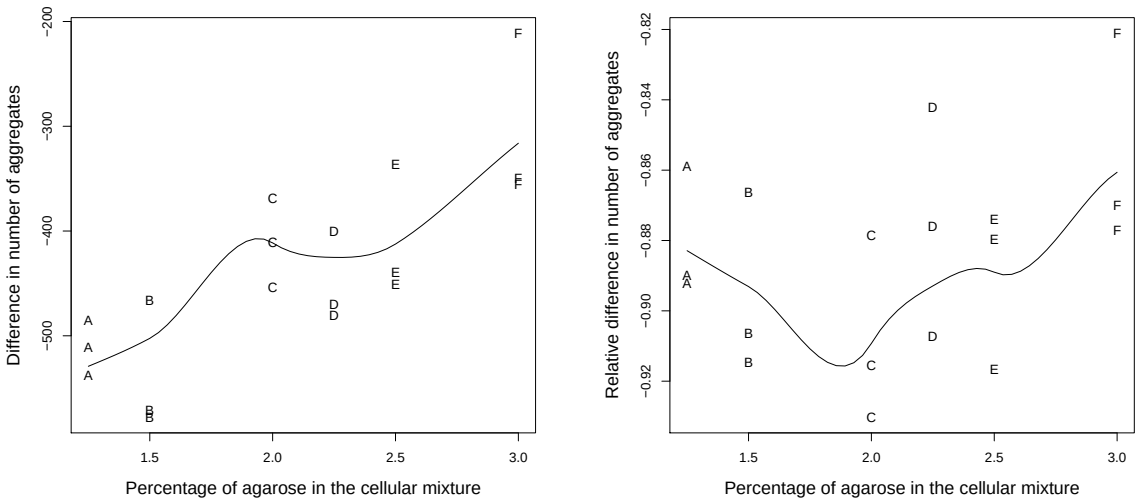

Figure 3.5: Total (left) and relative (right) difference in number of aggregates between day 6 and day 1.

to a normal distribution is better when using the transformation.

Days 6-8

Figure 3.7 shows the number of aggregates for all samples during days 6-8. With the exception of a few samples, there is a decrease in the number of aggregates during the time period. Furthermore, Figure 3.8 shows the total and relative difference respectively in number of aggregates between day 8 and day 6 for all samples. There is an obvious trend of both total and relative decrease in number of aggregates. The decrease is larger for samples in populations containing more agarose. Noticeable is that the relative decrease differs substantially between populations with low respectively high percentage of agarose in the cellular mixture.

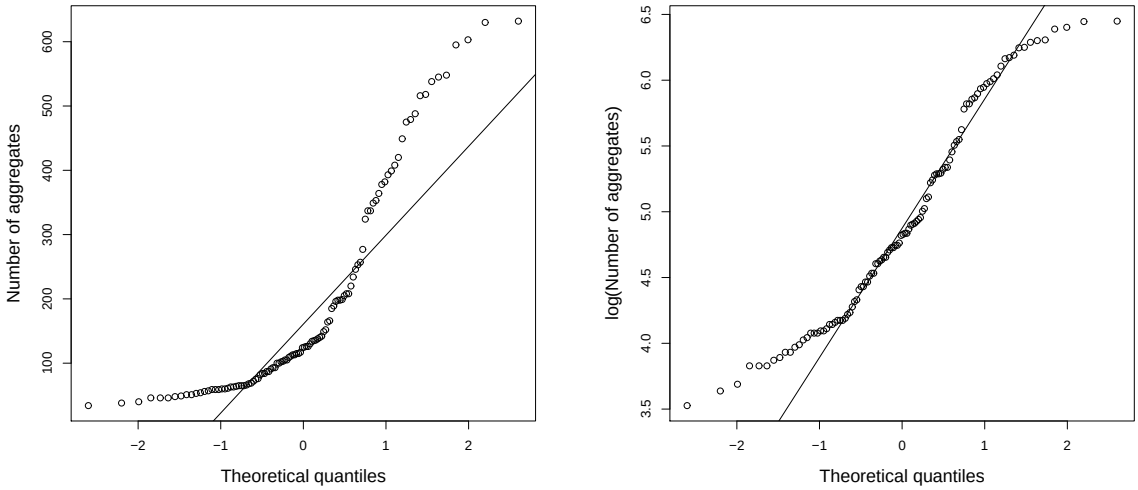

Figure 3.6: Normal Q-Q plots for $N_{i,j}$ (left) and $\log(Ni,j)$ (right) during the days 1-6.

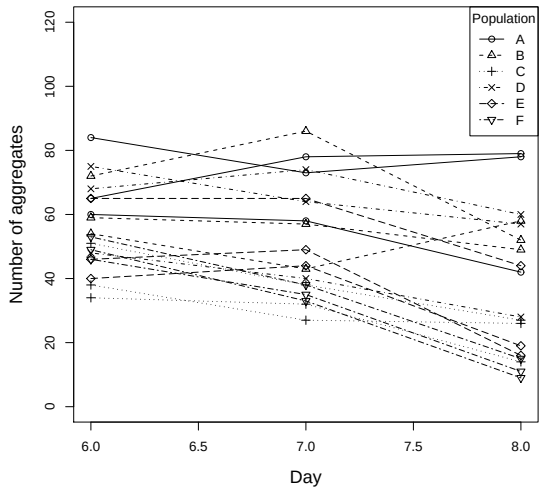

Figure 3.7: The number of aggregates for each sample during days 6-8.

No transformations will be used for the data due to that both the log and square root (common for count variables) transformations show a poor fit to a normal distribution.

3.2 Perimeter of aggregates

Days 1-8

Figure 3.9 shows the average aggregate perimeter, $\bar{P}_{i,j}$, for each sample during days 1-8. Over time there is an increase in the average perimeter values, however the rate of the increase seems to differ for different samples. Figure 3.10 shows the total and relative difference respectively in average perimeter between day 8 and day 1 for all samples in each population. For convenience there are again smoothers included in the figures. Apart from the relatively low and high values

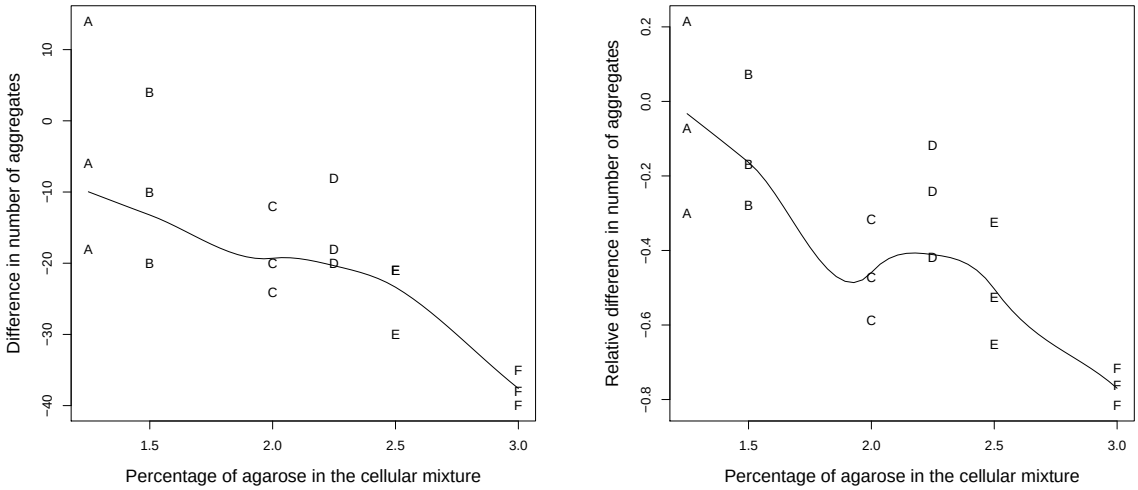

Figure 3.8: Total (left) and relative (right) difference in number of aggregates between day 8 and day 6.

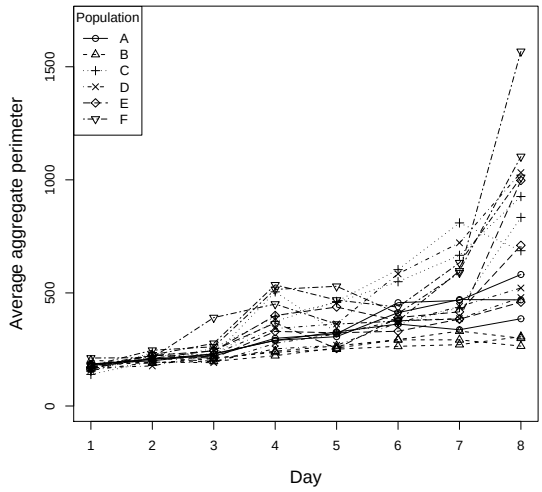

Figure 3.9: The average perimeter of aggregates, $\bar{P}_{i,j}$, for each sample during days 1-8.

of average perimeter for populations B and C respectively, the figures indicate an increasing average perimeter for populations containing more agarose.

A Shapiro-Wilk test of the observations gives a p-value < 0.0001, hence at a significance level $\alpha = 0.05$ it rejects the null hypothesis that the observations are normally distributed. This lack of fit from a normal distribution is further indicated by the normal Q-Q plot of the perimeter observations shown in Figure 3.11. The convex form of the data (with respect to the horizontal axis) indicates a distribution that is skewed to the right. p-values obtained for log and square root transformations of the data (0.0002 and < 0.0001 respectively) also indicate a lack of fit from a normal distribution. Figure 3.11 also shows the normal Q-Q plot for $\log(\bar{P}_{i,j})$. The log transformation gives a better fit to a normal distribution, although some skewness is still

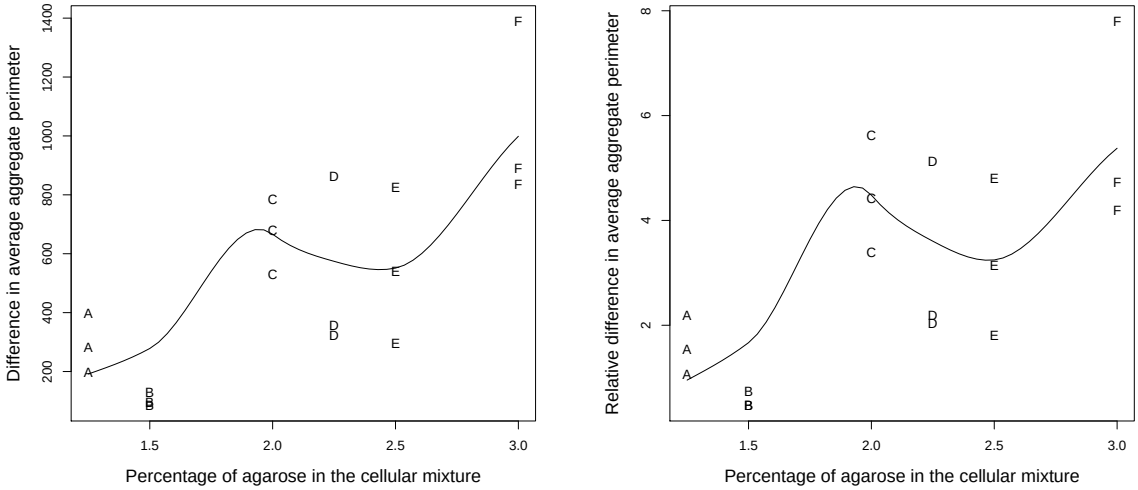

Figure 3.10: Total (left) and relative (right) difference in average perimeter $\bar{P}_{i,j}$ for each sample between day 8 and day 1.

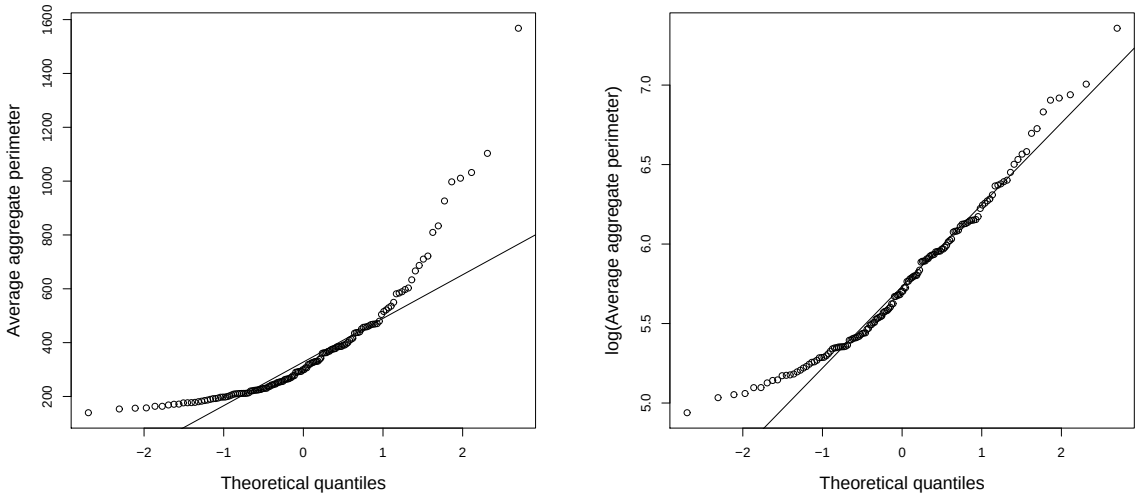

Figure 3.11: Normal Q-Q plots for $\bar{P}_{i,j}$ (left) and $\log(\bar{P}_{i,j})$ (right) for days 1-8.

present.

Days 6-8

Figure 3.12 shows $\bar{P}_{i,j}$ for each sample during days 6-8. Most noticeable is the high average perimeter value for one of the samples from population F. Moreover, there are a few samples that show very small or no increase at all in the average perimeter.

Figure 3.13 shows the total and relative difference respectively in average perimeter between day 8 and day 6 for each sample from populations A-F. Overall the pattern of increase in perimeter seems to coincide with the one for days 1-8, suggesting that a log transformation is needed for this time period as well. Figure 3.14 shows the normal Q-Q plots for the untransformed and

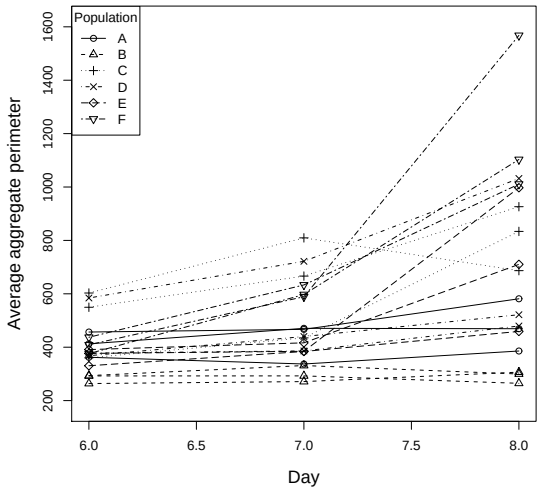

Figure 3.12: The average perimeter of aggregates for each sample during days 6-8.

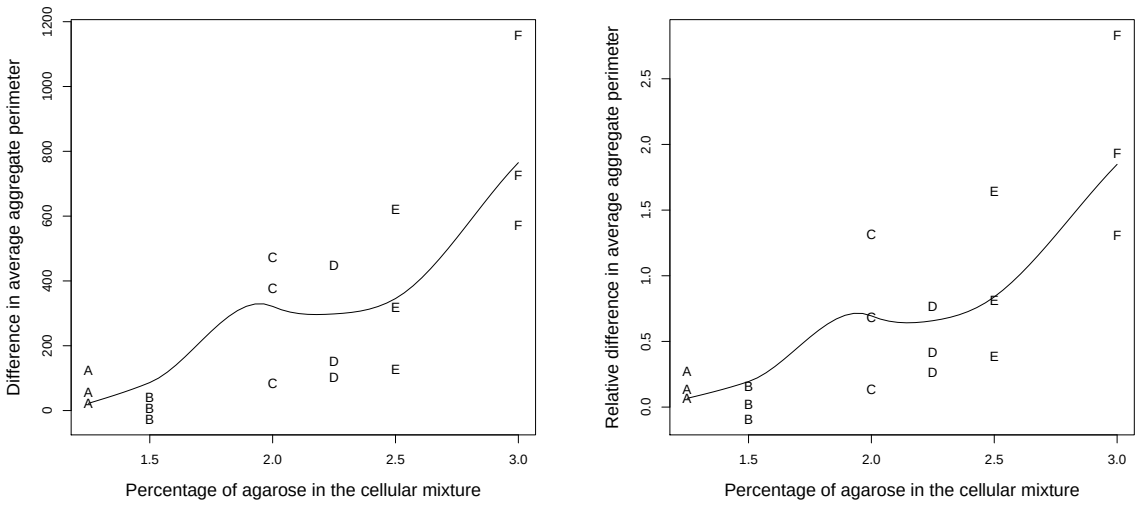

Figure 3.13: Total (left) and relative (right) difference in average perimeter $\bar{P}_{i,j}$ for each sample between day 8 and day 6.

log transformed perimeter data respectively for days 6-8. The log transformed data is a better fit to a normal distribution than the untransformed data, however the tails deviate from those of a normal distribution. A Shapiro-Wilks test (for the transformed data) gives a p-value of 0.01368.

3.3 Area of aggregates

Days 1-8

Figure 3.15 shows the total coverage area of the aggregates, $A_{i,j}$, for each sample and the mean total coverage area of the aggregates, $\hat{A}_{i,j}$, for each population respectively during days 1-8.

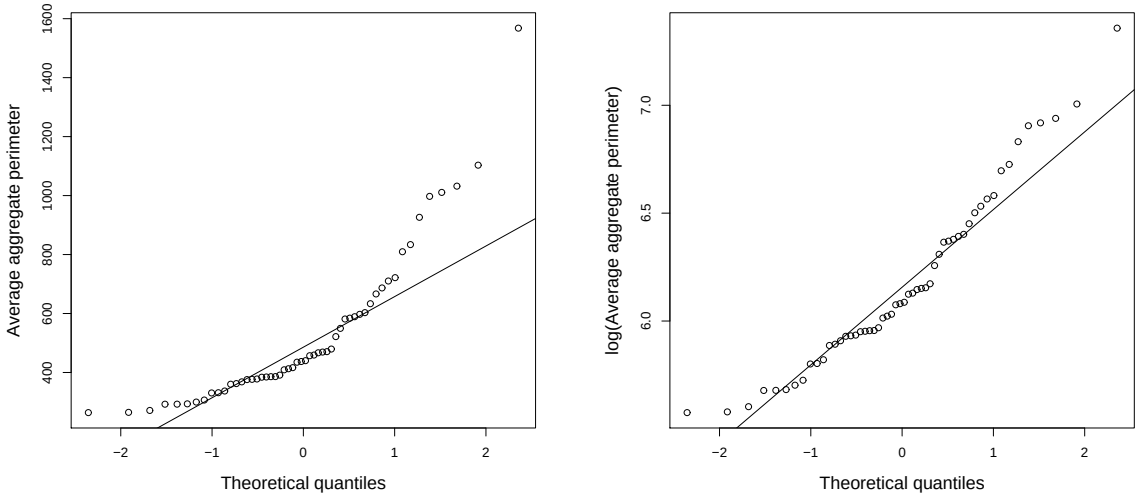

Figure 3.14: Normal Q-Q plots for $\bar{P}_{i,j}$ (left) and $\log(\bar{P}_{i,j})$ (right) for days 6-8.

It is observed that after day 3, all populations except B have an increase in their mean area.

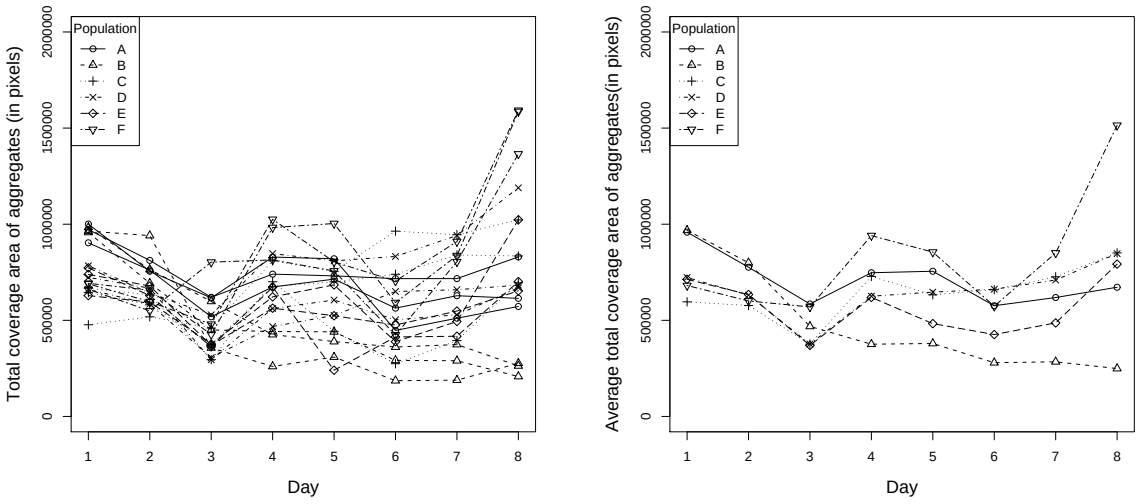

Figure 3.15: Total coverage area of aggregates, $A_{i,j}$, for each sample (left) and the mean total coverage area of aggregates, $\hat{A}_i$, for each population (right) during days 1-8.

Furthermore, several populations have a decrease in their mean between days 4 and 6, after which an increase is seen for all populations except B. These tendencies may be due to the additions to and replacement of the medium.

Figure 3.16 shows the total and relative difference in total coverage area respectively between day 8 and day 1 for the different samples. First, note that the total difference in coverage area is positive for some samples and negative for others. E.g. all samples from populations A and B have a decrease in their total coverage area from day 1 to 8 while all samples from populations C and F have increases in theirs. Moreover, note the large total increase of coverage area for population F and the almost as large total decrease for population B.

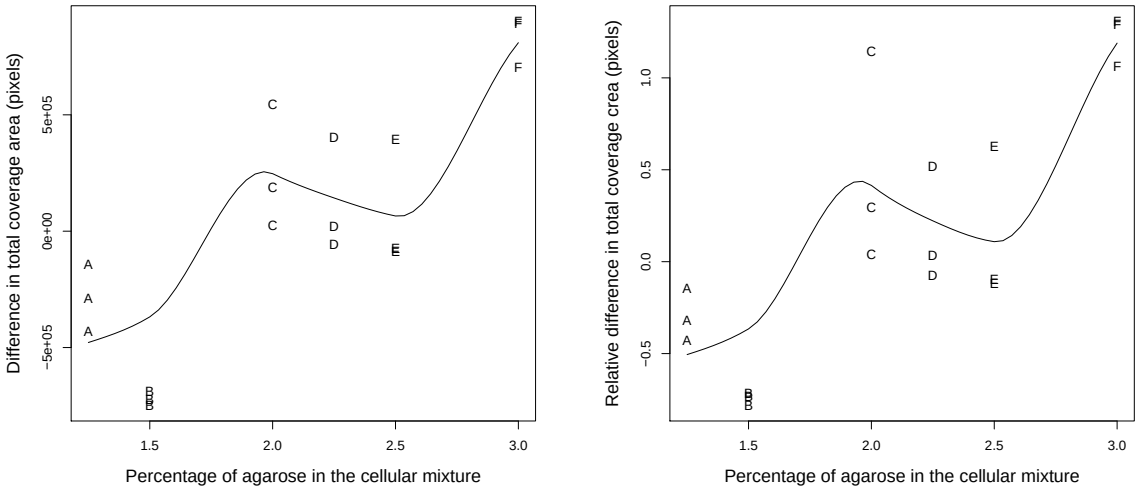

Figure 3.16: Total (left) and relative (right) difference in $A_{i,j}$ for each sample between days 8 and 1.

A Q-Q plot for the total coverage area of aggregates is shown in Figure 3.17. The plot

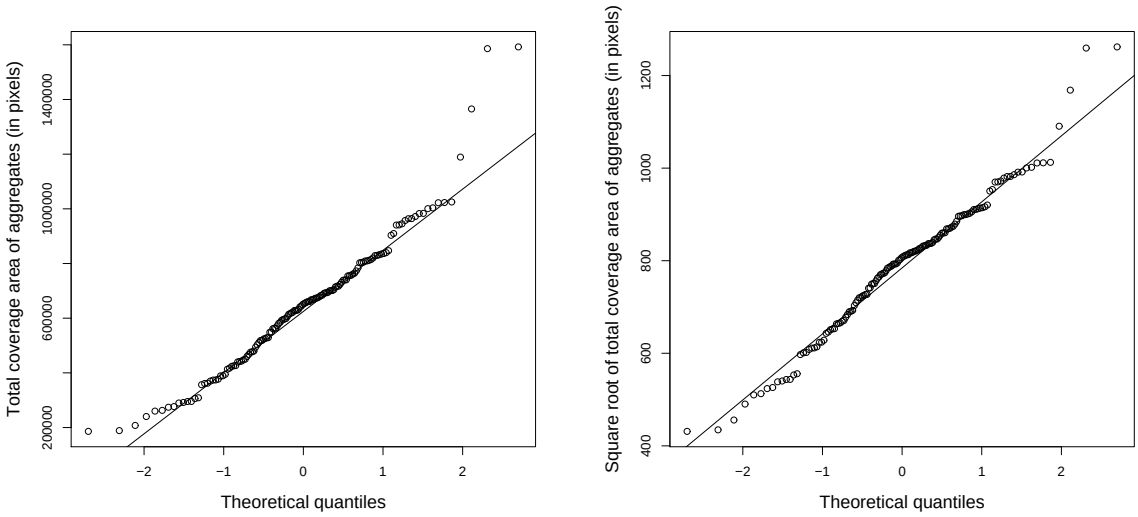

Figure 3.17: Normal Q-Q plots for (left) $A_{i,j}$ and (right) $\sqrt{A_{i,j}}$ for days 1-8.

shows a slightly heavier left tail and indicates the presence of a few outliers (particularly from population F). A Shapiro-Wilks test gives a p-value of $8.643 \cdot 10^{-5}$. Figure 3.17 also shows a Q-Q plot of the square root transformed area data. The outliers are still present but overall it seems to be a slightly better fit to a normal distribution, further concluded by a S-W test which gives a p-value of 0.0637.

Days 6-8

Figure 3.18 shows the total coverage area for each sample and the mean total coverage area for each population respectively during days 6-8. Notice especially the large increase for population

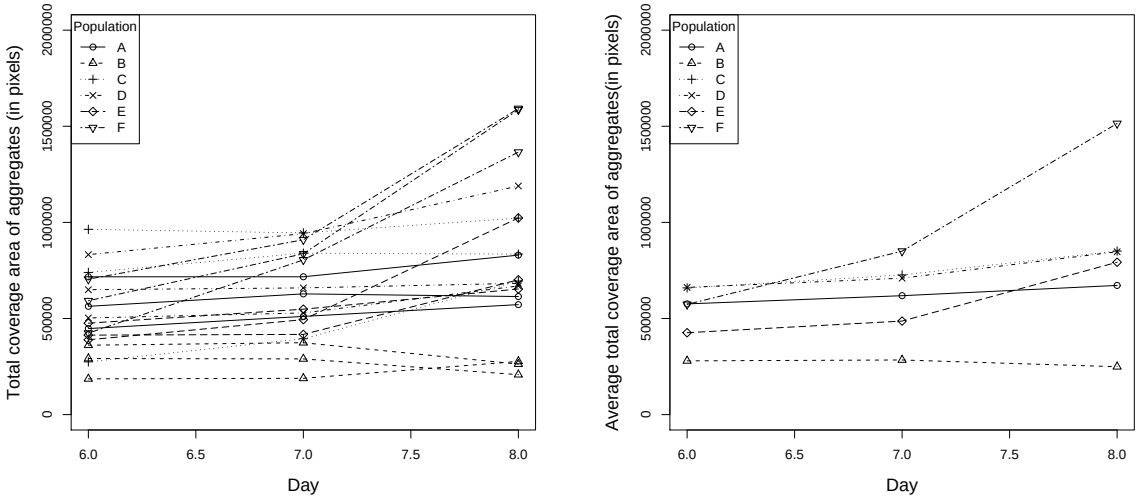

Figure 3.18: Total coverage area of aggregates, $A_{i,j}$, for each sample (left) and the mean total coverage area of aggregates, $\hat{A}_i$, for each population (right) during days 6-8.

F compared to the other populations. Studying the original and processed images (not shown here) at day 8 for the samples from population F, it is observed that the large increase may be due to the inclusion of "free space" mentioned in Chapter 2. This increase is further observed in Figure 3.19 which shows the difference (both total and relative) between days 8 and 6 for all populations. The figure indicates that the total coverage area during days 6-8 increases

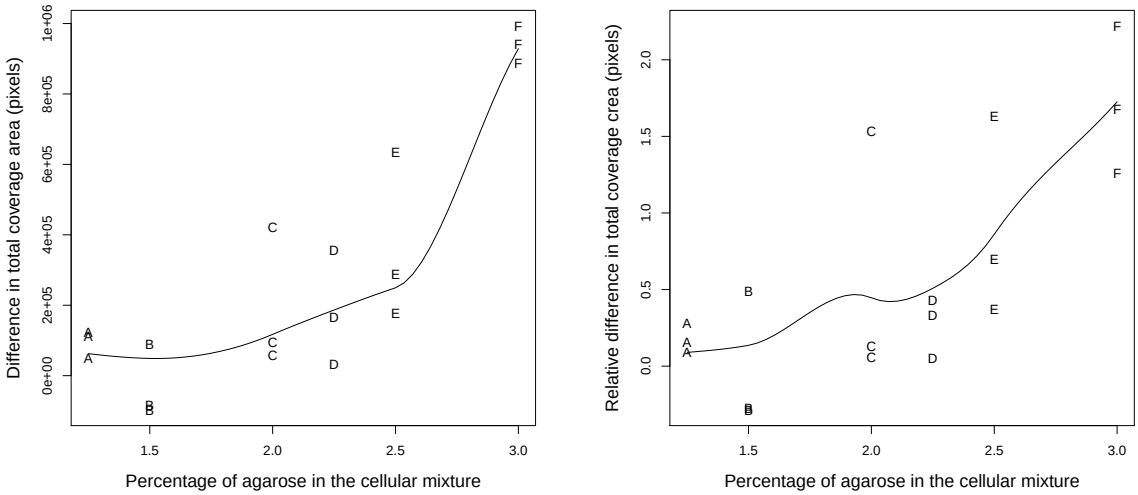

Figure 3.19: Total (left) and relative (right) difference in $A_{i,j}$ for each sample between days 8 and 6.

more for populations containing a higher percentage of agarose. Note that only two samples (population B) decreased their coverage area during days 6-8 (compared to the eight samples that were observed during days 1-8). Moreover, Figure 3.19 indicates the possible presence of outliers amongst populations, particularly for populations C and E.

A normal Q-Q plot for the total coverage area of aggregates during days 6-8 is shown in Figure 3.20. The plot is very similar to the corresponding one for days 1-8, with a slightly

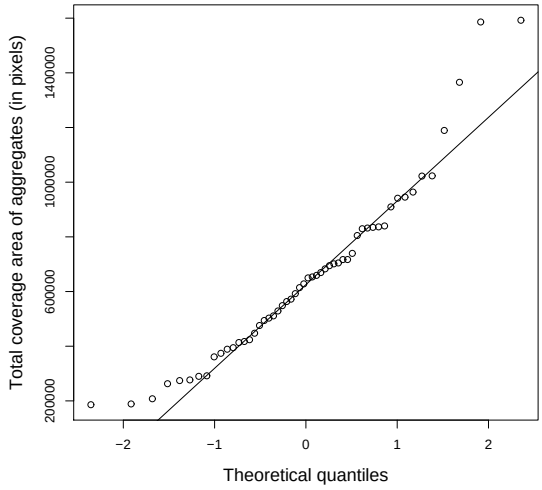

Figure 3.20: A normal Q-Q plot for $A_{i,j}$ for days 6-8.

heavier left tail and an indication of a few outliers. While a Shapiro-Wilk test rejects (at a significance level $\alpha = 0.05$) the hypothesis of normally distributed observations, tests of both log and square root transformations of the data give p-values > 0.05 (the latter transformation providing the best fit).

3.4 Circularity of aggregates

Days 1-8

The circularity, $C_{i,j}^k$, of an aggregate is defined in terms of its area and perimeter. The measure for a specific aggregate is between 0 and 1, with 1 meaning that the aggregate is shaped as a perfect circle. When the circularity value approaches 0 it refers to an increasingly elongated polygon. Figure 3.21 shows the average circularity of the aggregates for all samples ($\bar{C}_{i,j}$) and populations ($\hat{C}_{i,j}$) respectively during days 1-8. There seems to be a slight decrease in the circularity from day 1 to day 8, however no obvious trends with respect to populations are observed. The circularity ranges from 0.64 to 0.76 at day 1 and 0.52 to 0.77 at day 8.

Figure 3.22 shows the total and relative difference respectively in $\bar{C}_{i,j}$ for each sample between day 8 and day 1. Notice that (excluding population F) there are indications of a slightly larger decrease in circularity for populations with a higher percentage of agarose. However, the decrease (total as well as relative) for the samples is, for most populations, very scattered.

Figure 3.23 shows a Q-Q plot for the circularity data. The concave shape indicates that the data is skewed to the right. Note that neither a log nor a square root transformation of the data (not shown here) yields a better fit to a normal distribution.

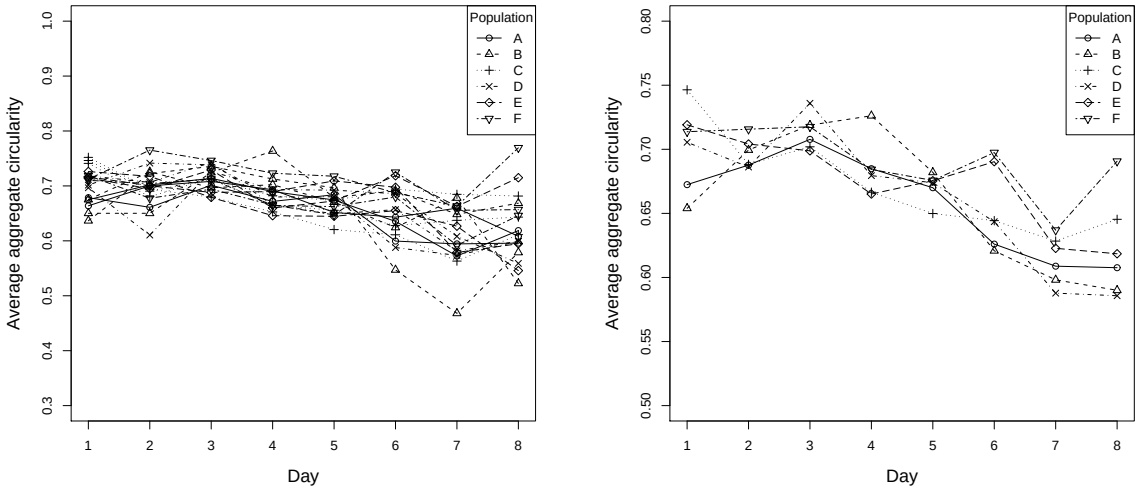

Figure 3.21: Average circularity of aggregates for samples ($\bar{C}_{i,j}$) and populations ($\hat{C}_{i,j}$) respectively during days 1-8.

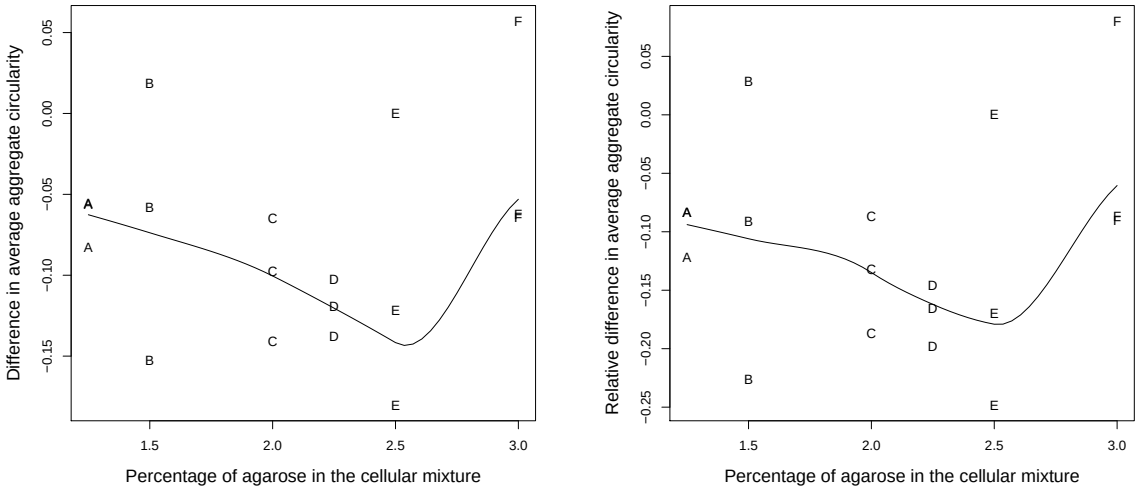

Figure 3.22: Total (left) and relative (right) difference respectively in $\bar{C}_{i,j}$ for each sample between day 8 and 1.

Days 6-8

The average circularity of aggregates for all samples ($\bar{C}_{i,j}$) and populations ($\hat{C}_i$) respectively for this time period are shown in Figure 3.24. For most samples and populations, there is a decrease between days 6 and 7. Between days 7 and 8, all populations except C and F have either a continued decrease or no change at all. For the different samples, no general trends can be observed; samples from the same population behave very differently during the time period. Figure 3.25 shows the total and relative difference in $\bar{C}_{i,j}$ during days 6-8. The observations are very scattered and samples from the same populations (except A and E) differ between increased and decreased average circularity from day 6 to 8. No obvious trends with respect to populations

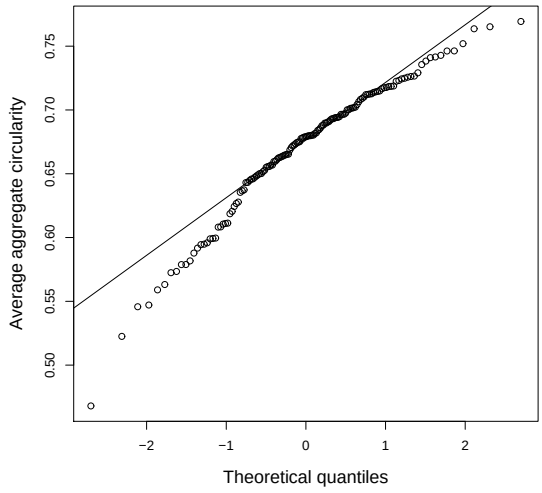

Figure 3.23: A normal Q-Q plot for $\bar{C}_{i,j}$ for days 1-8.

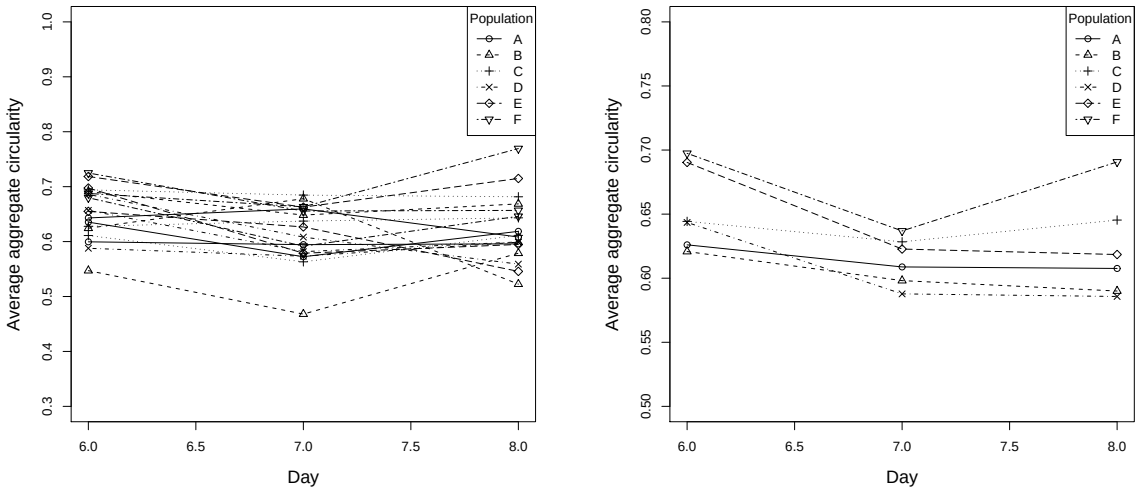

Figure 3.24: Average circularity of aggregates for samples ($\bar{C}_{i,j}$) and populations ($\hat{C}_{i,j}$) respectively during days 6-8.

are observed.

In Figure 3.26 a Q-Q plot for the average circularity of aggregates, days 6-8, is shown. The plot shows that the normal distribution is a good fit for the data. This is further indicated by a Shapiro-Wilk test (p-value 0.9446, significance level $\alpha = 0.05$).

3.5 Comments

Population B behaves different for all measures as compared to the rest of the populations; it tends to not follow trends that seem to hold for the others. An explanation for this is that cells have sunken through the cellular mixture upon which they have been placed, putting them out

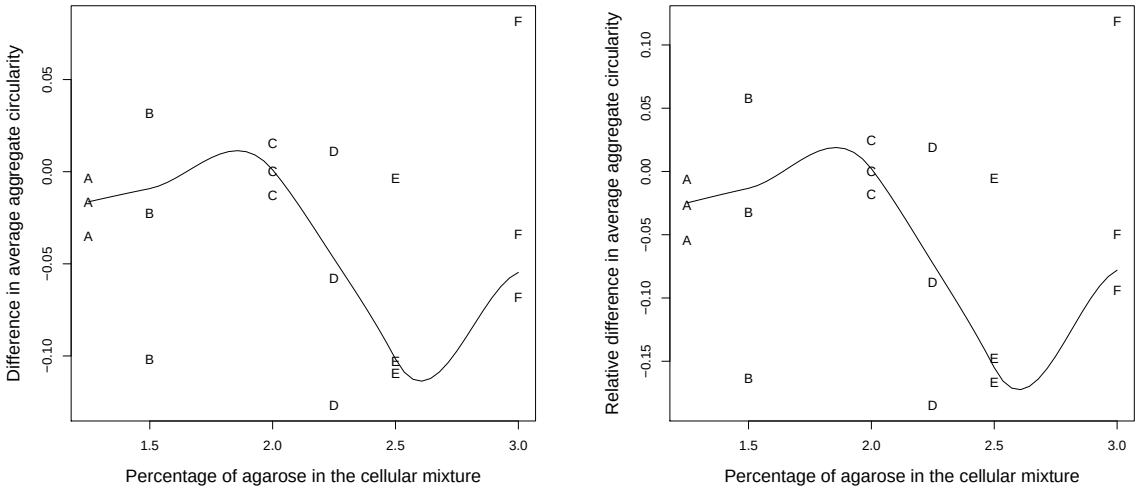

Figure 3.25: Total (left) and relative (right) difference respectively in $\bar{C}_{i,j}$ for each sample between day 8 and 6.

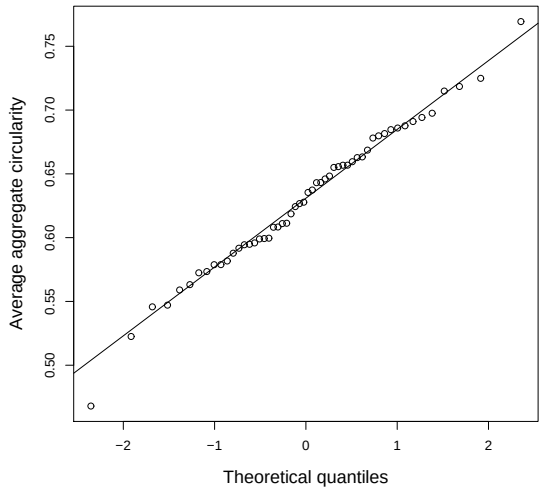

Figure 3.26: A normal Q-Q plot for the average circularity of aggregates for days 6-8.

of focus of the images and thus excluding them from the data. Such cells then reappear and disappear again during the time period of the experiment. This can be qualitatively observed in the original images for the samples of population B.

With few exceptions, the different data sets are not normally distributed. In some cases a transformation of the data yields a better fit, however still with indications of heavy tails, skewness etc.

The area data has been further investigated in [31] using regression analysis. Different models for total coverage area as a function of time and stiffness[1] are considered. Results indicate that a higher percentage of agarose in the cellular mixture implies a larger total coverage area. However,

[1]Here stiffness corresponds to what we have denoted as *population*

it is in [31] noted that the increase in total coverage area for wells with a high percentage of agarose is affected by the inclusion of "free space" when images are processed. This is consistent with what have been mentioned here. Due to this discrepancy for some populations and the previous investigation of the area data, we do not pursue any further analysis of this measure.

Moreover, due to the circularity measure being defined in terms of estimated area and perimeter, further investigations regarding this measure is omitted. For now focus is instead on the perimeter data and the analysis of the number of aggregates in each sample.

As mentioned in Chapter 2 the data obtained from the experiments contained the position of each aggregate for the different days and samples. Attempts to quantify this information have been made, however it has shown to be a task of great difficulty due to the fact that there is no possibility to follow a certain cell or particle over time. For conclusions regarding cell movement, it would be desirable to obtain data from tracking of distinct cells over a period of time, thus enabling cell paths to be obtained for the different samples. Mathematical models for single cell movement (e.g. [15], [26]) may perhaps be adjusted to fit the specific characteristics of breast cancer cells.

Chapter 4

Repeated measurements for longitudinal data

The concept of repeated measurements is found in a vast number of real life situations and applications. In general, the term repeated measurements refers to the case when a characteristic of a subject is observed a repeated number of times.

Repeated measurements data has certain characteristics that makes it hard to use conventional methods and models for analysis. Perhaps the most significant difficulty is that due to multiple observations being made on the same subject there is a dependence between observations. Therefore, to fully investigate the data and obtain inference on it, it is necessary to take this dependence into account. There are more complications that arise with repeated measurements (unbalanced or missing data being others), however it is this dependence that is of most concern for the experiments presented in this thesis.

Different methods have been developed or altered for use on repeated measurements data. Examples are univariate and multivariate, ANOVA and generalized linear model (GLM) methods. For an account on how they can be used in the setting of repeated measurements and what their advantages and disadvantages are, see e.g. [11]. The perhaps most common approach to analyze repeated measurements data is that of linear mixed models, methods from which are used in this thesis. Using linear mixed models (LMM), it is possible to quantify and model the dependence between observations on a specific subject.

In Section 4.1 the general linear mixed model is defined and the concept of covariance structures used in the model is discussed. Following mainly [11] and [29], in Section 4.2 it is briefly described how the linear mixed model is applied to repeated measurements data.

This is by no means an attempt to give a complete description of the model and its properties. The interested reader is recommended to view e.g. [22] or [29] for a good and thorough treatment of the theory of linear mixed models. [11] gives a more brief introduction to the theory and its use for repeated measurements, including several examples.

4.1 The linear mixed model

Suppose that n observations of some sort are available (e.g. responses from an experimental unit, measurements of some kind etc.). One of the usual statistical approaches for analyzing such experimental data is to try and fit a linear model of the form

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \tag{4.1}$$

Here $\mathbf{y}$ is an $n \times 1$ vector of the independent observations, β is a $m \times 1$ vector of unknown regression coefficients, $\mathbf{X}$ is an $n \times m$ matrix consisting of the proposed covariates and ε is an $n \times 1$ vector of independent errors. It is usually assumed that the components of ε are independent with zero mean and some constant variance σ^2. This is commonly known as a general linear model.

The linear *mixed* model can instead be formulated as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon. \tag{4.2}$$

Here, with the same notation as in Equation 4.1, $\mathbf{Z}$ is a $n \times p$ design matrix of known covariates. γ is a $p \times 1$ random vector of subject specific effects and ε is a random vector of residual components. While $\mathbf{X}$ is the design matrix for the fixed effects, $\mathbf{Z}$ is the design matrix for the random effects between subjects. In the linear mixed model, the following assumptions are made regarding the random vectors

$$\begin{aligned} \mathrm{E}[\gamma] &= \mathbf{0}_p, \\ \mathrm{E}[\varepsilon] &= \mathbf{0}_n, \\ \mathrm{Var}(\gamma) &= \mathbf{B}, \\ \mathrm{Var}(\varepsilon) &= \mathbf{W}, \end{aligned}$$

where $\mathbf{B}$ and $\mathbf{W}$ are some arbitrary covariance matrices. Furthermore γ and ε are assumed to be uncorrelated. With these assumptions on γ and ε it holds that

$$\begin{aligned} \mathrm{E}[\mathbf{y}] &= \mathbf{X}\beta, \\ \mathrm{Var}(\mathbf{y}) &= \mathbf{Z}\mathbf{B}\mathbf{Z}' + \mathbf{W}. \end{aligned}$$

A main feature of the model is that the components ε_i, $1 \leq i \leq n$, of ε are not necessarily assumed to be independent. Note that when they are, i.e. $\mathbf{W} = \sigma^2\mathbf{I}$, and $\mathbf{Z} = 0$, the mixed model becomes the standard linear model (4.1).

4.1.1 Covariance structures

Important properties of (4.2) are the covariance structures of γ and ε. Letting Σ denote the covariance matrix of some random vector $\alpha = [\alpha_1, ..., \alpha_n]$, the following are examples of some common covariance structures. Throughout, σ^2 represents some variance and ρ represents correlation.

The simplest covariance structure is the independent covariance model, i.e.

$$\Sigma = \sigma^2 \mathbf{I}.$$

This structure corresponds to the elements of α being uncorrelated.

The simplest covariance structure that takes correlation into account is known as the *compound symmetry model* (abbreviated CS). For this model (with the matrix symmetric with respect to the diagonal)

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & ... & \rho \\ & 1 & \rho & ... & \rho \\ & & 1 & ... & \rho \\ & & & ... & ... \\ & & & & 1 \end{pmatrix}.$$

Hence CS implies that the correlation will be the same for any pair of (different) elements of α.

The *first-order autoregressive*, AR(1), model is slightly more sophisticated, with the covariance matrix given by

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & ... & \rho^{n-1} \\ & 1 & \rho & ... & \rho^{n-2} \\ & & 1 & ... & \rho^{n-3} \\ & & & ... & ... \\ & & & & 1 \end{pmatrix}.$$

The difference between this model and CS is that it implies that elements of α that are close to each other (i.e. the difference $|i - j|$ of their indices is small) tend to be higher correlated than those farther apart. When α is a vector of observations over time, this corresponds to observations close in time being more highly correlated than those farther apart in time. There is a continuous time version of this structure as well, corresponding to letting the covariance structure be a continuous time AR(1) process and obtaining the covariance between observations according to such a process.

A covariance model similar to the AR(1) model is the *Toeplitz* model. The covariance matrix

for this model is

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ & 1 & \rho_1 & \dots & \rho_{n-2} \\ & & 1 & \dots & \rho_{n-3} \\ & & & \dots & \dots \\ & & & & 1 \end{pmatrix}.$$

Thus the Toeplitz model, like the AR(1), gives the same correlation for pairs of elements of α that are an equal distance apart. However, the Toeplitz model is more complex since there is no known function between the distance $|i-j|$ between elements, and the correlation $\rho_{|i-j|}$.

The AR(1) and Toeplitz models take into account that correlation is different for pairs of elements of α separated by different distances. Another property to model for is different variances for different elements. Examples of models that takes this into account are the *Heterogeneous AR(1)*, ARH(1), and *Heterogeneous Toeplitz*, TOEPH. Such models have covariance matrices in which the (i,j):th elements are $\sigma_i\sigma_j\rho^{|i-j|}$ and $\sigma_i\sigma_j\rho_{|i-j|}$ respectively. Thus the number of parameters is $n+1$ for the ARH(1) model and $2n-1$ for the TOEPH model.

The most complex structure available is the *unstructured* covariance model (UN). Each pair of elements has its own unique covariance, thus making the matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \dots & \sigma_{1,n} \\ & \sigma_2^2 & \sigma_{2,3} & \dots & \sigma_{2,n} \\ & & \sigma_3^2 & \dots & \sigma_{3,n} \\ & & & \dots & \dots \\ & & & & \sigma_n^2 \end{pmatrix}.$$

Therefore the UN model requires the estimation of $\frac{n(n+1)}{2}$ parameters.

The covariance structures described here are some examples of the most common structures, but there is a wide variety of others to use. In (4.2), the different Σ can be used for both $\mathbf{B}$ and $\mathbf{W}$. However, in most cases one of $\mathbf{B}$ and $\mathbf{W}$ is assigned an independent covariance structure, i.e. $\sigma^2\mathbf{I}$, and only one is assigned a more complex covariance structure. That is, only one of the vectors γ and ε has correlated elements. Which of $\mathbf{B}$ and $\mathbf{W}$ that is normally assigned $\sigma^2\mathbf{I}$ differs in reference literature, e.g. [25] uses $\mathbf{B} = \sigma^2\mathbf{I}$ in most cases while [29] uses $\mathbf{W} = \sigma^2\mathbf{I}$ as standard.

4.2 Application of LMM to repeated measurements data

Assume that we have longitudinal data for n subjects (e.g. patients). Let, for subject i,

$$\mathbf{y}_i = (y_{i,1},\dots,y_{i,t_i})',$$

be the $t_i \times 1$ vector of observations. Applying (4.2) to each subject gives the equations

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\gamma_i + \varepsilon_i, \qquad i = 1,2,...,n. \tag{4.3}$$

It is assumed that, with g the dimension of the γ_i:s,

- the γ_i vectors are $N_g(\mathbf{0}_g, \mathbf{B})$,
- the ε_i vectors are $N_{t_i}(\mathbf{0}_{t_i}, \mathbf{W}_i)$,
- γ_1, ..., γ_n, ε_1,...,ε_n independent.

Note that while $\mathbf{X}_i$, $\mathbf{Z_i}$ and $\mathbf{W}_i$ are subject-specific matrices, $\mathbf{B}$ is not. The assumptions give that $\mathbf{y}_1$,...,$\mathbf{y}_n$ are independent $N_{t_i}(\mathbf{X}_i\beta, \mathbf{V}_i)$, where

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{B}\mathbf{Z}_i' + \mathbf{W}_i. \tag{4.4}$$

The choice of structures for $\mathbf{B}$ and $\mathbf{W}_i$ is an essential part of the LMM. For longitudinal data the CS structure is often considered unrealistic. Structures that allow for a decrease in correlation as the distance (in time) between pairs of observations increase, such as AR(1) and Toeplitz, are often a better fit. Note however that both AR(1) and Toeplitz require that the time between adjacent observations is held constant for each subject (but can change between subjects).

For analyzing repeated measurements data using LMM methods, the **MIXED** procedure in SAS is often preferred. In particular, some 20 different covariance structures (including those in Section 4.1) are included in this procedure. See [25], [32] for more on how to use SAS.

Chapter 5

Modeling of count and perimeter data

In Chapter 3 it was concluded that the **number of aggregates** and **perimeter** data seem suitable for further analysis. Recalling the definition in Chapter 4, it is clear that the experimental data obtained as was described in Chapter 2 is of repeated measurements type. Therefore, it is appropriate to analyze the data using methods from linear mixed models.

As stated in Chapter 4, the LMM approach to data analysis is developed for normal data. However, in Chapter 3 it was concluded that even when using transformations, the data generally cannot be considered as normally distributed. We have here used the convention to apply normal-theory methods even when the assumption of a normal distribution does not hold.

Section 5.1 contains the analysis of the number of aggregates data and Section 5.2 the corresponding analysis for the perimeter data. Emphasis is put on finding appropriate covariance structures (following methods described in [25]) and drawing inference on any differences between the different populations. Note that Population is always considered as a categorical variable. Section 5.3 contains some comments about the modeling and the obtained results.

5.1 Analysis of aggregate count

Recall from Chapter 3 that the number of aggregates $N_{i,j}$ in each sample j decreased over time. We will with the results presented in this section try to determine whether the patterns of change of aggregates are different for different populations. Section 5.1.1 presents the modeling and selection of covariance structures for the different time periods of the data and Section 5.1.2 presents the results from regression analysis using the selected covariance structures.

5.1.1 Modeling the covariance structure

To select an appropriate covariance model the patterns of correlation between observations at different times have been examined, as well as information criteria that measure the fit of competing covariance models. In the following sections we model the covariance structures for different time periods of the data – days 1-8, days 1-6 and days 6-8.

Days 1-8

Patterns of correlation structure can be visualized by plotting changes in covariance and correlation among residuals on the same sample (here from the same well) at different times over distance between times of observation. Figure 5.1 shows such a covariance plot. The values

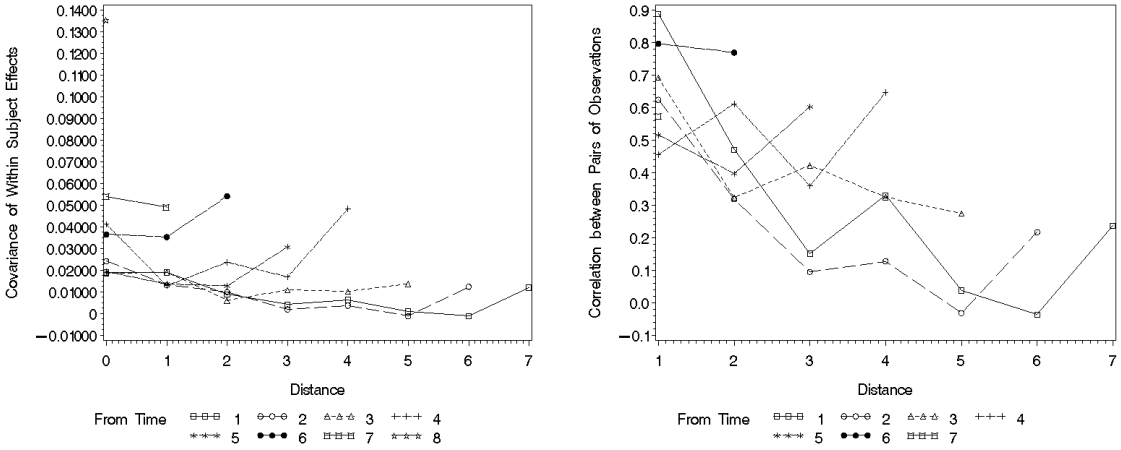

Figure 5.1: The covariance (left) and correlation (right) as a function of distance in time between pairs of observations for days 1-8.

plotted at the distance 0 are the variances among the observations for each of the eight days. Furthermore, the profile "From time 1" gives the covariance between pairs of measurements whose first observation occurred at Day 1. The covariance between days 1 and 2 is plotted at the distance 1, the covariance between days 1 and 3 at the distance 2 and so forth. The figure indicates that when the distance between pairs of observations increase the covariance tends to decrease. However, notice the increase in variance (i.e. the value at distance 0) among the observations for day 6, 7 and 8. This trend of increasing variance suggest that a heterogeneous covariance model may be the best fit to the data.

Figure 5.1 also shows the correlation of residuals on the same subject at different times over the distance between times of observation. The concept of the reference times is the same as for the covariance plot. At a distance 0 all correlations are 1 regardless of the variance and therefore this plot cannot be used to determine if a heterogeneous covariance model is the best fit. However, there is a trend of decreasing correlation when the distance between pairs of observations increases. Hence Figure 5.1 indicates that a good fit could be a covariance structure where adjacent observations are more correlated than observations farther apart.

Table 5.1 gives the Akaike information criterion (AIC) for the four covariance structures UN, CS, AR(1) and ARH(1); the model that minimizes the information criterion is the preferred one. If several models seem to be equally good, the simpler one is preferred. Comparing the complexity of the models, UN is more complex than AR(1) since the number of parameters needed to be estimated is far more. Furthermore, since a unique variance has to be estimated for each day in the ARH(1) model, this is obviously a more complex model than AR(1).

Table 5.1: Akaike information criterion for four plausible covariance structures for days 1-8.

Covariance structure	AIC
Unstructured model	-22.9
Compound symmetric model	5.1
First-order autoregressive model	-8.5
Heterogeneous first-order autoregressive model	-27.4

The smallest AIC is obtained for the heterogeneous first-order autoregressive model. Hence the suspicion about a trend among variances, seen in the covariance and correlation plots, is confirmed. Therefore in the following analysis, the ARH(1) model will be used to model the covariance structure for days 1-8.

Days 1-6

Figure 5.2 shows the covariance and correlation respectively among residuals from the same sample at different times over distance between times of observations for days 1-6. The concepts of reference times and distances are the same as in the previous section. Not surprisingly, when the distance between pairs of observations increases the covariance and correlation tends to decrease. However, looking at the correlation plot there does not seem to be a trend of increasing (or decreasing) variance with time of observation, suggesting that a model with constant variance over time would be adequate.

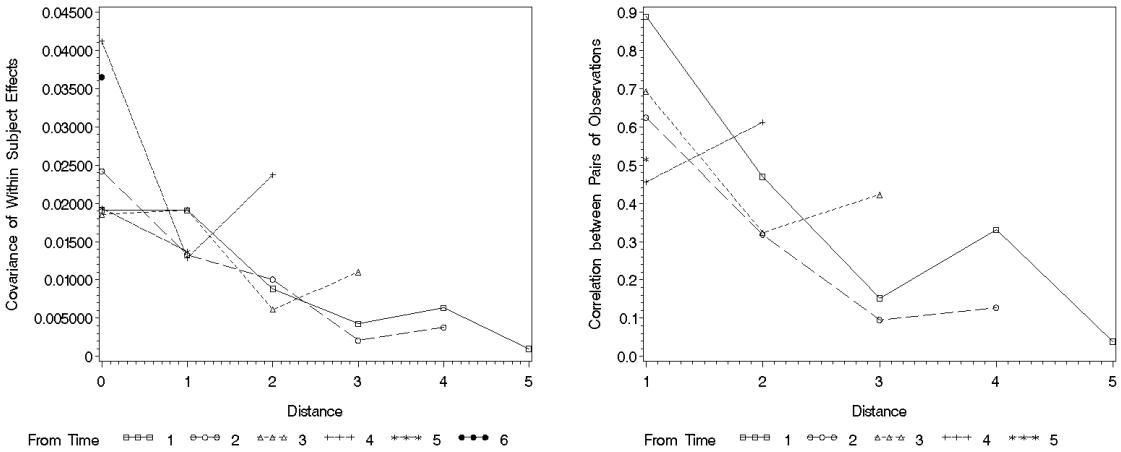

Figure 5.2: The covariance (left) and correlation (right) as a function of distance in time between pairs of observations for days 1-6.

Table 5.2 gives the Akaike information criterion (AIC) for the four covariance structures UN, CS, AR(1) and ARH(1). What was indicated from the covariance and correlation plots is also concluded by the AIC, i.e. an autoregressive covariance model is the best fit for the data. A heterogeneous model gives a slightly smaller AIC value than a homogeneous model, however

Table 5.2: Akaike information criterion for four plausible covariance structures for days 1-6.

Covariance structure	AIC
Unstructured model	-37.7
Compound symmetric model	-30.3
First-order autoregressive model	-38.0
Heterogeneous first-order autoregressive model	-39.1

this difference is small and therefore the complexity of the models should be considered. As mentioned above a heterogeneous model is far more complex than a homogeneous and therefore the covariance structure during days 1-6 will be modeled by an AR(1) model.

Days 6-8

Due to the short time period, covariance and correlation plots would provide very limited information about trends in covariance and correlation respectively and will therefore not be shown. However, Akaike's information criterion for different covariance models can still be considered, the result is shown in Table 5.3. The model with the smallest AIC value is CS, indicating equal

Table 5.3: Akaike information criterion for four plausible covariance structures for days 6-8.

Covariance structure	AIC
Unstructured model	295.7
Compound symmetric model	290.9
First-order autoregressive model	292.7
Heterogeneous first-order autoregressive model	293.7

correlation between pairs of measurements regardless of the distance between them. This is the model that will be used in future analysis. Note however that here all the covariance structures had AIC close to each other, making any one of them a valid choice – the CS is chosen due to its simple form.

5.1.2 Regression analysis

The main focus is to establish whether or not there are any significant differences in the patterns of change in number of aggregates for different populations at different days. The following sections present the results from regression analysis for the different time periods (days 1-8, days 1-6 and days 6-8) using the appropriate covariance structures modeled in previous sections. The model selection has mainly been based on AIC. The models that have been tested include Population and Day as covariates, and will consist of different combinations of Population, Day^k and interaction terms of the type Population $\times \text{Day}^k$, $k = 1,2$. As mentioned earlier the variable Population has been considered as categorical.

Days 1-8

Figure 5.3 shows the average of $\log(N_{i,j})$, for each population. As time goes, there is obviously

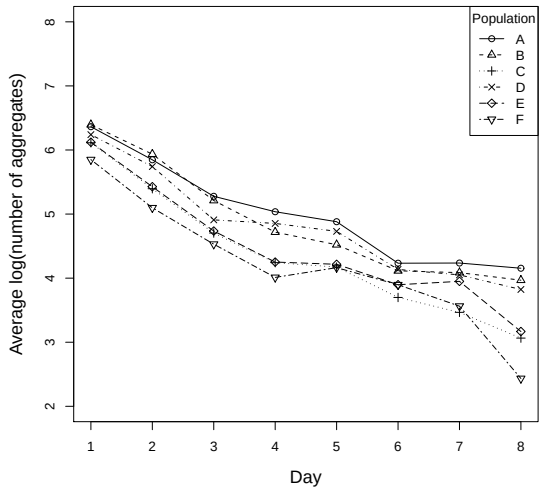

Figure 5.3: The average of $\log(N_{i,j})$ for each population (A-F) during days 1-8.

a decrease in $\log(N_{i,j})$ which seems to be linear with some curvature and almost the same for all populations. Regression indicates that the best fitted model for the fixed effects is

$$\log(N_{i,j}) = \beta_0 + \beta_{\text{Population}} + \beta_1 \times \text{Day} + \beta_2 \times \text{Day}^2, \tag{5.1}$$

where $\beta_{\text{Population}}$ is different for each population (i corresponding to the specified one). Notice that the model does not include an interaction term between Population and Day, i.e. there are no significant differences in the decrease of the response variable for different populations. This coincides with what can be seen in Figure 5.3: The lines corresponding to different populations appear to be, for the most part, parallel. The solution (estimates of coefficients and their standard errors, t-test statistics and corresponding p-values) for this particular model is given in Table 5.4.[2] Thus, e.g. for a sample from population F, the model for the fixed effects would be

$$\log(N_{6,j}) = \underbrace{6.672}_{\text{Intercept}} - 0.852\ \text{Day} + 0.0587\ \text{Day}^2.$$

Days 1-6

Figure 5.4 shows the average of $\log(N_{i,j})$ for each population i during days 1-6. As for the days 1-8, a decrease in $\log(N_{i,j})$ is observed. The decrease for all populations again seem to be linear with some curvature. This is not very surprising since we are now considering a fairly large subset of the observations for days 1-8. The model tests reveal that the best fit for the fixed

[2]The paired t-test for the populations compares the means of the fixed effects for two populations. Population F is taken to be the reference category. The p-value corresponding to the t-statistic indicate whether or not the difference in mean between the populations is statistically significant (<0.05)

Table 5.4: Solution for fixed effects of the model for days 1-8; estimates of coefficients and their standard error, t-test statistics and p-values.

Effect	Estimate	St. error	t value	Pr > |t|	
Intercept	6.6717	0.07780	85.75	<.0001	***
Population A	0.5292	0.8899	5.95	<.0001	***
Population B	0.4996	0.8899	5.61	0.0001	***
Population C	0.1141	0.8899	1.28	0.2239	NS
Population D	0.3247	0.8899	3.65	0.0033	**
Population E	0.2065	0.8899	2.32	0.0387	*
Population F	0	.	.	.	
Day	-0.8516	0.03518	-24.21	<.0001	***
Day^2	0.05873	0.004682	12.54	<.0001	***

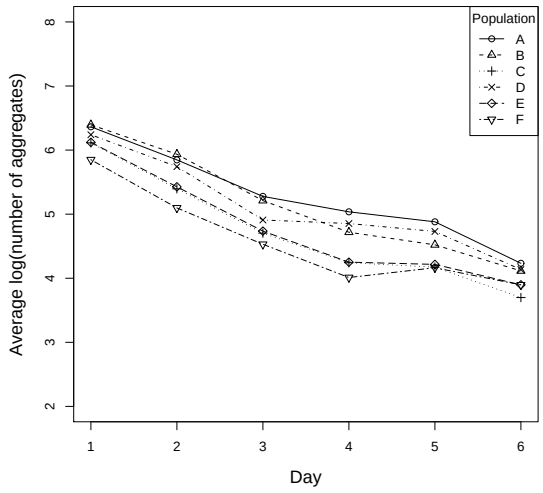

Figure 5.4: The average of $\log(N_{i,j})$ for each population (A-F) during days 1-6.

effects was

$$\log(N_{i,j}) = \beta_0 + \beta_{\text{Population}} + \beta_1 \times \text{Day} + \beta_2 \times \text{Day}^2. \tag{5.2}$$

Note that this corresponds to the model (5.1) that was selected for days 1-8. Again, the interaction term between Population and Day is left out, indicating no significant differences in the patterns of change in aggregates over time between different populations. The Day^2 term could explain the small curvature that is observed in Figure 5.4. This is further indicated by the positive estimate of the corresponding coefficient, given in Table 5.5 along with the rest of the solution for (5.2). As in the previous section, the estimates in Table 5.5 can be used to set up models of the fixed effects corresponding to different populations (with F the reference population).

Table 5.5: Solution for fixed effects of the model for days 1-6; estimates of coefficients and their standard error, paired t-test statistics and p-values.

Effect	Estimate	St. error	t value	Pr > |t|	
Intercept	6.5778	0.1054	62.41	<.0001	***
Population A	0.6523	0.09281	7.03	<.0001	***
Population B	0.5374	0.09281	5.79	<.0001	***
Population C	0.1182	0.09281	1.27	0.2268	NS
Population D	0.4879	0.09281	5.26	0.0002	***
Population E	0.1776	0.09281	1.91	0.0798	.
Population F	0	.	.	.	
Day	-0.7905	0.05584	-14.16	<.0001	***
Day^2	0.05229	0.007769	6.73	<.0001	***

Days 6-8

Figure 5.5 shows the average number of aggregates for each population. Note that in Figures 5.3-5.4 it was the average of $\log(N_{i,j})$ that was plotted, thus explaining the difference in the values on the y-axis. As opposed to the time periods 1-8 and 1-6, in Figure 5.5 no common

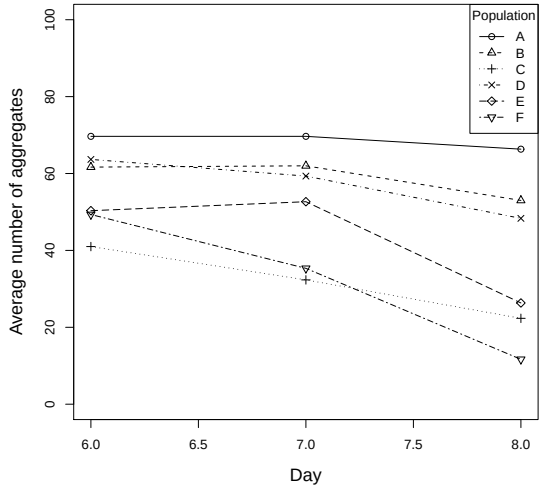

Figure 5.5: The average number of aggregates for each population (A-F) during days 6-8.

trends for the populations are observed, with the exception that they all exhibit small decreases. Instead the behavior over time is very different for different populations. The model yielding the best fit of the fixed effects to the data was

$$\begin{aligned} \mathrm{N}_{i,j} &= \beta_0 + \beta_{\text{Population}} + \beta_1 \times \text{Day} + \beta_2 \times \text{Day}^2 + \\ &\quad \beta_3 \times \text{Population} \times \text{Day} + \beta_4 \times \text{Population} \times \text{Day}^2. \end{aligned} \tag{5.3}$$

This model is very different from (5.1) and (5.2), this could be due to the limited number of observations. Noticeable is that, although included in the model, tests show that neither Population nor the two interaction terms are significant. This non-significance may arise from the small number of samples. According to the AIC the model including interaction terms explains the data the best (of those tested). The solution for the fixed effects obtained by SAS for this model is not shown here due to the large number of parameters, making an interpretation of the solution difficult.

5.2 Analysis of the perimeter of the aggregates

In Chapter 3, figures showed an obvious trend of increasing (average) perimeter over time. Moreover, the rate at which the perimeter increased seemed to be different for different populations. Further analysis of the perimeter data, presented in the following sections, will conclude whether or not this difference in rate between populations is statistically significant. In Section 5.2.1 the appropriate covariance structures for each time period are modeled and selected and in Section 5.2.2 the results of regression analysis of the data are presented. Note that it is the average perimeter $\bar{P}_{i,j}$ of all aggregates in a sample that is considered.

5.2.1 Modeling the covariance structure

As for the aggregate count, to select a covariance model for the perimeter data patterns of correlation between observations at different times are examined. Information criteria that measure the fit of competing covariance models have been used to quantitatively select the most appropriate model. As in Section 5.1.1 the correlation structure is visualized by plotting changes in covariance and correlation among residuals on the same sample (here from the same well) at different times over distance between times of observation. In the following sections we model the covariance structures for the different time periods of the data – days 1-8 and days 6-8.

Days 1-8

Figure 5.6 shows the aforementioned covariance and correlation plots for days 1-8. There seems to be a slight decrease in covariance when the distance between pairs of observation increases. Also, the covariance plot shows indications of unequal variances for the days of observations. Particulary, the variance tends to increase with day. However, no obvious patterns are observed for neither the covariance nor the correlation.

Table 5.6 gives AIC and BIC (Schwarz's bayesian information criterion) for plausible covariance structures for the perimeter data. From the information criteria we can conclude that:

- There exist correlation between observations (since the simple model yields a bad fit).
- The variances for different times of observation do not seem to be equal (the heterogeneous models yield a better fit than the homogeneous models).

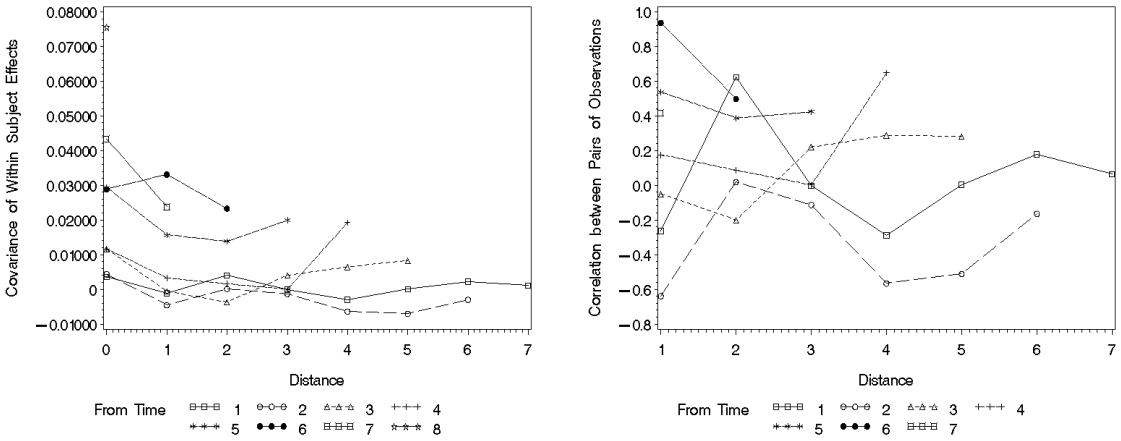

Figure 5.6: The covariance (left) and correlation (right) as a function of distance in time between pairs of observations for days 1-8.

Table 5.6: Akaike information criterion (AIC) and Schwarz's bayesian information criterion (BIC) for plausible covariance structures for days 1-8.

Covariance structure	AIC	BIC
Unstructured model	-72.5	-40.4
Compound symmetric model	-29.6	-27.8
First-order autoregressive model	-36.9	-35.1
Toeplitz model	-27.2	-20.1
Heterogeneous compound symmetric model	-53.0	-45.0
Heterogeneous first-order autoregressive model	-53.4	-45.4
Heterogeneous Toeplitz model	-45.1	-31.7
Variance structure (simple model)	-23.0	-22.1

These conclusions are supported by what was seen in Figure 5.6. Furthermore, ARH(1) and CSH give similar AIC and BIC values. However, there is a contradiction in whether one of these models or the UN model yield the better fit to the data. In [20] it is shown that AIC tends to choose more complex models than BIC. This is in agreement with what is seen in Table 5.6. Furthermore, selecting a too simple covariance structure increases the fixed effects type I error rate and selecting a model that is too complex sacrifices power. With this in mind we have chosen to use the heterogeneous first-order autoregressive covariance structure in future analysis of the data. Note that the same number of parameters is estimated for the CSH and ARH(1) models and they are thus equally complex.

Days 6-8

Covariance and correlation plots are omitted due to the limited number of observations. Instead the selection of covariance structure are based solely on the AIC and BIC. Table 5.7 gives the AIC and BIC for plausible covariance structures for days 6-8. The UN model is the model

Table 5.7: AIC and BIC for plausible covariance structures for days 6-8.

Covariance structure	AIC	BIC
Variance structure (simple model)	15.5	16.4
Unstructured model	-6.2	-0.8
Compound symmetric model	7.6	9.3
First-order autoregressive model	7.0	8.8
Toeplitz model	8.7	11.4
Heterogeneous compound symmetric model	5.0	8.6
Heterogeneous first-order autoregressive model	3.1	6.6
Heterogeneous Toeplitz model	5.1	9.5

that minimizes both the AIC and the BIC. However, the difference between a UN model and an ARH(1) model is rather small. By the same arguments as for Section 5.2.1 the ARH(1) covariance structure has been chosen for further modeling of the perimeter data.

5.2.2 Regression analysis

The following sections present the results from regression analysis of the perimeter data for the two time periods (days 1-8 and days 6-8) using the appropriate covariance structures modeled in the previous sections. The model selection has mainly been based on AIC. The models that have been tested included Population and Day as covariates, and consisted of different combinations of Population, Day^k and interaction terms of the type Population $\times$ Day^k, $k = 1,2$.

Days 1-8

Figure 5.7 shows the average of $\log(\bar{P}_{i,j})$ for each population for days 1-8. As mentioned earlier there is an increase in the average perimeter for each population over time. This increase seems to differ in rate for different populations. Regression indicates that the best fitted model for the data is

$$\log(\bar{P}_{i,j}) = \beta_0 + \beta_{\text{Population}} + \beta_1 \times \text{Day} + \beta_2 \times \text{Population} \times \text{Day}, \tag{5.4}$$

where $\beta_{\text{Population}}$ is different for each population. The interaction term included in the model concludes that the patterns of change for different populations are significantly different. Also, the lack of second degree terms indicate a linear increase in the average of the logarithm of average perimeter value. This is in agreement with what is seen in Figure 5.7. The solution (estimates of coefficients and their standard errors, t-test statistics and corresponding p-values) for this particular model is given in Table 5.8. Thus e.g. for a sample from population F, the model would be

$$\log(\bar{P}_{6,j}) = \underbrace{3.066}_{\text{Intercept}} + 0.490 \text{ Day}.$$

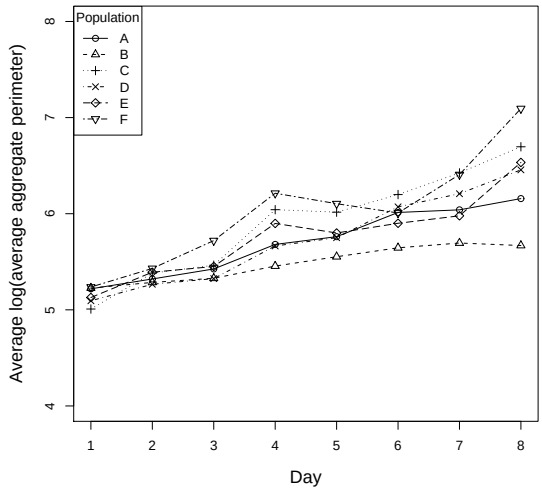

Figure 5.7: The average of $\log(\bar{P}_{i,j})$ for each population (A-F) during days 1-8.

Table 5.8: Solution for fixed effects of the model for days 1-8; estimates of coefficients and their standard error, t-test statistics and p-values.

Effect	Estimate	St. error	t value	Pr > |t|	
Intercept	5.0307	0.04256	118.19	<.0001	***
Population A	0.03435	0.06019	0.57	0.5788	NS
Population B	0.1129	0.06019	1.88	0.0852	.
Population C	-0.2323	0.06019	-3.86	0.0023	**
Population D	-0.1345	0.06019	-2.24	0.0452	*
Population E	-0.04130	0.06019	-0.69	0.5057	NS
Population F	0	.	.	.	
Day	0.2095	0.01363	15.37	<.0001	***
Population A × Day	-0.06953	0.01927	-3.61	0.0005	***
Population B × Day	-0.1332	0.01927	-6.91	<.0001	***
Population C × Day	0.03157	0.01927	1.64	0.1041	NS
Population D × Day	-0.02765	0.01927	-1.43	0.1541	NS
Population E × Day	-0.04694	0.01927	-2.44	0.0163	*
Population F × Day	0	.	.	.	

Days 6-8

Figure 5.8 shows the average of the logarithm of the average perimeter, at each day, for each population. The model yielding the best fit to the data was

$$\log(\bar{P}_{i,j}) = \beta_0 + \beta_{\text{Population}} + \beta_1 \times \text{Day} + \beta_2 \times \text{Population} \times \text{Day}, \tag{5.5}$$

i.e. the same model as for days 1-8. This is not surprising since the rate of increase in the perimeter for all populations seems to be steady over days 1-8 (except for the dip in average perimeter value for population F at day 6). The solution for this particular model is given in

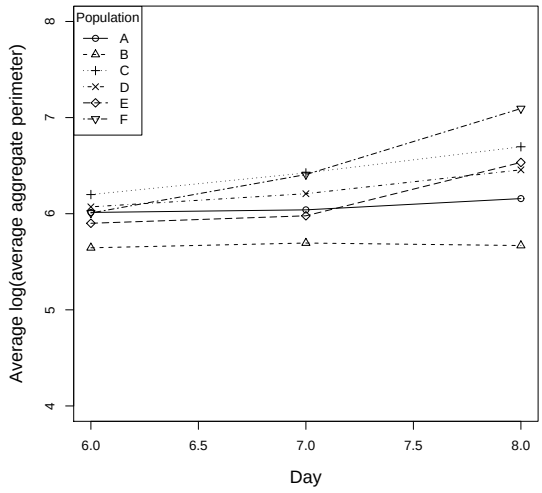

Figure 5.8: The average of $\log(\bar{P}_{i,j})$ for each population (A-F) during days 6-8.

Table 5.9. The estimates of the interaction term coefficients (except for Population C) indicate

Table 5.9: Solution for fixed effects of the model for days 6-8; estimates of coefficients and their standard error, t-test statistics and p-values.

Effect	Estimate	St. error	t value	Pr > |t|	
Intercept	3.0659	0.4926	6.22	<.0001	***
Population A	2.6141	0.6967	3.75	0.0028	**
Population B	2.4248	0.6967	3.48	0.0045	**
Population C	1.6936	0.6967	2.43	0.0317	*
Population D	1.9728	0.6967	2.83	0.0151	**
Population E	1.4620	0.6967	2.10	0.0577	.
Population F	0	.	.	.	
Day	0.4895	0.07938	6.17	<.0001	***
Population A × Day	-0.4343	0.1123	-3.87	0.0005	***
Population B × Day	-0.4636	0.1123	-4.13	0.0003	***
Population C × Day	-0.2497	0.1123	-2.22	0.0338	*
Population D × Day	-0.3177	0.1123	-2.83	0.0082	**
Population E × Day	-0.2621	0.1123	-2.33	0.0265	*
Population F × Day	0	.	.	.	

that the slope of the curves, corresponding to different populations, becomes steeper when the percentage of agarose in the cellular mixture increases. That is, the average aggregate perimeter in populations containing more agarose increases in size faster than in populations with less agarose.

5.3 Comments

Even though we have presented results for the time period 6-8 in both analysis, the number of observations during this period is so small that one should use them with caution. Results indicating non-significance may arise from the small number of both samples and observations for each sample. Moreover, especially in the analysis of the number of aggregates, the graphics corresponding to this time period as well as the models obtained are not easily interpreted. Still, the results for days 6-8 should not be overlooked, as this is perhaps the most interesting time period to observe due to no outside interference. Rather, more data for a corresponding time period would be recommended in order to have more certainty in the results.

The covariance structures chosen for days 1-8 and 1-6 in the aggregate count analysis and days 1-8 in the perimeter analysis, see Table 5.10, coincide well with what one would expect for a process of this kind, with high correlation between observations at adjacent time points. It is natural that a high aggregate count at time t_k would imply a high count at time t_{k+1} as

Table 5.10: Modeled covariance structures for the different time periods.

Analysis	Time period	Covariance structure model
Number of aggregates	1-8	Heterogeneous AR(1)
	1-6	AR(1)
	6-8	Compound symmetric
Perimeter of aggregates	1-8	Heterogeneous AR(1)
	6-8	Heterogeneous AR(1)

well. Moreover, the AR(1) structure suggest that this correlation decreases as the observations become farther apart in time, something that also coincides with what one would expect – e.g. the aggregate count at day one is not expected to have a great effect on the count at day six. Note that the different covariance structures have been evaluated in a purely statistical fashion. If there are any biological or physiological reasons or motivations for different structures, these should also be taken into consideration.

In the application of linear mixed model methods to repeated measurements one assumes that the response variable is continuous and normally distributed. In the analysis of number of aggregates the response is a count variable, this may have an effect on the results. However, the transformed data for days 1-8 and 1-6 is close to normal (ignoring the heavy tails for days 1-6), and the normal-theory methods applied here should therefore still be appropriate. Note that instead of the square root transformation often used for count variables, we have here used a log transformation which gave a better fit to a normal distribution. Furthermore, for both time periods in the perimeter data analysis the response variable $\bar{P}_{i,j}$, although transformed, did show a lack of fit to a normal distribution. Still we used the convention of applying normal-theory methods to the data, keeping in mind that it might effect the results.

The number of aggregates is not solely connected to the cells' tendency to aggregate. It

is of course also dependent on cell division and cell death - both processes obviously altering the number of aggregates in a sample. Furthermore, there is an uncertainty in the process of counting the number of aggregates in a sample, e.g. it can be unclear whether nearby cells should be regarded as distinct or part of an aggregate. Also, it may be that aggregates shapes differently depending on the composition of the cellular mixture. Thus to, from the data used here, explicitly draw any conclusions regarding the cells' tendency to aggregate in different populations, such things must also be taken into consideration.

Part II

Stochastic modeling of tumor growth

Chapter 6

A mathematical model for initial tumor growth

This chapter is meant as a summary of the model and the results in [28], which is the foundation for the work on modeling in this thesis. In the article by Sherratt and Nowak, a mathematical model for the initial growth of a tumor is developed. It is assumed that several regulatory chemicals have an impact on the growth rate of normal and mutant cells and mutations that either express an oncogene or causes the loss of an anti-oncogene are considered. We here present the model and some of the results obtained in [28].

In Section 6.1 the model that Sherrat and Nowak develop in [28] is discussed. The different types of mutations and how they are included in the model are explained and the partial differential equations that constitute the model are motivated . Some of the results obtained with the model are then presented in Section 6.2. Note that these correspond to the results in [28], thus ensuring that a successful scheme for obtaining numerical solutions has been implemented. For the interested reader, [28] contains additional results that are not mentioned here and the biology behind different parts of the model is discussed in a bit more detail. Also, [28] contains an extended list of references for further reading on related topics.

6.1 Model for initial tumor growth

The model developed by Sherratt and Nowak takes into account that several growth regulating chemicals will have an effect on the division of both normal and mutant cells. With an effect of crowding included, the division rate per normal cell is modeled as

$$R_0 r(n) s_1(c_1) ... s_j(c_j). \tag{6.1}$$

Here $n = n(\mathbf{x},t)$ is the normal cell density at space and time coordinates $\mathbf{x}$ and t respectively and $r(n)$ is a function reflecting the crowding effect. Furthermore, s_i is a function representing the effect on cell division by the chemical concentration $c_i = c_i(\mathbf{x},t)$, $1 \leq i \leq j$ (assuming j regulatory chemicals), and R_0 is the growth rate of normal cells in normal tissue when cell

density and chemical concentrations are in an equilibrium. For (6.1) to have a valid physical interpretation

- $r(n)$ should be a decreasing function; an increased cell density should decrease the division rate.
- for $1 \leq i \leq j$, $s_i(\cdot)$ should be an increasing or decreasing function depending on whether c_i is a mitotic activator or mitotic inhibitor respectively.

As in [28], the equilibrium state present in normal tissue is denoted by

$$n = n^e, \quad c_i = c_i^e, \quad 1 \leq i \leq j,$$

and it is assumed that in the equilibrium state the division rate per cell is equal to R_0, i.e.

$$r(n^e) = s_i(c_i^e) = 1, \quad 1 \leq i \leq j.$$

The model includes the following five types of mutations (referenced to by the corresponding roman numerals):

- Increased response to a mitotic activator (I).
- Decreased response to a mitotic inhibitor (II).
- Increased production of a mitotic activator (III).
- Decreased production of a mitotic inhibitor (IV).
- Escape from biochemical dependence (V).

Mutations of types I and II can be thought of as having the effect that mutant cells detect a concentration of ξ times the real concentration of the corresponding chemical. Depending on whether the mutation is of type I or II, ξ will be either > 1 or < 1. Mutations of types III and IV have in [28] been taken to only affect the conservation equations of the chemical concentrations. Mutations of type V are modeled by adding a constant term s_0 to the "chemical" part of the division rate per cell, i.e. the product $s_1(c_1)...s_j(c_j)$.

Denote the mutant cell density by $m(\mathbf{x}, t)$. When considering a population mixed of both normal and mutant cells, the function $r(\cdot)$ for crowding effects will depend on the sum $n(\mathbf{x}, t) + m(\mathbf{x},t)$. Therefore, using that the cells are subject to a first order death rate R_0, the conservation equations for normal and mutant cells become

$$\begin{aligned} \frac{\partial n}{\partial t} &= D\nabla^2 n + R_0 n r(n+m) s_1(c_1)...s_j(c_j) - R_0 n, \\ \frac{\partial m}{\partial t} &= D\nabla^2 m + R_0 m r(n+m)[s_0 + s_1(\xi c_1)...s_j(c_j)] - (R_0 + \delta)m. \end{aligned} \tag{6.2}$$

The first term of the RHS of these equations corresponds to cell migration (i.e. cell migration is assumed to follow a linear diffusion). The second term corresponds to the already introduced

biochemically regulated cell division and the third term is the above mentioned cell death. It is assumed that the effect of an immune response can be modeled by a first order representation, δ being the rate at which the immune response kills mutant cells. The immune response is then included by the $-\delta m$ term on the RHS of the conservation equation for $m(x,t)$. Moreover, note that it is also assumed that any mutations with respect to the response of a regulatory chemical (types I and II) occur with respect to chemical 1.

The regulatory chemicals can act in an either autocrine or paracrine way, or in a mix of the two. Simply put, this corresponds to the chemical being created by the cells themselves or by other cells respectively. For the general case, i.e. when chemical i acts in both an autocrine and paracrine way, suppose that the chemical is produced at a constant rate P_i (thus independent of cell density) and at a rate $p_i(n+m)$ per cell. Then, if the chemical decay is taken to be a first order process (rate d_i), the chemical conservation equations become

$$\frac{\partial c_i}{\partial t} = D_i \nabla^2 c_i + (n+m)p_i(n+m) + P_i - d_i c_i. \tag{6.3}$$

The first term corresponds to chemical diffusion (with $D_i > 0$ the chemical diffusion coefficient), the second term is the chemical production caused by cells, the third term the constant chemical production and the last term the chemical decay. If chemical i is produced in a purely autocrine way, then $P_i = 0$ since only the cell density is relevant. If it is instead produced in a purely paracrine way, the production is completely independent of cell density and $p_i \equiv 0$.

As previously mentioned, mutations of types III and IV have an effect on the conservation equation (Equation (6.3)). Such mutations are in [28] modeled by adding a term $Hmp_i(n+m)$ to the RHS of the equation. For a mutation of type III (i.e. an increased production of a mitotic activator), $H > 0$. Note that in the purely autocrine or paracrine case, this term is interpreted as the increased production of an autocrine factor and the triggering of the autoproduction of a chemical normally produced only by other cell types respectively. For a mutation of type IV, $-1 < H < 0$. Such a mutation is only relevant for chemicals that act in an autocrine way since it is not possible for the cells to produce less of a paracrine chemical.

Equations (6.2) and (6.3) constitute the model for initial tumor growth. The space dimension is taken to be one and the initial and boundary conditions are taken as follows, with $2L$ denoting a typical cell length,

$$\begin{aligned} n(x,0) &= \begin{cases} 0 & |x| < L \\ n^e & |x| > L \end{cases}, \quad m(x,0) = \begin{cases} n^e & |x| < L \\ 0 & |x| > L \end{cases}, \quad c_i(x,0) \equiv c_i^e,\ 1 \le i \le j, \\ n &= n^e, \quad m = 0, \quad c_i = c_i^e \ \text{ at } x = \pm\infty, \ \ 1 \le i \le j. \end{aligned} \tag{6.4}$$

The initial conditions correspond to a single cell mutation at the origin, whereas the boundary conditions correspond to no disturbances far from the site of mutation, i.e. the equilibrium state.

Taking the spatial domain to be infinite is appropriate due to the fact that the tissue will be much larger than any tumor that is considered with this model.

The conservation equations and their initial and boundary conditions are made dimensionless using different re-scalings (see [28] for the different scalings), resulting in the system

$$\begin{aligned}
\frac{\partial n}{\partial t} &= D\frac{\partial^2 n}{\partial x^2} + nr(n+m)s_1(c_1)...s_j(c_j) - n, \\
\frac{\partial m}{\partial t} &= D\frac{\partial^2 m}{\partial x^2} + nr(n+m)\left[s_0 + s_1(\xi c_1)s_2(c_2)...s_j(c_j)\right] - (1+\delta)m, \\
\frac{\partial c_1}{\partial t} &= D_1\frac{\partial^2 c_1}{\partial x^2} + P_1 + (n + m(H+1))\,p_1(n+m) - c_1(P_1 + p_1(1)), \\
\frac{\partial c_i}{\partial t} &= D_i\frac{\partial^2 c_i}{\partial x^2} + P_i + (n+m)\,p_i(n+m) - c_i(P_i + p_i(1)),\ 2 \le i \le j,
\end{aligned} \tag{6.5}$$

subject to the constraints

$$\begin{aligned}
n(x,0) &= \begin{cases} 0 & |x| < 1 \\ 1 & |x| > 1 \end{cases}, \quad m(x,0) = \begin{cases} 1 & |x| < 1 \\ 0 & |x| > 1 \end{cases}, \quad c_i(x,0) \equiv 1,\ 1 \le i \le j, \\
n &= 1, \quad m = 0, \quad c_i = 1 \ \text{ at } x = \pm\infty, \ \ 1 \le i \le j.
\end{aligned} \tag{6.6}$$

Note that although all the parameters have been re-scaled, the same notation as before is used in the system above (and henceforth). The following are used for the functions $r(\cdot)$, $s_i(\cdot)$ and $p_i(\cdot)$:

$$\begin{aligned}
r(n) &= \frac{N-n}{N-1}, \\
s_i(c_i) &= \begin{cases} \alpha_i + c_i(1-\alpha_i),\ \alpha_i \in (0,1) & \text{chem. } i \text{ a mitotic activator}, \\ \dfrac{k_i}{1 + c_i(k_i - 1)},\ k_i \in (1,\infty) & \text{chem. } i \text{ a mitotic inhibitor}, \end{cases} \\
p_i(n) &= \begin{cases} \dfrac{h_i(1+\beta_i)}{1+\beta_i n^2},\ h_i, \beta_i \in (0,\infty) & \text{chem. } i \text{ a mitotic activator}, \\ \dfrac{h_i(1+\beta_i n)}{1+\beta_i},\ h_i, \beta_i \in (0,\infty) & \text{chem. } i \text{ a mitotic inhibitor}. \end{cases}
\end{aligned}$$

Thus N becomes an upper bound for the cell densities. These specific functions were used in [28] due to earlier work by Sherratt and Nowak. The important conditions are that they satisfy the constraint $r(1) = s_i(1) = 1$ and that they behave in a way that is consistent with the properties of the corresponding chemicals. Other functions satisfying these conditions could be used as well.

6.2 Results for deterministic initial tumor growth

The following are reproductions of some of the results presented in [28]. As indicated above, the results shown are closely connected to the work in upcoming chapters.

First, the immune response is assumed to have no effect on the initial tumor growth (i.e. $\delta = 0$). It is in [28] concluded that mutations of types I, II and V give similar solutions to the system 6.5. Moreover, mutations of types III and IV give solutions of the same form, however very different from the other three mutations. Figures 6.1 and 6.2 show the change in normal and mutant cell densities in the case of mutation that combines type I and type V and a mutation of type III respectively. Here an equilibrium density $n_e = \frac{1}{2}$ has been used when obtaining a cell count from the cell densities[3]. Note the large difference in time scale for the two cases.

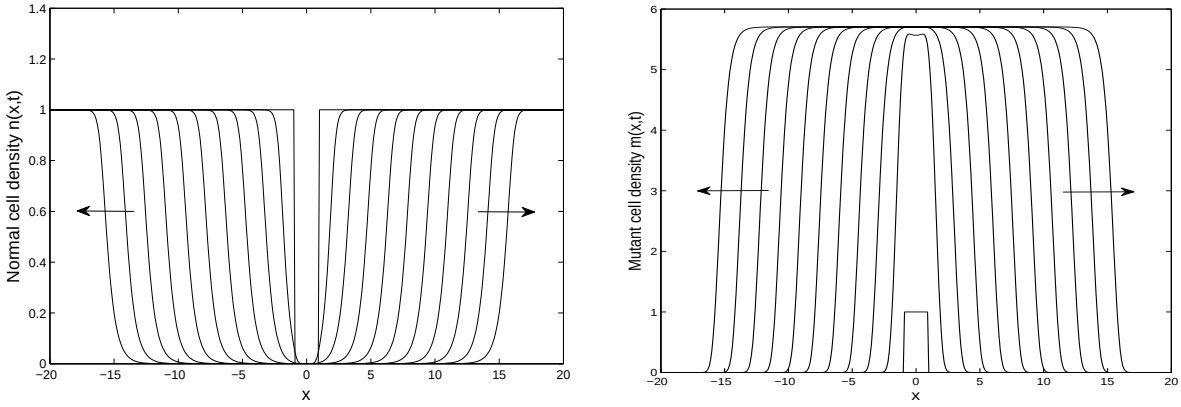

Figure 6.1: The initial growth of a tumor after a mutation that combines types I and V.

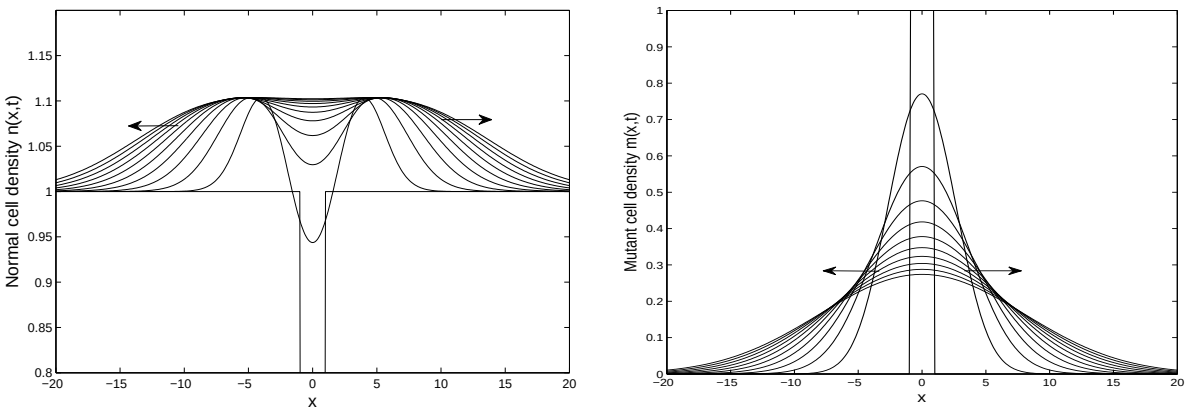

Figure 6.2: The initial growth of a tumor after a mutation of type III.

Mutations of types I, II and V give rise to an advancing wave of mutant cells and a receding wave of normal cells. For mutations of types III and IV, somewhat higher values of both $n(x,t)$ and $m(x,t)$ are observed near the origin followed by a gradual outspread of both cell types. Thus there are significant differences in initial tumor growth for different categories of

[3]This gives results similar to those in [28], thus enabling comparisons.

mutations. Figure 6.3 further illustrates this by showing how mutations of types I, II and V cause the normal cells to be replaced by mutant cells whereas for mutations III and IV both cell types increase slowly over time. Again, the large difference in observation time for the two cases should be pointed out.

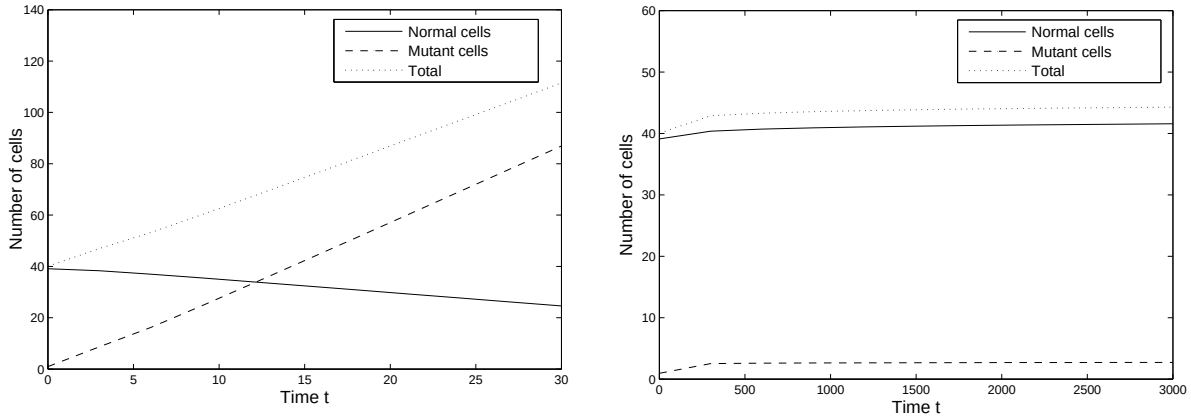

Figure 6.3: The change in normal, mutant and total number of cells for mutations of types I, II, V (left) and III, IV (right).

Now assume that the immune response has an effect on the mutant cells, i.e. $\delta > 0$. Figures 6.4-6.5 illustrate the solutions of (6.5) for mutations of types I, II and V. The difference between Figures 6.4 and 6.5 is the value of δ.

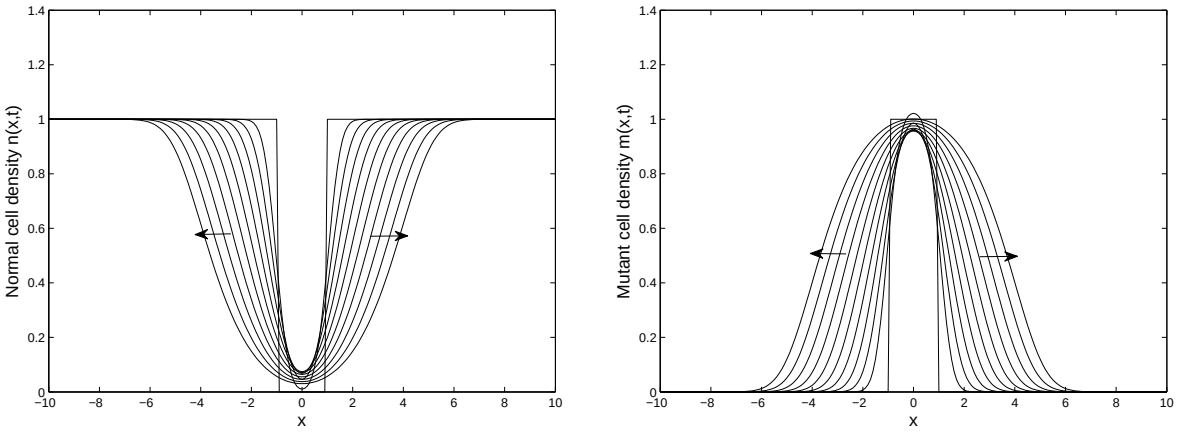

Figure 6.4: Normal and mutant cell densities respectively in the case of a combined mutation (types I and V). Parameter values used were $\xi = 2$, $s_0 = 2$ and $\delta = 2.8$. The non-dimensional time step was $t_{\text{step}} = 7.5$, with other parameters as before.

In Figure 6.4 an advancing wave of mutant cells (similar to what was seen for mutations of types I, II and V when no immune response was taken into consideration) is observed, suggesting that the immune response cannot suppress tumor growth. However in Figure 6.5 no such wave is observed for mutant cells. Instead, the mutant cell density decreases rapidly to zero and the normal cell density approaches one, thus filling in the "gap". The main result in [28] regarding the immune response is that for this model there exist a critical value δ_{crit} above which the

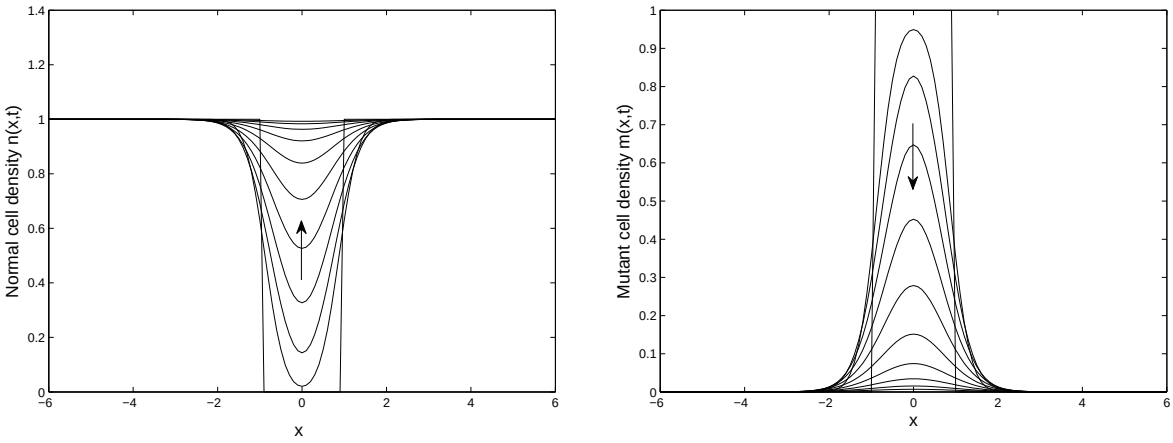

Figure 6.5: Normal and mutant cell densities respectively in the case of a combined mutation (types I and V). Parameter values used were $\xi = 2$, $s_0 = 2$ and $\delta = 3.0$. The non-dimensional time step was $t_{\text{step}} = 7.5$, with other parameters as before.

immune response is "strong enough" to suppress tumor growth. The expression for this value is (see the appendix of [28] for a derivation)

$$\delta_{\text{crit}} = s_1(\xi) + s_0 - 1. \tag{6.7}$$

The existence of a critical value is clearly hinted when comparing Figures 6.4 and 6.5, although δ is only changed a small amount the solutions look very different. In biological terms, (6.7) indicates that in order for the immune response to be able to suppress tumor growth it has to have a rate that is greater than the difference between the proliferation rate of mutant and normal cells. For the mutation used in Figures 6.4-6.5 the critical value is $\delta_{\text{crit}} = 2.9$.

6.3 Comments

The specific parameter values (h_i, β_i etc.), along with motivations for them, used to obtain the results here can be found in the article by Sherratt and Nowak. In order to obtain comparable results we will for the rest of this thesis continue to use these parameter values, should nothing else be stated.

In upcoming chapters, it will be mutations of types I and V that are of special interest due to the advancing wave of mutant cells that result from them. The time frame $t \in [0,30]$ was here used to study the mutations impact on cell densities. Henceforth, whenever a time interval $[0,T]$ is observed and nothing else is explicitly stated, $T = 30$ will be used for numerical computations. Moreover, the time step used will be $t_{\text{step}} = 3$ unless otherwise stated.

Chapter 7

Stochastic behavior in the model for tumor growth

In (6.5), i.e. the governing equations in the model for cell growth developed by Sherratt and Nowak, all parameters are regarded as to be of a deterministic nature. In particular, the parameters describing different mutations are all taken to be fixed constants. In this chapter different types of random behavior are introduced in the system, one being small random perturbations with respect to mutation parameters. Furthermore, more significant parameter randomness is considered in the sense that random variables (or rather random processes) are used to assign mutation parameters their values.

Recall mutations I-V in Chapter 6. The investigations in this chapter are restricted to mutations with respect to the response of a mitotic activator (I) and biochemical escape (V) respectively, due to their tendency to give rise to advancing waves of mutant cells as shown in [28] and illustrated in Figure 6.1. In Section 7.1 mutations of types I and V are studied with small stochastic perturbations in the characterizing parameter. In Section 7.2 the same types of mutations are studied when ξ and s_0 are random processes rather than fixed constants. The case when ξ is allowed to take on values on both sides of one is particularly studied. This corresponds to the mutant cells alternating between having an advantage and disadvantage respectively for proliferation compared to normal cells.

Throughout the chapter the immune response is discarded from the model (i.e. $\delta = 0$) in order to not have its impact on the tumor growth interfere with the impact of the random behavior of the mutation parameters.

7.1 Random perturbations

In this section the characterizing parameters of mutations are considered constant and the randomness in the system is due to small random perturbations, the perturbations taken with respect to the mutation parameters. Throughout the section all mutation parameters are considered as *system* parameters, i.e. one parameter is used for all mutant cells. If K represents the

parameter for a specific mutation, consider

$$K + \varepsilon\psi(t),$$

where K is a constant, $\varepsilon > 0$ a small number and $\psi(t)$ a random process of some sort. The smallness of the perturbation thus comes from ε rather than the actual process $\psi(t)$. A common case is to consider $\psi(t)$ a stationary Gaussian process [18]. Here, we consider $\psi(t)$ to be defined in terms of a discrete time Gaussian process in the following sense: Letting $Z_t,\ t \in T \subseteq \mathbb{N}$, be such that $Z_t \overset{\mathcal{D}}{=} N(0, \sigma^2)$ for every discrete time point t and Δt be the (desired) time between changes in $\psi(t)$, define

$$\psi(t) \triangleq \sum_{i=0}^{\lfloor t/\Delta t \rfloor} Z_i I\left(t \in [i\Delta t, (i+1)\Delta t)\right), \tag{7.1}$$

where $\lfloor t/\Delta t \rfloor \triangleq \sup_{n\in\mathbb{N}} n \leq t/\Delta t$. We here let the Z_i:s be iid. variables with some variance σ^2. Note that the stationarity of Z_t is lost for $\psi(t)$.

Define n^ε, m^ε and c_i^ε as the quantities of the system (6.5) but with a perturbation applied. Applying the above to the mutation parameter ξ changes the conservation equation for mutant cells to

$$\frac{\partial m^\varepsilon}{\partial t} = D\frac{\partial^2 m^\varepsilon}{\partial x^2} + n^\varepsilon r(n^\varepsilon + m^\varepsilon)\left[s_1((\xi + \varepsilon\psi(t))c_1)s_2(c_2)...s_j(c_j)\right] - m^\varepsilon(\delta + 1). \tag{7.2}$$

When instead small perturbations are considered with respect to s_0 the conservation equation for mutant cells becomes

$$\frac{\partial m^\varepsilon}{\partial t} = D\frac{\partial^2 m^\varepsilon}{\partial x^2} + n^\varepsilon r(n^\varepsilon + m^\varepsilon)\left[s_0 + \varepsilon\psi(t) + s_1(c_1)s_2(c_2)...s_j(c_j)\right] - m^\varepsilon(\delta + 1). \tag{7.3}$$

Furthermore, let $X^\varepsilon(t)$ denote the number of normal cells in the perturbed system at time t, $Y^\varepsilon(t)$ the corresponding number of mutant cells and $x(t)$, $y(t)$ the number of normal and mutant cells respectively in the deterministic system at time t. When instead the number of cells at a time T is of interest, $X_k^{\varepsilon,T}$ and $Y_k^{\varepsilon,T}$ are used to denote the number of normal and mutant cells respectively for each realization, k indicating which realization that is considered.

Realizations of $\{X^\varepsilon(t)\}_{t\in[0,T]}$ and $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ for a perturbation with respect to s_0 ($\varepsilon = 0.1$ and $\sigma^2 = 2$) are shown in Figure 7.1. Figure 7.2 shows a closer look of $Y^\varepsilon(t)$. It is clearly seen how $Y^\varepsilon(t)$ fluctuates around (the deterministic) $y(t)$, illustrating the effect a perturbation (with respect to s_0) has on the solution of the governing equations (6.5).

Consider a perturbation in the response to a mitotic activator. Figure 7.3 shows the variances of $\{X_i^{\varepsilon,T}\}_{i=1}^{40}$ and $\{Y_i^{\varepsilon,T}\}_{i=1}^{40}$ respectively for different ε and σ^2. Clearly, perturbations of the sizes here considered have little effect on the governing equations in terms of variability. Furthermore, the average numbers of normal and mutant cells were almost constant (differences of magnitude 0.01) for a particular ξ and different ε and σ^2. Figure 7.4 shows variances of $\{X_i^{\varepsilon,T}\}_{i=1}^{40}$ and $\{Y_i^{\varepsilon,T}\}_{i=1}^{40}$ when a perturbation with respect to s_0 is assumed. As opposed to a perturbation

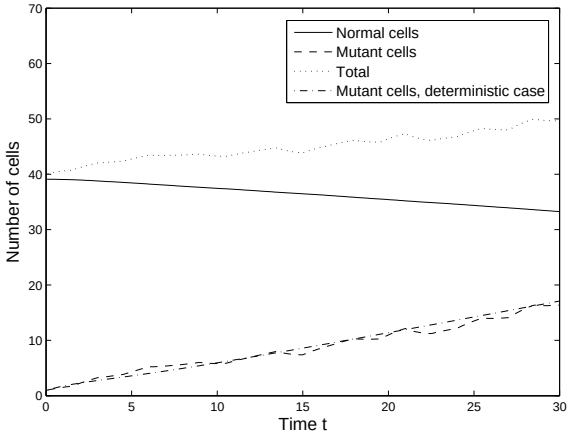

Figure 7.1: Realizations of $\{X^\varepsilon(t)\}_{t\in[0,T]}$ and $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ for a perturbation ($\varepsilon = 0.1$ and $\sigma^2 = 2$) with respect to s_0. Also included is $y(t)$ for the case when $s_0 = 1$, i.e. the value coincides with the expected value of $s_0 + \varepsilon\psi(t)$.

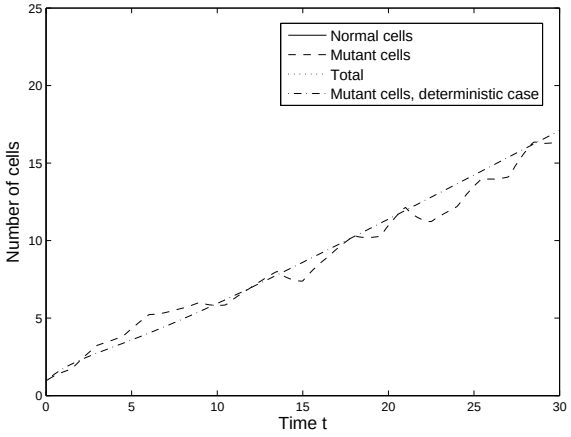

Figure 7.2: A closer look at $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ and $y(t)$ from Figure 7.1.

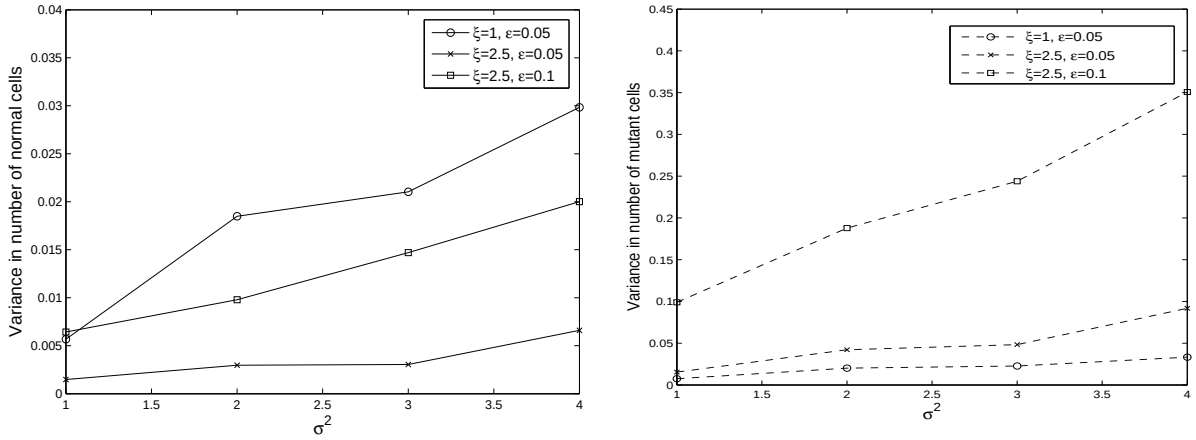

Figure 7.3: Variance in number of normal and mutant cells respectively at time T for different ξ and ε, with respect to σ^2.

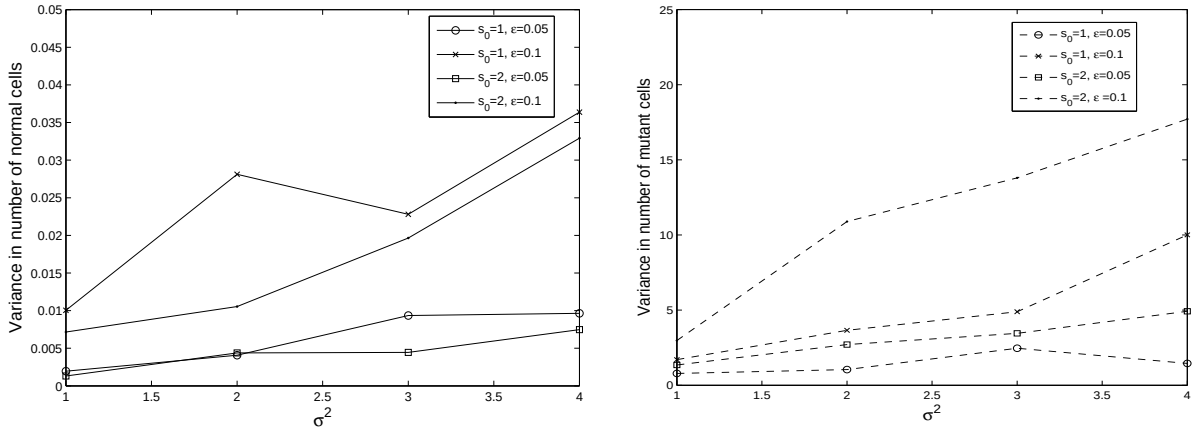

Figure 7.4: Variance in number of normal and mutant cells respectively at time T for different s_0 and ε, with respect to σ^2.

with respect to ξ, changes in ε and σ^2 now have relatively large effects on the variance in number of mutant cells at time T. The observed variance in number of normal cells is still negligible. Moreover, the average number of cells (normal as well as mutant) was close to constant for fixed s_0 with different values for ε and σ^2.

7.2 Random mutation parameters

The parameters characterizing mutations of types I and V are now subject to more significant parameter randomness than just small random perturbations. As for the case of perturbations, mutation parameters are considered as *system* parameters. Mutation parameter randomness is considered for time, rather than spatial coordinates. Suppose $t \in [0,T]$ for some $T < \infty$. In order for the physical interpretation of the mutations to be valid it must hold that

$$\xi(t),\ s_0(t) \geq 0, \quad \forall t \in [0,T].$$

For a mutation of type V, any such $s_0(t)$ will give an oncogenic mutation. However, for mutations of type I it is only when $\xi(t) > 1$ that mutant cells have a proliferation advantage compared to normal cells. $\xi(t) < 1$ instead gives the normal cells a proliferation advantage. If a decreased response to a mitotic inhibitor (mutation type II) is considered, it is values in [0,1) that give a proliferation advantage for mutant cells.

Consider the random process $\{\phi(t)\}_{t\in[0,T]}$ defined as

$$\phi(t) = \begin{cases} \phi_0 + Z_i & \text{if } t \in [t_i,t_{i+1}), \\ \phi_0 + Z_{n-1} & \text{if } t = T, \end{cases} \tag{7.4}$$

where $0 \leq i \leq n-1$ and the t_i:s are the times of change in $\phi(t)$, $t_0 = 0$, $t_n = T$. The times $t_1, ..., t_{n-1}$ may be fixed or random, with $0 = t_0 < t_1 < t_2 < ... < t_n = T$. The Z_i:s are some random variables in $\mathbb{R}$ and Z is used throughout the section to denote a generic random variable

of the same (in each case specified) type. $\{\phi(t)\}_{t\in[0,T]}$ can be written as

$$\phi(t) = \begin{cases} \phi_0 + \sum_{i=0}^{n-1} Z_i I\left(t \in [t_i, t_{i+1})\right) & \text{if } t \in [0,T), \\ \phi_0 + Z_{n-1} & \text{if } t = T. \end{cases} \tag{7.5}$$

Thus $\{\phi(t)\}_{t\in[0,T]}$ looks similar to a continuous-time jump process, with the differences that the jump at a time t_k is taken from ϕ_0 rather than the current value of $\phi(t_k^-)$ and $\phi(0)$ is not necessarily 0. This type of process is used to model random parameters for mutations of types I and V respectively.

The distributions here used for the Z_i:s in (7.5) are exponential and uniform and the Z_i:s will be considered as iid. Both fixed and random times of change are used. The fixed times are taken so that $|t_i - t_j| = \Delta t$ for any i,j such that $|i - j| = 1$, i.e. adjacent time points are a constant distance Δt apart. The case of random times is considered for exponentially distributed interarrival times, i.e. $(t_{i+1} - t_i) \overset{\mathcal{D}}{=} \text{Exp}(\lambda)$ for some $\lambda_t > 0$. Figure 7.5 shows four realizations of $\{\phi(t)\}_{t\in[0,T]}$ when $Z \overset{\mathcal{D}}{=} \text{Exp}(\lambda)$ and $Z \overset{\mathcal{D}}{=} \text{Uni}([a,b])$, two for fixed times and two for random times with $\lambda_t = 1$. Throughout the section $Z \overset{\mathcal{D}}{=} \text{Exp}(\lambda_\phi)$ is referred to as the

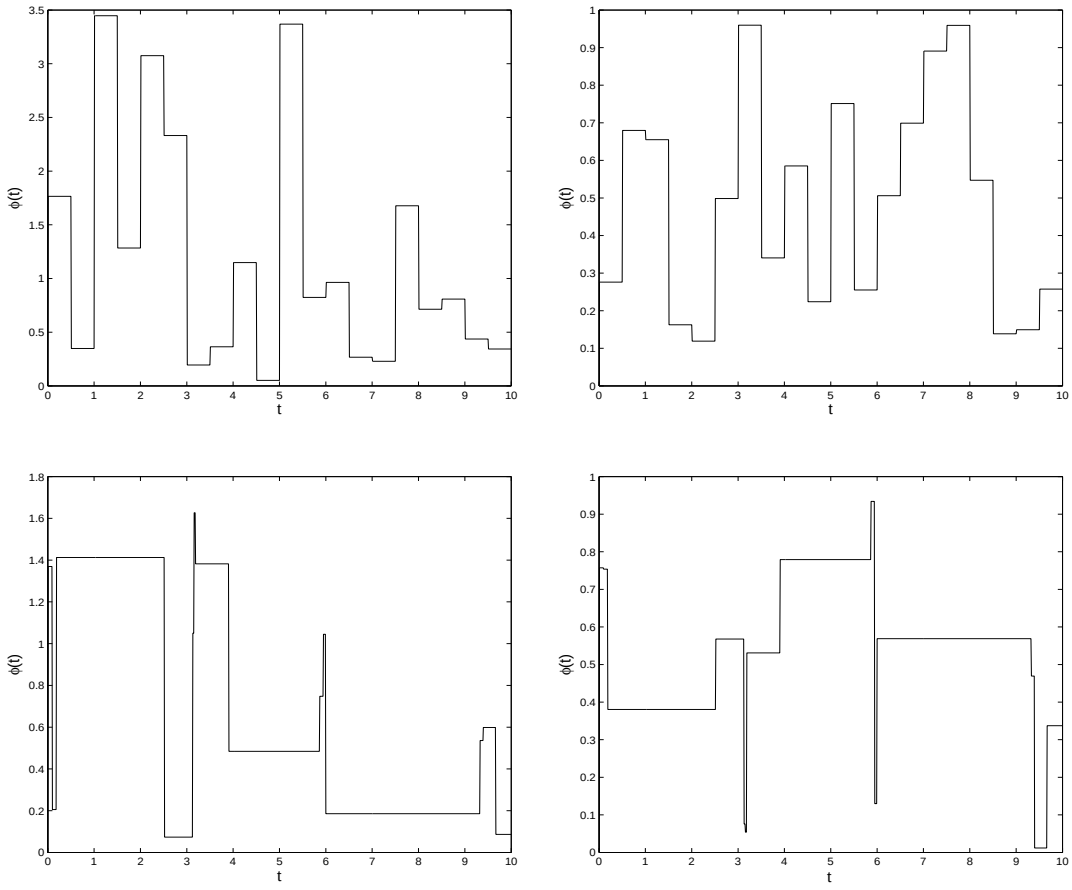

Figure 7.5: Realizations of $\{\phi(t)\}_{t\in[0,10]}$, $\phi_0 = 0$, when $Z \overset{\mathcal{D}}{=} \text{Exp}(\lambda)$ (left column) and $Z \overset{\mathcal{D}}{=} \text{Uni}([a,b])$ (right column). The top row shows realizations for fixed times, $\Delta t = 0.5$, and the bottom row for random times, $\lambda_t = 1$.

exponential case with λ_ϕ and $Z \overset{\mathcal{D}}{=} \text{Uni}([a,b])$ is referred to as the *uniform case with* $I^\phi_{a,b}$, where ϕ is used to identify the relevant mutation parameter. When a realization of $\phi(t)$ is used for a mutation parameter, $X^{(i)}(t)$ and $Y^{(i)}(t)$ are used to denote the number of normal and mutant cells respectively at time t, $i \in \{1,2\}$ with 1 referring to the exponential case and 2 to the uniform case. When instead the number of cells at a time T is of interest, $X_k^{(i),T}$ and $Y_k^{(i),T}$ are used to denote the number of normal and mutant cells respectively for each realization, i again denoting the type of distribution that has been used for the Z_j:s and k indicates which realization that is considered. Thus $\{X_k^{(i),T}\}_{k=1}^l$ and $\{Y_k^{(i),T}\}_{k=1}^l$ are random sequences (l the number of realizations in a particular case). Similarly, in the case of one realization on $[0,T]$, $\{X^{(i)}(t)\}_{t\in[0,T]}$ and $\{Y^{(i)}(t)\}_{t\in[0,T]}$ are random processes describing the number of normal and mutant cells respectively. Note that how the sequences are produced (i.e. which distribution parameters have been used and what types of times of change) is not indicated in the notations.

7.2.1 Increased response to a mitotic activator

Consider a mutation of type I, i.e. an increased response to a mitotic activator. Use (7.5) for $\xi = \xi(t)$:

$$\xi(t) = \begin{cases} \xi_0 + \sum_{i=0}^{n-1} X_i I\left(t \in [t_i, t_{i+1})\right) & \text{if } t \in [0,T), \\ \xi_0 + X_{n-1} & \text{if } t = T, \end{cases} \tag{7.6}$$

with everything as previously defined. First, consider fixed times of change. Table 7.1 shows the different values of λ_ξ, ξ_0 and Table 7.2 shows the values of a,b used in simulations for the exponential and uniform cases respectively; $\xi_0 = 0$ for the latter. Note that $\xi(t)$ is always $\geq \xi_0$

Table 7.1: Combinations of λ_ξ and ξ_0 used for simulations of the exponential case together with the corresponding $\text{E}[\xi(t)]$ and $\text{Var}(\xi(t))$. Note that some cases satisfy the $\text{E}[\xi(t)] = 1$ "normal" condition.

λ_ξ	ξ_0	$\text{E}[\xi(t)]$	$\text{Var}(\xi(t))$
2	1/2	1	1/4
4/3	1/4	1	9/16
8/7	1/8	1	49/64
1	0	1	1
1/2	0	2	4
1/3	0	3	9
1/4	0	4	16

but allowed to take values on both sides of one due to the properties of the exponential and uniform distributions. Moreover, some cases satisfy the $\text{E}[\xi(t)] = 1$ "normal condition", i.e. the system looks normal (referring to the case $\xi(t) = \xi = 1$) on average.

Figure 7.6 shows the mean number of normal and mutant cells for the different λ_ξ in Table 7.1 and Figure 7.7 shows the corresponding variances. Figures 7.8-7.9 are the corresponding plots for the uniform cases with $I^\xi_{a,b}$ as in Table 7.2. For both the exponential and uniform

Table 7.2: $I^{\xi}_{a,b}$ used for simulations of the uniform case together with corresponding $\mathrm{E}[\xi(t)]$ and $\mathrm{Var}(\xi(t))$. Note that some cases satisfy the $\mathrm{E}[\xi(t)] = 1$ "normal" condition.

$I^{\xi}_{a,b}$	$\mathrm{E}[\xi(t)]$	$\mathrm{Var}(\xi(t))$
[0.75,1.25]	1	1/48
[0.5,1.5]	1	1/12
[0.25,1.75]	1	3/16
[0,2]	1	1/3
[0,3]	1.5	3/4
[0,4]	2	4/3
[0,6]	3	3

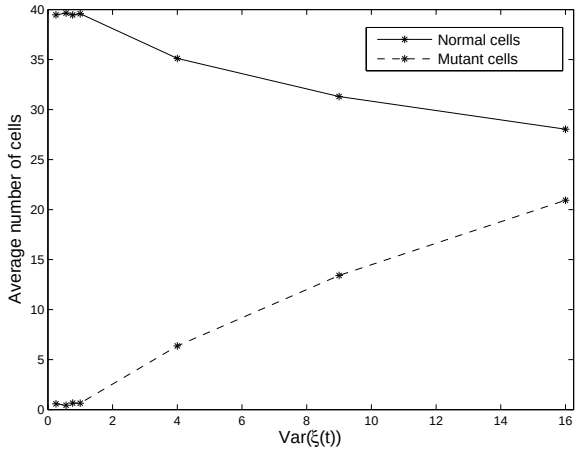

Figure 7.6: Average number of normal and mutant cells with respect to $\mathrm{Var}(\xi(t))$ for the exponential case with λ_ξ as in Table 7.1.

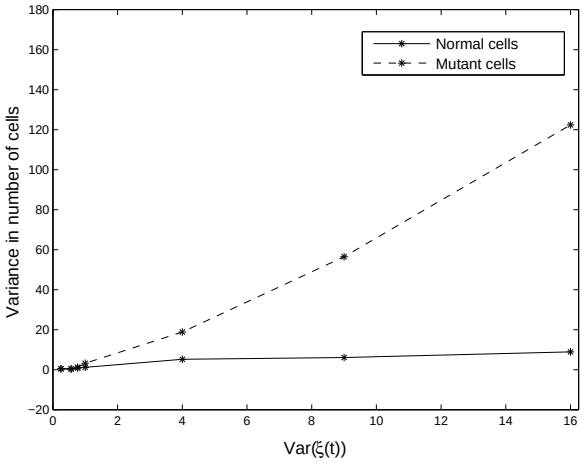

Figure 7.7: Variance in number of normal and mutant cells respectively with respect to $\mathrm{Var}(\xi(t))$ for the exponential case with λ_ξ as in Table 7.1.

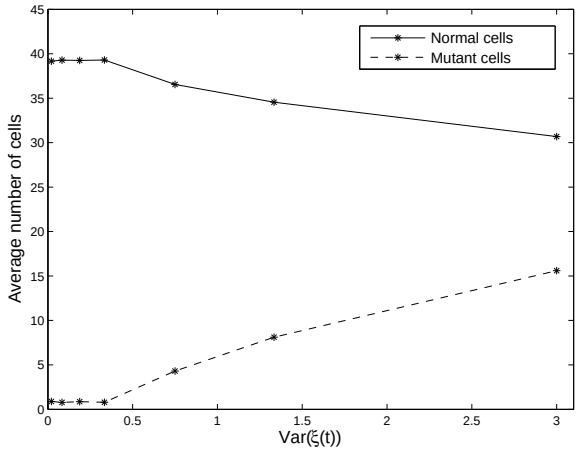

Figure 7.8: Average number of normal and mutant cells with respect to $\mathrm{Var}(\xi(t))$ for the uniform case with $I^{\xi}_{a,b}$ as in Table 7.2.

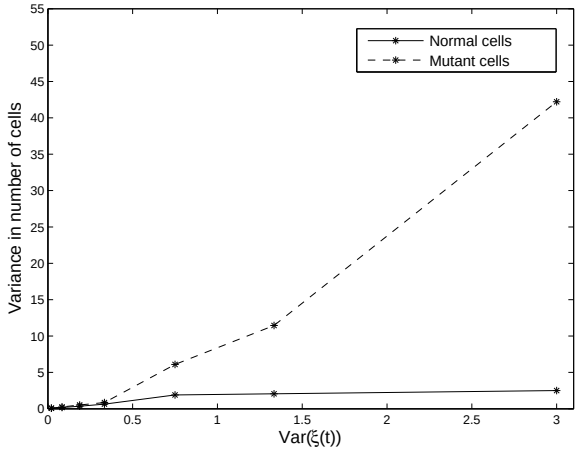

Figure 7.9: Variance in number of normal and mutant cells with respect to $\mathrm{Var}(\xi(t))$ for the uniform case with $I^{\xi}_{a,b}$ as in Table 7.2.

case, it is clearly seen that when $\mathrm{E}[\xi(t)] = 1$ there are only small differences between the different $\hat{X}^{(i),T}$ and $\hat{Y}^{(i),T}$ (the means of the random sequences arising from numerical solutions of (6.5)). However, as the expected value and $\mathrm{Var}(\xi(t))$ increase, there is a rapid change in the observed average and variance for both normal and mutant cells.

Using stepwise regression analysis for the simulated data, models (7.7) are obtained for the number of normal ($\alpha^{(i)}(T)$) and mutant ($\beta^{(i)}(T)$) cells at T when $Z \stackrel{\mathcal{D}}{=} \mathrm{Exp}\,(\lambda_\xi)$ $(i = 1)$ and $Z \stackrel{\mathcal{D}}{=} \mathrm{Uni}\,([a,b])$ $(i = 2)$. The explanatory variables were taken to be $\mathrm{E}[\xi(t)]$ and $\mathrm{Var}(\xi(t))$.

$$\begin{aligned}
\alpha^{(1)}(T) &= 44.53 - 5.28\ \mathrm{E}[\xi(t)]) + 0.2891\ \mathrm{E}[\xi(t)]^2, \\
\beta^{(1)}(T) &= -6.25 + 6.67\ \mathrm{E}[\xi(t)], \\
\alpha^{(2)}(T) &= 45.15 - 6.42\ \mathrm{E}[\xi(t)] + 0.54\ \mathrm{E}[\xi(t)]^2, \\
\beta^{(2)}(T) &= -6.57 + 7.37\ \mathrm{E}[\xi(t)].
\end{aligned} \tag{7.7}$$

Furthermore, models (7.8) were obtained when the type of distribution also was used as an explanatory variable.

$$\begin{aligned} \alpha(T) &= 44.92 - 5.69\ \mathrm{E}[\xi(t)] - 0.43\ \mathrm{Dist} + 0.37\ \mathrm{E}[\xi(t)]^2, \\ \beta(T) &= -4.66 + 4.81\ \mathrm{E}[\xi(t)] - 0.71\ \mathrm{Dist} + 1.03\ \mathrm{E}[\xi(t)] \times \mathrm{Dist} + 0.39\ \mathrm{E}[\xi(t)]^2. \end{aligned} \tag{7.8}$$

Here Dist is defined as

$$\mathrm{Dist} = \begin{cases} 0 & \text{if } Z \overset{\mathcal{D}}{=} \mathrm{Exp}(\lambda_\xi)\,, \\ 1 & \text{if } Z \overset{\mathcal{D}}{=} \mathrm{Uni}([a,b])\,. \end{cases}$$

$\mathrm{Var}(\xi(t))$ is not included in any model whereas $\mathrm{E}[\xi(t)]$ is included in all, suggesting that it is the expected value of the parameter that has the most impact on the outcome. Moreover, the inclusion of the categorical variable corresponding to the type of distribution (Dist) is consistent with the very different characteristics of the exponential and uniform probability distributions. The R^2 values for the models are given in Table 7.3, being consistently higher for models for the number of normal cells. This may be explained by the randomness being explicitly introduced

Table 7.3: R^2 values for the models in equations (7.7) and (7.8). The last two entries correspond to models in which Dist was used as an explanatory variable.

Cell type	Distribution	R^2
Normal	$\mathrm{Exp}(\lambda)$	0.859
Mutant	$\mathrm{Exp}(\lambda)$	0.665
Normal	$\mathrm{Uni}([a,b])$	0.897
Mutant	$\mathrm{Uni}([a,b])$	0.760
Normal	-	0.872
Mutant	-	0.698

in the conservation equation for mutant cells. The randomness should lower the possibility to explain the outcome with a deterministic model, hence lowering the R^2 values.

A realization of $\{Y_k^{(1),T}\}_{k=1}^{40}$, $\lambda_\xi = \frac{1}{2}$ and $\xi_0 = 0$, is shown in Figure 7.10. An interest is here put on the empirical cumulative distribution functions (CDF's) of $\{X_k^{(i),T}\}_{k=1}^{40}$ and $\{Y_k^{(i),T}\}_{k=1}^{40}$. The cases in Tables 7.1-7.2 give rise to empirical CDF's and the Kolmogorov distance can be used to test how well theoretical CDF's fit them.

Definition 7.2.1 *The Kolmogorov distance between an empirical cumulative distribution function F_{emp} and a theoretical cumulative distribution function F is*

$$d_K = \max_{x \in \mathbb{R}} |F_{emp}(x) - F(x)|.$$

Figure 7.11 shows the empirical CDF for the $\{Y_k^{(1),T}\}_{k=1}^{40}$ in Figure 7.10. Table 7.4 shows the result of comparisons between the fit of different theoretical CDF's to the empirical CDF's of the different $\{X_k^{(i),T}\}_{k=1}^{40}$ and $\{Y_k^{(i),T}\}_{k=1}^{40}$ using the Kolomogorov distance d_K as test statistic.

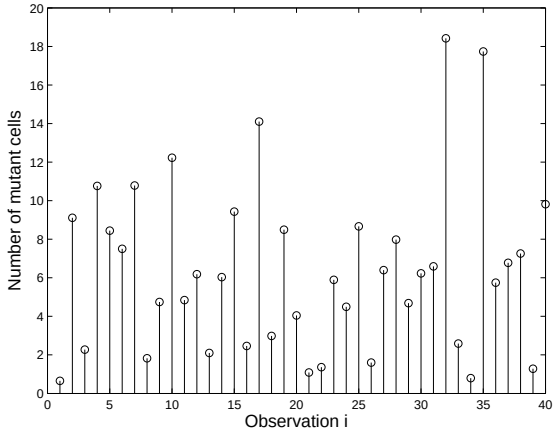

Figure 7.10: A realization of the random sequence $\{Y_k^{(1),T}\}_{k=1}^{40}$ for $\lambda_\xi = \frac{1}{2}$ and $\xi_0 = 0$.

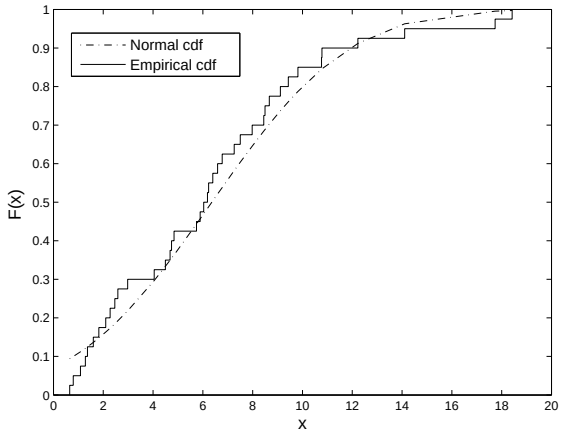

Figure 7.11: Empirical CDF for $\{Y_k^{(1),T}\}_{k=1}^{40}$, $\lambda_\xi = \frac{1}{2}$ and $\xi_0 = 0$.

Note that for the exponential case with $\lambda_\xi = \frac{1}{2}$ the Weibull distribution was also a (relatively) good fit. Hence the Weibull distribution was a good fit for the distribution of mutant cells in each exponential case. Moreover, for normal cells several cases had relatively large values of d_K for all tested distributions.

Now let the t_i:s in (7.6) (except for $t_0 = 0$ and $t_n = T$) be random according to an "exponential clock", i.e. $(t_{i+1} - t_i) \overset{\mathcal{D}}{=} \text{Exp}(\lambda_t)$. Figures 7.12-7.13 show the mean of $\{X_k^{(1),T}\}_{k=1}^{40}$, $\{Y_k^{(1),T}\}_{k=1}^{40}$ for three λ_ξ and $\lambda_t = 0.5, 1, 2$. It is observed that, with the exception of $\lambda_\xi = 1$, a decrease in λ_t (hence an increase in variance of interarrival times) causes a decrease and an increase in the average number of normal and mutant cells respectively. Paired t-tests are used to test the hypothesis that the true mean for normal and mutant cells respectively (for some λ_ξ) are equal for different λ_t, i.e. the tests are used to conclude whether or not the different expected interarrival times (λ_t^{-1}) have any significant effects on the number of normal and mutant cells. Table 7.5 shows the resulting p-values. For $\lambda_\xi = 1$ different values of λ_t does not give rise to any significant differences in the average number of cells. However, for $\lambda_\xi = 0.25, 0.5$ there are

Table 7.4: Theoretical CDF's which fit the different $\xi(t)$'s the best according to the Kolmogorov distance.

Distribution of X	ξ_0	Distribution with lowest d_K	
		Normal	Mutant
Exp(2)	1/2	Weibull	Weibull
Exp(4/3)	1/4	Weibull	Weibull
Exp(8/7)	1/8	Weibull	Weibull
Exp(1)	0	Weibull	Weibull
Exp(1/2)	0	Log-normal	Normal
Exp(1/3)	0	Normal	Weibull
Exp(1/4)	0	Normal	Weibull
Uni([0.75,1.25])	0	Weibull	Log-normal
Uni([0.5,1.5])	0	Weibull	Weibull
Uni([0.25,1.75])	0	Weibull	Normal
Uni([0,2])	0	Weibull	Weibull
Uni([0,3])	0	Log-normal	Normal
Uni([0,4])	0	Log-normal	Weibull
Uni([0,6])	0	Log-normal	Normal

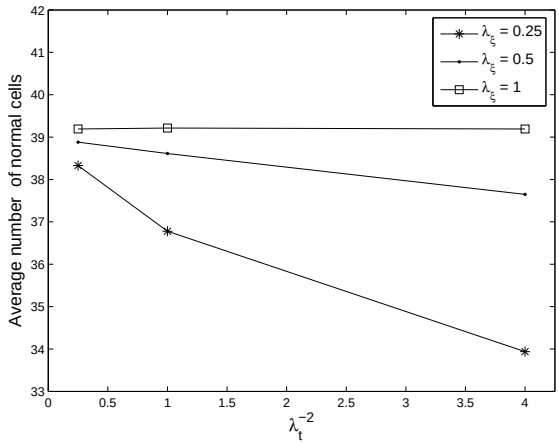

Figure 7.12: Mean of a realization $\{X_k^{(1),T}\}_{k=1}^{40}$ for three different λ_ξ when random times for $\xi(t)$ are considered.

Table 7.5: p-values from paired t-tests with the hypothesis that for different λ_t the average number of normal and mutant cells respectively are equal.

	$\lambda_\xi = 0.25$		$\lambda_\xi = 0.5$		$\lambda_\xi = 1$	
	Normal	Mutant	Normal	Mutant	Normal	Mutant
$\lambda_t = 0.5$ vs. $\lambda_t = 1$	<0.0001	0.0139	0.00867	0.0340	0.775	0.485
$\lambda_t = 0.5$ vs. $\lambda_t = 2$	<0.0001	<0.0001	0.000274	0.00309	0.994	0.680
$\lambda_t = 1$ vs. $\lambda_t = 2$	<0.0001	0.00187	0.0517	0.170	0.753	0.177

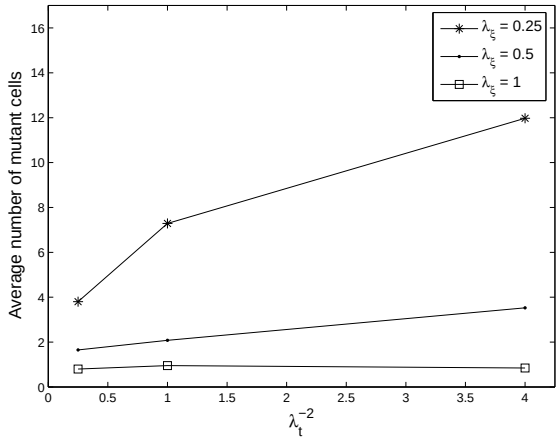

Figure 7.13: Mean of a realization $\{Y_k^{(1),T}\}_{k=1}^{40}$ for three different λ_ξ when random times for $\xi(t)$ are considered.

significant differences in the average number of cells for almost every choice of λ_t.

To investigate any difference in number of cells at time T between fixed and random times, realizations of $\{X_k^{(1),T}\}_{k=1}^{40}$ and $\{Y_k^{(1),T}\}_{k=1}^{40}$ for the two cases with $\lambda_t^{-1} = \Delta t$ are used. Figure 7.14 shows an example of a scatter plot of the residuals of two realizations of $\{X_k^{(1),T}\}_{k=1}^{40}$, one with fixed times of change and one with random, $\lambda_\xi = 1$. Figure 7.15 shows the corresponding scatter plot for mutant cells. The figures show relatively small differences, indicated by the

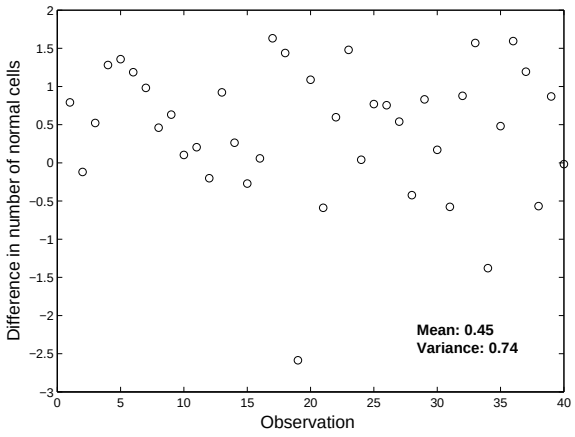

Figure 7.14: Residuals from comparison of fixed and random times for realizations of $\{X_k^{(1),T}\}_{k=1}^{40}$ when $\lambda_t = \Delta t = 1$ and $\lambda_\xi = 1$.

mean of the residuals. The residuals seem to be evenly distributed around the mean (with one possible outlier) and according to a Lilliefors test they are normally distributed for the case of normal cells. Table 7.6 gives the p-values obtained from paired t-tests regarding equal mean for realizations with fixed and random times respectively. The p-values are all significant, hence the hypothesis that realizations from fixed and random times have equal means is rejected.

Mutations considered up to this point have not been purely oncogenic, i.e. mutant cells have not necessarily had a proliferation advantage compared to normal cells at all times t. Therefore,

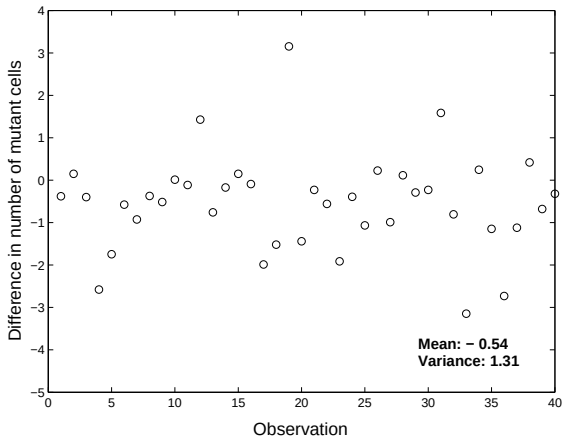

Figure 7.15: Residuals from comparison of fixed and random times for realizations of $\{Y_k^{(1),T}\}_{k=1}^{40}$ when $\lambda_t = \Delta t = 1$ and $\lambda_\xi = 1$.

Table 7.6: p-values from paired t-tests of equal mean for realizations with fixed and random times respectively.

	Normal	Mutant
$\lambda_\xi = 0.25$	<0.0001	<0.0001
$\lambda_\xi = 0.5$	<0.0001	<0.0001
$\lambda_\xi = 1$	0.0021	0.0047

the case with fixed times and $\xi_0 = 1$ is now considered, making all values of $\xi(t)$ give the mutant cells a proliferation advantage. Figure 7.16 shows the mean number of normal and mutant cells for realizations of the purely oncogenic case with different λ_ξ and Figure 7.17 shows the corresponding variances. Clearly, the average number of mutant cells is increased compared

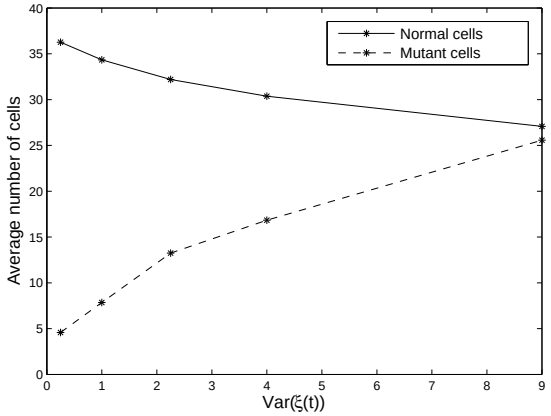

Figure 7.16: Average number of normal and mutant cells respectively with respect to $\mathrm{Var}(\xi(t))$, for purely oncogenic mutations with $\xi_0 = 1$.

to the case where $\xi(t)$ can take values on both sides of one. The variance in number of mutant

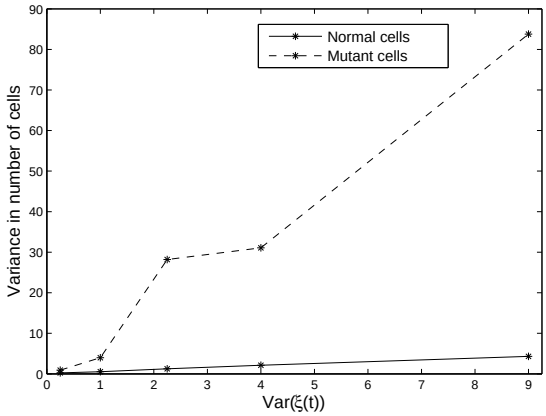

Figure 7.17: Variance in number of normal and mutant cells respectively with respect to $\mathrm{Var}(\xi(t))$, for purely oncogenic mutations with $\xi_0 = 1$.

cells is also increased for the purely oncogenic case. However the number of normal cells is less affected, being similar to the previous case.

7.2.2 Escape from chemical control

Applying (7.5) to the parameter for a type V mutation yields

$$s_0(t) = \begin{cases} \sum_{i=0}^{n-1} Z_i I\left(t \in [t_i, t_{i+1})\right) & \text{if } t \in [0,T), \\ Z_{n-1} & \text{if } t = T, \end{cases} \tag{7.9}$$

with everything as previously defined. No deterministic component is included due to the mutation giving mutant cells a proliferation advantage at time t if $s_0(t) > 0$, which is guaranteed by the Z_i:s for exponential and uniform distributions.

Figure 7.18 shows the average number of normal and mutant cells for the different λ_{s_0} in Table 7.7 and Figure 7.19 shows the corresponding averages for the uniform cases in Table 7.8.

In both cases the average number of mutant cells increases quite rapidly as $\mathrm{Var}(s_0(t))$ is

Table 7.7: Combinations of λ_{s_0} used for simulations of the exponential case together with the corresponding $\mathrm{E}[s_0(t)]$ and $\mathrm{Var}(s_0(t))$.

λ_{s_0}	$\mathrm{E}[s_0(t)]$	$\mathrm{Var}(s_0(t))$
4	1/4	1/16
2	1/2	1/4
1	1	1
3/4	4/3	16/9
2/3	3/2	9/4

increased, whereas the average number of normal cells exhibits a slow decrease of less than ten cells. For the exponential case, the variance in number of normal and mutant cells respectively

Table 7.8: $I_{a,b}^{s_0}$ used for simulations of the uniform case together with corresponding $\mathrm{E}[s_0(t)]$ and $\mathrm{Var}(s_0(t))$.

$I_{a,b}^{\xi}$	$\mathrm{E}[s_0(t)]$	$\mathrm{Var}(s_0(t))$
[0,0.5]	0.25	1/48
[0,1]	0.5	1/12
[0,1.5]	0.75	3/16
[0,2]	1	1/3
[0,2.5]	1.25	25/48
[0,3]	1.5	3/4
[0,4]	2	4/3

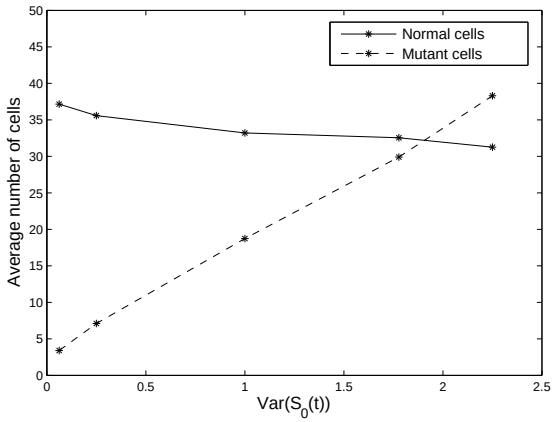

Figure 7.18: Average number of normal and mutant cells respectively with respect to $\mathrm{Var}(s_0(t))$, for the exponential case with λ_ξ as in Table 7.7.

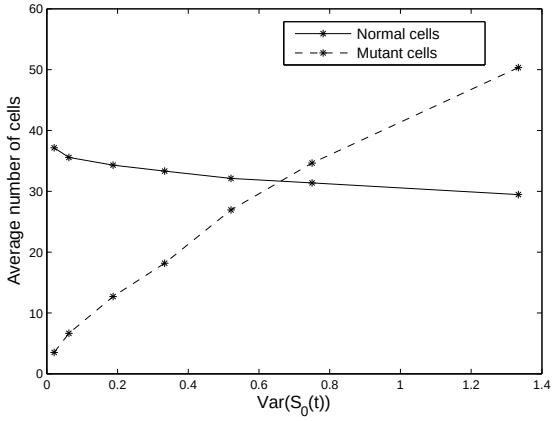

Figure 7.19: Average number of normal and mutant cells respectively with respect to $\mathrm{Var}(s_0(t))$, for the uniform case with $I_{a,b}^{s_0}$ as in Table 7.8.

were in the (approximate) ranges $0-2$ and $0-500$. The corresponding ranges for the uniform cases were $0-0.7$ and $0-500$.

As for mutations with respect to the response of a mitotic activator, stepwise regression

analysis is used to obtain models for the number of normal and mutant cells respectively at time T. With $\mathrm{E}[s_0(t)]$ and $\mathrm{Var}(s_0(t))$ as explanatory variables, the following models were obtained (with notations as in the case of a type I mutation)

$$\begin{aligned}
\alpha^{(1)}(T) &= 39.79 - 11.89\ \mathrm{E}[s_0(t)] + 6.60\ \mathrm{Var}(s_0(t)) - 1.07\ \mathrm{Var}(s_0(t))^2, \\
\beta^{(1)}(T) &= -0.99 + 16.60\ \mathrm{E}[s_0(t)] + 2.83\ \mathrm{Var}(s_0(t))^2, \\
\alpha^{(2)}(T) &= 38.88 - 7.52\ \mathrm{E}[s_0(t)] + 5.97\ \mathrm{Var}(s_0(t)) - 1.31\ \mathrm{Var}(s_0(t))^2, \\
\beta^{(2)}(T) &= 2.57 + 52.59\ \mathrm{Var}(s_0(t)) - 12.60\ \mathrm{Var}(s_0(t))^2.
\end{aligned} \tag{7.10}$$

When distribution is also used as an explanatory variable, the models for normal and mutant cells respectively become

$$\begin{aligned}
\alpha(T) &= 39.73 - 11.62\ \mathrm{E}[s_0] + 6.34\ \mathrm{Var}(s_0) - 0.80\ \mathrm{Dist} - 1.03\ \mathrm{Var}(s_0)^2 \\
&\quad +3.93\ \mathrm{E}[s_0] \times \mathrm{Dist} - 0.41\ \mathrm{Var}(s_0)^2 \times \mathrm{Dist}, \\
\beta(T) &= 2.35 + 0.41\ \mathrm{Dist} + 17.19\ \mathrm{E}[s_0]^2 - 0.66\ \mathrm{Var}(s_0)^2 - 11.23\ \mathrm{Var}(s_0)^2 \times \mathrm{Dist}.
\end{aligned} \tag{7.11}$$

Note that each model contains $\mathrm{Var}(s_0(t))$ in some form (first or second order), different than for a mutation with respect to the response of a mitotic activator. Also, due to the lack of deterministic components in the process for s_0, for the exponential case it holds that $\mathrm{E}[s_0(t)]^2 = \mathrm{Var}(s_0(t))$.

Table 7.9 shows the R^2 value for each of the models. The same behavior in R^2 as for a mutation of type I is observed, the values being consistently higher for models for the number of normal cells.

Table 7.9: R^2 values for the models in Equations (7.10)-(7.11), obtained from stepwise regression analysis for $\{X_j^{(i),T}\}_{j=1}^{40}$ and $\{Y_j^{(i),T}\}_{j=1}^{40}$ as randomness is introduced in s_0, i = 1,2. The last two entries correspond to models in which Dist was used as an explanatory variable.

Cell type	Distribution	R^2
Normal	Exp(λ)	0.840
Mutant	Exp(λ)	0.465
Normal	Uni([a,b])	0.949
Mutant	Uni([a,b])	0.639
Normal	-	0.910
Mutant	-	0.567

Distributions that best fit the empirical CDF's obtained from repeated numerical solutions of (6.5) are found using the Kolmogorov distance d_K. Table 7.10 shows the distribution with the lowest d_K for different distributions of Z when compared to the empirical CDF's. For a mutation of type I (exponential case) a Weibull distribution was usually the best fit for the empirical CDF corresponding to the number of mutant cells . As seen in Table 7.10 this is not the case for a type V mutation, where instead a log-normal distribution tends to have the

Table 7.10: Theoretical CDF's that best fit the different empirical CDF's, when $s_0(t)$ is random, according to the Kolmogorov distance.

Distribution of Z	Distribution with lowest d_K	
	Normal	Mutant
Exp(4)	Normal	Log-normal
Exp(2)	Normal	Log-normal
Exp(1)	Log-normal	Log-normal
Exp(3/4)	Normal	Log-normal
Exp(2/3)	Log-normal	Weibull
Uni([0,0.5])	Log-normal	Log-normal
Uni([0,1])	Log-normal	Log-normal
Uni([0,1.5])	Log-normal	Log-normal
Uni([0,2])	Normal	Log-normal
Uni([0,2.5])	Log-normal	Weibull
Uni([0,3])	Normal	Weibull
Uni([0,4])	Normal	Normal

smallest d_K (for the number of mutant cells in the exponential case). Also, whereas a Weibull distribution was the most common one in general when considering a mutation of type I, here it is instead a log-normal distribution that is most common.

Now consider the interarrival times to be exponentially distributed with parameter λ_t. Figure 7.20 shows numerical solutions of (6.5) for three different λ_ξ for $\lambda_t = 0.5, 1, 2$. The non-constant

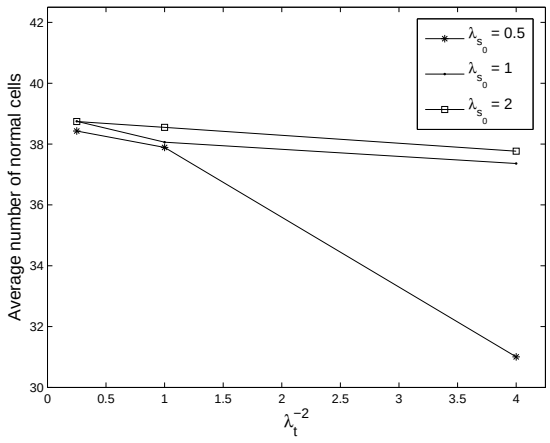

Figure 7.20: Mean of a realization $\{X_k^{(1),T}\}_{k=1}^{40}$ for three different λ_ξ when random times for $s_0(t)$ are considered.

averages (with respect to λ_t^{-2}) for different λ_{s_0} observed in the figures indicate that the value of λ_t has an effect on the average number of cells. p-values for paired t-tests of this effect are shown in Table 7.11. The results further indicate that the expected interarrival time indeed has an effect on the average number of cells (two cases excluded, consistent with what is seen in the

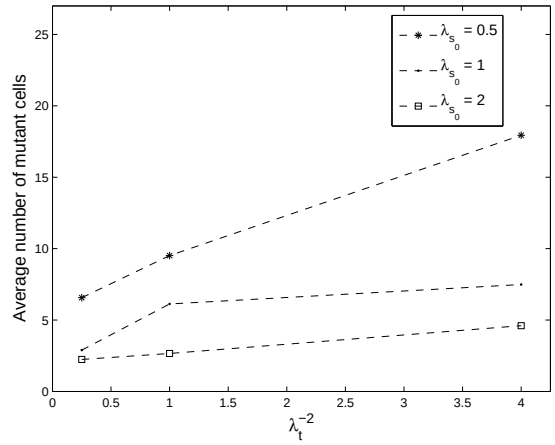

Figure 7.21: Mean of a realization $\{Y_k^{(1),T}\}_{k=1}^{40}$ for three different λ_ξ when random times for $s_0(t)$ are considered.

Table 7.11: p-values for paired t-tests for the hypothesis that for different λ_t the average number of normal and mutant cells respectively are equal.

	$\lambda_{s_0} = 0.5$		$\lambda_{s_0} = 1$		$\lambda_{s_0} = 2$	
	Normal	Mutant	Normal	Mutant	Normal	Mutant
$\lambda_t = 0.5$ vs. $\lambda_t = 1$	0.00074	0.00068	$<$0.0001	0.29	$<$0.0001	0.012
$\lambda_t = 0.5$ vs. $\lambda_t = 2$	0.00031	$<$0.0001	$<$0.0001	0.00030	$<$0.0001	0.0043
$\lambda_t = 1$ vs. $\lambda_t = 2$	$<$0.0001	0.0090	$<$0.0001	$<$0.0001	0.00015	0.22

figures). Next, compare for random and fixed times, $\Delta t = \lambda_t^{-1}$, realizations of $\{X_k^{(1),T}\}_{k=1}^{40}$ and $\{Y_k^{(1),T}\}_{k=1}^{40}$ (for $\lambda_{s_0} = \frac{1}{2}$,1,2). Figures 7.22-7.23 show examples of scatter plots resulting from such comparisons, $\lambda_{s_0} = 1$. Figure 7.22 indicates a difference in average number of normal

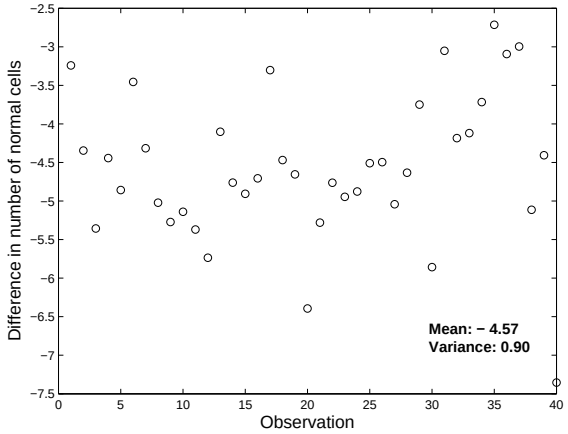

Figure 7.22: Difference in normal cells when the times of change in s_0 are considered as fixed and random respectively, with $\lambda_t^{-1} = \Delta t = 1$ and $\lambda_{s_0} = 1$.

cells when times of change are fixed as compared to random, with the variance of the residuals

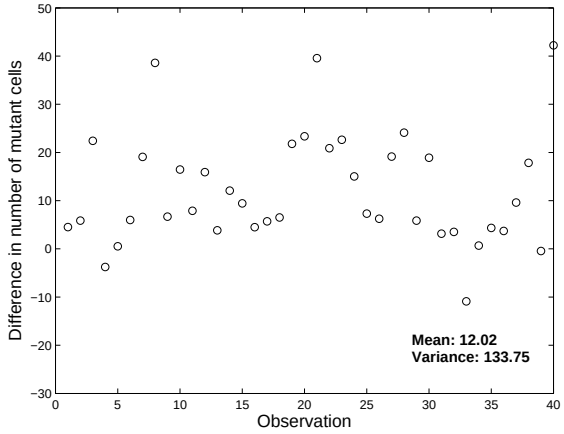

Figure 7.23: Difference in mutant cells when the times of change in s_0 are considered as fixed and random respectively, with $\lambda_t^{-1} = \Delta t = 1$ and $\lambda_{s_0} = 1$.

being very small. The small variance suggest that the number of normal cells is fairly constant in each case. Figure 7.23 indicates the same thing for mutant cells, although now the variance of the residuals is very large. Furthermore, the residuals were normally distributed (according to a Lilliefors test) for normal cells when $\lambda_{s_0} = 1{,}2$ and for mutant cells when $\lambda_{s_0} = 2$. Paired t-tests are used to test whether or not there are significant differences between the average number of cells (normal and mutant) for fixed and random times for different λ_{s_0}; the resulting p-values are presented in Table 7.12. Tests conclude that there are significant differences for both normal and mutant cells for the λ_{s_0} considered.

Table 7.12: p-values from paired t-tests regarding equal mean number of cells when times of change are fixed and random respectively.

	Normal	Mutant
$\lambda_{s_0} = 0.5$	<0.0001	<0.0001
$\lambda_{s_0} = 1$	<0.0001	<0.0001
$\lambda_{s_0} = 2$	<0.0001	<0.0001

7.3 Comments

The case of random perturbations with respect to mutation parameters is in accordance with how one usually adds a random component to an otherwise deterministic system of equations, and has a straightforward biological interpretation. Random mutation parameters are perhaps not as easily to motivate from a biological perspective. However, it is a fact that cancer cells behave in very "strange" ways and this randomness may be used to exhibit such "strangeness" in the model [9]. Moreover, it is interesting to see how the model equations are affected by a random component that is of greater magnitude than the small perturbations first considered,

thus motivating the study of the impact of the type of processes here being used for the mutation parameters.

The exponential and uniform distributions were used for random mutation parameters due to them being easy to realize while still satisfying the conditions imposed by biological interpretations of parameter values. More exotic distributions that satisfy such conditions could also be used, should there be any biological reasons for it.

Chapter 8

Comparison of random and deterministic system

In Chapter 7 random behavior was introduced in the model for initial tumor growth developed in [28]. Here, comparisons are made between the stochastic case and the deterministic approach originally used. In particular, large deviations theory is used to conclude on the asymptotic behavior of probabilities of some improbable events for the model that includes some random component. From this, approximate probabilities regarding initial tumor growth in the random case are obtained.

Section 8.1 presents a short introduction to the mathematical framework of large deviations. Although it is but a very brief introduction to the theory, it is presented in a technical way. In Section 8.2 the theory is used to analyze the case of small random perturbations in the system, a particular case of perturbations with respect to s_0 is shown. Section 8.3 gives comparisons between the stochastic and deterministic system from a large deviations perspective, as well as numerical comparisons of sample paths, as mutation parameters are defined using a stochastic process as in Section 7.2. Here, a particular case of a random ξ is shown.

For those interested in further reading on large deviations [14] is recommended. Being a very technical and somewhat dense account of the theory, readers more interested in the applications of large deviations may instead want to consult [33]. Moreover, [21] gives a very nice account of large deviations theory, perhaps somewhat more easily accessible than [14]. A good account on how large deviation techniques can be used in the context of randomness in dynamical systems is given in [18], some of which has inspired the work in this chapter.

8.1 Introduction to large deviations theory

Large deviations theory is a part of probability theory and deals with so called rare events and their probabilities. The importance of the subject is reflected by the Abel prize awarded to Professor S.R.S. Varadhan in 2007 for [24]

"his fundamental contributions to probability theory and in particular for creating a unified theory of large deviations."

To get an idea of what a rare event is, let $X_1, X_2, ... \in [-1,1]$ be iid. random variables and define $S_n = \frac{1}{2}\sum_{i=1}^{n} X_i$. By the weak law of large numbers (WLLN) a well known fact is

$$S_n \to 0 \text{ in probability,}$$

i.e. $\forall \delta > 0$, $P(S_n > \delta)$ and $P(S_n < -\delta)$ both tend to zero as $n \to \infty$. Hence in this sense, the events $\{S_n > \delta\}$, $\{S_n < -\delta\}$ are rare when n is large. Moreover, Hoeffding's inequality gives that $\forall n \geq 1$

$$P(S_n \geq \delta),\ P(S_n \leq -\delta) \leq e^{-n(\delta^2/2)}.$$

Thus $\{S_n > \delta\}$, $\{S_n < -\delta\}$ are even exponentially rare, i.e. for a fixed δ their probabilities decrease exponentially as n grows. This gives a hint of what a rare event is and also a hint of the context of the theory of large deviations.

Henceforth, let $\mathcal{X}$ be a topological space. The following concepts are essential for large deviations theory.

Definition 8.1.1 *A function $f : \mathcal{X} \to \mathbb{R}$ is lower semicontinuous if $\forall \alpha \in [0,\infty)$, the level set $\{x : f(x) \leq \alpha\}$ is closed. In particular, if $\mathcal{X}$ is a metric space f is lower semicontinuous if and only if $x_n \to x$ implies $\liminf f(x_n) \geq f(x)$.*

Definition 8.1.2 *A function $I : \mathcal{X} \to \mathbb{R}$ is a rate function if it is ≥ 0 and lower semicontinuous. The effective domain of I is defined as $D_I = \{x : I(x) < \infty\}$.*

Now let $\mathcal{B}$ be the Borel σ-algebra on $\mathcal{X}$. Consider probability measures $\mu_1, \mu_2, ...$ on $(\mathcal{X}, \mathcal{B})$. For some set A, let $\overline{A}$ be the closure of A and A^o the interior of A. The backbone of large deviations theory is the so-called *large deviations principle*, given in Definition 8.1.3.

Definition 8.1.3 *The family of probability measures $\{\mu_n\}$ satisfies the large deviations principle (LDP) with rate function I if:*
(i) For all closed sets $F \subseteq \mathcal{X}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) \leq - \inf_{x \in F} I(x).$$

(ii) For all open sets $G \subseteq \mathcal{X}$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(F) \geq - \inf_{x \in G} I(x).$$

An equivalent definition is that a family $\{\mu_n\}$ of probability measures satisfy LDP with rate function I if

$$- \inf_{x \in A^o} I(x) \leq \liminf \frac{1}{n} \log \mu_n(A) \leq \limsup \frac{1}{n} \log \mu_n(A) \leq - \inf_{x \in \bar{A}} I(x)$$

for all $A \subseteq \mathcal{B}$.

It can be seen that the LDP characterizes the family $\{\mu_n\}$ of probability measures in the sense that it states how their limit will behave for different types of sets. It is now easy to understand why it is called a rate function - it basically determines the rate of the exponential decrease for the probability measures. However, one should note that although stating how the μ_n:s will behave as n increases, the LDP gives no hint of how to find the rate function I. In certain settings the rate function can be determined by using Cramer's theorem, which is here stated for the $\mathbb{R}^1$-case. For this some additional definitions are in place. Especially essential is the *Fenchel-Legendre transformation* in Definition 8.1.5.

Definition 8.1.4 *For any law μ (here on $\mathbb{R}$) and values $\lambda \in \mathbb{R}$ where it is defined and finite,*

$$M(\lambda) = \int_{-\infty}^{\infty} e^{\lambda x} d\mu(x),$$

is called the moment generating function of μ. The logarithmic moment generating function associated with the law μ is then defined as

$$\Lambda(\lambda) = \log M(\lambda).$$

Let $\mathcal{D}_\Lambda \triangleq \{\lambda : \Lambda(\lambda) < \infty\}$. Since $\Lambda(0) = 0$, $\mathcal{D}_\Lambda$ is never empty.

Definition 8.1.5 *The Fenchel-Legendre transformation of $\Lambda(\cdot)$ is defined as*

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}} \left(\lambda x - \Lambda(\lambda)\right).$$

Furthermore, $\mathcal{D}_{\Lambda^*} \triangleq \{x : \Lambda^*(x) < \infty\}$.

Both $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ have some interesting properties, see e.g. [14] for more on this.

Now consider $X_1, X_2, ... \in \mathbb{R}$ iid. $X_i \sim \mu$ and let $S_n = \sum_{i=1}^n X_i$. Furthermore, let $\mu_n = \mathcal{L}(S_n)$ denote the probability law of S_n. Cramer's theorem (in $\mathbb{R}$) is stated in Theorem 8.1.6.

Theorem 8.1.6 (Cramer's theorem in $\mathbb{R}$) *In the above setting, $\{\mu_n\}$ satisfies the LDP with convex rate function $\Lambda^*(\cdot)$, i.e.*
(i) for all closed sets $F \subseteq \mathbb{R}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) \leq - \inf_{x \in F} \Lambda^*(x).$$

(ii) for all open sets $G \subseteq \mathbb{R}$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq - \inf_{x \in G} \Lambda^*(x).$$

Cramer's theorem thus gives a way of finding the rate function for the empirical mean of iid. random variables. Moreover, one obtains Corollary 8.1.7.

Corollary 8.1.7 *For any* $y \in \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n ([y,\infty)) = -\inf_{x\geq y} \Lambda^*(x).$$

8.2 Random perturbations

Random perturbations with respect to ξ and s_0 were considered in Section 7.1. The resulting perturbed system is now compared to the deterministic system. As in the previous chapter, mutation parameters are always considered as system parameters. Adopting the ideas of Section 7.1, if K represents a mutation parameter, consider

$$K + \varepsilon\psi(t), \tag{8.1}$$

where $\varepsilon > 0$ is a small number and $\psi(t)$ is derived from a Gaussian process as previously described. Recall that applying this to the mutation parameter characterizing either a type I or a type V mutation changes the equation for mutant cell density to

$$\frac{\partial m^\varepsilon}{\partial t} = D\frac{\partial^2 m^\varepsilon}{\partial x^2} + n^\varepsilon r(n^\varepsilon + m^\varepsilon)\left[s_1((\xi + \varepsilon\psi(t))c_1)s_2(c_2)...s_j(c_j)\right] - m^\varepsilon(\delta + 1), \tag{8.2}$$

for the case of a mutation that alters the response to a mitotic chemical and

$$\frac{\partial m^\varepsilon}{\partial t} = D\frac{\partial^2 m^\varepsilon}{\partial x^2} + n^\varepsilon r(n^\varepsilon + m^\varepsilon)\left[s_0 + \varepsilon\psi(t) + s_1(c_1)s_2(c_2)...s_j(c_j)\right] - m^\varepsilon(\delta + 1), \tag{8.3}$$

for a mutation causing escape from biochemical control.

It has been numerically observed that the stochastic system can be compared to the deterministic version with the corresponding empirical mean $K + \varepsilon\bar{\psi}$ as mutation parameter value. By the law of large numbers, $\bar{\psi} \to E[Z] = 0$ almost surely as $n \to \infty$. With the above, this implies that the stochastic system tends to the deterministic (with mutation parameter value K) as $\Delta t \to 0$ and/or $T \to \infty$. Using large deviation techniques, it is possible to obtain the rate of the decrease in probability of the rare events for $\bar{\psi}$ causing a noticeable difference between the stochastic and deterministic versions. It should be noted that the estimate using $\bar{\psi}$ holds for all $t = k\Delta t$, $k \in \mathbb{N}$ sufficiently large, if the empirical mean is adjusted accordingly.

Consider a realization of (8.1) for [0,T]. The empirical mean of the process up to a time $n\Delta t$ is

$$S_n = \frac{1}{n}\sum_{j=1}^{n}(K + \varepsilon Z_j) = K + \frac{1}{n}\sum_{j=1}^{n}\varepsilon Z_j, \tag{8.4}$$

where the Z_i:s are the random variables of Section 7.1 and n represents the number of changes in $\psi(t)$. Cramer's theorem (or rather Corollary 8.1.7) is well suited to give the limiting behavior of $\mu_n \triangleq \mathcal{L}(S_n - K)$. For $Z \overset{\mathcal{D}}{=} N(0,\sigma^2)$ the moment generating function becomes

$$M(\lambda) = e^{\lambda^2(\sigma^2/2)}.$$

Thus the log moment generating function is

$$\Lambda(\lambda) = \frac{\lambda^2\sigma^2}{2}.$$

Recalling Definition 8.1.5 for the Fenchel-Legendre transformation, it holds that

$$\Lambda^*(x) = \sup_{\lambda\in\mathbb{R}} \left(\lambda x - \frac{\lambda^2\sigma^2}{2}\right).$$

Letting $g(\lambda) \triangleq \lambda x - \lambda^2\sigma^2/2$ yields

$$\frac{d}{d\lambda}g(\lambda) = x - \lambda\sigma^2 \Longrightarrow \frac{d}{d\lambda}g(\lambda) = 0 \Leftrightarrow \lambda = \frac{x}{\sigma^2}.$$

Moreover, $(d^2/d\lambda^2)g(\lambda) = -\sigma^2 < 0$ and thus $\lambda = x/\sigma^2$ yields a maximum of $g(\lambda)$ $\forall x \in \mathbb{R}$. Therefore the explicit expression for the Fenchel-Legendre transformation in this case is

$$\Lambda^*(x) = \frac{x^2}{2\sigma^2}.$$

Applying Corollary 8.1.7 for the set $[y,\infty)$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \in [y,\infty)\right) = -\inf_{x\geq y} \frac{x^2}{2\sigma^2}.$$

With $f(x) \triangleq x^2/2\sigma^2$, f is clearly increasing in x ($x \geq 0$ only interesting) and thus

$$\inf_{x\geq y} \frac{x^2}{2\sigma^2} = \frac{y^2}{2\sigma^2}.$$

Hence for $n \in \mathbb{N}$ sufficiently large,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \in [y,\infty)\right) \approx e^{-n(y^2/2\sigma^2)}, \tag{8.5}$$

which gives the approximate rate (holds asymptotically) of the exponential decrease of the probability of $\{\bar{Z} \in [y,\infty)\}$, where $\bar{Z}$ is the mean of $Z_1,...,Z_n$. Hence it gives the rate at which the random system approaches, in the sense discussed earlier, the deterministic one with mutation parameter K.

The above deals with the stochastic systems tendency to deviate from the deterministic system at specific times $n\Delta t$. Next, consider instead the sample paths (adopting the notation of Chapter 7) $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ and $\{y(t)\}_{t\in[0,T]}$ for the random and deterministic system respectively. A realization of $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ ($\varepsilon = 0.1$ and $\sigma^2 = 2$), when the perturbation is with respect to $s_0 = 1$, is shown in Figure 8.1. The corresponding $\{y(t)\}_{t\in[0,T]}$ is also included, as are $\{X^\varepsilon(t)\}_{t\in[0,T]}$ and $\{(X^\varepsilon + Y^\varepsilon)(t)\}_{t\in[0,T]}$. Figure 8.2 shows a closer look of $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ and $\{y(t)\}_{t\in[0,T]}$. It is clearly seen how $Y^\varepsilon(t)$ fluctuates around $y(t)$, illustrating the effect a

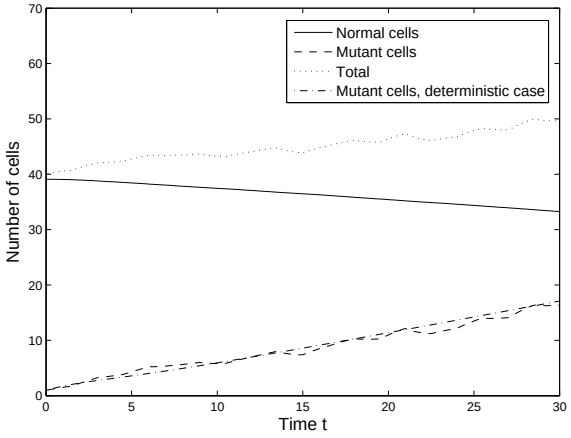

Figure 8.1: A realization of $\{X^\varepsilon(t)\}_{t\in[0,T]}$ and $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ for a perturbation ($\varepsilon = 0.1$ and $\sigma^2 = 2$) with respect to $s_0 = 1$. The corresponding trajectory for $\{y(t)\}_{t\in[0,T]}$ is also included.

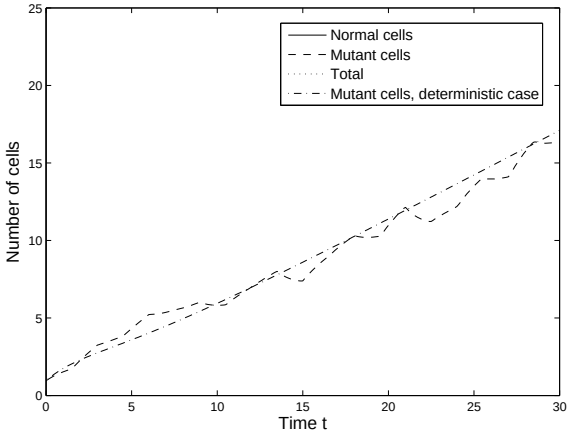

Figure 8.2: A closer look at $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ and $\{y(t)\}_{t\in[0,T]}$ from Figure 8.1.

perturbation with respect to s_0 has on the system (6.5) of governing equations. An interesting property of the realization of the stochastic system is the maximum "distance" (in number of cells) between it and the deterministic system, determined by

$$||Y^\varepsilon - y|| = \sup_{t\in[0,T]} |Y^\varepsilon(t) - y(t)|. \tag{8.6}$$

For the particular realization shown in Figure 8.1, $||Y^\varepsilon - y|| = 1.57$. Now consider 40 realizations of $\{Y^\varepsilon(t)\}_{t\in[0,T]}$, each labeled as Y_i^ε, $1 \le i \le 40$. Here

$$\sup_{1\le i\le 40} ||Y_i^\varepsilon - y|| = 6.41,$$

a rather large difference since $y(T) = 17.10$. Figure 8.3 shows a scatter plot of $||Y_i^\varepsilon - y||$, $1 \le i \le 40$, thus indicating the maximal distance between the random and deterministic systems for different sample paths. The average difference is 2.98 and the observations have a variance

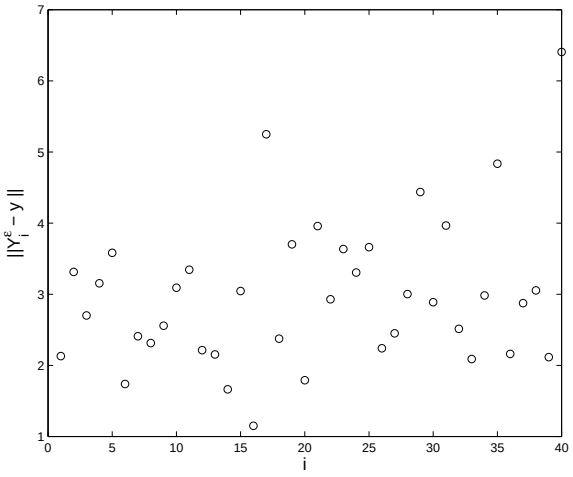

Figure 8.3: Scatter plot of $||Y_i^\varepsilon - y||$, $1 \leq i \leq 40$ for a perturbation with respect to s_0; $s_0 = 1$, $\varepsilon = 0.1$ and $\sigma^2 = 2$.

1.06. Moreover, Figure 8.4 shows the difference $Y_i^\varepsilon(t) - y(t)$ for all i, $1 \leq i \leq 40$ and all $t \in [0,T]$ (an observation was made every $\Delta t = 0.1$). This clearly shows how the randomness due to the

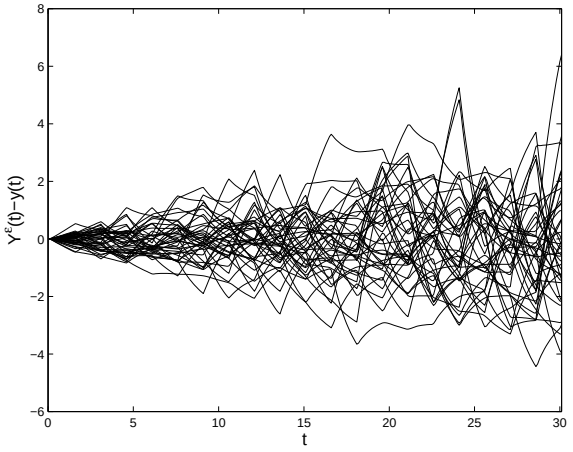

Figure 8.4: Sample paths for the difference $Y_i^\varepsilon(t) - y(t)$, $1 \leq i \leq 40$, for a perturbation with respect to s_0; $s_0 = 1$, $\varepsilon = 0.1$ and $\sigma^2 = 2$.

perturbation has a greater effect on the number of mutant cells as t is increased. The difference between $Y^\varepsilon(t)$ and $y(t)$ is sometimes as large as 40% (compared to the corresponding $y(t)$), indicating the possibility of a sample path of $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ that is quite different from that of $\{y(t)\}_{t\in[0,T]}$.

Remark One should note that due to the properties of normal random variables, the sum of the Z_i:s is a normal random variable as well. Hence it is possible to use standard techniques to obtain the probability that $\frac{1}{n}\sum_{j=1}^n \varepsilon Z_i$ belongs to some set. However, the large deviations approach shows a different route to obtain bounds on the probabilities of certain sets without having to estimate error functions and similar.

8.3 Random mutation parameters

In Chapter 7, a process $\{\phi(t)\}_{t\in[0,T]}$ was defined according to (7.5) and was used for the characterizing parameters for mutations of types I and V. Suggested distributions for the Z_i:s were exponential and uniform. Here it is investigated how the resulting stochastic system compares to the deterministic.

As for random perturbations, it has been numerically observed that the empirical mean obtained from a realization of $\{Y^{(i)}(t)\}_{t\in[0,T]}$ can be used to represent the random system. Adapting the notation of Chapter 7 and letting $\bar{\phi}$ denote the empirical mean of the specified mutation parameter, the $y(T)$ produced with mutation parameter value $\bar{\phi}$ gives a good estimate of the corresponding $Y^{(i)}(T)$. As in the previous section, this holds not only for time $t = T$, but also for any $t = k\Delta t$ where $k \in \mathbb{N}$ is sufficiently large.

Consider $Z \overset{\mathcal{D}}{=} \text{Exp}\,(\lambda)$. The moment generating function for Z is

$$M(\alpha) = \int_0^\infty e^{\alpha t}\lambda e^{-\lambda t}dt = \begin{cases} +\infty & \alpha = \lambda, \\ \frac{\lambda}{\lambda-\alpha} & \text{otherwise.} \end{cases}$$

Omitting the case $\alpha = \lambda$, the log moment generating function is

$$\Lambda(\alpha) = \log M(\alpha) = \log(\lambda) - \log(\lambda - \alpha),$$

yielding a Fenchel-Legendre transformation

$$\Lambda^*(x) = \sup_{\alpha\in\mathbb{R}} \left(\alpha x - \log(\lambda) + \log(\lambda - \alpha)\right).$$

Letting $g(\alpha) \triangleq (\alpha x - \log(\lambda) + \log(\lambda - \alpha))$, for $x \geq 0$

$$\frac{d}{d\alpha}g(\alpha) = x - \frac{1}{\lambda - \alpha} \Longrightarrow \frac{d}{d\alpha}g(\alpha) = 0 \Leftrightarrow \alpha = \lambda - \frac{1}{x}.$$

Moreover $(d^2/d\alpha^2)g(\alpha) = -1/(\alpha - \lambda)^2 < 0$, $\alpha = \lambda - 1/x$ thus yielding a maximum for $g(\alpha)$ for $x \geq 0$. Therefore the explicit expression for the Fenchel-Legendre transform is

$$\Lambda^*(x) = \begin{cases} \lambda x - 1 - \log(\lambda x) & \text{if } x > 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Applying Corollary 8.1.7 for the set $[y, \infty)$, some $y > \lambda^{-1}$ (for other y the result holds but is uninteresting), then gives

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{i=1}^n Z_i \in [y,\infty)\right) = -\inf_{x\geq y} \left(\lambda x - 1 - \log(\lambda x)\right).$$

Moreover, it is easily derived that

$$\inf_{x\geq y}(\lambda x - 1 - \log(\lambda x)) = \begin{cases} \frac{1}{\lambda} + \log(\lambda) - 1 & \text{if } \frac{1}{\lambda^2} \geq y, \\ \lambda y - 1 - \log(\lambda y) & \text{if } \frac{1}{\lambda^2} < y. \end{cases}$$

Thus for $n \in \mathbb{N}$ large

$$P\left(\frac{1}{n}\sum_{i=1}^{n} Z_i \in [y,\infty)\right) \approx \begin{cases} e^{-n(\frac{1}{\lambda}+\log(\lambda)-1)} & \text{if } \frac{1}{\lambda^2} \geq y, \\ e^{-n(\lambda y - 1 - \log(\lambda y))} & \text{if } \frac{1}{\lambda^2} < y, \end{cases} \tag{8.7}$$

which gives the rate of the exponential decrease of the probability of $\left\{\bar{Z} \in [y,\infty)\right\}$ for different y:s.

As for random perturbations, the sample paths of realizations of $\{\{X^{\varepsilon}(t)\}_{t\in[0,T]}$ and $\{\{Y^{\varepsilon}(t)\}_{t\in[0,T]}$ are compared to those of $\{x(t)\}_{t\in[0,T]}$, $\{y(t)\}_{t\in[0,T]}$. Figure 8.5 shows a realization of the sample paths of $\{X^{(1)}(t)\}_{t\in[0,T]}$ and $\{Y^{(1)}(t)\}_{t\in[0,T]}$ for a mutation of type I with $\xi_0 = 1$, $\lambda_\xi = \frac{2}{3}$ (thus comparable to the deterministic case $\xi = 2.5$ used in [28] and also included in the figure). Compared to the case of random perturbations, fluctuations in $Y^{(1)}(t)$ are now more obvious.

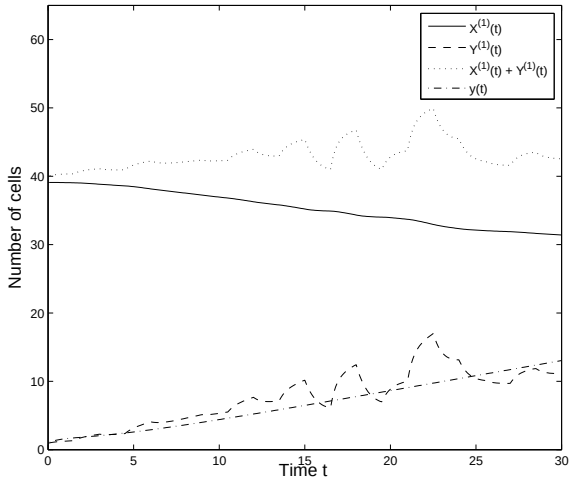

Figure 8.5: Sample paths for $X^{(1)}(t)$, $Y^{(1)}(t)$ and $y(t)$ ($\xi = 2.5$) for $t \in [0,T]$; $\xi_0 = 1$ and $\lambda_\xi = \frac{2}{3}$.

For this sample path $||Y^{(1)} - y|| = 7.24$, a large increase compared to Section 8.2 (especially since $y(T)$ is now approximately four units smaller). Figure 8.6 shows a scatter plot of $||Y_i^{(1)} - y||$, $1 \leq i \leq 40$ representing different realizations. For the $\{Y_i^{(1)}(t)\}_{t\in[0,T]}$:s observed here

$$\sup_{1\leq i\leq 40} ||Y_i^{(1)} - y|| = 14.75.$$

The observations in Figure 8.6 have a mean 6.12 and a variance 5.24. In Figure 8.7 all sample paths $Y_i^{(1)}(t) - y(t)$, $1 \leq i \leq 40$, are shown. The impact of the randomness becomes stronger as t is increased and the largest difference at any t is 70% of $y(t)$. Figure 8.7 indicates, just as Figure 8.4 did for the random perturbations, that a sample path of $\{Y^{(1)}(t)\}_{t\in[0,T]}$ can differ

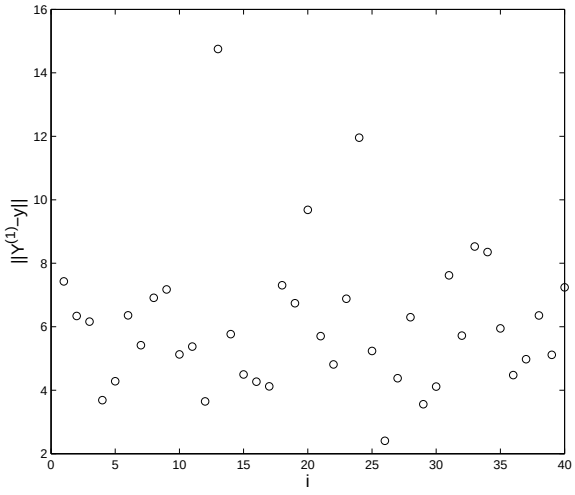

Figure 8.6: $||Y_i^{(1)} - y||$, $1 \leq i \leq 40$, for the case of random mutation parameter with $\xi_0 = 1$ and $\lambda_\xi = \frac{2}{3}$.

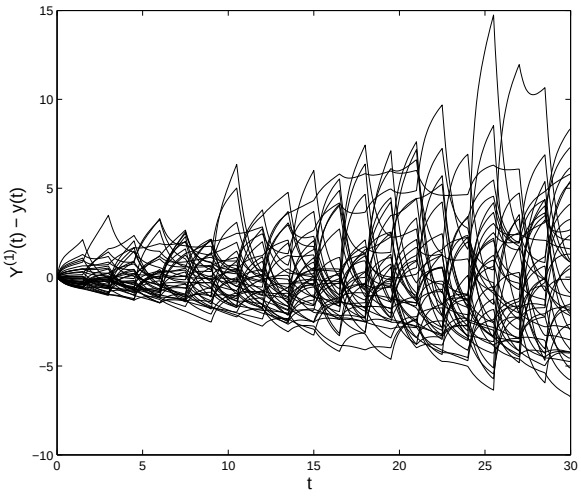

Figure 8.7: Sample paths for the difference $Y_i^{(1)}(t) - y(t)$, $1 \leq i \leq 40$, for the case of random mutation parameter with $\xi_0 = 1$ and $\lambda_\xi = \frac{2}{3}$.

significantly from that of the corresponding $\{y(t)\}_{t\in[0,T]}$.

8.4 Comments

Numerical solutions show that the estimate of $X^\varepsilon(T)$ and $Y^\varepsilon(T)$ using $x(T)$ and $y(T)$ respectively are better when the mutation parameter is set to the corresponding empirical mean rather than the expected value of the stochastic process. Furthermore, the estimations improve when Δt is decreased and/or T is increased. Such changes will decrease the relative time any "extreme" values can affect the system, possibly explaining the more accurate approximations.

The results obtained using large deviation theorems are asymptotic in n. Thus for small

values of n the approximate probabilities will not necessarily hold. However, it has been observed that n does not need to increase too much before the approximate probabilities obtained will differ very little from those estimated from realizations of $\frac{1}{n}\sum_{i=1}^{n} Z_i$. For example for the probability of $\{\bar{Z} \in [y, \infty)\}$, most tried combinations of $\bar{Z}$, y and λ start to give close results for numerical estimations and approximate probabilities when n is in the range $30 - 80$.

An alternative approach to studying randomness would be to add a (scaled) Brownian motion $B(t)$ to the mutation parameters. This would give a continuous change in the parameter value instead of the jumps considered here. Results such as Schilder's theorem (see e.g. [14]) can be used to conclude on the probability of the process crossing certain thresholds related to tumor growth. Instead of the empirical mean one would then be interested in the occupation time (see e.g. [23]) of certain sets and results such as those due to Budhiraja and Dupuis [7] could perhaps be used to conclude on the related probabilities.

Chapter 9

Conclusions

In Part I, experimental data related to the dependence of breast cancer cell behavior on surrounding tissue stiffness is analyzed. An exploratory analysis of the data shows the characteristics of the aggregates and their morphological parameters. While the number of aggregates in each sample decreases with time, the aggregates average perimeter is found to increase. Moreover, different populations seem to have different patterns of change. The rates of change seem to be either increasing or decreasing with respect to the percentage of agarose in the cellular mixture.

The number of aggregates and perimeter data has been analyzed from a repeated measurements perspective, using methods from the theory of linear mixed models. Using the obtained covariance structures, the data is modeled for the different time periods days 1-8, 1-6 and 6-8 using regression analysis. For days 1-8 and 1-6, there are significant differences in the logarithm of the number of aggregates for different populations, but no significant differences in the patterns of change. Furthermore, models suggest significant differences in the logarithm of the aggregates average perimeter as well as the patterns of change of this logarithm for different populations.

The data is still too preliminary for a definite conclusion regarding the aggregates tendency to cluster. Biological aspects such as the rate of cell mitosis and death rate must be considered. In addition, more experimental data from a time period with no outside interference is desirable due to the suspected large impact of media addition/replacement on aggregate formation.

In Part II, stochastic modeling related to a mathematical model for initial tumor growth (Sherratt and Nowak, [28]) is considered. It is investigated numerically how the model responds to stochastic behavior of the parameters defining mutation characteristics. First, small random perturbations are introduced, causing small fluctuations around a specific parameter value. Then a more pronounced randomness of parameters is studied by letting stochastic processes represent the mutation parameters.

The model for tumor growth is observed to be rather stable with respect to small random perturbations. For the case of significant parameter randomness, the average number of cells (normal and mutant) at a time T is highly dependent on the expected value of the stochastic process representing the corresponding parameter value. Especially the number of mutant cells is greatly affected by changes in the expected value and variability of the process, whereas the

number of normal cells is less sensitive. Randomness in s_0 is shown to affect the model more than randomness in ξ. When random jump times are considered, results indicate that the expected value of the interarrival times has a significant effect on the average number of cells at time T. Furthermore, there is a significant difference between random and fixed jump times in the sense of average number of cells.

It is studied how the stochastic system compares to the deterministic one. In particular, large deviations theory is used to obtain asymptotic probabilities for certain rare events connected to tumor growth, describing the convergence of the stochastic system to the deterministic case.

What is presented in Chapters 7 and 8 is meant as a "survey" of what happens to numerical solutions of the governing equations that constitute the model for initial tumor growth, as different types of randomness are included. Rather than to give very specific results, an attempt has been made to try and characterize the behavior that the model exhibits. As experiments become more and more specific and/or relatable to the model, the specific cases one wants to consider can more easily be expressed and the analysis can be more focused on them.

Future work Experiments regarding normal and cancer cells' response to oxygen levels are currently being developed at Clemson University. The aim is to adapt the model by letting oxygen take the place of a biochemical and relate this to experimental results. To that end, a first step is to simplify the model to account for the rather coarse measurements that will be available at first and the fact that oxygen behaves differently from a biochemical. Furthermore, new versions of the experiments described in Section 2, complying with our suggestions for improvements, have recently been finished and data should be available for analysis in the near future.

List of figures

2.1 Images taken from one of the wells at day 2 (left) and at day 7 (right). 6
2.2 Image taken at day 2 from Figure 2.1 and the corresponding processed image. . . 6
2.3 Original image (left) and processed image (right) for the same sample as in Figures 2.1 and 2.2 for day 12. 7

3.1 The number of aggregates, $N_{i,j}$, for each sample during days 1-8. 10
3.2 Total (left) and relative (right) difference in number of aggregates, $N_{i,j}$, between day 8 and day 1. 10
3.3 Normal Q-Q plots for $N_{i,j}$ (left) and $\log(N_{i,j})$ (right) during the days 1-8. 11
3.4 The number of aggregates for each sample during days 1-6. 12
3.5 Total (left) and relative (right) difference in number of aggregates between day 6 and day 1. 12
3.6 Normal Q-Q plots for $N_{i,j}$ (left) and $\log(Ni,j)$ (right) during the days 1-6. 13
3.7 The number of aggregates for each sample during days 6-8. 13
3.8 Total (left) and relative (right) difference in number of aggregates between day 8 and day 6. 14
3.9 The average perimeter of aggregates, $\bar{P}_{i,j}$, for each sample during days 1-8. 14
3.10 Total (left) and relative (right) difference in average perimeter $\bar{P}_{i,j}$ for each sample between day 8 and day 1. 15
3.11 Normal Q-Q plots for $\bar{P}_{i,j}$ (left) and $\log(\bar{P}_{i,j})$ (right) for days 1-8. 15
3.12 The average perimeter of aggregates for each sample during days 6-8. 16
3.13 Total (left) and relative (right) difference in average perimeter $\bar{P}_{i,j}$ for each sample between day 8 and day 6. 16
3.14 Normal Q-Q plots for $\bar{P}_{i,j}$ (left) and $\log(\bar{P}_{i,j})$ (right) for days 6-8. 17
3.15 Total coverage area of aggregates, $A_{i,j}$, for each sample (left) and the mean total coverage area of aggregates, $\hat{A}_i$, for each population (right) during days 1-8. . . . 17
3.16 Total (left) and relative (right) difference in $A_{i,j}$ for each sample between days 8 and 1. 18
3.17 Normal Q-Q plots for (left) $A_{i,j}$ and (right) $\sqrt{A_{i,j}}$ for days 1-8. 18
3.18 Total coverage area of aggregates, $A_{i,j}$, for each sample (left) and the mean total coverage area of aggregates, $\hat{A}_i$, for each population (right) during days 6-8. . . . 19

3.19 Total (left) and relative (right) difference in $A_{i,j}$ for each sample between days 8 and 6. 19
3.20 A normal Q-Q plot for $A_{i,j}$ for days 6-8. 20
3.21 Average circularity of aggregates for samples ($\bar{C}_{i,j}$) and populations ($\hat{C}_{i,j}$) respectively during days 1-8. 21
3.22 Total (left) and relative (right) difference respectively in $\bar{C}_{i,j}$ for each sample between day 8 and 1. 21
3.23 A normal Q-Q plot for $\bar{C}_{i,j}$ for days 1-8. 22
3.24 Average circularity of aggregates for samples ($\bar{C}_{i,j}$) and populations ($\hat{C}_{i,j}$) respectively during days 6-8. 22
3.25 Total (left) and relative (right) difference respectively in $\bar{C}_{i,j}$ for each sample between day 8 and 6. 23
3.26 A normal Q-Q plot for the average circularity of aggregates for days 6-8. 23

5.1 The covariance (left) and correlation (right) as a function of distance in time between pairs of observations for days 1-8. 32
5.2 The covariance (left) and correlation (right) as a function of distance in time between pairs of observations for days 1-6. 33
5.3 The average of $\log(N_{i,j})$ for each population (A-F) during days 1-8. 35
5.4 The average of $\log(N_{i,j})$ for each population (A-F) during days 1-6. 36
5.5 The average number of aggregates for each population (A-F) during days 6-8. . . 37
5.6 The covariance (left) and correlation (right) as a function of distance in time between pairs of observations for days 1-8. 39
5.7 The average of $\log(\bar{P}_{i,j})$ for each population (A-F) during days 1-8. 41
5.8 The average of $\log(\bar{P}_{i,j})$ for each population (A-F) during days 6-8. 42

6.1 The initial growth of a tumor after a mutation that combines types I and V. . . . 51
6.2 The initial growth of a tumor after a mutation of type III. 51
6.3 The change in normal, mutant and total number of cells for mutations of types I, II, V (left) and III, IV (right). 52
6.4 Normal and mutant cell densities respectively in the case of a combined mutation (types I and V). Parameter values used were $\xi = 2$, $s_0 = 2$ and $\delta = 2.8$. The non-dimensional time step was $t_{\text{step}} = 7.5$, with other parameters as before. . . . 52
6.5 Normal and mutant cell densities respectively in the case of a combined mutation (types I and V). Parameter values used were $\xi = 2$, $s_0 = 2$ and $\delta = 3.0$. The non-dimensional time step was $t_{\text{step}} = 7.5$, with other parameters as before. . . . 53

7.1 Realizations of $\{X^\varepsilon(t)\}_{t\in[0,T]}$ and $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ for a perturbation ($\varepsilon = 0.1$ and $\sigma^2 = 2$) with respect to s_0. Also included is $y(t)$ for the case when $s_0 = 1$, i.e. the value coincides with the expected value of $s_0 + \varepsilon\psi(t)$. 57
7.2 A closer look at $\{Y^\varepsilon(t)\}_{t\in[0,T]}$ and $y(t)$ from Figure 7.1. 57

7.3 Variance in number of normal and mutant cells respectively at time T for different ξ and ε, with respect to σ^2. 57
7.4 Variance in number of normal and mutant cells respectively at time T for different s_0 and ε, with respect to σ^2. 58
7.5 Realizations of $\{\phi(t)\}_{t\in[0,10]}$, $\phi_0 = 0$, when $Z \overset{\mathcal{D}}{=} \mathrm{Exp}(\lambda)$ (left column) and $Z \overset{\mathcal{D}}{=} \mathrm{Uni}([a,b])$ (right column). The top row shows realizations for fixed times, $\Delta t = 0.5$, and the bottom row for random times, $\lambda_t = 1$. 59
7.6 Average number of normal and mutant cells with respect to $\mathrm{Var}(\xi(t))$ for the exponential case with λ_ξ as in Table 7.1. 61
7.7 Variance in number of normal and mutant cells respectively with respect to $\mathrm{Var}(\xi(t))$ for the exponential case with λ_ξ as in Table 7.1. 61
7.8 Average number of normal and mutant cells with respect to $\mathrm{Var}(\xi(t))$ for the uniform case with $I^{\xi}_{a,b}$ as in Table 7.2. 62
7.9 Variance in number of normal and mutant cells with respect to $\mathrm{Var}(\xi(t))$ for the uniform case with $I^{\xi}_{a,b}$ as in Table 7.2. 62
7.10 A realization of the random sequence $\{Y_k^{(1),T}\}_{k=1}^{40}$ for $\lambda_\xi = \frac{1}{2}$ and $\xi_0 = 0$. 64
7.11 Empirical CDF for $\{Y_k^{(1),T}\}_{k=1}^{40}$, $\lambda_\xi = \frac{1}{2}$ and $\xi_0 = 0$. 64
7.12 Mean of a realization $\{X_k^{(1),T}\}_{k=1}^{40}$ for three different λ_ξ when random times for $\xi(t)$ are considered. 65
7.13 Mean of a realization $\{Y_k^{(1),T}\}_{k=1}^{40}$ for three different λ_ξ when random times for $\xi(t)$ are considered. 66
7.14 Residuals from comparison of fixed and random times for realizations of $\{X_k^{(1),T}\}_{k=1}^{40}$ when $\lambda_t = \Delta t = 1$ and $\lambda_\xi = 1$. 66
7.15 Residuals from comparison of fixed and random times for realizations of $\{Y_k^{(1),T}\}_{k=1}^{40}$ when $\lambda_t = \Delta t = 1$ and $\lambda_\xi = 1$. 67
7.16 Average number of normal and mutant cells respectively with respect to $\mathrm{Var}(\xi(t))$, for purely oncogenic mutations with $\xi_0 = 1$. 67
7.17 Variance in number of normal and mutant cells respectively with respect to $\mathrm{Var}(\xi(t))$, for purely oncogenic mutations with $\xi_0 = 1$. 68
7.18 Average number of normal and mutant cells respectively with respect to $\mathrm{Var}(s_0(t))$, for the exponential case with λ_ξ as in Table 7.7. 69
7.19 Average number of normal and mutant cells respectively with respect to $\mathrm{Var}(s_0(t))$, for the uniform case with $I^{s_0}_{a,b}$ as in Table 7.8. 69
7.20 Mean of a realization $\{X_k^{(1),T}\}_{k=1}^{40}$ for three different λ_ξ when random times for $s_0(t)$ are considered. 71
7.21 Mean of a realization $\{Y_k^{(1),T}\}_{k=1}^{40}$ for three different λ_ξ when random times for $s_0(t)$ are considered. 72
7.22 Difference in normal cells when the times of change in s_0 are considered as fixed and random respectively, with $\lambda_t^{-1} = \Delta t = 1$ and $\lambda_{s_0} = 1$. 72

7.23 Difference in mutant cells when the times of change in s_0 are considered as fixed and random respectively, with $\lambda_t^{-1} = \Delta t = 1$ and $\lambda_{s_0} = 1$. 73

8.1 A realization of $\{X^{\varepsilon}(t)\}_{t\in[0,T]}$ and $\{Y^{\varepsilon}(t)\}_{t\in[0,T]}$ for a perturbation ($\varepsilon = 0.1$ and $\sigma^2 = 2$) with respect to $s_0 = 1$. The corresponding trajectory for $\{y(t)\}_{t\in[0,T]}$ is also included. 80

8.2 A closer look at $\{Y^{\varepsilon}(t)\}_{t\in[0,T]}$ and $\{y(t)\}_{t\in[0,T]}$ from Figure 8.1. 80

8.3 Scatter plot of $||Y_i^{\varepsilon} - y||$, $1 \leq i \leq 40$ for a perturbation with respect to s_0; $s_0 = 1$, $\varepsilon = 0.1$ and $\sigma^2 = 2$. 81

8.4 Sample paths for the difference $Y_i^{\varepsilon}(t) - y(t)$, $1 \leq i \leq 40$, for a perturbation with respect to s_0; $s_0 = 1$, $\varepsilon = 0.1$ and $\sigma^2 = 2$. 81

8.5 Sample paths for $X^{(1)}(t)$, $Y^{(1)}(t)$ and $y(t)$ ($\xi = 2.5$) for $t \in [0,T]$; $\xi_0 = 1$ and $\lambda_{\xi} = \frac{2}{3}$. 83

8.6 $||Y_i^{(1)} - y||$, $1 \leq i \leq 40$, for the case of random mutation parameter with $\xi_0 = 1$ and $\lambda_{\xi} = \frac{2}{3}$. 84

8.7 Sample paths for the difference $Y_i^{(1)}(t) - y(t)$, $1 \leq i \leq 40$, for the case of random mutation parameter with $\xi_0 = 1$ and $\lambda_{\xi} = \frac{2}{3}$. 84

List of tables

2.1 Agarose content in the cellular mixture for the different populations. 7

5.1 Akaike information criterion for four plausible covariance structures for days 1-8. 33
5.2 Akaike information criterion for four plausible covariance structures for days 1-6. 34
5.3 Akaike information criterion for four plausible covariance structures for days 6-8. 34
5.4 Solution for fixed effects of the model for days 1-8; estimates of coefficients and their standard error, t-test statistics and p-values. 36
5.5 Solution for fixed effects of the model for days 1-6; estimates of coefficients and their standard error, paired t-test statistics and p-values. 37
5.6 Akaike information criterion (AIC) and Schwarz's bayesian information criterion (BIC) for plausible covariance structures for days 1-8. 39
5.7 AIC and BIC for plausible covariance structures for days 6-8. 40
5.8 Solution for fixed effects of the model for days 1-8; estimates of coefficients and their standard error, t-test statistics and p-values. 41
5.9 Solution for fixed effects of the model for days 6-8; estimates of coefficients and their standard error, t-test statistics and p-values. 42
5.10 Modeled covariance structures for the different time periods. 43

7.1 Combinations of λ_ξ and ξ_0 used for simulations of the exponential case together with the corresponding $\mathrm{E}[\xi(t)]$ and $\mathrm{Var}(\xi(t))$. Note that some cases satisfy the $\mathrm{E}[\xi(t)] = 1$ "normal" condition. 60
7.2 $I^\xi_{a,b}$ used for simulations of the uniform case together with corresponding $\mathrm{E}[\xi(t)]$ and $\mathrm{Var}(\xi(t))$. Note that some cases satisfy the $\mathrm{E}[\xi(t)] = 1$ "normal" condition. . 61
7.3 R^2 values for the models in equations (7.7) and (7.8). The last two entries correspond to models in which Dist was used as an explanatory variable. 63
7.4 Theoretical CDF's which fit the different $\xi(t)$'s the best according to the Kolmogorov distance. 65
7.5 p-values from paired t-tests with the hypothesis that for different λ_t the average number of normal and mutant cells respectively are equal. 65
7.6 p-values from paired t-tests of equal mean for realizations with fixed and random times respectively. 67

7.7 Combinations of λ_{s_0} used for simulations of the exponential case together with the corresponding $\mathrm{E}[s_0(t)]$ and $\mathrm{Var}(s_0(t))$. 68
7.8 $I_{a,b}^{s_0}$ used for simulations of the uniform case together with corresponding $\mathrm{E}[s_0(t)]$ and $\mathrm{Var}(s_0(t))$. 69
7.9 R^2 values for the models in Equations (7.10)-(7.11), obtained from stepwise regression analysis for $\{X_j^{(i),T}\}_{j=1}^{40}$ and $\{Y_j^{(i),T}\}_{j=1}^{40}$ as randomness is introduced in s_0, $i = 1,2$. The last two entries correspond to models in which Dist was used as an explanatory variable. 70
7.10 Theoretical CDF's that best fit the different empirical CDF's, when $s_0(t)$ is random, according to the Kolmogorov distance. 71
7.11 p-values for paired t-tests for the hypothesis that for different λ_t the average number of normal and mutant cells respectively are equal. 72
7.12 p-values from paired t-tests regarding equal mean number of cells when times of change are fixed and random respectively. 73

Bibliography

[1] Adam, J.A. (1988). A mathematical model of tumor growth by diffusion, *Mathematical and Computer Modelling* **11** 455–456.

[2] Adam, J.A. (1986). A simplified mathematical model of tumor growth, *Mathematical Biosciences* **81** 229–244.

[3] Bell, G.I. (1976). Models of carcinogenesis as an escape from mitotic inhibitors, *Science* **192** 569–572.

[4] Billingsley, P. (1995). *Probability and Measure, 3rd Ed.* John Wiley & Sons.

[5] Bland, E. (2009). Substrate stiffness vs. cell migration followup test, *Technical report, Department of bioengieering, Clemson University.*

[6] Brockwell, P. Davis R.A. (1998). *Time Series: Theory and Methods, 2nd Ed.* Springer.

[7] Budhiraja, A. Dupuis, P. (2003). Large deviations for the empirical measures of reflecting brownian motion and related constrained processes in $\mathbb{R}_+$, *Electronic Journal of Probability* **8**.

[8] Burg, K.J.L. (2008). EFRI-CBE: Emerging frontiers in 3-D breast cancer tissue test systems, *NSF grant proposal* (unpublished material).

[9] Burg, K.J.L. *Private communication.*

[10] Crooke, P.S.
http://www.math.vanderbilt.edu/ pscrooke/CancerModeling.pdf [Last visited 2009-12-04].

[11] Davis, C.S. (2003). *Statistical Methods for the Analysis of Repeated Measurements* Springer.

[12] DeLisi, C., Rescigno, A. (1977). Immune surveillance and neoplasia - 1: A minimal mathematical model, *Bulletin of Mathematical Biology* **39** 201–221.

[13] DeLisi, C., Rescigno, A. (1977). Immune surveillance and neoplasia - II: A two stage mathematical model, *Bulletin of Mathematical Biology* **39** 487–497.

[14] Dembo, A., Zeitouni, O. (1998). *Large Deviations Techniques and Applications, 2nd Ed.* Springer.

[15] Dickinson, R.B. Tranquillo, R.T. (1993). A stochastic model for adhesion-mediated cell random motility and haptotaxis, *Journal of Mathematical Biology* **31** 563–600.

[16] Dickinson, R.B. McCarthy, J.B. and Tranquillo, R.T. (1993). Quantitative characterization of cell invasion *in vitro*: Formulation and validation of a mathematical model of the collagen gel invasion assay, *Annals of Biomedical Engineering* **21** 679–697.

[17] Dudley, R. (2002). *Real Analysis and Probability, 2nd Ed.* Cambridge University Press.

[18] Freidlin, M.I., Wentzell, A.D. (1998). *Random Perturbations of Dynamical Systems, 2nd Ed.* Springer.

[19] Gard, T.C. (1987). *Introduction to Stochastic Differential Equations,* Marcel Dekker Inc.

[20] Guerin, L., Stroup, W.W. (2000). *A Simulation Study to Evaluate PROC MIXED Analysis of Repeated Measures Data,* Proceedings of the 12th Annual Conference on Applied Statistics in Agriculture.

[21] den Hollander, F. (2000). *Large Deviations* American Mathematical Society (Fields Institute).

[22] Jiang, J. (2007). *Linear and General Linear Mixed Models and Their Applications* Springer.

[23] Karatzas, I., Shreve, S. (1991). *Brownian Motion and Stochastic Calculus, 2nd Ed.* Springer.

[24] Lindstrom, T. (2007). *S.R.S. Varadhan*, http://www.abelprisen.no/nedlastning/2007/varadhan_en.pdf [Last visited 2009-12-14].

[25] Littell, R.C. Stroup, W.W. and Freund, R.J. (2002). *SAS® For Linear Models, 4th Ed.* SAS Institute Inc.

[26] Moghe, P.V., Tranquillo, R.T. (1994). Stochastic model of chemoattractant receptor dynamics in leukocyte chemosensory movement, *Bulletin of Mathematical Biology* **56** 1041–1093.

[27] National Cancer Institute
http://www.cancer.gov [Last visited 2009-12-03].

[28] Sherratt, J.A., Nowak, M.A. (1992). Oncogenes, anti-oncogenes and the immune response to cancer: A mathematical model, *Proceedings: Biological Sciences* **248** 261–271.

[29] Verbeke, G., Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data* Springer.

[30] Verbeke, G., Molenberghs, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach* Springer.

[31] Reinhold, D., Budhiraja, A. and Leadbetter, M.R. (2009). Total coverage area and related measures of breast cancer cell development – 1: Early stiffness experiments and analysis, *EFRI Project Report UNCSTAT #1* (in preparation).

[32] SAS Institute Inc. (2008). *SAS/STAT 9.2 User's Guide*

[33] Shwartz, A., Weiss, A. (1995). *Large Deviations for Performance Analysis* Chapman & Hall.

[34] Tranquillo, R.T., Durrani, M.A. and Moon, A.G. (1992). Tissue engineering science: Consequences of cell traction force, *Cytotechnology* **10** 225–250.