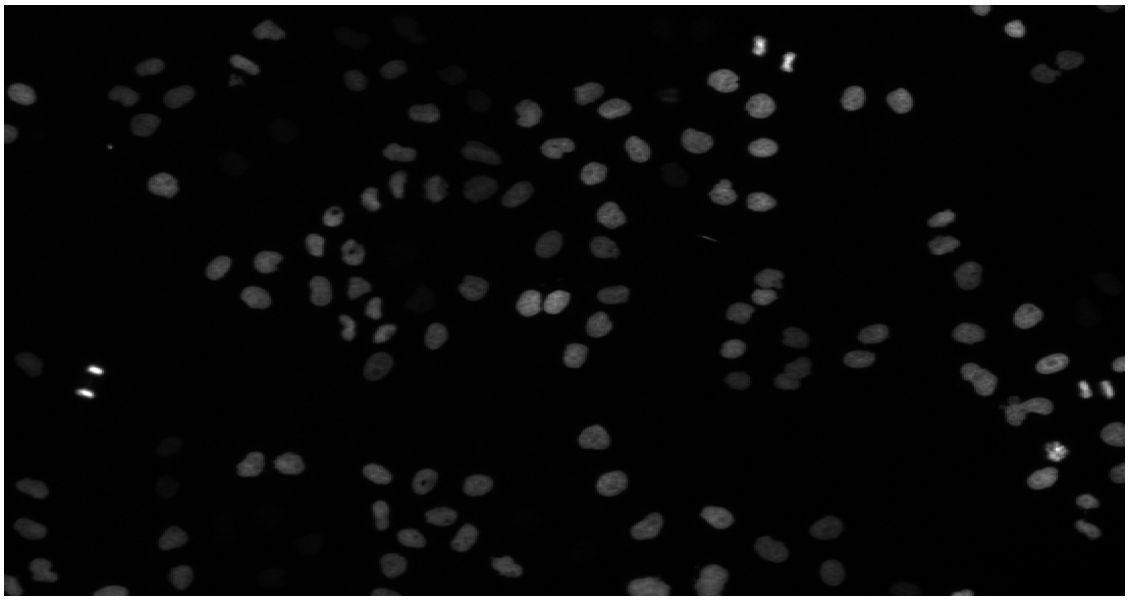


Non-parametric methods for quantifying the Allee effect among cancer cell populations

Master's thesis in Engineering Mathematics and Computational Science

ANNA KÄLLSGÅRD

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Non-parametric methods for quantifying the Allee effect among cancer cell populations

Anna Källsgård

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
Gothenburg, Sweden 2023

Non-parametric methods for quantifying the Allee effect among cancer cell populations
Anna Källsgård

Supervisor: Philip Gerlee, Department of Mathematical Sciences, Chalmers University of Technology
Supervisor: Gustav Lindwall, Department of Mathematical Sciences, Chalmers University of Technology
Examiner: Philip Gerlee, Department of Mathematical Sciences, Chalmers University of Technology

Master's Thesis 2023
Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Microscopic imaging of HeLa cells stably expressing H2B-GFP [1]

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Non-parametric methods for quantifying the Allee effect among cancer cell populations

Anna Källsgård
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

The Allee effect is a phenomenon that has been extensively studied in ecology and is characterized by the per-capita growth rate of a population being low, zero, or even negative in small populations. Recent studies have shown the effect being present within human cancer cell populations. This thesis formulates two non-parametric methods to quantify the Allee effect among cancer cell populations. An agent-based model for cell population dynamics was formulated and used to generate simulated datasets, and microscopic images of cervical cancer cells constitute an in vitro dataset. Promising results were achieved for large simulated datasets. The in vitro dataset considered for this thesis was not of sufficient size for reliable inference, implying the need for large, high-quality datasets for future studies.

Acknowledgements

First and foremost, I would like to thank my supervisors, Gustav Lindwall and Philip Gerlee, for giving me the opportunity to complete my studies with an interesting project. Their knowledgeable advice and continuous support throughout our time together has not only made this thesis a wholesome experience, but also furthered my journey of recovery in personal matters.

Further, I want to express my gratitude to friends and family, especially my father Bosse Källsgård, for showing an interest in what I do and listening when I need to express my thoughts.

Finally, I would like to give a special thanks to Per and Irma, Gustav's feline companions, for allowing me to pet them and for listening to my procrastinational research of Roman history.

Anna Källsgård, Gothenburg, February 2023

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AMISE	Asymptotic Mean Integrated Squared Error
CTC	Cell Tracking Challenge
HGCC	Human Glioblastoma Cell Culture
ISE	Integrated Squared Error
KDE	Kernel Density Estimation
MISE	Mean integrated square error
SSE	Sum of Squared Errors

Notation

Below is the nomenclature of indices, sets, parameters, and variables that have been used throughout this thesis.

Indices

i	Index for cells.
k	Index for images.
j	Index for bins.

Intervals

$[t_k, t_{k+1})$	Time interval between images.
$[r_j, r_{j+1})$	Bin j in partition of density range $[0, 1]$.

Parameters

α	Scaling parameter for density kernel.
r_0	Equilibrium distance for Morse potential.
a	Well steepness for Morse potential.
D_e	Well depth for Morse potential.
σ	Diffusion coefficent for cells.
λ_0	Parameter for cell growth model.
λ_1	Parameter for cell growth model.
ω	Parameter for cell growth model.

Variables

t	Time.
ρ	Local density.
β	Number of cell divisions.
R	Matrix containing all cell densities in all images.
B	Matrix containing the number of cell divsions of all cells in all images.

Contents

List of Figures

List of Tables

1

Introduction

Cancer cell populations at low cell densities are commonly assumed to have an exponential increase in cell number and first at higher densities experience a decrease in growth rate as a response to limited resources, such as space and nutrients. Deviations from the low-density exponential growth have, however, been observed in cancer cell populations at limited densities both cultured in *in vivo*, within their natural context, xenotransplanted cancer cells in mice and *in vitro*, outside of their natural context. Taken together, this suggests an Allee effect [2, 3].

The Allee effect is a phenomenon that emerged in the field of ecology and is characterized by the per-capita growth rate of a population being low, zero, or even negative at small population sizes. Mechanisms suggested to contribute to the effect are cooperative behaviours such as the need for a cooperative food gathering, defence against predators, cooperative breeding, and limited choice of a mating partner. Usually, one differentiates between two subcategories of the Allee effect: weak and strong. One observes a weak Allee effect when the per-capita growth rate increases for low densities but remains positive, whereas one observes a strong Allee effect when the per-capita growth rate is negative for sufficiently low densities. Thus if the population density is below a certain critical value, it may lead to the extinction of the population. [4]

A cooperative behaviour that could be a mechanism behind the Allee effects among cancer cells is autocrine growth factor signalling, which is when cancer cells release signalling molecules that bind to surface receptors of other cancer cells and, in turn, increases the rate of cell division [5]. Autocrine signalling has recently been explored and shown to give rise to Allee effects in an *in vitro* system of glioblastoma cells [6].

In a clinical setting, the lower limit of tumour detection is about 1 million cells [7], which does not generally capture the low-density growth dynamics and implies that it is currently impossible to directly measure the Allee effect in patients. However, in modelling of post-resection recurrence of glioblastomas, a malignant form of brain tumour, the post-resection growth of the tumour was better explained by assuming a weak Allee effect rather than the typically assumed logistic growth [8].

1.1 Aim

A better understanding of cancer cell population dynamics at low densities is important for successful treatment and eradication of cancer in patients. If an Allee effect is present and it could be exploited, it may be used to aid and improve treatment and treatment possibilities.

One possible way to investigate low-density dynamics is by examining time-lapse microscopic images of cancer cells. With advanced technology it is possible to obtain clear images of cells, which, through segmentation and tracking of cells, their positions and mitotic events, can be used to investigate any relation between local density and cell division rate.

Another way to investigate low-density dynamics and the possible presence of Allee effects is through *in silico* modelling, that is, via computer simulation. Here it is possible to set up mechanistic rules and investigate how observable dynamics may emerge from single-cell behaviour. In this vein, we will formulate a cell population dynamics model through single-cell rules for cell migration, division, and death. The model will be implemented and used to generate synthetic data where the cell division rate has different types of dependencies on the local cell density, representing the Allee effects.

This project aims to quantify the impact of local cell density on the rate of cell division among cancer cell populations through two non-parametric approaches, in order to investigate the presence of an Allee effect. The first method utilises histograms together with maximum likelihood estimation to build an estimate of the rate of cell division as a function of local cell density. The second method utilises kernel density estimation, a continuous analogue to histograms, as its main building block for the same purpose. Both approaches will be applied on synthetic data for evaluation and thereafter they will be applied to experimental data in the form of microscopic images and then evaluated.

1.2 Limitations

To have meaningful comparisons between real and simulated data, in relation to the project's aim, the dependencies used will be limited to those corresponding to exponential growth, logistic growth, and the presence of weak and strong Allee effects.

This project does not aim to explain any mechanism behind any potential Allee effect that is observed. However, discussions about possible mechanisms and suggestions for further explorations may be presented.

A limitation that stems from the origin of the data is that any results may be valid for *in vitro* conditions but not necessarily for *in vivo* conditions. The cells in the microscopy data are placed in a well with a heterogeneous, nutritious liquid which may not reflect the conditions in a more complex setting, such as within the human body.

2

Theory

This chapter contains the relevant theoretical background needed to formulate the two inference approaches in this project. The first section presents a summary of previous works related to the Allee effect among cancer cells, It is followed by a section containing a brief recapitulation of maximum likelihood estimation, a section with kernel density estimation theory and lastly a section describing a calculate confidence intervals.

2.1 Previous work

People have been modelling tumours since the 1930s and since the millennia, mathematical modelling approaches in cancer research have steadily grown in both complexity and number [9, 10]. They have been used to better understand cancer's driving mechanisms and processes, make predictions and systematically evaluate assumptions through the quantitative descriptions of cancer mathematical modelling. [11].

It is relatively recently that the Allee effect in a cancer setting gained traction. In 2017, Neufeld et al. presented a model at tissue scale, investigating how tumor invasion fronts are delayed by resection [8]. This model did not show any of the clinically observed delay of cancer remission after tumour resection. Following *in vitro* experiments with glioblastoma cells seeded at low densities, that exhibited an Allee effect, a modified growth model was proposed, which could accommodate for the delay observed in patients.

In 2019, Johnson et al. used a stochastic modelling framework for describing cancer cell population growth[2]. From the stochastic model they derive a master equation, from which they compute the first two moments, the mean and variance, that are subsequently used for parameter inference. The models were validated on simulated data and applied to time-lapse cell proliferation data of cells from a type of breast cancer, which was best described by using a model that considers the Allee effect for small population sizes.

A recent contribution to the topic was presented by Gerlee et al. in 2022, where they investigated the role of autocrine signalling factors in relation to the Allee effect [6]. By combining an on-lattice agent-based model describing the cells as discrete entities and a continuous field describing secreted growth factors, a mean-

field ordinary differential equation model for the cell density was derived. Fitting this model to *in vitro* growth data of glioblastoma cell culture showed the presence of an Allee effect. Further, the model showed that autocrine signalling suffices to cause both weak and strong Allee effects, and whether it leads to the former or latter depends directly on the ratio of cell death to proliferation and indirectly on cellular dispersal.

2.2 Maximum Likelihood Estimation

Maximum likelihood estimation is a widely used method for estimating parameters of an assumed probability distribution, given observed data [12]. The estimator is obtained by formulating a likelihood function and then maximising it with respect to the parameters.

Let $x_1, x_2, \ldots, x_n \in \mathbb{R}$ be a sample of independent, identically distributed random variables from a population with probability density function $f(x|\theta)$ with a single parameter $\theta \in \Theta$, where Θ is the domain of the parameter, then the likelihood function is defined by

$$L\left(\theta|\boldsymbol{x}\right) = L\left(\theta|x_1, \ldots, x_n\right) = \prod_{i=1}^{n} f(x_i|\theta)$$

and the maximum likelihood estimator is

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} L\left(\theta|\boldsymbol{x}\right).$$

The maximum is usually found through looking at local extrema, where the first derivative of the likelihood function is zero. In most cases it is easier to consider the natural logarithm of the likelihood, the *log-likelihood* $l(\theta|\boldsymbol{x}) = \ln L\left(\theta|\boldsymbol{x}\right)$, which share local extrema with the likelihood for $\theta \in (0, \infty)$.

A confidence interval for the MLE $\hat{\theta}_n$ of the true, but unknown, parameter $\theta_0 \in \Theta$ can be constructed using the Fisher information, which can be defined as

$$\mathcal{I}_n(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} l(\theta|\boldsymbol{X})\right)^2\right].$$

It can be shown, through asymptotic normality of the MLE, that

$$\hat{\theta}_n \xrightarrow{d} \mathcal{N}(\theta_0, \mathcal{I}_n(\theta_0)^{-1})$$

where $\xrightarrow{d}$ denotes convergence in distribution. This allows us, for sufficiently large n, not only to use the standard normal deviate $z_{1-\alpha/2}$ to construct a confidence interval of appropriate level but also to replace θ_0 with $\hat{\theta}_n$ in the Fisher information. A $1-\alpha$ confidence level interval for $\hat{\theta}$ is then

$$\hat{\theta}_n \pm \frac{z_{1-\alpha/2}}{\sqrt{\mathcal{I}(\hat{\theta}_n)}}$$

with the Fisher information as defined above.

2.3 Kernel Density Estimation

Kernel density estimation (KDE) is a method of estimating a probability density function given a sample from from said density. KDE does not require any assumption about any parametric family to which the underlying density may belong, rather it learns the shape of the data automatically by placing a kernel on every data point and then averaging out the mass.

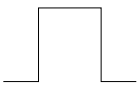
Consider a sample of independent, identically distributed random variables $X_1, X_2, \ldots, X_n \in \mathbb{R}$ from a population with unknown distribution with density function p. Then the kernel density estimator for p is

$$\hat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.1}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a kernel function and $h > 0$ is the smoothing bandwidth.

Some common choices for K can be found in Table 2.1. The effect on the estimation error by choice of kernel is a shift by a generally small constant, allowing a choice of kernel with other properties in mind. [13]

Table 2.1: Common kernel choices for KDE.

Epanechnikov	$\begin{cases} \frac{3}{4}\left(1 - \frac{1}{5}x^2\right)\sqrt{5} \text{ for } \lvert x\rvert < \sqrt{5} \\ 0 \text{ otherwise} \end{cases}$	
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2}$	
Rectangle	$\begin{cases} \frac{1}{2} \text{ for } \lvert x\rvert < 1 \\ 0 \text{ otherwise} \end{cases}$	
Triangle	$\begin{cases} 1 - \lvert x\rvert \text{ for } \lvert x\rvert < 1 \\ 0 \text{ otherwise} \end{cases}$	

2.3.1 Error

Two types of estimation errors will be helpful to consider: the uniform error and the mean integrated square error (MISE). For this purpose, let $\hat{p}_h(x)$ be the KDE defined by (2.1) of the true density $p(x)$.

The uniform error U is defined as the maximum difference between the KDE and

the true density,

$$\begin{aligned} \mathrm{U} &= \sup_x |\hat{p}_h(x) - p(x)| \\ &= \sup_x |\underbrace{\mathbb{E}[\hat{p}_h(x)] - p(x)}_{\text{Bias}} + \underbrace{\hat{p}_h(x) - \mathbb{E}[\hat{p}_h(x)]}_{\text{Stochastic variation}}| \end{aligned} \tag{2.2}$$

This error provides control over the error of the entire support and can be used to construct confidence bands for $\hat{p}_h(x)$.

The mean integrated squared error is defined as

$$\mathrm{MISE}\,(\hat{p}_h) = \mathbb{E}\int (\hat{p}_h(x) - p(x))^2 dx.$$

The integrand is positive, and thus one can change the order of integration and expectation, making the integrand the mean squared error (MSE), which in turn, by properties of mean and variance, can be rewritten as the sum of the squared bias and variance of $\hat{p}_h(x)$

$$\begin{aligned} \mathrm{MISE}\,(\hat{p}_n(x)) &= \int \mathbb{E}(\hat{p}_h(x) - p(x))^2 dx \\ &= \int \mathrm{MSE}_x(\hat{p}_h) dx \\ &= \int \left(\mathbb{E}[\hat{p}_h(x)] - p(x)\right)^2 dx + \int \mathrm{Var}[\hat{p}_h(x)] dx. \end{aligned}$$

The MISE is then the sum of the integrated squared bias and the integrated variance of the estimator. This error is used for obtaining an optimal smoothing bandwidth, which adjust this common bias-variance trade-off. [13]

Substituting expressions for the bias and variance of the KDE to obtain exact expressions for the error appears straightforward, but more often than not the calculations are intractable. Under suitable conditions on the kernel and the density to be estimated, approximate expressions for bias and variance can replace the exact ones. Let us now consider such conditions. In the following, assume that the kernel K is symmetric and

$$\int K(x)dx = 1, \quad \int xK(x)dx = 0, \quad \int x^2 K(x)dx = k_1 \neq 0 < \infty.$$

Further assume that the unknown density p has continuous derivatives of at least order two. With these assumptions, the bias can be shown to equal

$$\mathbb{E}[\hat{p}_h(x)] - p(x) = \frac{h^2}{2} \cdot k_1 \cdot p''(x) + O(h^2)$$

and the variance can be shown to be

$$\mathrm{Var}[\hat{p}_h(x)] = \frac{1}{nh} \cdot k_2 \cdot p(x) + O\left(\frac{1}{nh}\right).$$

where $k_2 = \int K^2(t)dt$. Note that the bias does not depend on the sample size n.

With the above calculations the uniform error can be bounded

$$\begin{aligned} \mathrm{U} &= \sup_x |\hat{p}_h(x) - p(x)| \\ &= O(h^2) + O\left(\sqrt{\frac{\log(n)}{nh}}\right) \end{aligned}$$

where $\log(n)$ arises from empirical process theory [14]. The MISE can be written

$$\begin{aligned} \mathrm{MISE}\left(\hat{p}_h(x)\right) &= \int \left(\mathbb{E}[\hat{p}_h(x)] - p(x)\right)^2 dx + \int \mathrm{Var}[\hat{p}_h(x)]dx \\ &= \frac{h^4}{4}k_1^2 \int |p''(x)|^2 dx + \frac{k_2}{nh} + O(h^4) + O\left(\frac{1}{nh}\right) \end{aligned} \tag{2.3}$$

where the dominating terms are called the asymptotic mean integrated squared error (AMISE) and retain the bias-variance trade-off. A small value for bandwidth h is desirable to keep the bias, the first term, low, increasing the variance in the second term. The choice of bandwidth is then an important task to consider, which will be done in the following section.

2.3.2 Bandwidth selection

The choice of smoothing bandwidth h is critical. If it is too small, the estimation will be raggedy, and if it is too large, it might miss essential features. This balance is illustrated in Figure 2.1, where examples of oversmoothing, undersmoothing and correct amount of smoothing are shown. There are several approaches to choosing a reasonable smoothing bandwidth, such as the rule of thumb, cross-validation approaches, plug-in methods and even "by eye". [13, 14]

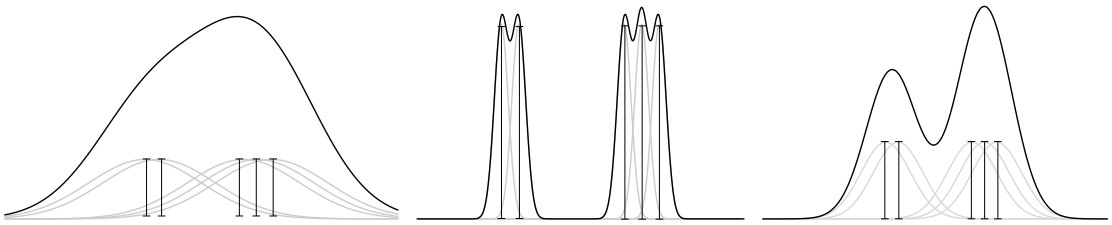


Figure 2.1: Kernel density estimation of the same dataset using different smoothing bandwidth, showing oversmoothing (left), undersmoothing (middle) and correct amount of smoothing (right).

The main idea of most approaches is to choose a bandwidth such that the AMISE is minimised, wherein the difference of the approaches lies in the choice of the estimator to the AMISE (see (2.3)) [13]. If h_{opt} denotes the optimal bandwidth, that is, it minimises

$$\frac{h^4}{4}k_1^2 \int |p''(x)|^2 dx + \frac{k_2}{nh}$$

then it can be shown through calculus that

$$h_{\mathrm{opt}} = k_1^{\frac{1}{5}} k_2^{-\frac{2}{5}} \left\{ \int p''(x)^2 dx \right\} n^{-\frac{1}{5}}. \tag{2.4}$$

As the estimated density is unknown, the question is what to replace the term $\int p''(x)^2 dx$ with. A possible approach is to make a subjective choice regarding the term such that it makes the estimate agree with one's prior beliefs about the density. Another approach is to use a standard family of distributions to assign a value to the unknown term, an approach in which the "rule of thumb" originated. Using a normal distribution with variance σ^2 the unknown term becomes

$$\int p''(x)^2 dx = \frac{3}{8}\pi^{-\frac{1}{2}}\sigma^{-5} \approx 0.212\sigma^{-5}.$$

One can estimate σ^2 from the data and then use the above value. When the underlying distribution is unimodal, this method works well but tends to oversmooth when the underlying distribution is multimodal, as the estimated variance will be higher. If we let the kernel be Gaussian, as defined in Table 2.1, the optmial bandwidth becomes

$$h_{\text{opt}} = 1.06\sigma n^{-\frac{1}{5}} \tag{2.5}$$

with σ estimated by the sample standard deviation or a more robust estimate of σ. [14]

2.3.3 Confidence band and bias handling

A confidence band of a density function is a random interval $C_{1-\alpha}$ such that it covers the true value of the density $p(x)$ with probability $1-\alpha$. That is, it satisfies

$$P\left(p(x) \in C_{1-\alpha}(x)\ \forall x \in \mathbb{K}\right) \geq 1-\alpha.$$

for the domain $\mathbb{K}$ of the density. A confidence band is called asymptotically valid if it has coverage $1-\alpha+o(1)$. In the following, we wish to find such an interval and we also want to correct for the bias of the KDE. These two quests nicely coincide in a single solution presented in the end of this section.

Recall the uniform error defined in (2.2) and let $G(t)$ be its cumulative distribution function, i.e.

$$G(t) = P\left(\sup_x |\hat{p}_h(x) - p(x)| < t\right).$$

and let $c_{1-\alpha} = G^{-1}(1-\alpha)$ be the $1-\alpha$ quantile. Then,

$$C(x) = [\hat{p}_h(x) - c_{1-\alpha}, \hat{p}_h(x) + c_{1-\alpha}]$$

can be shown to be an asymptotically valid confidence band for $\hat{p}_n(x)$. But, as it has been shown, the error contains a bias part that needs to be attended to, and there are various approaches to bias handling. When possible, focusing on the expected value of the estimator and ignoring the bias is an option. Otherwise, undersmoothing, i.e. choosing h such that the bias converges faster than the stochastic variation and thus making the latter dominate the errors, is a common approach in KDE. Another approach is explicitly correcting the bias and then constructing a confidence region with the bias-corrected KDE.

A combination of the two approaches was presented in [15], resulting in an technique that has the advantage of needing to select a single bandwidth through one of the

conventional selection approaches and still have a asymptotically valid confidence band. To explicitly correct the bias, the second derivative $p''(x)$ in the asymptotic bias (see (2.3)) is estimated with the second derivate of a KDE $\hat{p}_b(x)$ using bandwidth b with $\tau = \frac{h}{b}$ for a fixed $\tau \in (0, \infty)$. The bias-corrected KDE is then

$$\begin{aligned}\hat{p}_{\tau,h}(x) &= \hat{p}_h(x) - \frac{h^2}{2} \cdot k_1 \cdot \hat{p}_b^{(2)}(x) \\ &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) - \frac{h^2}{2} \cdot k_1 \cdot \frac{1}{nb^3} \sum_{i=1}^{n} K^{(2)}\left(\frac{x - X_i}{b}\right) \\ &= \frac{1}{nh} \sum_{i=1}^{n} M_\tau\left(\frac{x - X_i}{h}\right)\end{aligned} \tag{2.6}$$

where

$$M_\tau(x) = K(x) - \frac{1}{2} \cdot k_1 \cdot \tau^3 \cdot K^{(2)}(\tau \cdot x)$$

is called the bias-corrected kernel and $K^{(2)}(x)$ is the second derivative of the kernel function. When choosing the optimal bandwidth in (2.5), the estimator $\hat{p}_b^2(x)$ for $p''(x)$ may not be consistent but it is unbiased in the limit. It can also be shown that this choice turns out to result in undersmoothing for the bias-corrected KDE, as the bias of the debiased kernel is on a higher order than the regular KDE while stochastic variation is of the same order as in the regular KDE. A summary of the technique presented can be found in Figure 2.2. Further, it can be shown that approximating the quantile through this bootstraping procedure will retain the asymptotic validity.

Confidence Band of Density Function

(1) Choose an optimal smoothing bandwidth h_{opt} by a standard approach.

(2) Compute the bias-corrected KDE $\hat{p}_{\tau,h_{\text{opt}}}(x)$ with a fixed value τ (generally $\tau = 1$).

(3) Bootstrap the original sample m times and compute the bias-corrected KDE

$$\hat{p}^{*(1)}_{\tau,h_{\text{opt}}}(x), \ldots, \hat{p}^{*(m)}_{\tau,h_{\text{opt}}}(x).$$

(4) Compute the quantile

$$\overline{c}_{1-\alpha} = \hat{G}^{-1}(1-\alpha), \quad \hat{G}(t) = \frac{1}{m} \sum_{j=1}^{m} I\left(\sup_x |\hat{p}^{*(j)}_{\tau,h_{\text{opt}}}(x) - \hat{p}_{\tau,h_{\text{opt}}}(x)| < t\right).$$

(5) Output the confidence band

$$\overline{C}_{1-\alpha}(x) = \left[\hat{p}_{\tau,h_{\text{opt}}}(x) - \overline{c}_{1-\alpha}, \hat{p}_{\tau,h_{\text{opt}}}(x) + \overline{c}_{1-\alpha}\right].$$

Figure 2.2: Construction of an asymptotically valid confidence band of the density function as presented by [15].

One also has to consider boundary bias at the edge of the support. Boundary bias can be reduced through a mirroring approach, which can be viewed as moving the mass of the KDE that falls outside the boundary back into the support by reflection at the boundary. Kernel density with the mirroring approach will have consistent boundaries, but the bias there is of order $O(h)$. A schematic view of the process can be found in Figure 2.3. Other approaches with boundary kernels, such as generalised jack-knifing, can be constructed to have bias of order $O(h^2)$, but requires more effort to be implemented. [16]

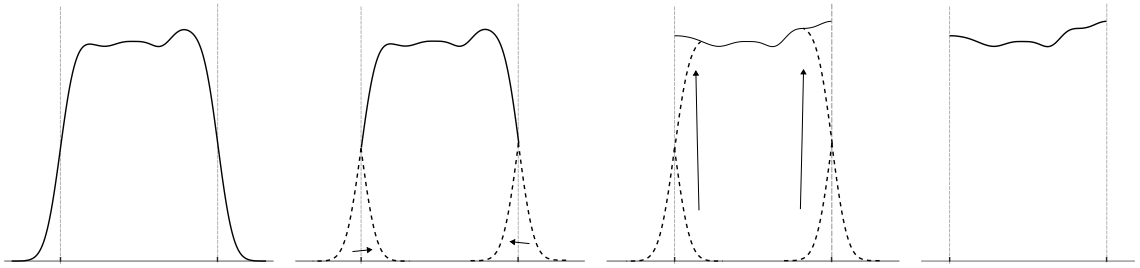


Figure 2.3: A schematic view of the mirroring process used in this project to decrease boundary bias.

2.4 Fieller's theorem

This section will present some theory for calculation of a confidence interval when the variable of interest is a ratio of two means. We will need this result when constructing confidence intervals for the estimate in one of our estimation methods later in this thesis. The following is a special case of Fieller's theorem [17] as presented in [18].

Let A and B be the mean of two samples from normal distributions with expectations μ_A and μ_B, respectively, and variances σ_A^2 and σ_B^2, respectively. Further, assume that we are interested in the quotient $Q = \frac{A}{B}$ and a corresponding confidence interval. Let S_A^2 and S_B^2 be unbiased estimators of σ_A^2 and σ_B^2. Then, calculate an intermediate variable g

$$g = \left(t_{r,\alpha} \cdot \frac{S_b}{B}\right)^2 \tag{2.7}$$

where $t_{r,\alpha}$ is the α level deviate from the Student's t-distribution with r degrees of freedom. In this setting r is the total number of observations used in calculating both means minus two. Under the assumption that A and B are not paired, an $1-\alpha$ confidence level interval for Q can be constructed as

$$\frac{Q}{1-g} \pm t_{r,\alpha} \cdot \frac{Q}{1-g} \cdot \sqrt{(1-g)\frac{S_A^2}{A^2} + \frac{S_B^2}{B^2}}. \tag{2.8}$$

Two pitfalls must be addressed when using this formula. The first one occurs $g > 1$, which is the case when S_B is very large in relation to B. The factor that the mean

quotient Q and its standard deviation are multiplied with then becomes negative and the confidence interval cannot be calculated. The second one occurs when the expression within the square root becomes negative, which is the case when

$$t_{r,\alpha}^2 > \frac{S_B^2}{B^2} + \frac{S_A^2}{A^2}.$$

Again, when this is the case, the confidence interval cannot be calculated as the standard deviation for the mean quotient has become complex.

3

Materials and Methods

In this chapter the material, methods and data used in this project will be presented. First, we give a brief description of the experimental data, followed by a presentation of the mathematical model for cell dynamics and how the model is applied to produce synthetic data. In the last section we describe the two methods for inference in this project.

3.1 Experimental data

The experimental data used in this project is provided by Cell Tracking Challenge (CTC), an initiative with focus on objective comparison of cell segmentation and tracking algorithms. The challenge was initially hosted under the auspices of the IEEE International Symposium on Biomedical Imaging but remains open for online submission. The CTC provides a host of 2D and 3D time-lapse video sequences from various microscopy modalities along with reference annotations for the training datasets. [19]

One such dataset is denoted "Fluo-N2DL-HeLa" and was provided to the CTC by the Mitocheck Consortium. HeLa cells come from the oldest immortalised human cell line which was obtained from a cervical cancer patient in the 1950s. Since then they have had an enormous effect on biomedical research [20]. The dataset contains two experiments, each consisting of a sequence of fluorescent microscopic images of HeLa cells stably expressing H2B-GFP, an alteration to its genetic code to produce green fluorescent protein. The cells are placed on a flat glass substrate and images have been taken at 30-minute intervals for 45 hours. The top row of Figure 3.1 shows the first and last image of the experiment which we will denote the first experiment/dataset. In the bottom row we find the same for the second experiment/dataset.

The reference annotation provided for the training dataset considered in this thesis are obtained as a consensual or majority opinion of several human experts and will be used as ground truth. By extracting the relevant information from the ground truth and thereafter apply a normalisation procedure, to make the cells have a diameter of approximately one, the reference annotations will be used as experimental data. [1]

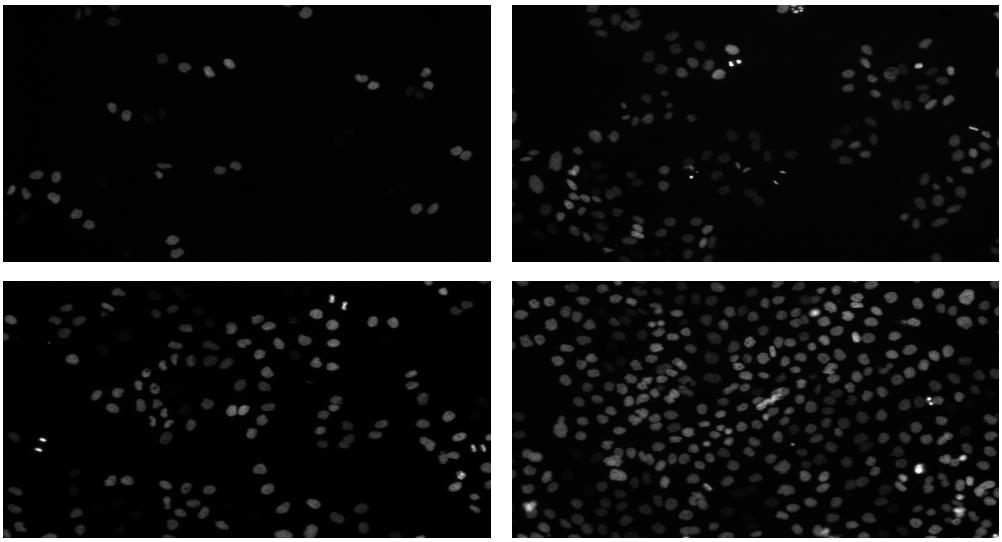


Figure 3.1: First and last images of the two experiments of the "Fluo-N2DL-HeLa" dataset. Top row: the first experiment. Bottom row: the second experiment.

3.2 Cell population dynamics model

In the following sections the various aspects of the mathematical model for cell dynamics will be presented. Some decisions are made with respect to the experimental data and some are made for mathematical convenience. An initial choice is to use an off-lattice model rather than an on-lattice. As the resolution of the spatial evolution of the cells is high, continuous spatial movement seems appropriate.

3.2.1 Calculation of local cell density

To calculate the local density for a cell, some assumptions and decisions must be made. For convenience, the cells are assumed to be circular and have a diameter of unit length.

Let $\boldsymbol{x}_i(t) \in \mathbb{R}^2$ denote the position of cell i at time t. If all cell positions are known at time t, the distribution of cell locations can be defined as the empirical measure

$$\mu^{N_t}(t) = \sum_{i=1}^{N_t} \delta_{\mathbf{x}_i(t)}(\mathbf{x})$$

where $\delta_{\mathbf{x}_i(t)}(\mathbf{x})$ is the Dirac measure and N_t is the number of cells at time t. Local density can then be defined through convolution with a density kernel, which should be a rapidly decreasing function describing the influence of neighbouring cells on cell i. Thus, let the local cell density of cell i at time t be defined as

$$\rho_i\left(\mu^{N_t}(t)\right) = \int_{-\infty}^{\infty} w\left(||\mathbf{y} - \mathbf{x}_i(t)||\right) d\mu^{N_t}(t) = \sum_{j \neq i} w\left(||\mathbf{x}_i(t) - \mathbf{x}_j(t)||\right) \tag{3.1}$$

where $w(\boldsymbol{x})$ is a density kernel. Let the density kernel be defined as

$$w(\mathbf{x}) = e^{-\alpha||\mathbf{x}||} \tag{3.2}$$

where $\alpha > 0$ is a scaling constant such that the local density ranges between zero and one with zero i.e.

$$0 \leq \rho_i \left(\mu^{N_t}(t)\right) \leq 1.$$

Here 0 indicates that the cell is alone and 1 indicates a complete hexagonal grid of neighbouring cells as far as their non-existing eyes can reach. It should be noted that the local density may temporarily exceed one, for instance at the moment immediately after cell division.

To determine an appropriate value for the scaling parameter α in the density kernel one can look to the steady-state distribution of an infinite number of cells on a hexagonal grid, which should correspond to a local density of 1 for a particular cell. An expression for the distance between any two cells in the grid aids in this quest. Consider a coordinate system as in Figure 3.2 and the distance between the center points of the cell at the origin and a cell at the place (n, m) in the coordinate system. Using the law of cosines, the distance between them is $d\sqrt{n^2 + m^2 + nm}$, where d is the cell diameter. The calculation is also illustrated in Figure 3.2. With this neat expression, the equation for maximum local density becomes

$$\rho_{max} = 6 \sum_{n=1}^{\infty} \sum_{m=0}^{\infty} e^{\left(-\alpha d\sqrt{n^2+m^2+nm}\right)} = 1.$$

Through an iterative procedure and using $d = 1$ to solve the equation, the scaling parameter becomes $\alpha \approx 2.1402$.

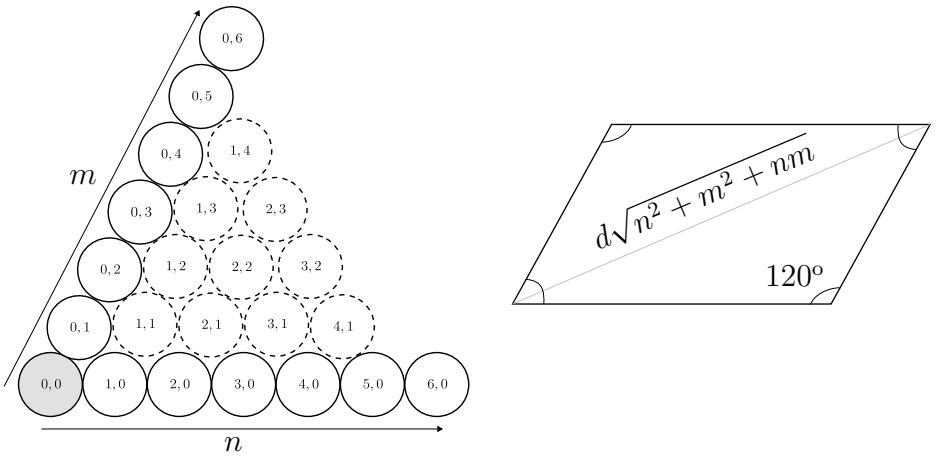


Figure 3.2: One sixth of a completely full hexagonal grid with internal coordinate system (left) and layout for computing the distance between cell at place $(0, 0)$ and cell at place (n, m) (right).

3.2.2 Proliferation and death

For modelling cell proliferation and death we turn to reliability theory, which is a branch of statistics that concerns itself with life length phenomena. Its focus on the duration until a certain event occurs, usually denoted death or failure, provides

a framework where the dependency on local cell density for cell division can be incorporated.

For cell division, consider the survival equation

$$h_i(t) = \frac{B_i'(t)}{1 - B_i(t)}, \quad B_i(0) = 0 \tag{3.3}$$

where $B_i(t)$ denotes the probability that cell i has divided at time t after either its birth or last division and $h_i(t)$ is the instantaneous cell division rate for cell i [21]. With $h_i(t) = \lambda_0$ a constant the growth will be exponential and is not affected by local cell density. With $h_i(\rho) = \lambda_0 \max\{0, 1 - \rho\}$, the cell division rate is inhibited as the local density around cell i increases, and this is known as logistic growth (the max function is to ensure non-negativity, as ρ temporarily may exceed one). With $h_i(\rho) = (\lambda_0 + \lambda_1\rho)\max\{0, 1 - \rho\}$ for low values of ρ the instantaneous cell division rate is increasing, yet still positive, which is known as a weak Allee effect.

For cell death, again consider the survival equation

$$\omega = \frac{D_i'(t)}{1 - D_i(t)}, \quad D_i(0) = 0 \tag{3.4}$$

where $D_i(t)$ denotes the probability that cell i has died at time t after either its birth and ω is a positive constant. This is equivalent to the cell's life length being modelled by an exponentially distributed random variable with rate ω [21]. Note that the event of a cell dying is independent of the event of a cell dividing. Together with cell division rate representing Allee effect, the death rate can lower the net proliferation to being zero or even negative for low values of ρ, which is known as a strong Allee effect.

The different models of cell division and death rates considered in this thesis is found in Table 3.1 and illustrated in Figure 3.3.

Table 3.1: Cell division models considered.

Name	$h(\rho)$	ω
Exponential	λ_0	≥ 0
Logistic	$\lambda_0 \max\{0, 1 - \rho\}$	≥ 0
Allee effects	$(\lambda_0 + \lambda_1\rho)\max\{0, 1 - \rho\}$	≥ 0

3.2.3 Migration and interaction

Vast simplifications and assumptions can be made for cell migration and interaction due to both homogeneity of the environment and due to the fact that cell division in this model not being directly dependent on the manner of movement. Even though in some experimental data cells appear to be more motile than others, the cells are assumed to be of same motility and thus have the same diffusion coefficient. When a cell is alone we assume Brownian motion, whereas in a group of cells we want

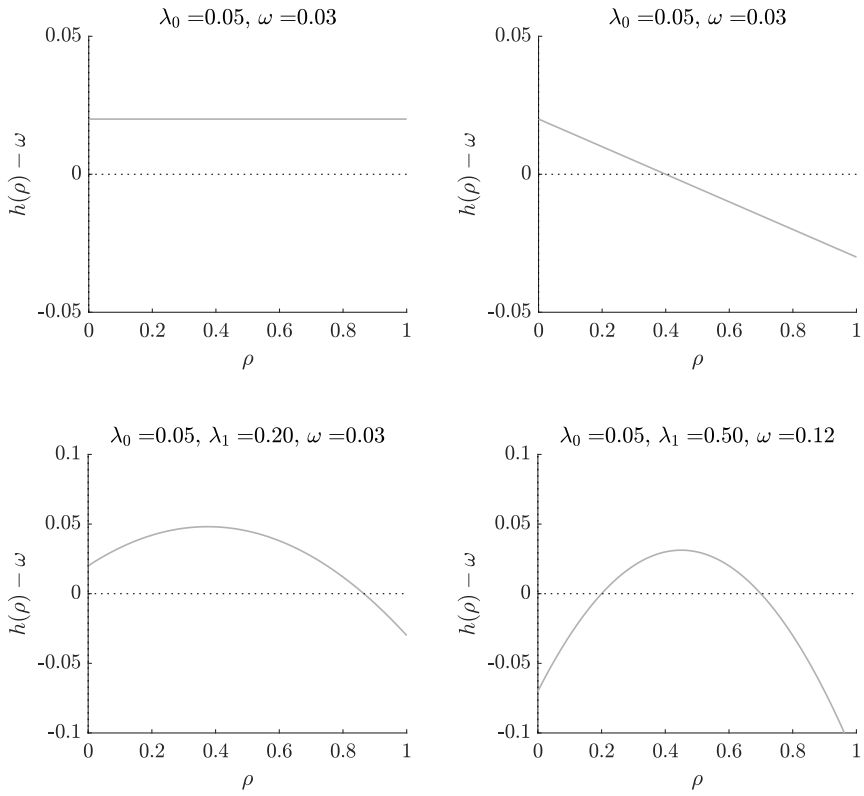


Figure 3.3: The growth rate as a function of local cell density. Exponential (top left), logistic (top right), weak Allee effect (bottom left) and strong Allee effect (bottom right). Values of λ_0, λ_1 and ω are as indicated above the graphs.

to assume some attraction-repulsion force between them. Cells in the experimental data appear to pull at each other once close and then to "stick to each other" somewhat.

With this in regard, cell movement and interaction will be modelled by a set of interacting stochastic differential equations with isotropic diffusion. At a given time t, the system evolves according to

$$d\boldsymbol{x}_i = -\nabla V(\boldsymbol{x}_i, t) + \sigma dW_i(t) \tag{3.5}$$
$$V(\boldsymbol{x}, t) = \sum_{j=1}^{N_t} U(||\boldsymbol{x} - \boldsymbol{x}_j||)$$
$$U(r) = D_e \left(1 - e^{-a(r-r_0)}\right)^2$$

where σ is the diffusion coefficient for the cells, $W(t)$ is the Wiener process and $U(r)$ the Morse potential. Here r is the distance between the midpoints of two cells and r_0 is the equilibrium bond distance which will be set to the cell diameter. Cells whose midpoints are closer than one cell diameter, under the assumption that cells have the ability to deform somewhat, will experience a repulsive force, the size of which depends on the well steepness a. Cells whose midpoints are at distance just above the cell diameter will experience a small attraction to each other, effectively

sticking to each other at various levels largely decided by the well depth D_e. A typical profile of the Morse potential can be seen in Figure 3.4 with the attraction and repulsion forces indicated with arrows.

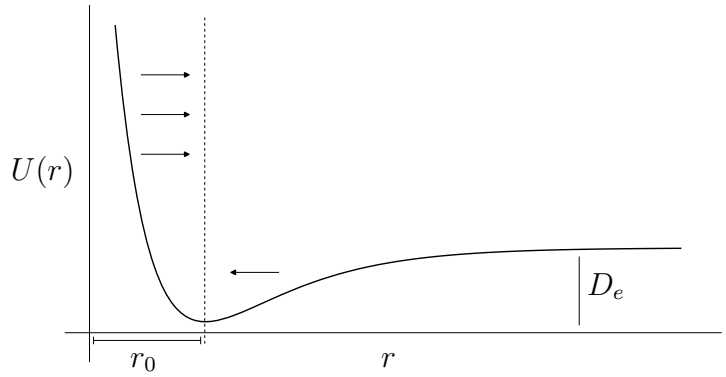


Figure 3.4: A typical profile of the Morse potential with well depth D_e, equilibrium distance r_0 and well steepness a, with attraction and repulsion forces indicated by arrows.

3.3 Generation of synthetic data

This section will describe how the model presented in Section 3.2 is implemented to produce synthetic data. It will be explained in the setting of a single "experiment" which is repeated L times.

In an experiment, the *in silico* time resolution is one second, i.e. $\delta t = 1$, and will progress until some time T, usually corresponding to several days. During regular intervals Δt, typically ranging from 10 minutes up to one hour, a record of cell locations will be made.

An initial number of cells N_0 will be placed uniformly on a circular disc representing the well. With all cell positions known, the local density can be calculated. Recall Equation (3.1) and the density kernel (3.2), yielding local density for cell i at time t

$$\rho_i(t) = \sum_{j \neq i} e^{-\alpha \cdot ||\boldsymbol{x}_i(t) - \boldsymbol{x}_j(t)||}. \tag{3.6}$$

Thereafter the cells will evolve spatially according system (3.5), which is simulated using an Euler-Maruyama scheme,

$$\boldsymbol{x}_i(t + \delta t) = \boldsymbol{x}_i(t) - \nabla V(\boldsymbol{x}_i(t), t) \cdot \delta t + \sigma \sqrt{\delta t} \cdot Z_i \tag{3.7}$$

where $Z_i \sim N(0, 1)$.

Simultaneously cell division and death for each cell will be simulated. Recall the equations (3.3) and (3.4), which both are ordinary differential equations and can be solved using en Euler scheme

$$\begin{aligned} B_i(t + \delta t) &= B(t) + \delta t \Big[h_i(t)(1 - B_i(t)) \Big] \\ D_i(t + \delta t) &= D(t) + \delta t \Big[\omega(1 - D_i(t)) \Big]. \end{aligned} \tag{3.8}$$

Two thresholds b_i and d_i, standard uniformly distributed, accompany each cell. When $B_i(t) > b_i$, a division is triggered. $B_i(t)$ returns to zero and a new b_i is generated. When $D_i(t) > d_i$, a death is triggered and the cell is removed.

Parameter values common for all synthetic data are found in Table 3.2 and parameter values related to different growth models are found in Table 3.3. The parameters mentioned in Table 3.4 were varied, one at the time, assuming the values presented. When a parameter was not varied, it assumed the value given in boldface in the table. For the exponential and logistic growth models, there are no datasets corresponding to the value $N_0 = 2000$ as it is computationally heavy and not the focus of this thesis.

A summary of the procedure, in the form of pseudocode, can be found in Algorithm 1 in Appendix A and a snapshot of a realisation of the algorithm can be seen in Figure 3.5.

Table 3.2: Parameter values common for all synthetic data.

Parameter	Value	Explanation
α	2.1402	Scaling parameter for density kernel
r_0	1	Equilibrium distance for Morse potential
a	3	Well steepness for Morse potential
D_e	$5 \cdot 10^{-5}$	Well depth for Morse potential
σ	$1 \cdot 10^{-2}$	Diffusion coefficient for cells

Table 3.3: Parameter values for synthetic data of different growth models. All values are per hour (h^{-1}).

Parameter	Exponential	Logistic	Weak Allee	Strong Allee
ω	0.03	0.03	0.03	0.12
λ_0	0.05	0.05	0.05	0.05
λ_1	-	-	0.20	0.50

Table 3.4: Parameters with varying values for synthetic data.

Parameter	Values
N_0	63, 125, 250, **500**, 1000, 2000
T (h)	12, 24, 36, 48, 60, **72**
Δt (min)	10, **20**, 30, 40

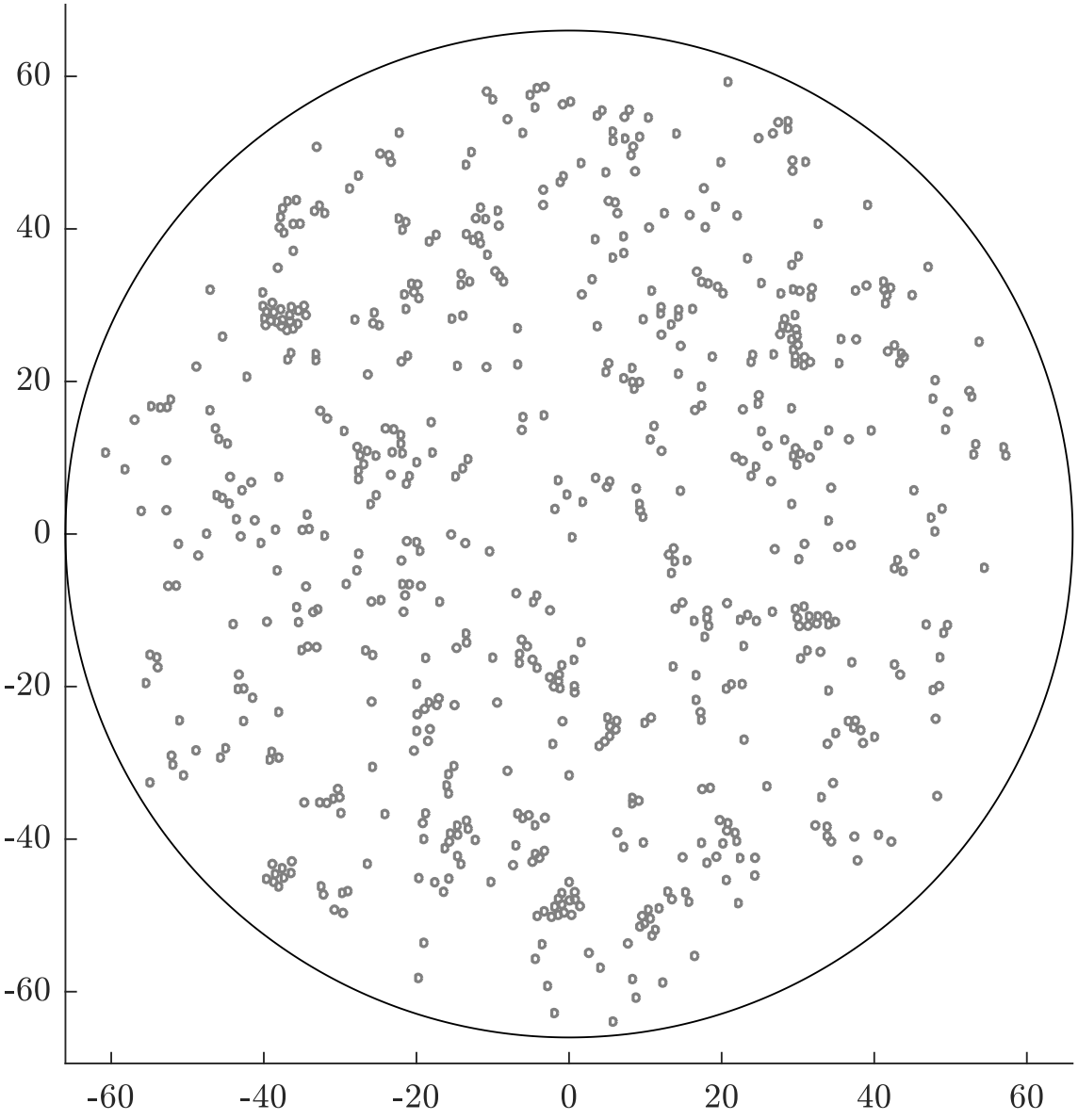


Figure 3.5: A snapshot of a synthetic dataset generated by Algorithm 1.

3.4 Methods for inference

This section will provide a guide through the two manners this project goes about to estimate the cell division rate as a function of local cell density. First, we present some common notation and assumptions, which will be followed by two sections in which the two approaches of estimating cell division intensity will be explained and thereafter a section that explains the approach based in kernel density estimation.

Let us now consider some of the notation and assumptions common to both methods. First, the K denote the number of images taken during the experiment and denote the time at image k by t_k. Then consider the interval between two consecutive images $[t_k, t_{k+1})$. First, note that $\Delta t = t_{k+1} - t_k$ for any appropriate k. Then, in accordance with Itô calculus, we replace the empirical process in every such interval with its leftmost value, that is

$$\rho_i(\mu^{N_t}(t)) = \rho_i(\mu^{N_{t_k}}(t_k)) = \rho_{ik} \text{ for } t \in [t_k, t_{k+1})$$

Let the number of cell divisions for cell i during the time interval $[t_k, t_{k+1})$ be denoted by β_{ik}. Further, we let R be the matrix containing the local densities of every cell at every time. That is, the element at row i and column k is ρ_{ik} as defined above. Finally, let the matrix B contain the number of cell divisions observed for every cell at every image. That is, the element at row i and column k is β_{ik} as defined above. Note that the entries of B rarely exceed one, as the intensity for the time intervals we consider is low. For ease of reading, a list of the notations with explanations is provided in Table 3.5.

Table 3.5: Variables and their explanations as notational aid for inference sections.

Notation	Explanation
t_k	Time in image k.
ρ_{ik}	Local density for cell i during $[t_k, t_{k+1})$.
β_{ik}	Number of cell divisions for cell i during $[t_k, t_{k+1})$.
R	Matrix containing all cell densities in all images.
B	Matrix containing the number of cell divisions of all cells in all images.
$[r_j, r_{j+1})$	Bin j in partition of density range $[0, 1]$.
β_j	Number of cell division in bin j.
$\hat{p}_h(x)$	Kernel density estimate with bandwidth h.
$\overline{p}_h(x)$	Number density made from $\hat{p}_h(x)$.

3.4.1 Maximum Likelihood Estimation Method

A natural assumption to make is that the number of cell divisions in a given time interval, for a given local density is Poisson distributed and that all β_{ik} are independent of each other.

$$\beta_{ik} \sim \text{Poisson}(h(\rho_{ik})\Delta t). \tag{3.9}$$

In this setting it is reasonable to consider, ideally small, ranges of values for the densities and consider cells with local density within that range together. Thus, we partition the total range of local densities into N equal subintervals of width $\frac{1}{N}$

$$0 = r_0 < r_1 < \ldots < r_N = 1$$

and denote the interval $[r_j, r_{j+1})$ for $j \in \{0, 1, \ldots, N-2\}$ and $[r_j, r_{j+1}]$ for $j = N-1$ the j^{th} bin. The j^{th} interval and the j^{th} bin refer to the same interval and can be used interchangeably. A cell i, for any time interval, is said to belong to the j^{th} bin when $\rho_{ik} \in [r_j, r_{j+1})$ (and the closed interval for $j = N-1$). Since the sum of several Poisson distributed random variables is itself a Poisson distributed random variable, we can reformulate (3.9). Let the number of cell divisions β_j in the j^{th} bin during any given time interval be

$$\beta_j \sim \text{Poisson}(h(\rho_j)\Delta t) \tag{3.10}$$

where ρ_j is the local density in the bin. For simplicity and as the bins are ideally small, we choose ρ_j to be the midpoint of the interval. Now with maximum likelihood estimation we can obtain an estimate for $h(\rho_j)\Delta t$.

To refresh our recollection, recall the theory in Section 2.2 and consider the case when $x_1, x_2, \ldots, x_n$ are independent and identically Poisson(λ) distributed random variables. The log-likelihood and its first derivative are then

$$l\left(\lambda|\boldsymbol{x}\right) = \sum_{i=1}^{n} \left(x_i \ln\left(\lambda\right) - \lambda - \ln\left(x_i!\right)\right)$$
$$\frac{\partial}{\partial\lambda} l(\lambda|\boldsymbol{x}) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n$$

which, by setting to zero and solving for λ, gives the MLE

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the Fisher information

$$\mathcal{I}_n\left(\lambda\right) = \mathbb{E}\left[\left(\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n\right)^2\right] = \frac{n}{\lambda}.$$

An $1-\alpha$ confidence level for $\hat{\lambda}$ is then found in

$$\hat{\lambda} \pm \frac{z_{1-\alpha/2}}{\sqrt{n/\hat{\lambda}}}$$

where $z_{1-\alpha/2}$ is the standard normal deviate.

Let us return to our bins and introduce some more notation. Denote n_j the total number of cells belonging to bin j and b_j the total number of cell division occurring

in bin j. That is,

$$n_j = \sum_k^K \sum_i^{N_{t_k}} \mathbb{I}_j(\rho_{ik})$$
$$b_j = \sum_k^K \sum_i^{N_{t_k}} \beta_{ik} \cdot \mathbb{I}_j(\rho_{ik}),$$

where $\mathbb{I}_j(x)$ is the indicator function for the j^{th} bin. The MLE $\hat{\lambda}_j$ for the rate in (3.10) and corresponding 95% confidence interval for $h(\rho_j)\Delta t$ of bin j are then given by

$$\hat{\lambda}_j = \frac{b_j}{n_j}$$
$$\hat{\lambda_j} \pm \frac{z_{0.975}}{\sqrt{n_j/\hat{\lambda}}}.$$

Naturally, if $n_j = 0$ we cannot estimate $\hat{\lambda_j}$. After division with Δt value, the estimates for all bins make up a discretised estimation of the cell division intensity as a function of the local cell density.

Note that the Poisson assumption and using MLE lets us consider the amount of cell divisions at a certain local density given the amount of opportunities at that density. This is something we need to consider in the coming approach.

3.4.1.1 An example

Consider the simple example in Figure 3.6, which spans over the two leftmost images. In all images, notice that the cells are grouped by a dotted box, a dashed box and free cells outside of these boxes. The three boxes groups cells with local density in the same range, with the dotted box having highest range, the free cells have lowest range and the dashed box have range in between. Going from one image to the next, we notice that some cells have divided and given rise to newborn cells, indicated in gray.

Counting the total number of cell divisions within the different regions, that is, counting the number of gray cells within each region, we see that that there is 1 in the dotted region, 3 in the dashed region and 5 in the remaining region. Counting the total number of cells within each region, we consider only the two leftmost images, as the third image serves only to show newborn cells. We then obtain the respective MLE and 95% confidence levels

$$\text{Dotted: } \frac{1}{6} \pm \frac{1.96}{6}$$
$$\text{Dashed: } \frac{3}{7} \pm \frac{1.96 \cdot \sqrt{3}}{7}$$
$$\text{Remaining: } \frac{5}{14} \pm \frac{1.96 \cdot \sqrt{5}}{14}.$$

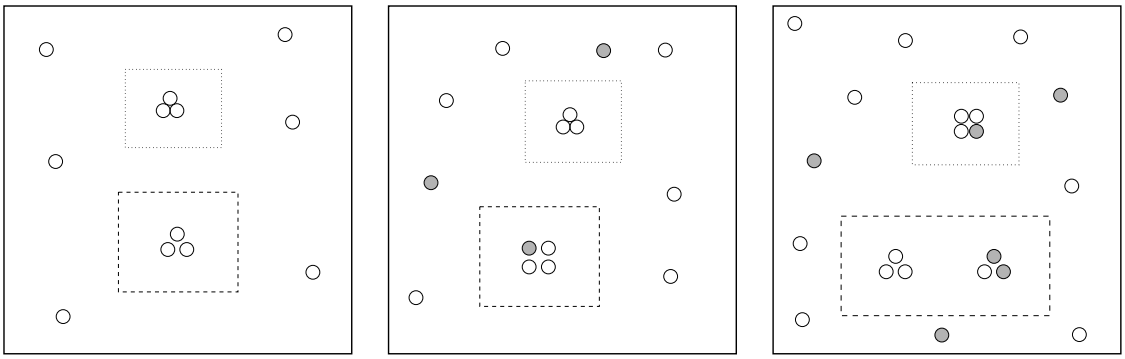


Figure 3.6: A simple example of a simulation spanning over the two leftmost images. Three regions representing different bins of local cells densities are indicated by dotted, dashed and no lines. Newborn cells are indicated in gray.

Of course, there are too few observations for asymptotic normality of the MLE to kick in, but it is a simple example that is easy to parse. Another way of viewing the inference problem is shown in Figure 3.7, which shows two set of histograms, one of R and one the elements of R, for which the corresponding elements in B are non-zero. Division of the counts of the latter by the counts of the former yields the MLE, which can be visualised in a bar graph, see the rightmost graph in Figure 3.7.

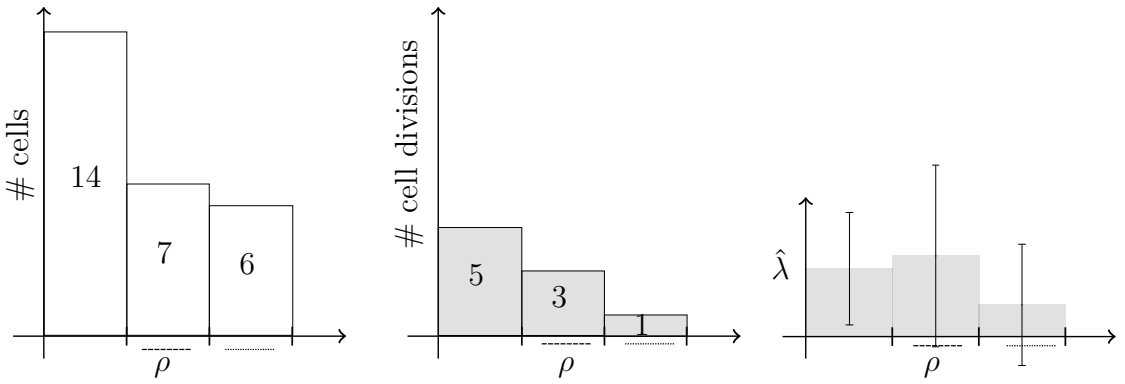


Figure 3.7: Histograms of the simple example presented in Figure 3.6 and a corresponding bar chart of the MLEs of their corresponding bins and 95% confidence interval.

3.4.2 Kernel Density Estimation Method

Kernel density estimation is mainly used for, as the name implies, to estimate densities. This is, as shown in Section 2.3, done through placing a kernel on each data point and then average over all kernels. If one did not average, one would obtain a number density, a quantity that describes the concentration of countable objects per unit of measure. Using number densities, we may look at the situation through the same divisions-per-opportunity ratio glasses as in the previous approach through a ratio of number densities. A matter that needs attending to is how to measure the certainty for said ratio. Let us first formalise the number densities and their confidence bands and thereafter return to this matter.

Recall matrices R and B and their elements, ρ_{ik} and β_{ik}, respectively, as defined in Section 3.4. Using the debiased kernel $M_\tau(x)$ defined in (2.6) with $\tau = 1$ and Gaussian underlying kernel as defined in Table 2.1, consider the following KDEs

$$\hat{p}_{h_\rho}(x) = \frac{1}{nh_\rho} \sum_{k=1}^{K} \sum_{i=1}^{N_{t_k}} M_\tau \left(\frac{x - \rho_{ik}}{h_\rho} \right)$$

and

$$\hat{p}_{h_\beta}(x) = \frac{1}{bh_\beta} \sum_{k=1}^{K} \sum_{i=1}^{N_{t_k}} M_\tau \left(\frac{x - \rho_{ik}}{h_\beta} \right) \beta_{ik}$$

where h_ρ and h_β are the respective bandwidths, chosen according to (2.5), and

$$n = \sum_{k}^{K} N_{t_k}$$

$$b = \sum_{k=1}^{K} \sum_{i=1}^{N_{t_k}} \beta_{ij}.$$

Now define the following

$$\overline{p}_{h_\rho}(x) = n \cdot \hat{p}_{h_\rho}(x) \tag{3.11}$$

$$\overline{p}_{h_\beta}(x) = b \cdot \hat{p}_{h_\beta}(x). \tag{3.12}$$

We can interpret (3.11) as the number density of all local densities and (3.12) as the number density of local densities at cell divisions.

To obtain a confidence band for the number densities, we turn to the process presented in Figure 2.2. In this process, exchange the debiased KDE with number densities (3.11) and (3.12). That is, compute $\overline{p}_{h_\rho}(x)$ and $\overline{p}_{h_\beta}(x)$, bootstrap the original sample m times and compute

$$\overline{p}^{*(1)}_{h_\rho}(x), \ldots, \overline{p}^{*(m)}_{h_\rho}(x)$$

$$\overline{p}^{*(1)}_{h_\beta}(x), \ldots, \overline{p}^{*(m)}_{h_\beta}(x)$$

from which we compute 95% confidence level bands as described. In this project we let $m = 50$.

Returning to the ratio we are interested in, which can now be defined as

$$\frac{\overline{p}_{h_\beta}(x)}{\overline{p}_{h_\rho}(x)} \tag{3.13}$$

for all $x \in [0, 1]$ where the denominator is not zero, where, again, the ratio cannot be estimated. With the Gaussian kernel it ideally never is, but using the debiased kernel there is a slight possibility. After division with Δt, this quotient makes up a continuous estimation of the cell division intensity as a function of the local cell density.

With R and B approximately representing the simple example presented in Figure 3.6 and using optimal bandwidth, the number densities and their resulting ratio can be seen in Figure 3.8. It is a basic example, and as with the histogram, asymptotics has not kicked in.

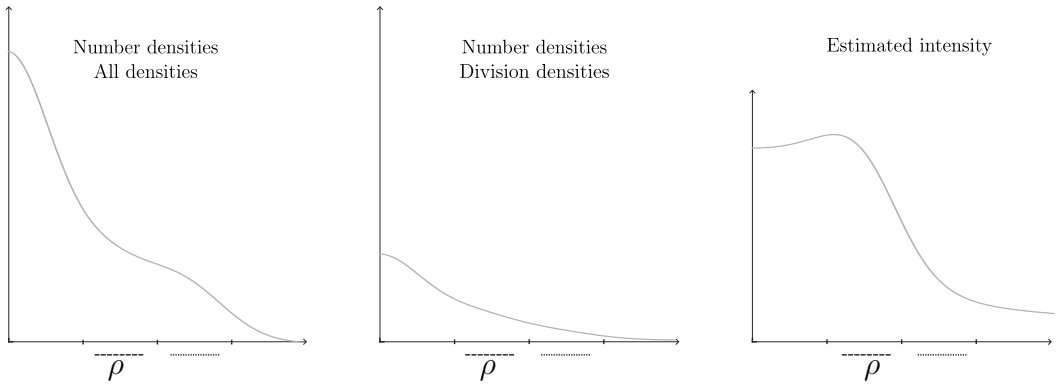


Figure 3.8: Number densities and their ratio of the simple example presented in Figure 3.6.

Now consider how to obtain a measure of certainty for the quotient. Through the bootstrapping procedure we obtain a sample of size m of $\overline{p}_{h_\rho(x)}$ and $\overline{p}_{h_\beta(x)}$ for each point x considered. In this project, we assume that $\overline{p}_{h_\rho(x)}$ and $\overline{p}_{h_\beta(x)}$ are normally distributed as visual inspection does not immediately dismiss that. However, we will not provide a proof of normality, or the potential error introduced. Considering the mean of each sample then, one can invoke the theory presented in Section 2.4, and obtain a pointwise confidence interval. Let A_x and B_x denote sample means

and $S^2_{A_x}$ and $S^2_{B_x}$ denote sample variances at x. That is,

$$A_x = \frac{1}{m}\sum_{i=1}^{m} \overline{p}^{*(i)}_{h_\beta}(x)$$
$$B_x = \frac{1}{m}\sum_{i=1}^{m} \overline{p}^{*(i)}_{h_\rho}(x)$$
$$S^2_{A_x} = \frac{1}{m-1}\sum_{i=1}^{m} (\overline{p}^{*(i)}_{h_\beta}(x) - A_x)^2$$
$$S^2_{B_x} = \frac{1}{m-1}\sum_{i=1}^{m} (\overline{p}^{*(i)}_{h_\rho}(x) - B_x)^2.$$

The 95% confidence level interval for $Q_x = A_x/B_x$ is then computed as in (2.8). As a rule of thumb, when the degrees of freedom is more than 30, one can exchange the Student's t statistic with the same from the standard normal distribution. In this case, the degrees of freedom is $2m - 2$ with $m = 50$ and thus the standard normal deviate will be used in (2.8).

3.5 Evaluation

This section presents the metric with which the approaches are evaluated. Let $h(\rho)$ denote the true intensity and $\hat{h}(\rho)$ denote the estimated intensity obtained by each model.

For the approach presented in Section 3.4.1 we would like to use the sum of squared error

$$SSE = \sum_{j=1}^{N} \left(\hat{h}(\rho_j) - h(\rho_j)\right)^2 \tag{3.14}$$

where ρ_j denote the midpoint of each interval in the partition of the interval $[0, 1]$. This presents, however, a problem when it comes to bins for which there is no estimate. We will therefore give a replacement value for such a bin with the value of a linear interpolation between the midpoints of the nearest two bins with estimates, evaluated at the midpoint of the bin with no estimate. If the bin without estimate is one or more of the rightmost bins the interpolation will be between the midpoint of the first bin with estimate to the left and the value zero at local density $\rho = 1$, which it necessarily must be in this model. If the bin without estimate is one or more of the leftmost bins, we extend the linear interpolation between the first two bins with estimate until it meets the y-axis. A figure illustrating these amendments can be seen in Figure 3.9.

For the approach presented in Section 3.4.2 we will use the integrated squared error

$$ISE = \int_0^1 \left(\hat{h}(\rho) - h(\rho)\right)^2 d\rho. \tag{3.15}$$

Each experiment has been repeated $L = 3$ times and so the error will be averaged over the experiments.

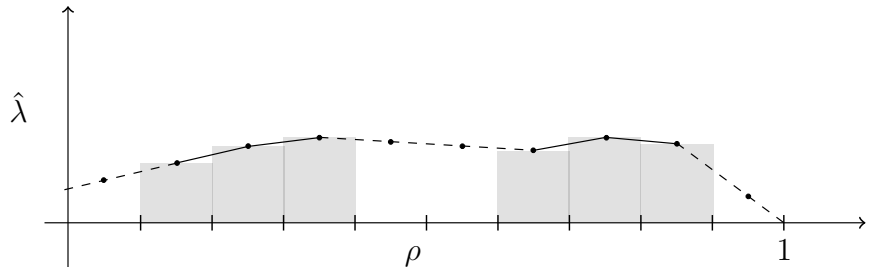


Figure 3.9: Linear interpolation for missing estimates.

4

Results

In the following we present the results from applying the two described methods to synthetic and experimental data. The section pertaining to the synthetic data is structured around the different growth models and the experimental data consist of two datasets, which is the focus the last section.

4.1 Synthetic data

This section contains results from applying the estimation methods to the synthetic data. We present the estimation errors and the best performing estimates from both methods for each growth model.

4.1.1 Exponential growth model

Recall that for exponential growth, the per capita growth rate is a constant. That is, the theoretical cell division intensity is

$$h(\rho) = 0.05$$

for $\rho \in [0, 1]$.

In Figure 4.1, we find figures with plots of the estimation errors as functions of N_0, Δt, and T, respectively. Looking at top left and bottom figures, it shows that as N_0 and T are increasing, the errors decrease for both methods. There is a slight increase, however, in error as we go from $N_0 = 500$ to $N_0 = 1000$. With increasing Δt, the error for the MLE method remains roughly constant, whereas the KDE method decreases, in stark contrast to the expected behaviour. Following, in Figure 4.2, we present the best performing estimate for both methods and leave the worst performing estimates in Appendix B.

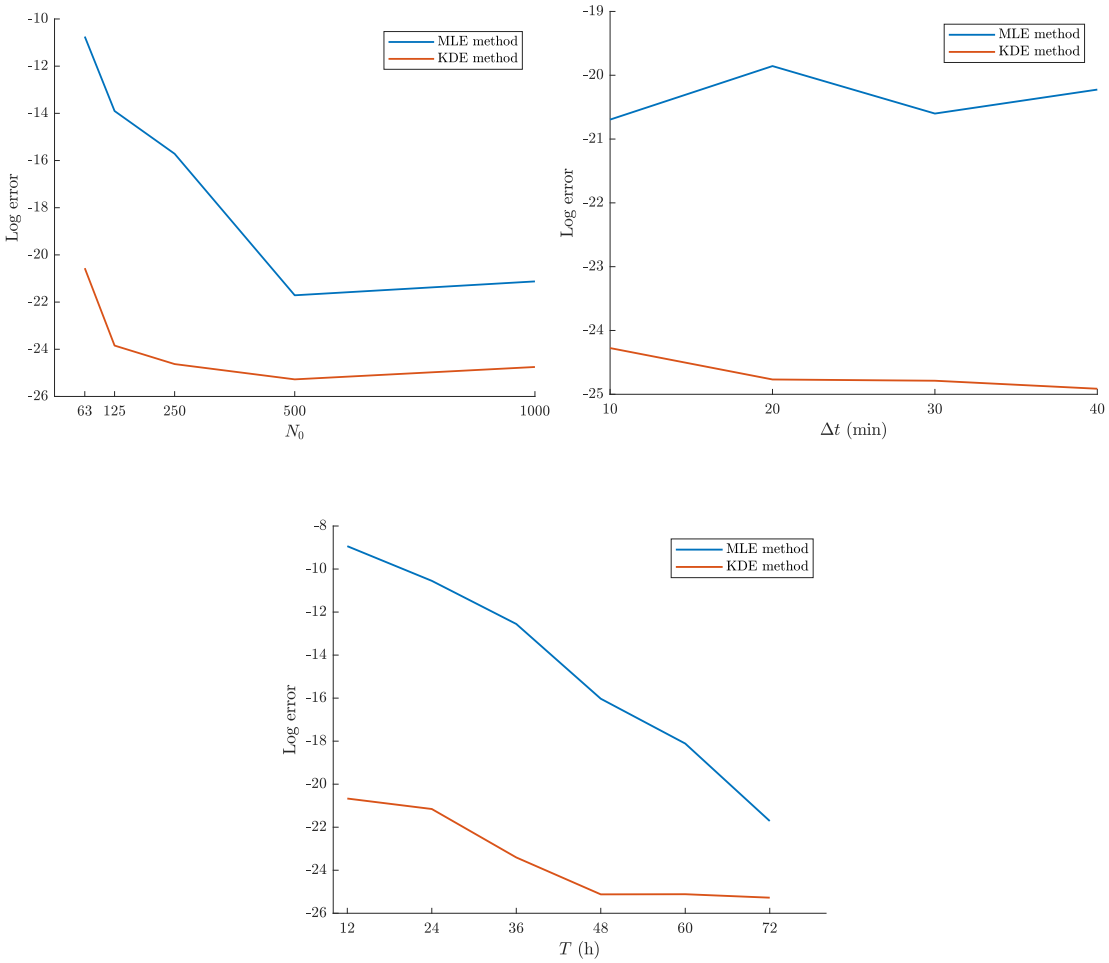


Figure 4.1: SSE (blue) as defined in (3.14) and ISE (red) as defined in (3.15), as functions of N_0, Δt and T for exponential growth model.

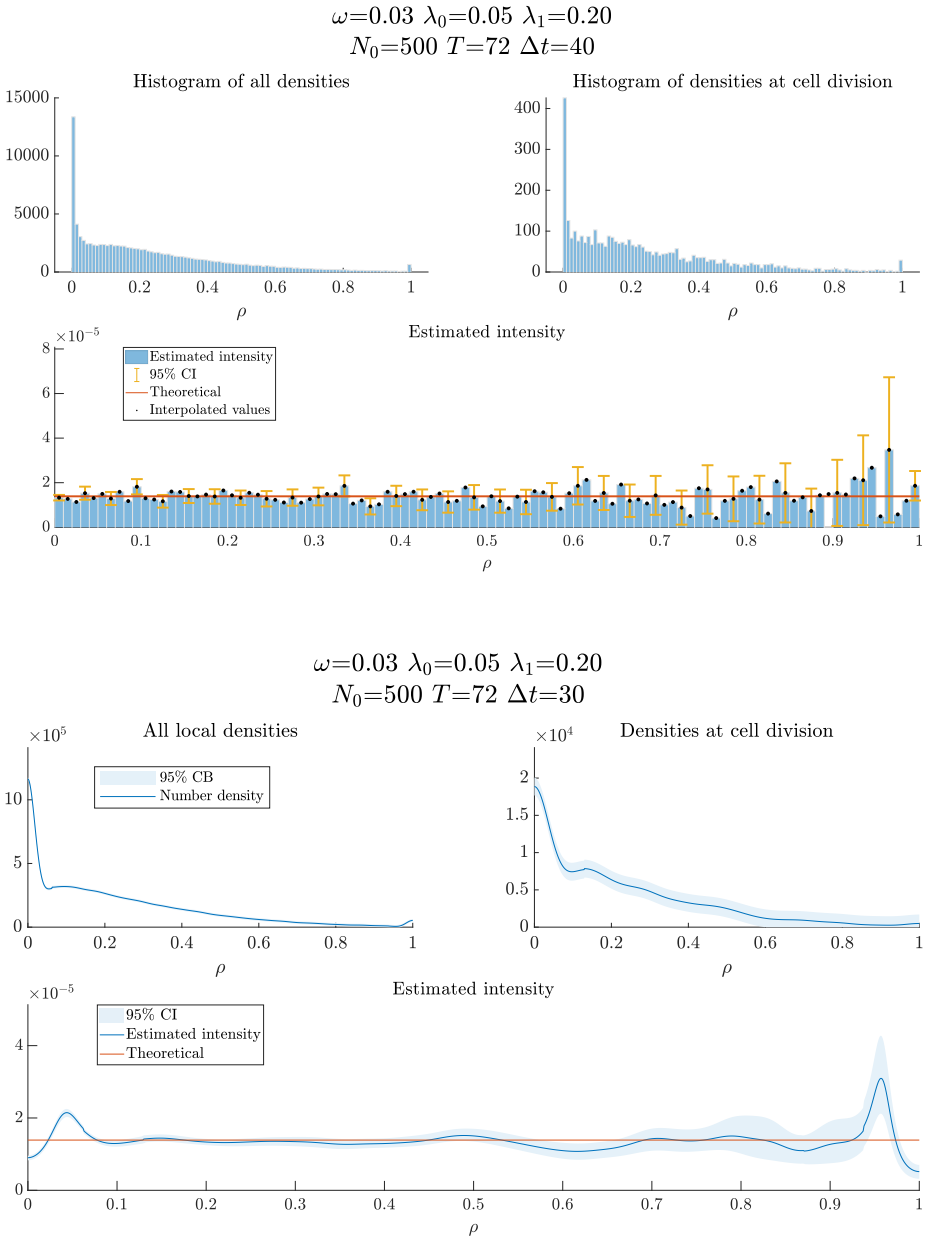


Figure 4.2: The best performing estimates of $h(\rho)$ for the exponential growth model.

4.1.2 Logistic growth model

Recall that logistic growth is a function of local density such that when the population reaches its carrying capacity, the growth rate decreases to zero. In this thesis, for the logistical growth, the theoretical cell division intensity is

$$h(\rho) = 0.05 \cdot (1 - \rho)$$

for $\rho \in [0, 1]$.

Figure 4.3 contain the estimation errors as functions of N_0, Δt and T, respectively. A new oddity is found in the bottom figure, where both methods perform better at $T = 12$ rather than $T = 24$. Once again, we present the best performing estimates for both methods in Figure 4.4 and leave the worst performing estimates in Appendix B.

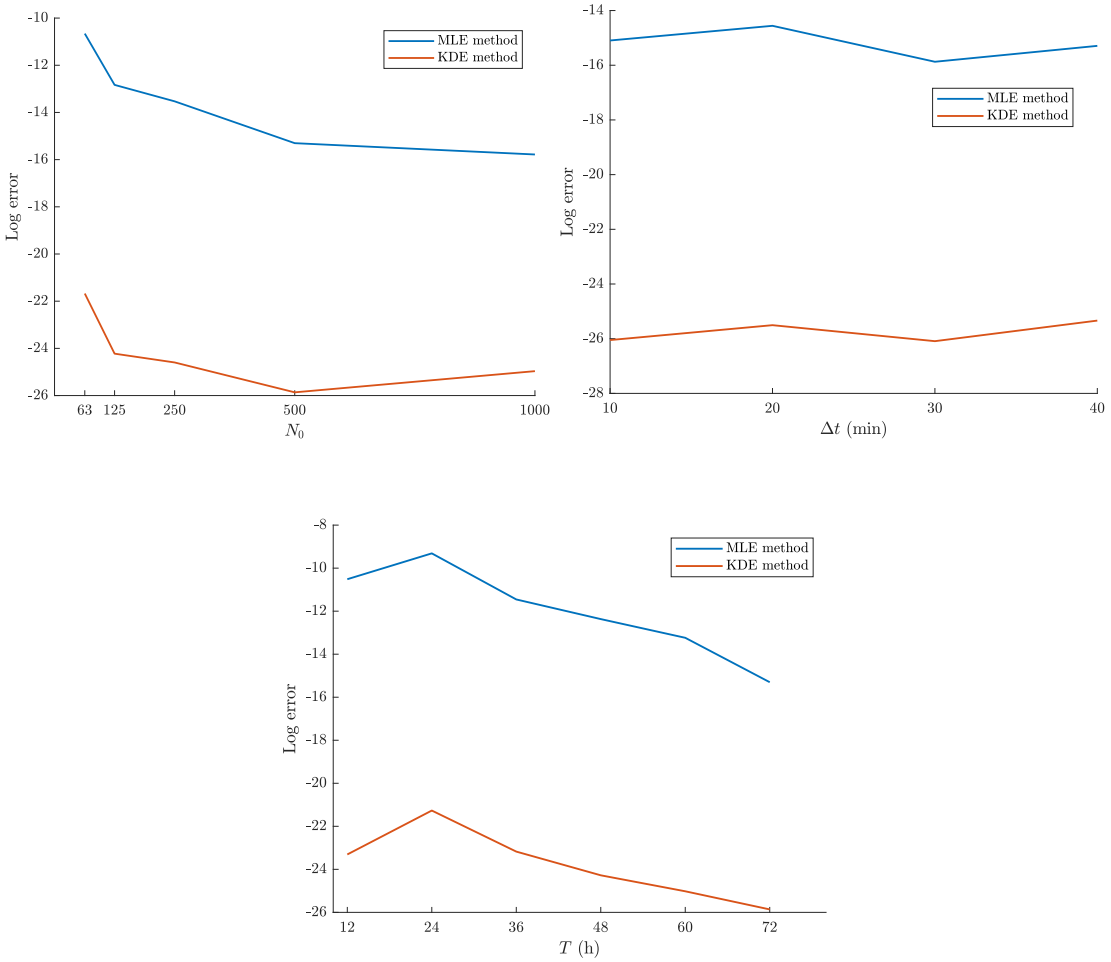


Figure 4.3: SSE (blue) as defined in (3.14) and ISE (red) as defined in (3.15), as functions of N_0, Δt and T for logistic growth model.

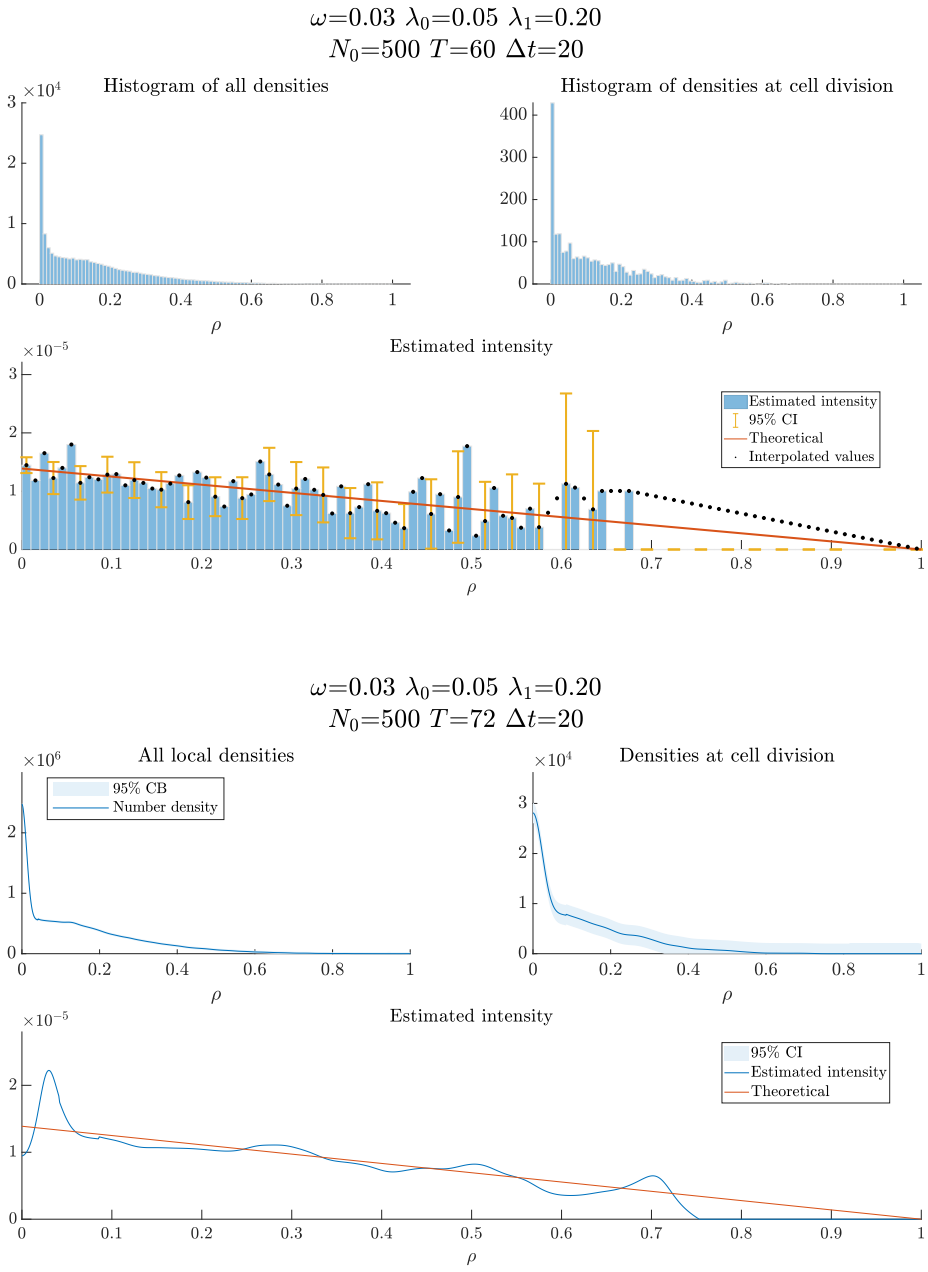


Figure 4.4: The best performing estimates of $h(\rho)$ for the logistic growth model.

4.1.3 Allee effect growth model

Recall that for a population exhibiting an Allee effect, the growth rate for low densities is increasing and as the density increases, the growth rate eventually reaches a peak and then decreases. For both categories of Allee effects, the theoretical cell division intensity given by

$$h(\rho) = (\lambda_0 + \lambda_1 \rho) \cdot (1 - \rho)$$

for $\rho \in [0, 1]$ with values for either category being chosen from Table 3.3.

The following two sections will first present the estimation errors for both methods and thereafter one of the best estimates from each method. A more complete presentation of estimates can be found in Appendix B.

4.1.3.1 Weak Allee effect

The estimation errors of both methods are found in Figure 4.5. In the top left figure, note that the error for the KDE method, as a functions of N_0, behave similarly as in the exponential and logistic case whereas the MLE method has had some kind of hiccup at $N_0 = 500$. With increasing Δt, see the top right figure, the error for the MLE method decreases until $\Delta t = 30$ to then increase a little whereas the error for the KDE method increases. The bottom figure shows that the error decreases for both methods as T increases, which is the expected behaviour.

Figure 4.6 presents one of the best performing estimates for each estimation method, both of which are obtained from experiments with a high number cells initially. Note the high concentration of cells with local density near 1, indicating a crowded situation for many cells. That is, cells with many other cells in its vicinity. The estimate are more accurate for values of ρ with more data, in both cases for $\rho > 0.5$ approximately. For the KDE method, observe the two ”horns” near the boundaries. These are present when the concentration of cells near the boundaries are considerably higher or lower than the concentration levels inside.

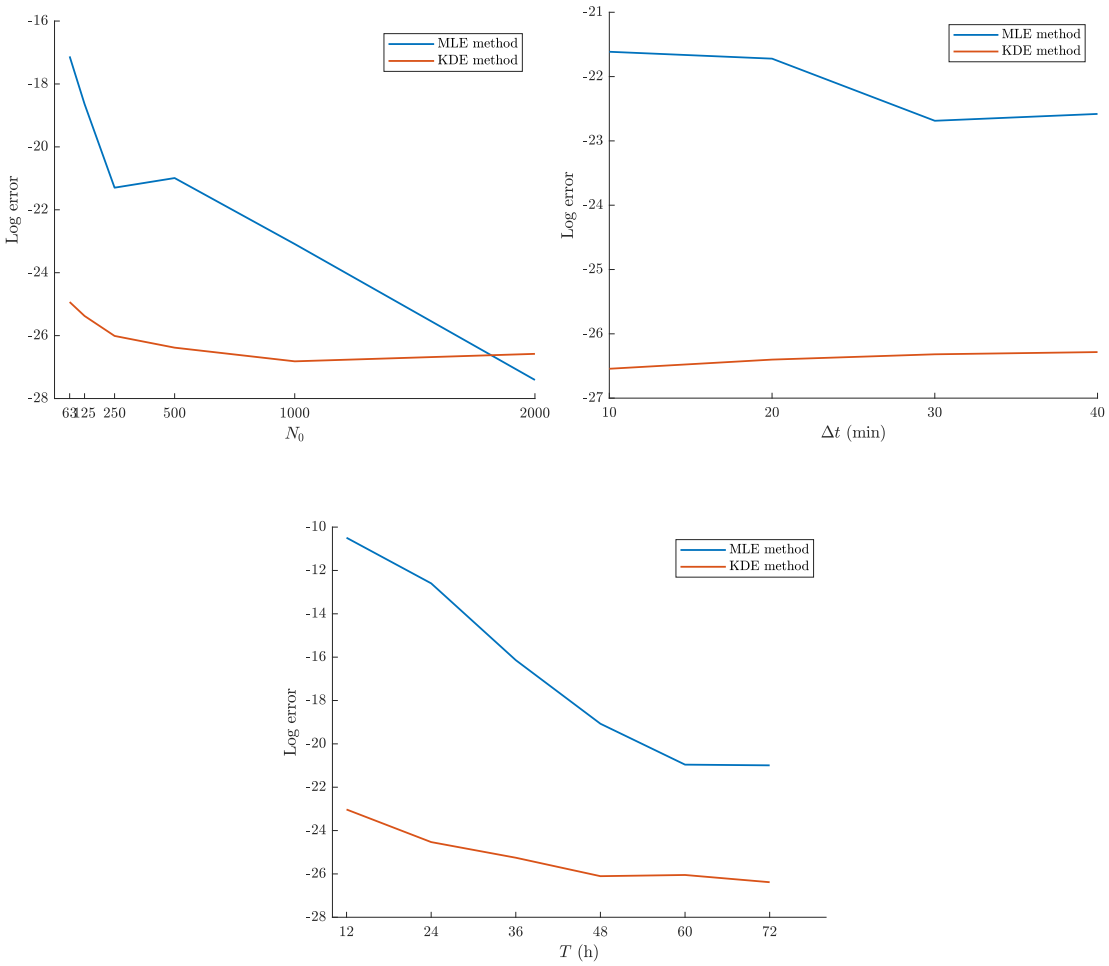


Figure 4.5: SSE (blue) and ISE (red) as a functions of N_0, Δt and T for weak Allee effect growth model.

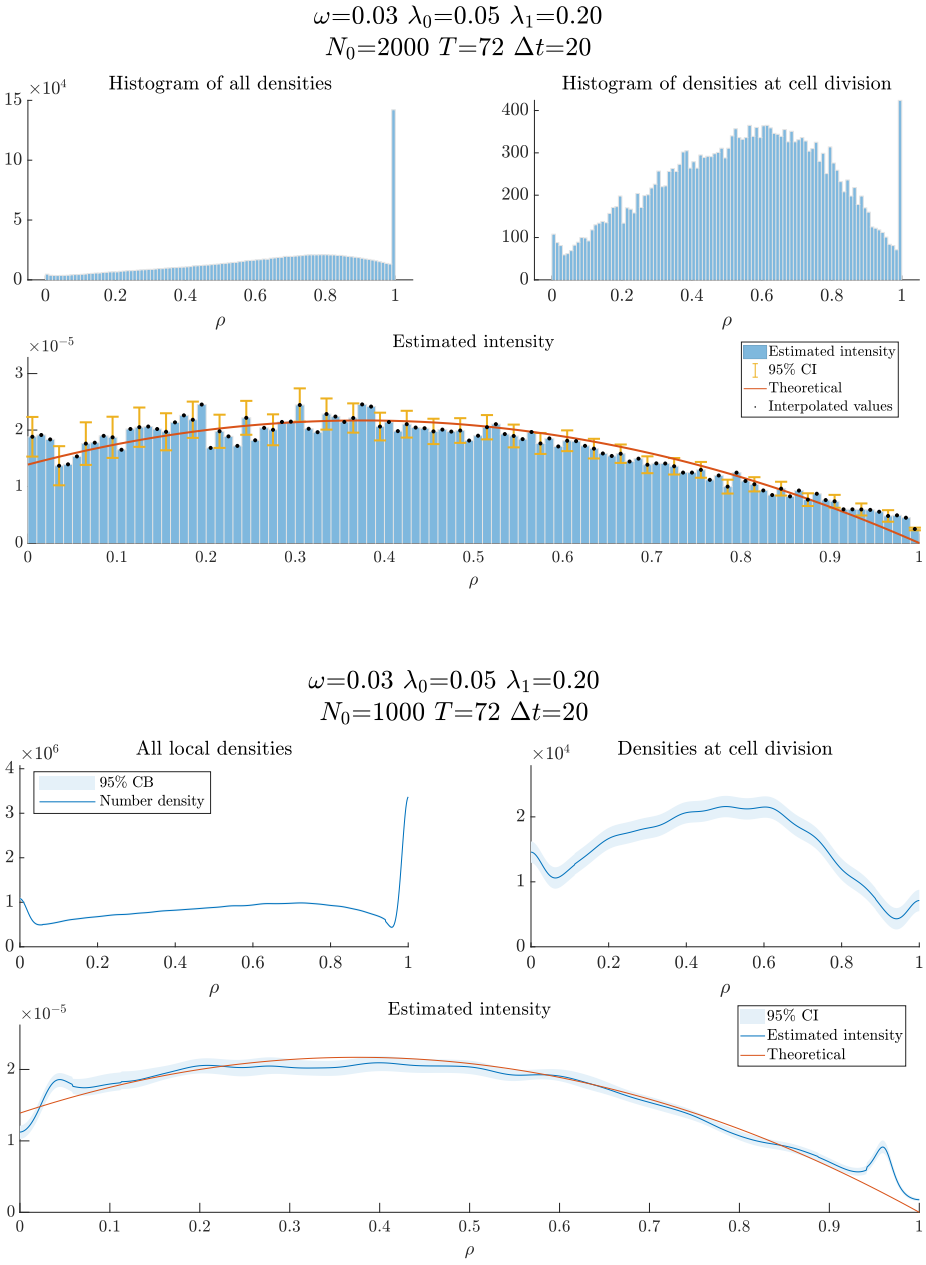


Figure 4.6: The best performing estimates of $h(\rho)$ for the weak Allee effect growth model.

4.1.3.2 Strong Allee effect

The estimation errors for the KDE method are found in Figure 4.7. In the top left figure, note the general decrease in error and then note the temporary increase in KDE method error from $N_0 = 63$ and $N_0 = 125$ and that those errors are larger than the MLE method error for the same N_0. For the case when $N_0 = 63$, the population in one out of the three experiments went extinct and for the case when $N_0 = 125$, the population in two out of the three experiments went extinct. With increasing Δt, see the top right figure, the errors are slightly increasing. The bottom figure shows that the error decreases for both methods as T increases.

Figure 4.8 presents one of the best performing estimates for each estimation method, both obtained from experiments with a high number cells initially. Here there is a relatively high concentration of cells with local density near 0, indicating a situation with few and far neighbours to those cells. The estimates are more accurate for values of ρ with more data, in both cases for $\rho < 0.5$ approximately. For the KDE method, again observe the two "horns" near the boundaries.

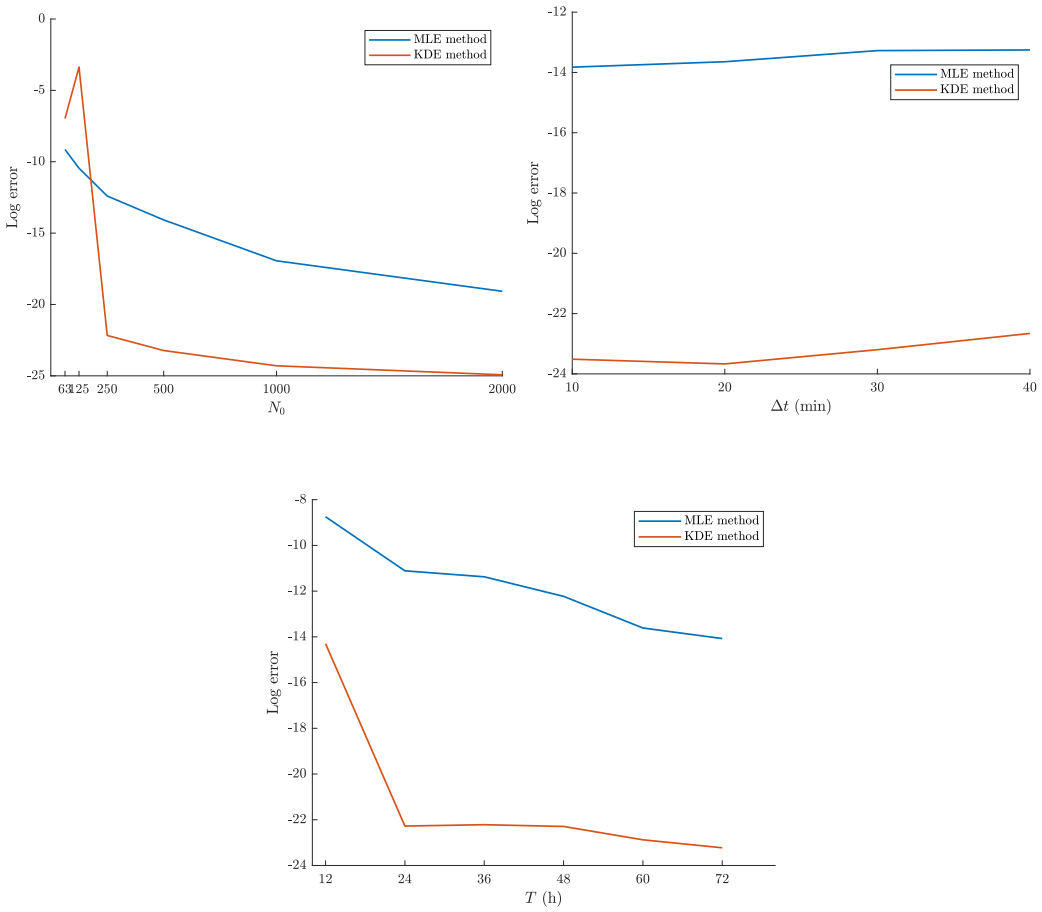


Figure 4.7: SSE (blue) and ISE (red) as a functions of N_0, Δt and T for strong Allee effect growth model.

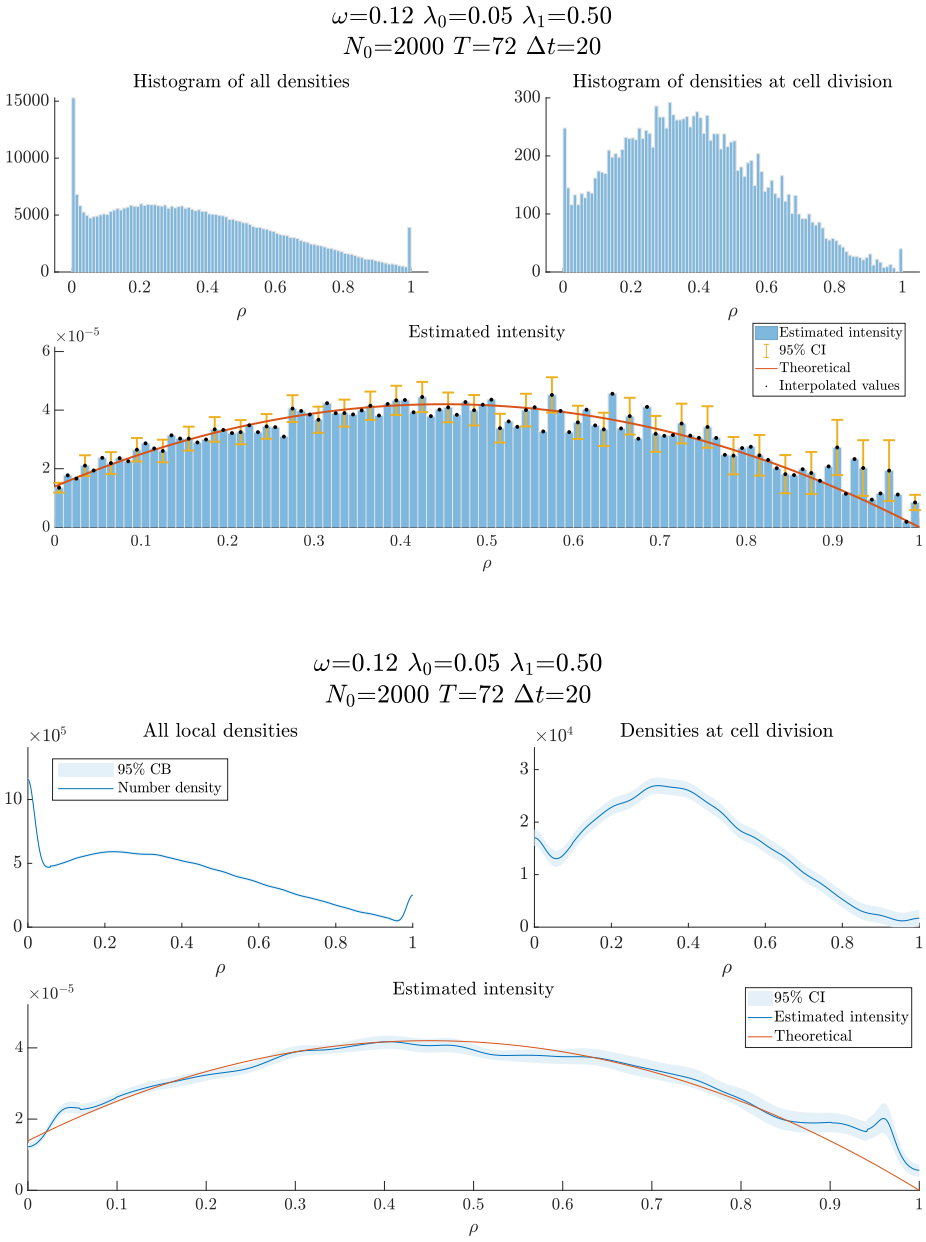


Figure 4.8: The best performing estimates of $h(\rho)$ for the weak Allee effect growth model.

4.2 Experimental data

Both experiments lasted 45 hours and images were taken at 30-minute intervals. The first experiment was initiated with 43 cells whereas the second experiment was initiated with 125 cells according to the reference annotation.

In Figure 4.9 the estimates from each method based on the first experiment are presented. Inspecting the histograms and number densities, there is a slight increase in concentration of cell with local density near 1 but the main body of the data is in the inner part and has a peak at about $\rho = 0.2$.

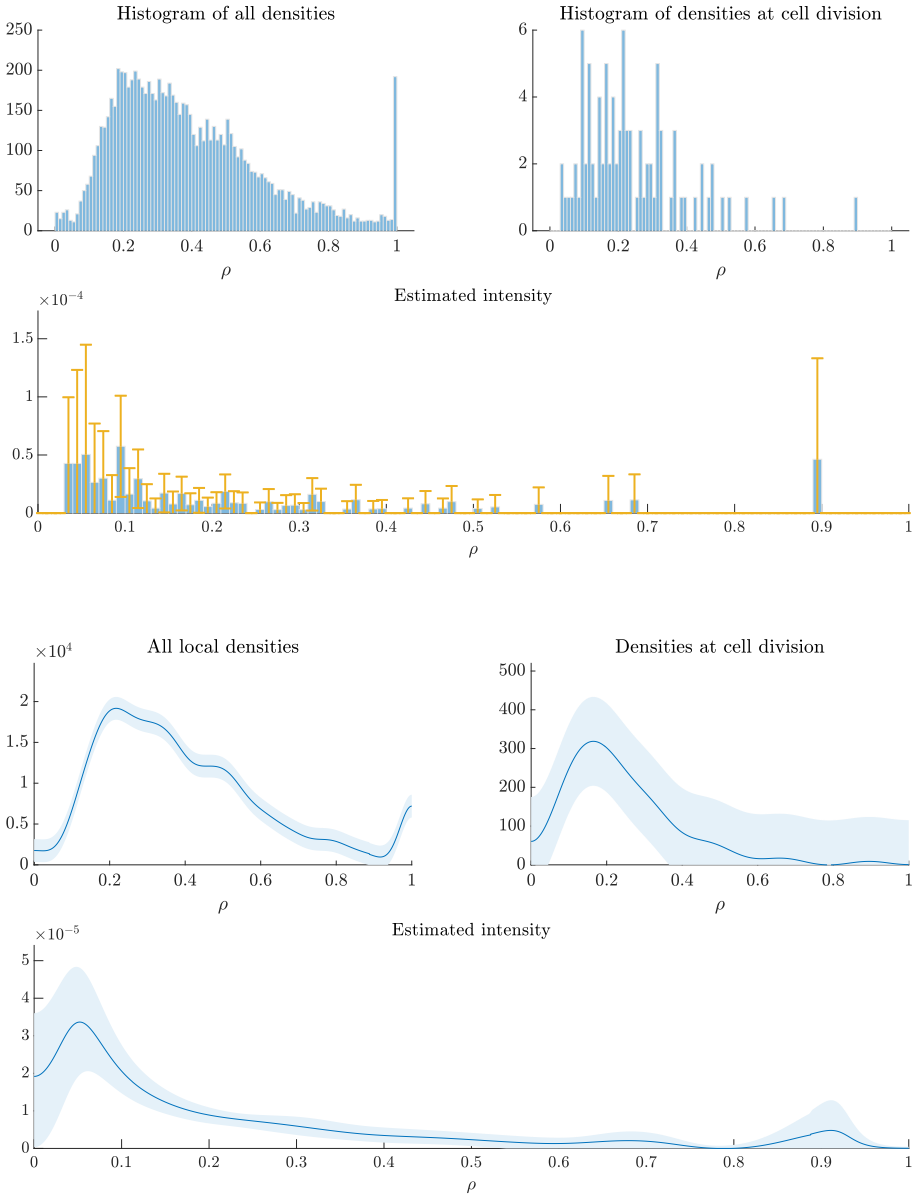


Figure 4.9: Estimates of $h(\rho)$ based on the first dataset.

In the second experiment, presented in Figure 4.10, the majority of of cells are have a local density near 1. Considering the KDE method estimate this causes the horn

at the right, with its peak at 0.4187 falling outside of the figure. As the scale of the figures do not allow it to be shown, the relative steep turn towards the origin gives the little horn to the left.

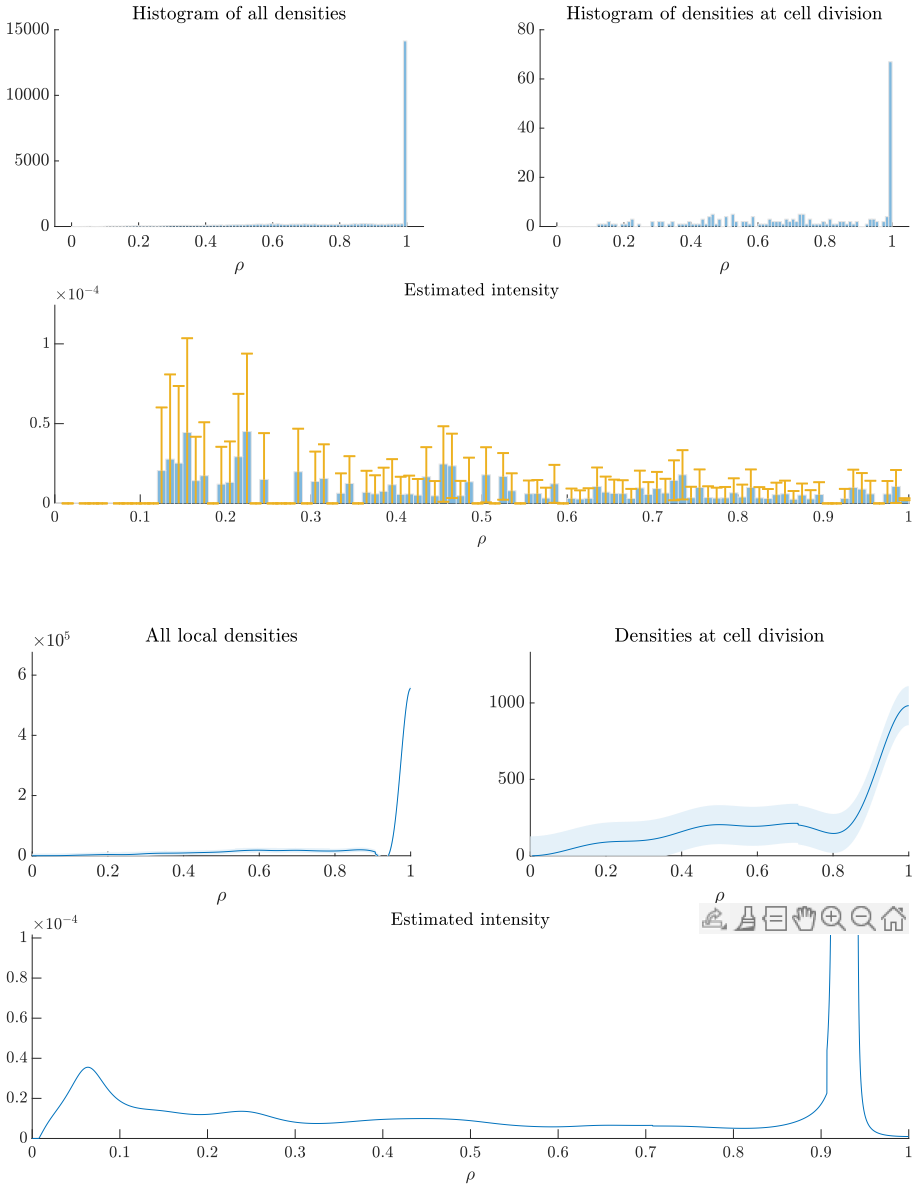


Figure 4.10: Estimates of $h(\rho)$ based on second dataset.

The choice of bin width for the MLE method was in this thesis set to 0.01, based in a required resolution and then remained so. When a histogram has the look of a comb with broken teeth, as seen in Figure 4.9, it is usually a sign of too small bin width. Therefore we also present some estimates from the MLE method with larger bin width in Figures 4.11.

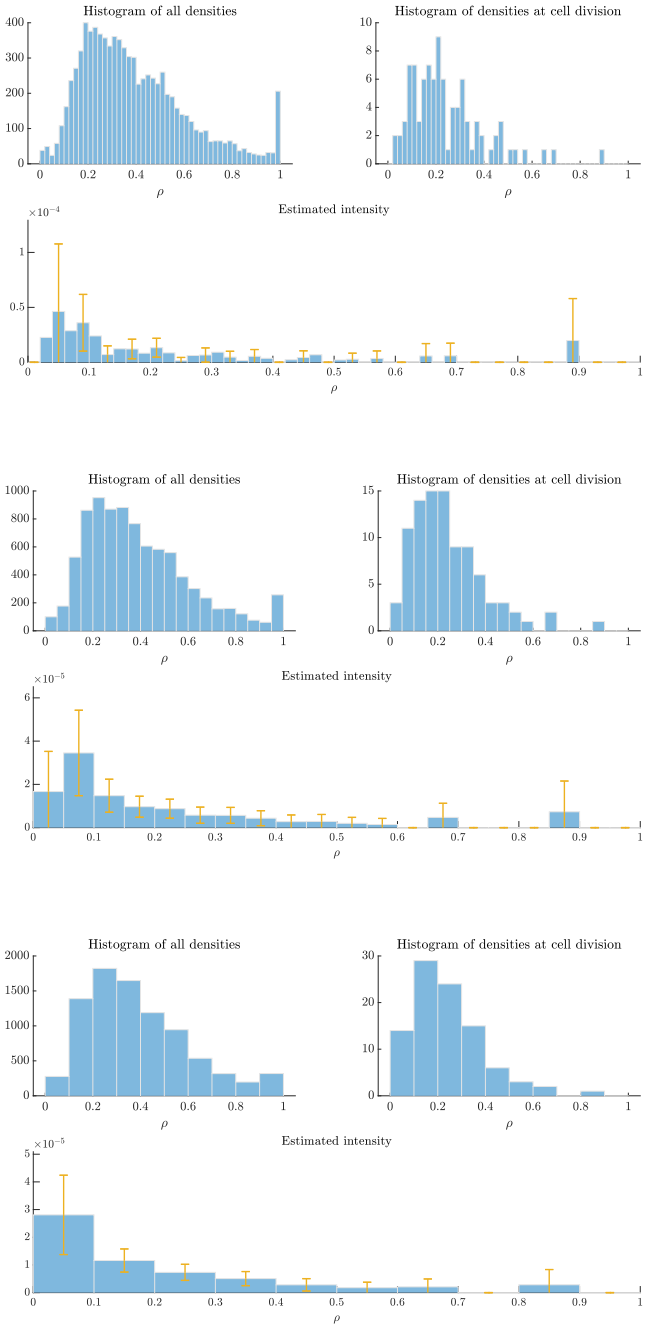


Figure 4.11: Estimates of $h(\rho)$ based on the first dataset using the MLE method and bin width 0.02, 0.05 and 0,1.

5

Discussion

This section will contain discussions of the material, methods and results. Various advantages and disadvantages of the data the two methods will be discussed and suggestions for improvement will be suggested.

5.1 Cell Population Dynamics Model

The model formulated was designed to be general and the focus was on incorporating the growth models, which was the most relevant aspect to this thesis. Another approach for modelling proliferation and death considered was through the Doob-Gillespie algorithm, a Monte Carlo method often used to simulate biochemical systems of reaction. There was not, however, a reasonable method of incorporating the different growth models.

Generating data was computationally heavy as one had to calculate the pairwise distances between all alive cells at each *in silico* second. Depending on parameters an experiment could take between some hours to some days. The majority of cells are, however, so far apart from each other that they do not affect each other with any forces. There is room for improvement by keeping some sort of neighbouring list that can be updated with some interval and thus trade larger bookkeeping for shorter run time.

5.2 Maximum Likelihood estimation method

Under certain conditions, the MLE method is able to generally capture the dependency of local cell density on the rate of cell division. Such conditions can be translated to conditions that allows for ”much data to be generated”, such as more cells being present and for longer time. When that is the case, the figures from the synthetic data show that the estimate and its confidence interval, visually speaking, can be rather accurate. With less data the estimate decreases in precision but accuracy seems to remain.

A weakness of the density domain in this method is that, mainly in low data situations, there may be no estimate at all for portions density domain. This problem was circumvented by interpolation of values based on our beliefs of the cell division rate. Other possible solutions to this problem could have been to add a punishment

term in the error or to use adaptive bin widths to insure estimates for the entire domain of ρ.

The parameters of the experimental data, especially N_0 and T, when compared with the performance of the method on the synthetic data do not seem to be suitable enough, and the estimate resembles the worst performing estimates of the synthetic data. Initially, the same bin width as for the synthetic data was used, and then larger widths was used due to the broken comb look of the first experiment. The more increased width, the more the estimates transformed into something similar of logistic growth.

This leads us to a second weakness of this approach, which is a weakness of histograms in general. The choice of bin widths can alter the face of the distribution in ways that are not necessarily descriptive of the underlying distribution. This was illustrated in Figure 4.11, as we increased the bin with for the first experiment.

5.3 Kernel density estimation method

The KDE method, also under certain conditions, is able to generally capture the dependency of local cell density. However, as seen in most estimates, it is sensitive to changes in slopes in the number densities, especially as densities accumulate at the boundaries of the domain of ρ. A possible explanation may be that the two number densities use different optimal bandwidths, which is something that could be further explored. Another possible explanations may be that it is an effect of boundary bias, as it is of a lower order than the bias in the inner parts, yet this seems unlikely. There are ”horns” in the inner domain of ρ (see Figure B.8) and there are boundaries without horns (see B.4). Adaptive bandwidth, that is, as a function ρ can also be implemented for this method and may yield improvements.

There is no guarantee that one obtains a confidence interval for the KDE method estimate through the Fieller theorem, a significant disadvantage. An investigation of the normality should ideally be done for any further endeavours with this approach. Further, one can observe some precise confidence intervals for the estimates, yet the accuracy of the estimate may be very low. It is likely a consequence of the issue discussed in the previous paragraph. The confidence intervals, for estimates of both methods, are smaller in regions with much data, and perhaps suggests there is a systematic error in the creation of the estimate.

5.4 Experimental and synthetic data

There are various differences between the experimental and synthetic data. As mentioned earlier, the initial number of cells and duration of the experiment are two parameters that may limit the ”generation of data”. Another such difference is the space in which the cell reside. Considering the last images shown in Figure 3.1, especially the second experiment, the area is saturated with cells. In comparison with Figure 3.5, the experimental data simply does not have the space to generate

the same amount of data as in the synthetic case.

The two datasets considered are, as mentioned, perhaps not suitable enough for the two methods presented in this thesis. There are, naturally, other datasets to use. In fact, this thesis was initially intended to use microscopy images of cells derived from patients with glioblastoma, an aggressive form of brain cancer, obtained from the biobank Human Glioblastoma Cell Culture (HGCC) [22]. The synthetic data was created to align with the experiments that generated the glioblastoma images. The underlying reason as to the change of experimental data was that we were unable to obtain satisfactory segmentation and tracking of the glioblastoma cells. The data provided by the CTC is annotated by experts and as such it was simply a matter of extracting the relevant data. If the segmentation and tracking issues are solved, it would be of interest to apply both methods to datasets obtained by the HGCC.

An advantage with the chosen experimental data is that HeLa cells are, as seen, quite round in their shape, matching the assumption in our cell based model, whereas the glioblastoma cells are not as regular, presenting its own set of issues.

5.5 Societal and ethical aspects

There are two sources of data in this thesis: simulated and experimental. The former is a commendable resource if done efficiently and accurately. A further commendable aspect of the former source, although not entirely related to this thesis, is that with ever increasing computational capabilities and improved models, *in silico* modelling could become an alternative to animal testing, lowering the amount of suffering in the name of science.

The second source, microscopic images of HeLa cells, has a sordid ethical history which demands further inspection and reflection. The name ”HeLa” originates from the initial letters of Henrietta Lacks, the African American woman whose cervical cancer is the source of all HeLa cells. She was a 31-year old mother of five and passed away in 1951, merely a few months after diagnosis and treatment at the John Hopkins hospital in Baltimore. As previously mentioned, HeLa cells are the first immortalised cell line and has had immense biomedical consequences. They have been instrumental for researchers to study cancer and genetics, and aided the creation of the polio vaccine and drugs for herpes, leukaemia, influenza, haemophilia and Parkinson’s disease among many more examples of scientific successes. Henrietta was neither informed, nor gave her consent for the collection and, thereafter, use of her biological tissue. The rules in place at that time did not require consent, as discarded biological tissue was the property of the treating doctor or medical institution. The cells were initially donated to the benefit of research, but were later commercialised. The Lacks family was not informed about the existence of the cells until almost 25 years after Henrietta’s death, when researchers needed their DNA solve a major contamination issue. Not only was the fact that Henrietta’s cells were still alive a shock to the family, but the reason for collecting their DNA was not made clear to them, as researchers did not try to bridge the gap in education through clear communication. Further grievance for the family was the fact that biological

material stemming from Henrietta was bought and sold, yet no compensation was given to the family, many members of which lived in poverty and had no medical insurance. [20]

In 2010, Rebecca Skloot published "The immortal life of Henrietta Lacks" [20] after decades of research and contact with the Lacks family. It received widespread attention and incited a public discussion on informed consent. Following that, in 2013, a German team of researchers published the complete genome sequence of a strain of HeLa cells online (which is often required by funding sources) without any contact with the Lacks family. This was not against any laws or rules, but as the identity behind the acronym was widely known, it was highly criticised and launched a further discussion on informed consent and privacy (due to genetic information being viewed as probabilistic medical information about living Lacks family members) resulting in the publication being taken down. In relation with this, the National Institutes of Health launched an access-controlled database, where NIH-funded researchers are expected to place genomic sequence data obtained from HeLa cell, and access to which is given on a case-by-case level by committee containing two Lacks family members. [23]

The history of Henrietta Lacks culminates in two ethical aspects highly present in biomedical research: informed consent and privacy. The practises regarding collecting and using biological tissue during the last 70 years have changed, and should be continuously evaluated as science progresses. The evaluation of this thesis depends on the study that generated the data, which was published during 2010 and investigated genes related to cell division [1]. The phenotypic profiling used a short interfering RNA library designed to suppress the expression of genes, and thus affecting various protein levels. Around 21,000 genes were targeted, after which images of affected cells were taken and protein levels were estimated, generating a large set of data. The dataset was, after its use in the study, published online to be of use for other studies and is still available. As the data does not disclose any genomic information, the study avoids the criticised misstep regarding privacy made by the German research team. The widespread use of HeLa cells in research makes informed consent highly impractical but not something to ignore. With the fact that the study was performed before the two highly publicised discussions, any critique directed at the study might be lessened. No rules or norms were transgressed, but then again, so weren't any back in 1951. Turning to Henrietta's children, her son Sonny told Skloot "[...] I'm proud of my mother and what she done for science. I just hope Hopkins and some of the other folks who benefited off her cells will do something to honor her and make right with the family" [20]. To honor Henrietta, we continue to tell Henrietta's story and to make right with the family we acknowledge the historic wrongdoings and aim to continue the discussion on informed consent and privacy.

6

Conclusion

The aim of this thesis was to quantify the impact of local cell density on the rate of cell division among cancer cell populations in order to investigate the presence of an Allee effect. To this end, a cell based model for population dynamics was formulated and implemented, using different growth models, to generate synthetic data. Experimental data in the form of microscopic images with reference annotations was obtained from the CTC. Two non-parametric estimation methods, the MLE method and KDE method, were formulated and applied to both kinds of data. The results from applying the methods to synthetic data show that, under conditions that supports a high level of proliferation at various levels of local cell density, the estimation errors decrease and one can visually discern the proliferation rate as an Allee effect. The KDE method presents, however, some issues possibly related to the choice of bandwidth, that affect the accuracy. Under less ideal conditions, both methods present some issues. The MLE method can yield portions of local cell density with no estimate for cell division rate and the KDE method may yield odd peaks and portions with no confidence interval for the estimate. When applying the methods to experimental data, all the aforementioned issues are present and there is no visual, definitive conclusion about the growth model. As the synthetic data was generated to relatively match the experiments that yielded other datasets, the glioblastoma data from HGCC is of interest in further investigations.

This thesis has demonstrated some potential for the two non-parametric methods to estimate the cell division rate, yet further improvement is needed and some possible issue-alleviating suggestions have been provided for this aim.

Bibliography

[1] B. Neumann, T. Walter, J. K. Hériché *et al.*, "Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes," *Nature*, vol. 464, pp. 721–727, 4 2010.

[2] K. E. Johnson, G. Howard, W. Mo *et al.*, "Cancer cell population growth kinetics at low densities deviate from the exponential growth model and suggest an allee effect," *PLoS Biology*, vol. 17, 8 2019.

[3] M. Janiszewska, D. P. Tabassum, Z. Castaño *et al.*, "Subclonal cooperation drives metastasis by modulating local and systemic immune microenvironments," *Nature Cell Biology*, vol. 21, pp. 879–888, 7 2019.

[4] F. Courchamp, L. K. Berec, and J. Gascoigne, *Allee Effects in Ecology and Conservation*. Oxford University Press, 9 2008.

[5] I. Nazarenko, S. M. Hede, X. He *et al.*, "Pdgf and pdgf receptors in glioma," *Uppsala Journal of Medical Sciences*, vol. 117, pp. 99–112, 5 2012.

[6] P. Gerlee, P. M. Altrock, A. Malik *et al.*, "Autocrine signaling can explain the emergence of allee effects in cancer cell populations," *PLoS Computational Biology*, vol. 18, pp. 1–15, 2022. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1009844

[7] H. Kobayashi, M. Ohkubo, A. Narita *et al.*, "A method for evaluating the performance of computeraided detection of pulmonary nodules in lung cancer CT screening: Detection limit for nodule size and density," *British Journal of Radiology*, vol. 90, no. 1070, 2017.

[8] Z. Neufeld, W. von Witt, D. Lakatos *et al.*, "The role of allee effect in modelling post resection recurrence of glioblastoma," *PLoS Computational Biology*, vol. 13, 11 2017.

[9] R. P. Araujo and D. L. McElwain, "A history of the study of solid tumour growth: The contribution of mathematical modelling," *Bulletin of Mathematical Biology*, vol. 66, no. 5, pp. 1039–1091, 2004.

[10] A. R. Anderson and P. K. Maini, "Mathematical Oncology," *Bulletin of Mathematical Biology*, vol. 80, no. 5, pp. 945–953, 2018. [Online]. Available: https://doi.org/10.1007/s11538-018-0423-5

[11] P. M. Altrock, L. L. Liu, and F. Michor, "The mathematics of cancer: Integrating quantitative models," *Nature Reviews Cancer*, vol. 15, no. 12, pp. 730–745, 2015. [Online]. Available: http://dx.doi.org/10.1038/nrc4029

[12] G. Casella and R. L. Berger, *Statistical inference.*, ser. Duxbury advanced series. Duxbury, 2002.

[13] Y. C. Chen, "A tutorial on kernel density estimation and recent advances," *Biostatistics and Epidemiology*, vol. 1, no. 1, pp. 161–187, jan 2017.

[14] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Springer, 1986. [Online]. Available: https://doi.org/10.1007/978-1-4899-3324-9

[15] G. Cheng and Y. C. Chen, "Nonparametric inference via bootstrapping the debiased estimator," *Electronic Journal of Statistics*, vol. 13, pp. 2194–2256, 2019.

[16] M. Jones, "Simple boundary correction for kernel density estimation," *Statistics and Computing*, vol. 3, pp. 135–146, 1993. [Online]. Available: https://doi.org/10.1007/BF00147776

[17] E. Fieller, "A fundamental formula in the statistics of biological assay, and some applications," *Quarterly Journal of Pharmacy and Pharmacology*, vol. 17, pp. 117–173, 1944.

[18] H. Motulsky, *Intuitive biostatistics*. Oxford University Press, 1995.

[19] V. Ulman, M. Maška, K. E. Magnusson *et al.*, "An objective comparison of cell-tracking algorithms," *Nature Methods*, vol. 14, pp. 1141–1152, 12 2017.

[20] R. Skloot, *The Immortal Life of Henrietta Lacks*. Broadway Books, 9 2010.

[21] G. Rodriguez, "Lecture notes on generalized linear models," 2007, last accessed 7 January 2023. [Online]. Available: https://grodri.github.io/glms/notes/c7.pdf

[22] Y. Xie, T. Bergström, Y. Jiang *et al.*, "The human glioblastoma cell culture resource: Validated cell models representing all molecular subtypes," *EBioMedicine*, vol. 2, pp. 1351–1363, 10 2015.

[23] L. M. Beskow, "Lessons from HeLa Cells: The Ethics and Policy of Biospecimens," *Annual Review of Genomics and Human Genetics*, vol. 17, pp. 395–417, 2016.

A

Algorithm for generation of synthetic data

Algorithm 1 An algorithm for generating synthetic data.

Require: N_0 cells with placement $\boldsymbol{x}(0)$, T and Δt.
Require: ω, λ_0, λ_1 and $h(t)$ according to one of the models in Table 3.3.
Require: α, r_0, a, D_e, σ according to Table 3.2.

```
Initialise b_i and d_i to be standard uniformly distributed.
Initialise B_i(0) = D_i(0) = 0.
for k = 1, ..., T do
    Let A be the index set of alive cells.
    for i ∈ A do
        Calculate ρ_i(k) according to (3.6).
        Calculate B_i(k) and D_i(k) according to (3.8).
        if B_i(k) > b_i then
            Reset B_i and generate new b_i.
            Add new cell index j to A .
            Initialise b_j and d_j to be standard uniformly distributed.
            Initialise B_j(k + 1) = D_j(k + 1) = 0.
            Record division in a list.
        end if
        if D_i(k) > d_i then
            Remove i from A
        end if
        Update x_i(t) according to (3.7).
        if k is multiple of Δt then
            Record x(k)
        end if
    end for
end for
Return record of cell positions.
Return list of cell divisions.
```

B

Complementary results

In Sections B.1 and B.2 we present the worst estimates of the per capita cell division rate for the exponential and logistic growth models. In Sections B.3 and B.4 we present a more wholesome repertoire of per capita cell division rate estimates for varying values of N_0, Δt and T.

B.1 Exponential growth model

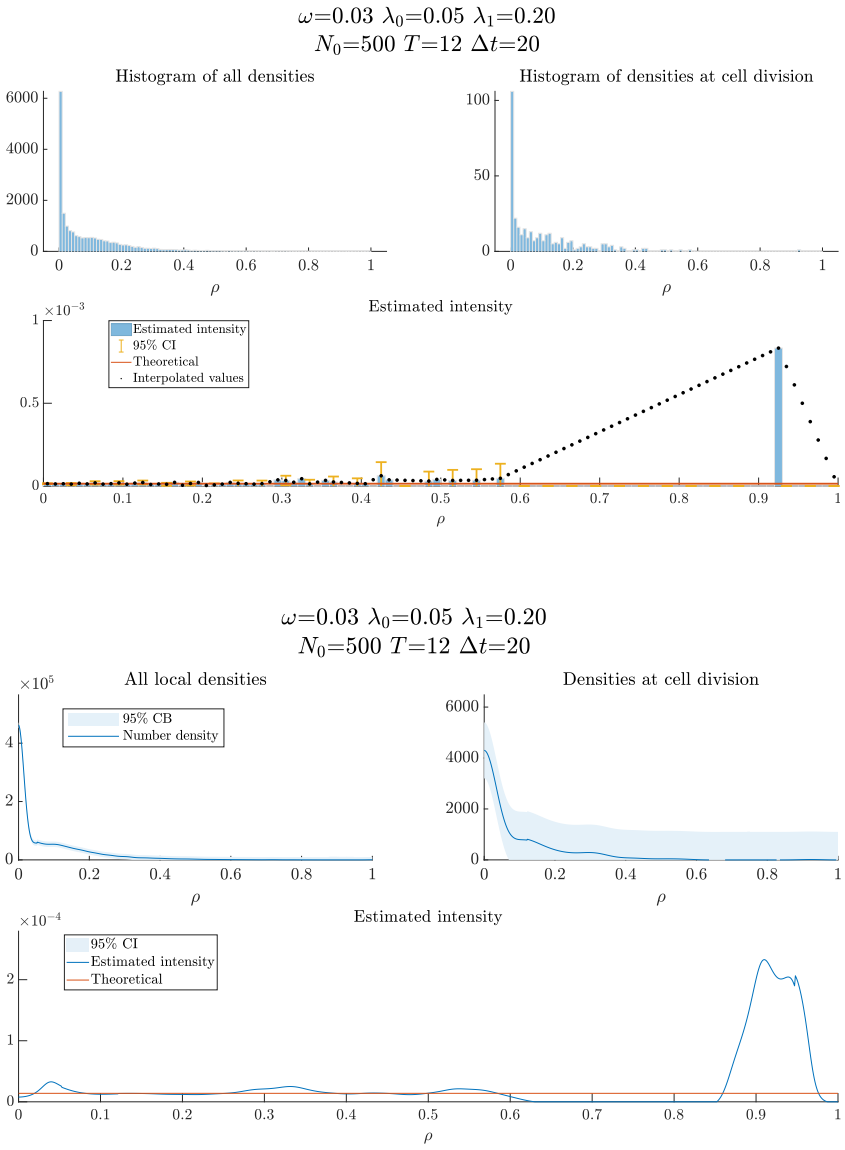


Figure B.1: The worst performing estimates of $h(\rho)$ for the exponential growth model.

B.2 Logistic growth model

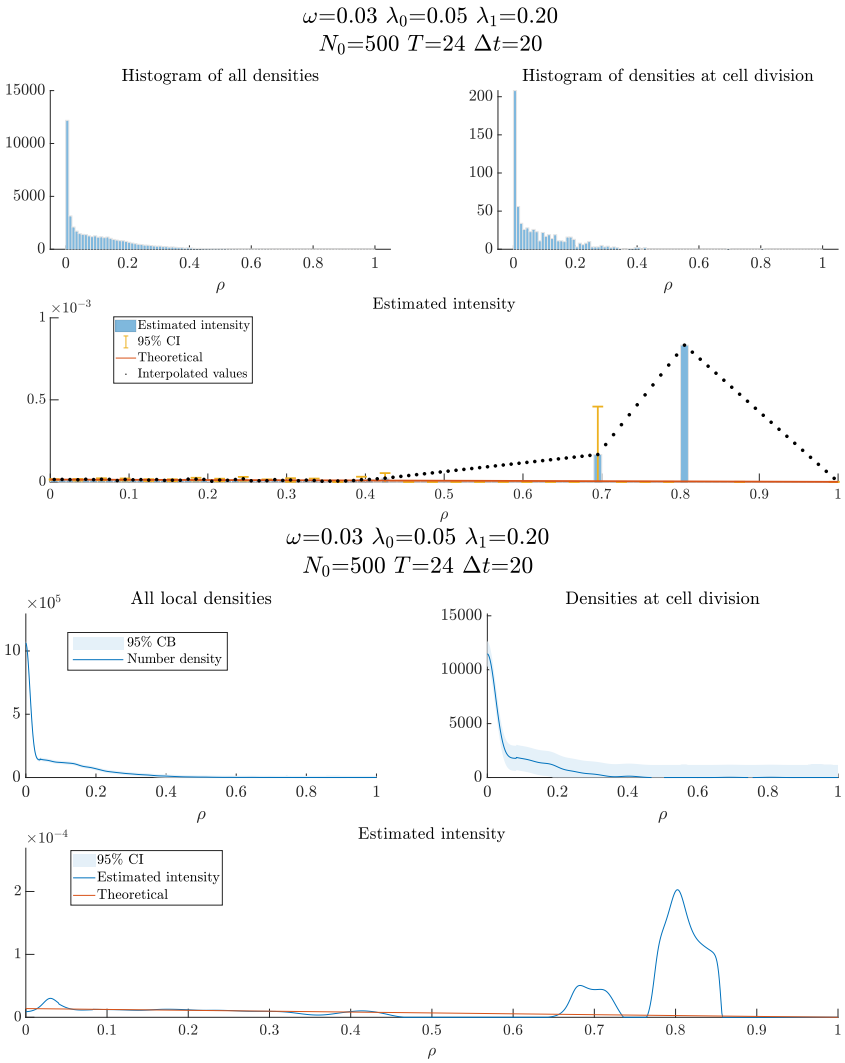


Figure B.2: The worst performing estimates of $h(\rho)$ for the logistic growth model.

B.3 Weak Allee effect growth model

Figures B.3 and B.4 present estimates of $h(\rho)$ from the MLE method and KDE method estimates, respectively, for various values of N_0. Turning our eyes to the smaller figures, we notice that, from top to bottom, the cell densities go from knocking on the door at $\rho = 0$, to knocking at both $\rho = 0$ and $\rho = 1$, to finally merely knocking at $\rho = 1$. That is, the cells go from sparser to more and more crowded situations. Further, compare with the corresponding areas for the estimates and note that we seem to have an increase in performance at areas with high counts in the histogram or high values in the number densities. This is expected, as more data for each estimate should make it a better estimate. Inspecting the MLE method estimate corresponding to $N_0 = 500$ in Figure B.3, together with two estimate not shown and the three estimates corresponding to $N_0 = 250$, there is nothing odd explaining the hiccup. The best performing estimate for the KDE method is not found through the dataset with $N_0 = 2000$, but rather $N_0 = 1000$. Looking at the middle group of figures in Figure B.4, the slopes in the rightmost number density are more amplified where the leftmost number density is relatively low. In all, it appears that the KDE method is more affected by changes in slope of the number densities than the parallel in the MLE method. Note here that the KDE estimates are clearly affected in this manner near the boundaries, forming little ”horns” when there are clear changes in slopes when cell densities accumulate there.

In Figures B.5 and B.6 one finds the MLE method and KDE method estimates, respectively, for various values of Δt. Again, top to bottom indicate worst to best. In the MLE method, the estimates are similar in visual representation, and seem not to vary in any distinguishable way. With increasing Δt the error is expected to increase, however, and with merely three experiments for each value of Δt, randomness can certainly affect the result. The way we interpolate values for missing estimates does not affect this result as there are no missing estimates in any of the experiments. The KDE method estimates are also visually similar, but it is easier to interpret a decrease in accuracy with increasing Δt. Note the little slope related horns are present here as well.

Finally, Figures B.7 and B.8 present MLE method and KDE method estimates, respectively, for various values of T. Once again, consider the histograms and number densities, in which we notice the same shift in mass from lower densities to higher densities with increasing T, as noticed with increasing N_0. If the population size is initially small and grows over time, making it more crowded, then an experiment that runs for a longer time will likely cover larger ranges of cell densities. The slope related horns remains in the KDE method estimates here as well.

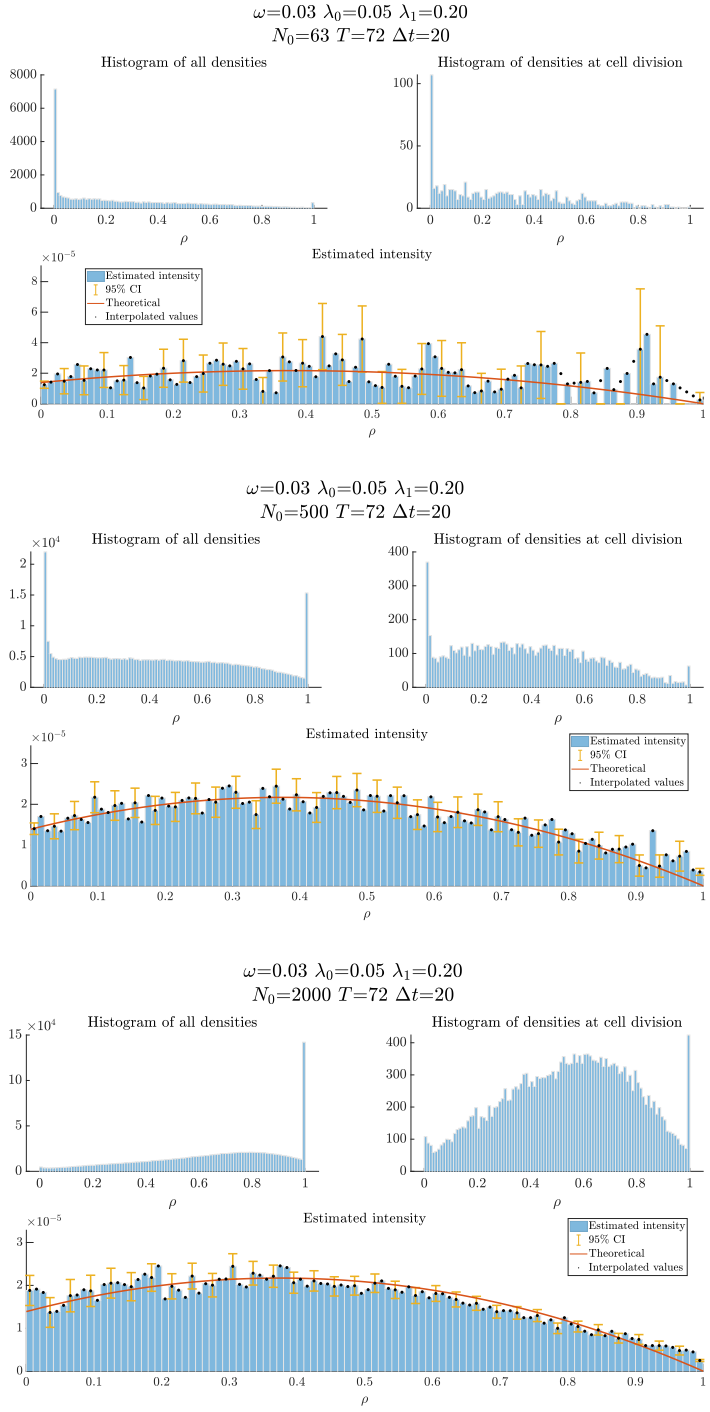


Figure B.3: Estimates of $h(\rho)$ for different values of N_0 through the MLE method for weak Allee effect growth model. Top to bottom, worst to best.

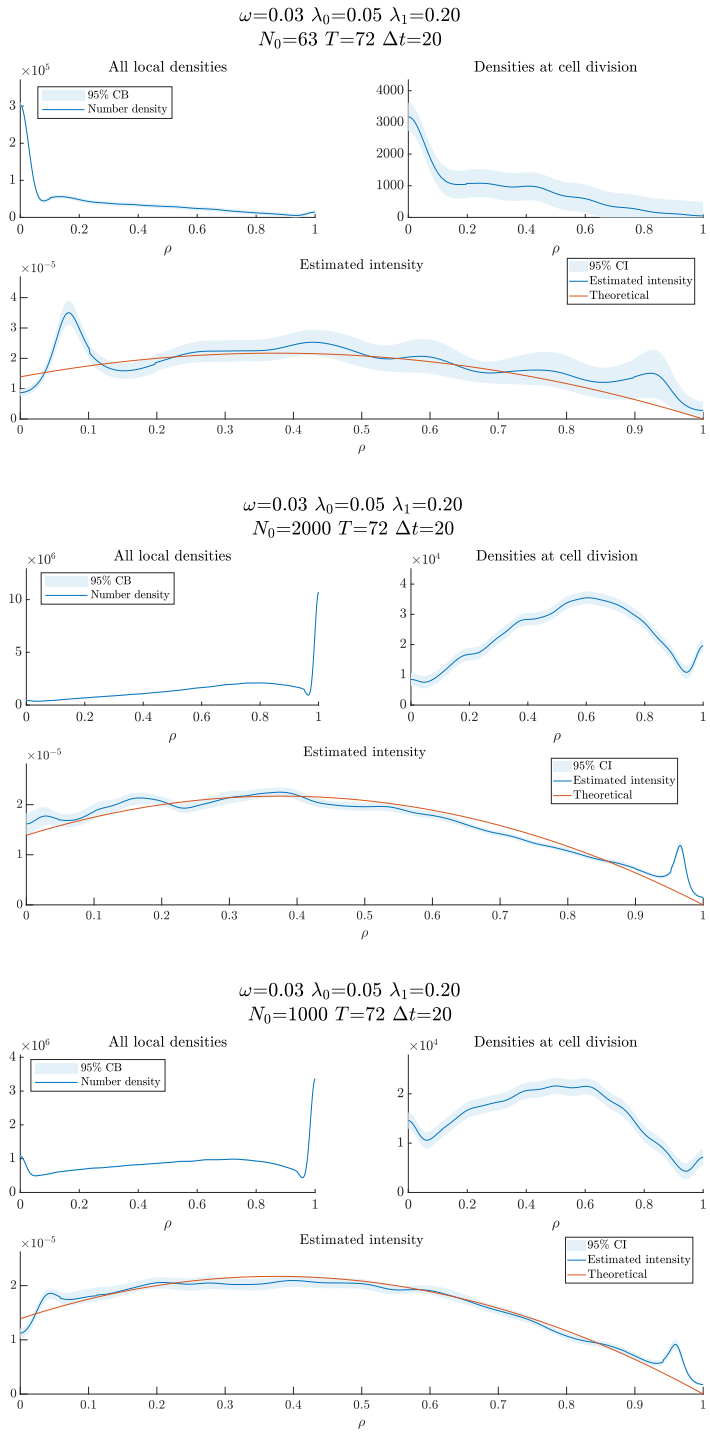


Figure B.4: Estimates of $h(\rho)$ for different values of N_0 through the KDE method for weak Allee effect growth model. Top to bottom, worst to best.

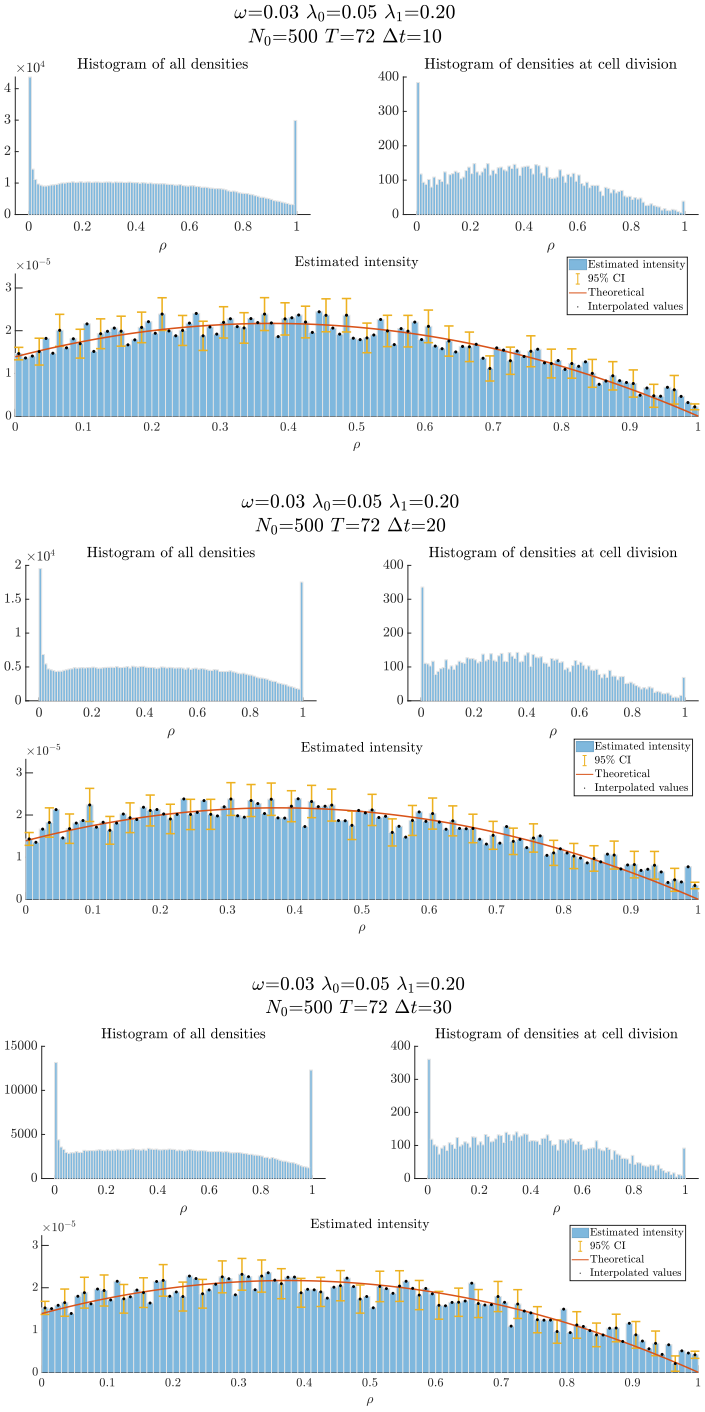


Figure B.5: Estimates of $h(\rho)$ for different values of Δt through the MLE method for weak Allee effect growth model. Top to bottom, worst to best.

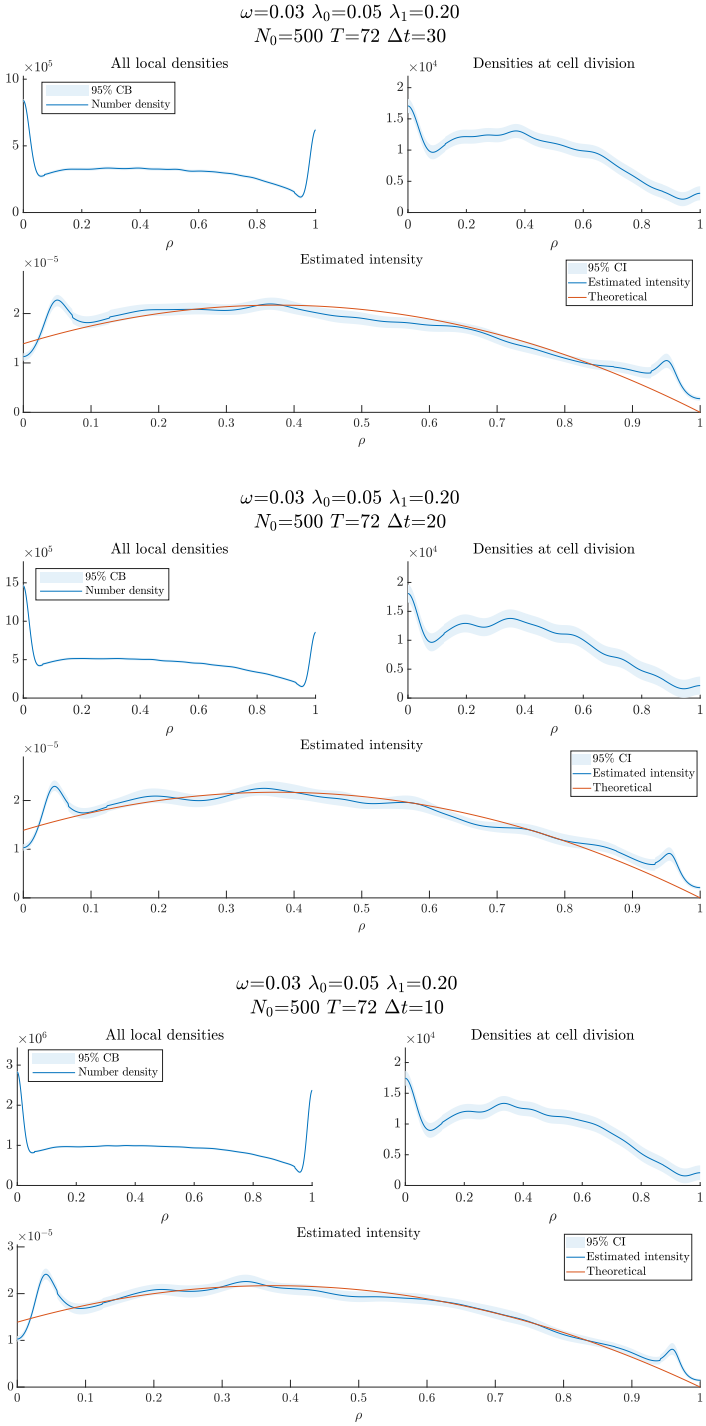


Figure B.6: Estimates of $h(\rho)$ for different values of Δt through the KDE method for weak Allee effect growth model. Top to bottom, worst to best.

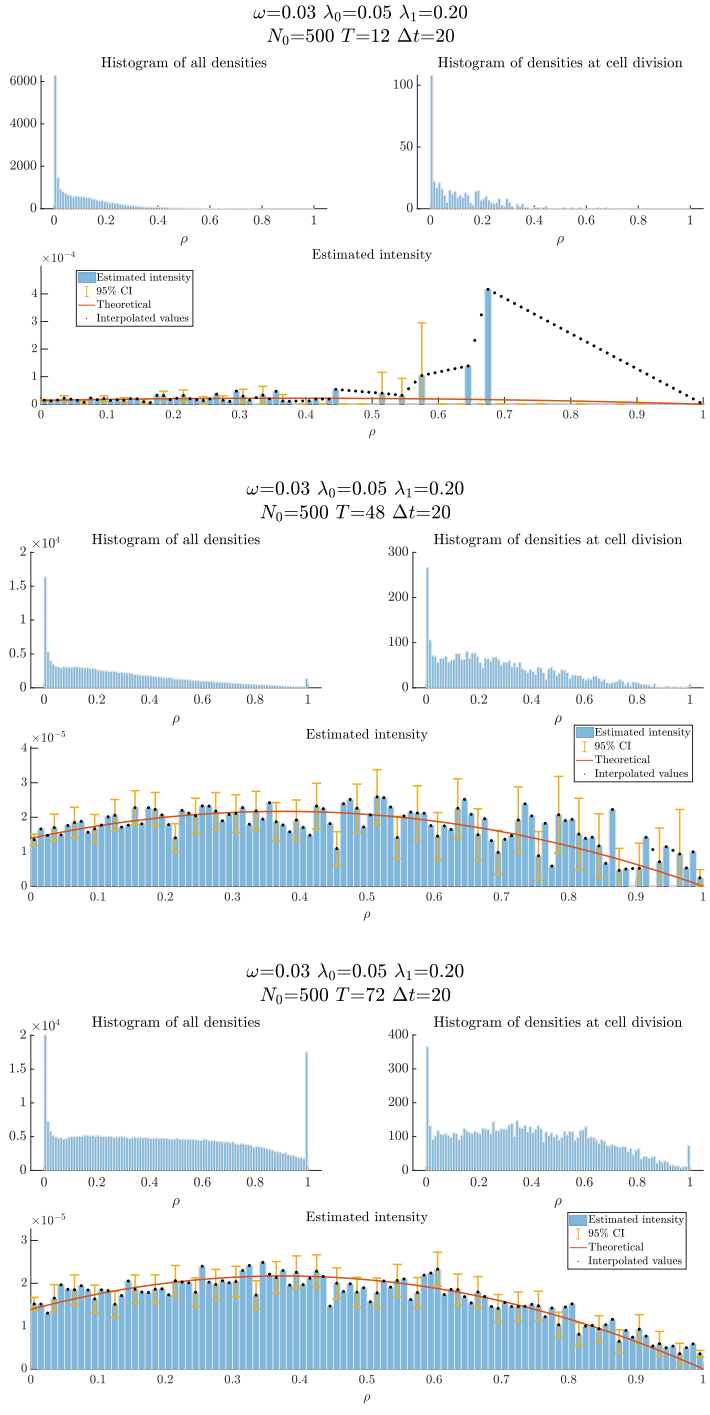


Figure B.7: Estimates of $h(\rho)$ for different values of T through the MLE method for weak Allee effect growth model. Top to bottom, worst to best.

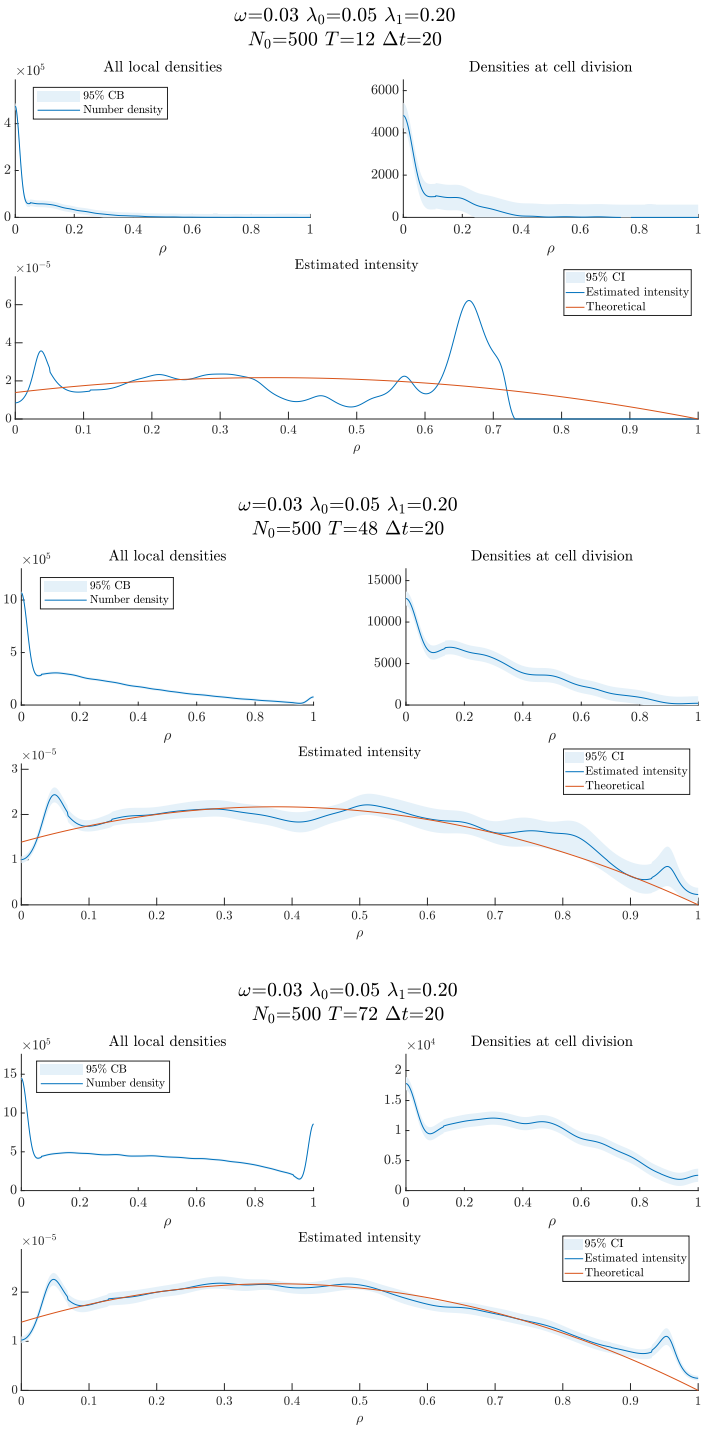


Figure B.8: Estimates of $h(\rho)$ for different values of T through the KDE method for weak Allee effect growth model. Top to bottom, worst to best.

B.4 Strong Allee effect growth model

Figure B.9 and B.10 presents MLE method and KDE method estimates, respectively, for different values of N_0. Consider the groups of figures corresponding to the $N_0 = 63$ MLE method estimate and $N_0 = 125$ KDE method estimate. It is evident that the data is significantly different from all data we have seen so far. The current two datasets are two of three in which the cell population went extinct, which occurred once with $N_0 = 63$ and twice for $N_0 = 125$. This should be related to the errors shown in 4.7. In the KDE method estimate that there is a spike which continues past boundaries of the figure, giving its peak value about 10, giving rise to quite large errors. The slope sensitivity seem to go to extremes with small datasets which can directly be linked with the error observed in Figure 4.7.

The bottom two groups of figures in Figure B.9 are more familiar to us, and one should compare it with the bottom two figures in Figure B.3 from the weak Allee effect section, as they have the same values for N_0, Δ_t and T. In a strong Allee effect setting, the instantaneous division rate is lower and the instantaneous death rate higher, yielding less data for us. We see this as the estimate in the middle group of figures is not as precise as its parallel. This little more of a struggle for the cells to create crowded situations for themselves is also visible when one compare the histograms for the estimates corresponding to $N_0 = 2000$.

Similarly, we recognise the bottom two groups of figures in Figure B.10. The slope related horns are present in the middle group but surprisingly enough they are incredibly mild in the bottommost group.

Figure B.11 and B.12 present the MLE method and KDE method estimates, respectively, for different values of Δt. In the former, it is visually clear that with increasing Δt the error increases whereas in the latter, in stark contrast, it is not visually clear as both the worst and best estimates are obtained with $\Delta t = 40$.

Finally, Figures B.13 and B.14 presents MLE method and KDE method estimates, respectively, for different values of T. Again, the estimates are clearly improving with increased T, however not in the same amount as in the weak Allee effect case. Comparing the smaller histograms and number densities with their parallels in Figures B.7 and B.8, again shows that in the strong Allee effect growth model, one obtains less data. The peak that passes borders of the figure in Figure B.14 has peak value 0.015.

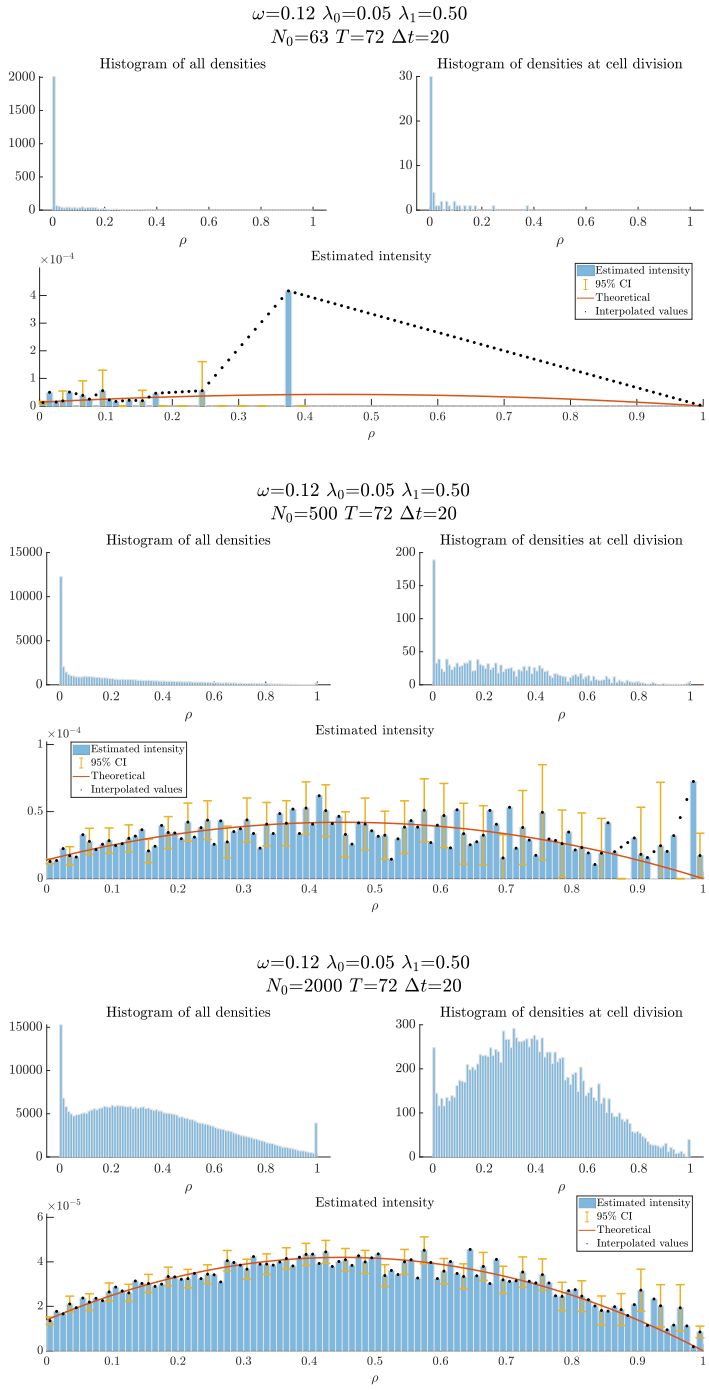


Figure B.9: Estimates of $h(\rho)$ for different values of N_0 through the MLE method for strong Allee effect growth model. Top to bottom, worst to best.

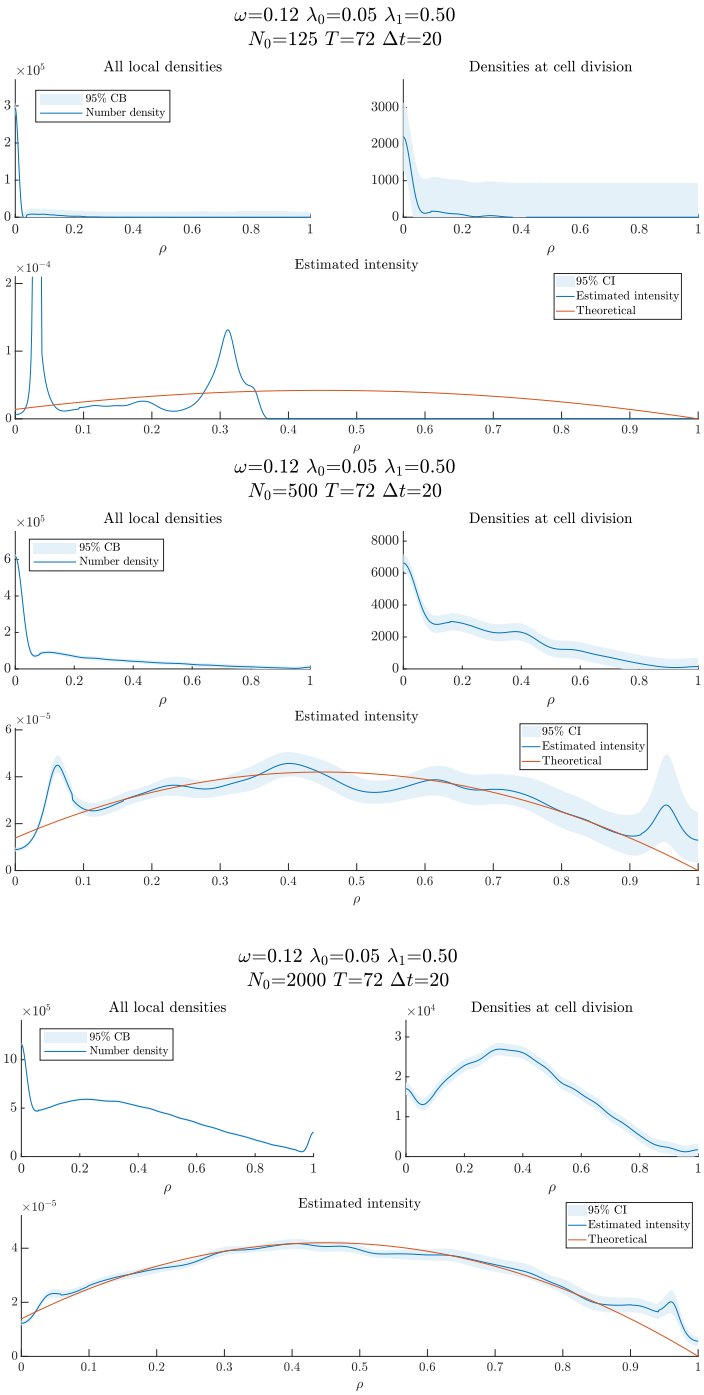


Figure B.10: Estimates of $h(\rho)$ for different values of N_0 through the KDE method for strong Allee effect growth model. Top to bottom, worst to best.

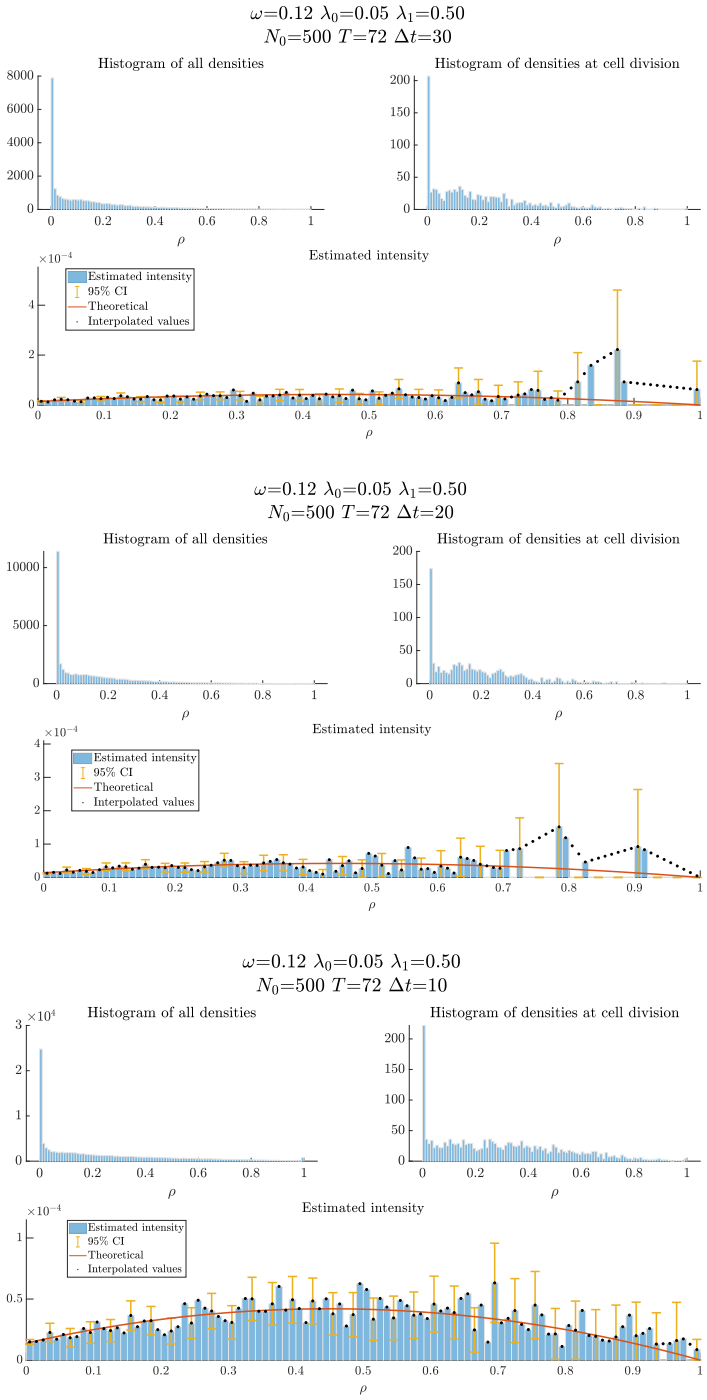


Figure B.11: Estimates of $h(\rho)$ for different values of Δt through the MLE method for strong Allee effect growth model. Top to bottom, worst to best.

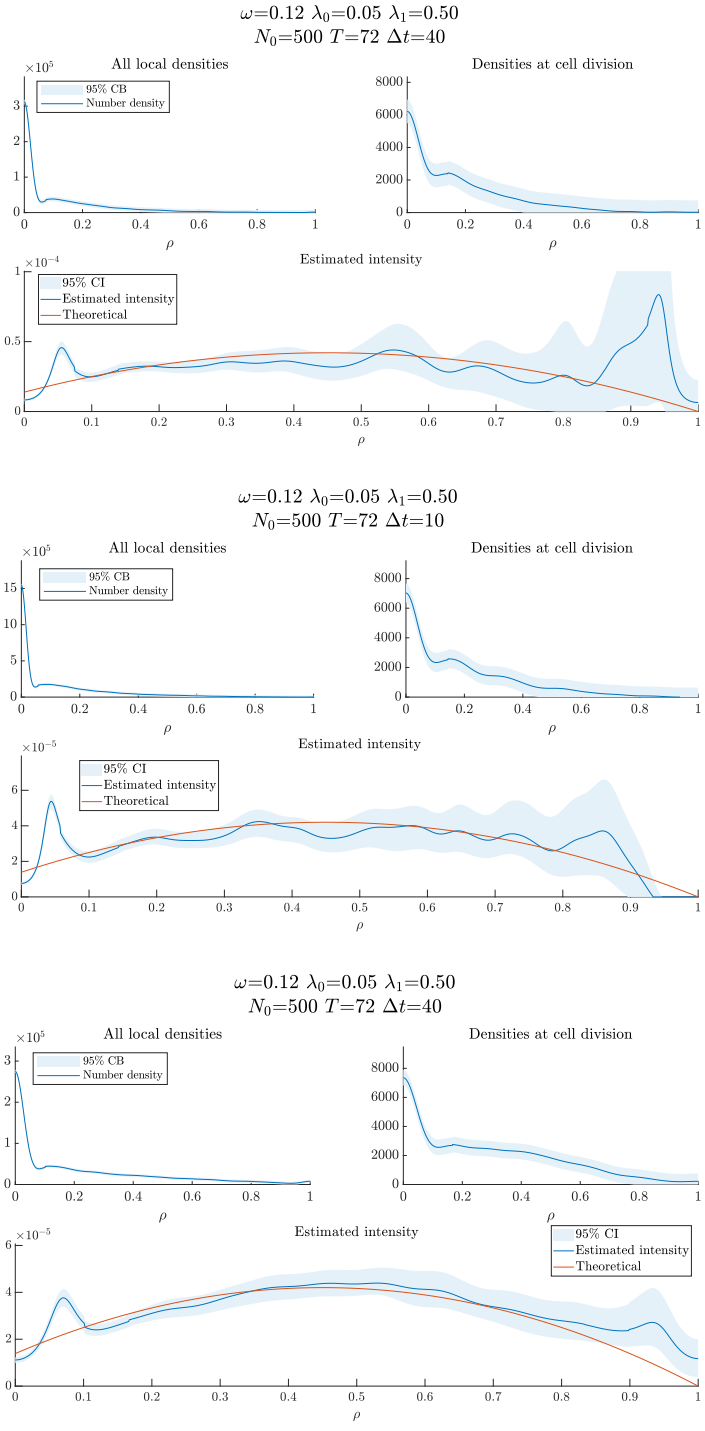


Figure B.12: Estimates of $h(\rho)$ for different values of Δt through the KDE method for strong Allee effect growth model. Top to bottom, worst to best.

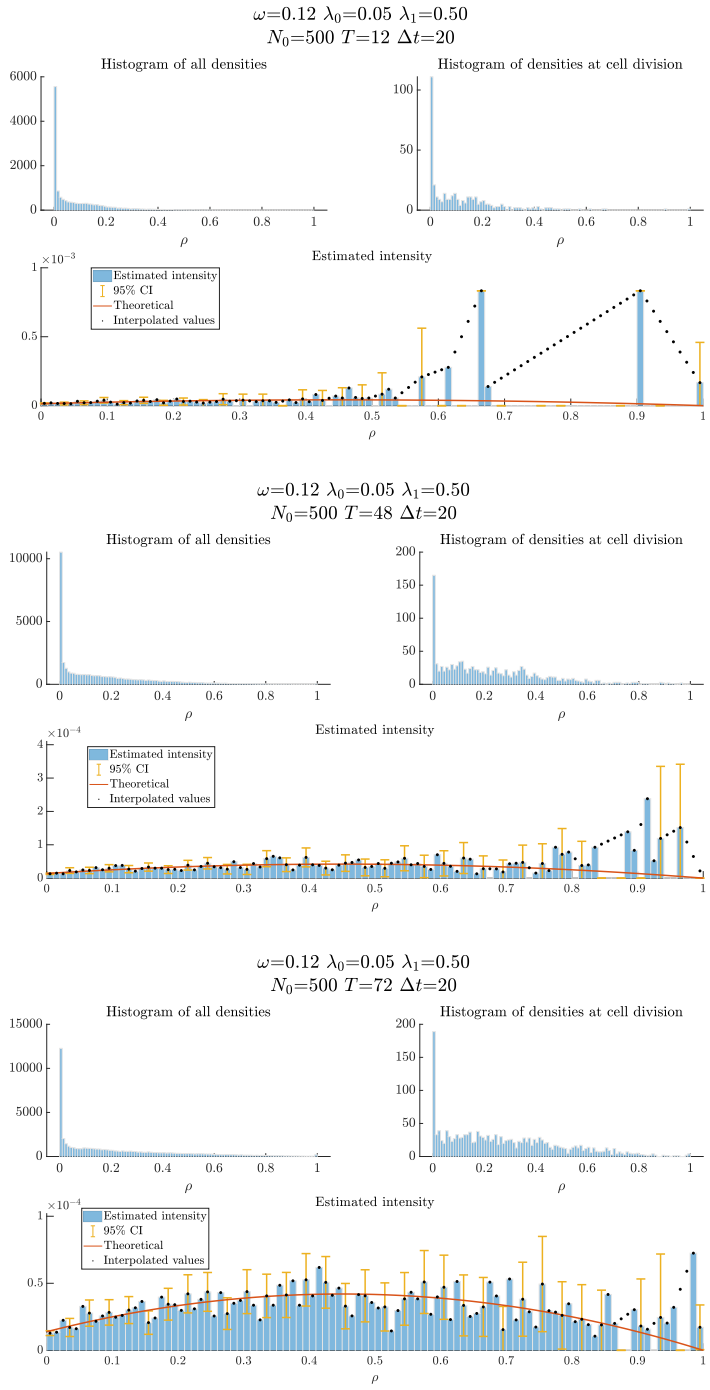


Figure B.13: Estimates of $h(\rho)$ for different values of T through the MLE method for strong Allee effect growth model. Top to bottom, worst to best.

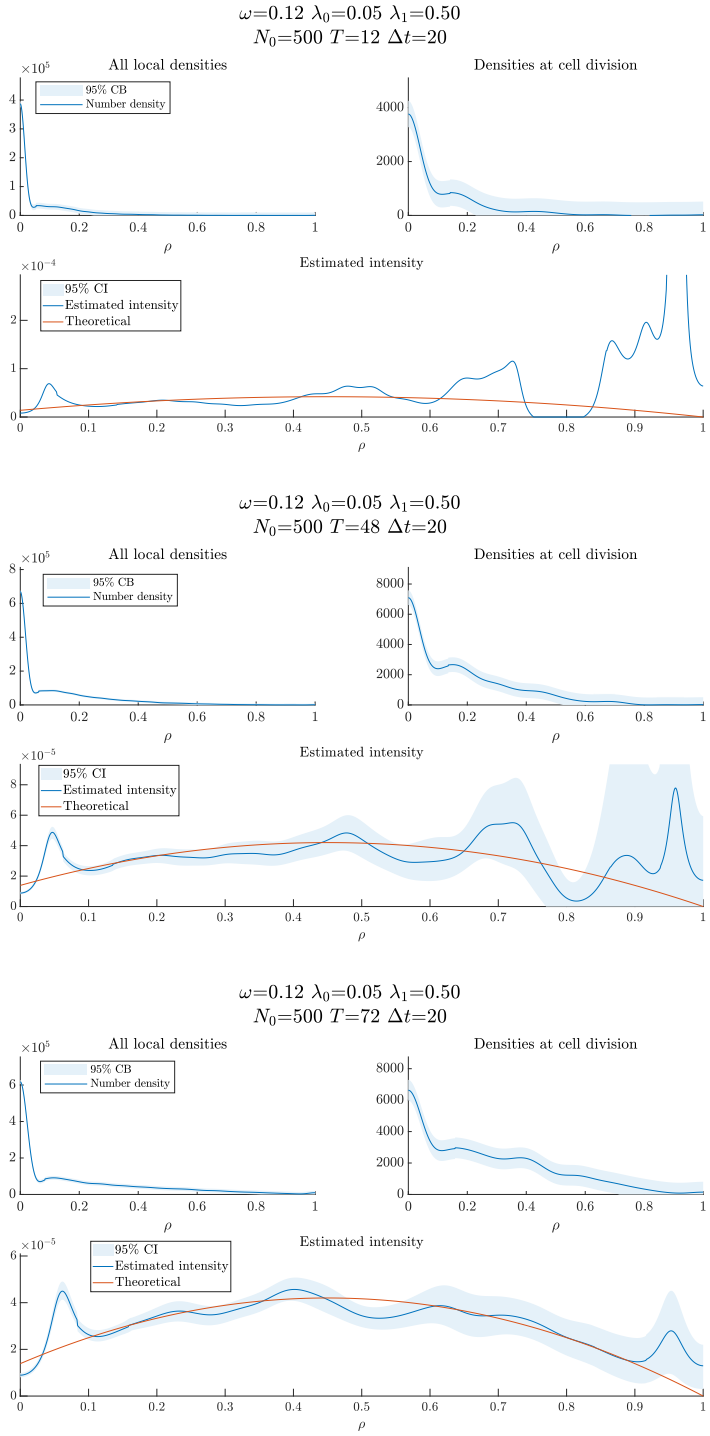


Figure B.14: Estimates of $h(\rho)$ for different values of T through the KDE method for strong Allee effect growth model. Top to bottom, worst to best.

B.5 Experimental data.

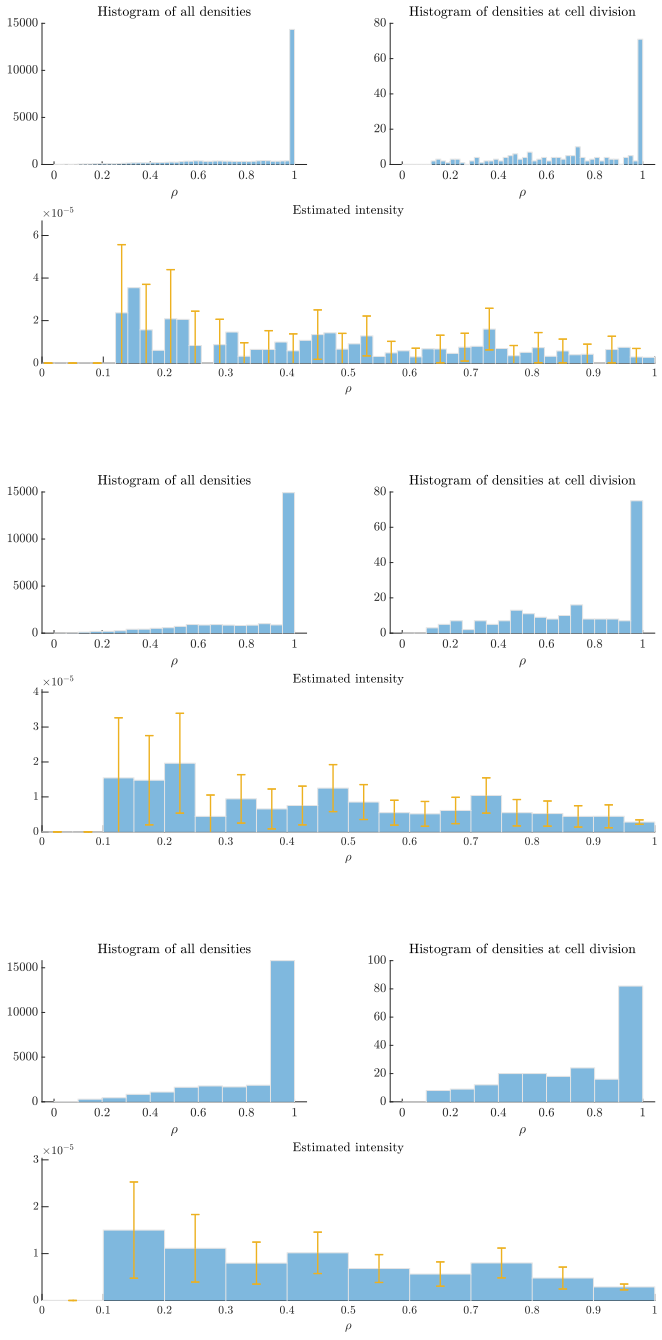


Figure B.15: Estimates of $h(\rho)$ based on the second dataset using the MLE method and bin width 0.02, 0.05 and 0,1.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY