Stochastic mathematical models of cell proliferation assays

by

Alexander P Browning
Bachelor of Mathematics

School of Mathematical Sciences
Faculty of Science and Engineering
Queensland University of Technology

A dissertation submitted in fulfilment
of the requirements for the degree of
Master of Philosophy

2017

Keywords: Approximate Bayesian computation; Cell proliferation assay; Individual based model; Logistic growth; Generalised logistic growth; Model calibration; Parameter estimation; Experimental design.

In accordance with the requirements of the degree of Master of Philosophy in the School of Mathematical Sciences, Faculty of Science and Engineering, I present the following thesis entitled,

Stochastic mathematical models of cell proliferation assays.

This work was performed under the supervision of Professor Matthew Simpson and Professor Scott McCue. I declare that the work submitted in this thesis is my own, except as acknowledged in the text and footnotes, and has not been previously submitted for a degree at Queensland University of Technology or any other institution.

Signed,

Alexander P Browning
25 August 2017

Acknowledgements

First and foremost, I would like to thank my principal supervisor, Professor Matthew Simpson. Without your guidance and encouragement this thesis could not exist. I am grateful for the time you have committed to further my education, and for giving me the opportunity to work with both experimental data, as well as mathematical modelling tools.

Thanks must go to my co-supervisor, Professor Scott McCue, for your support, guidance, encouragement and insights into the problems we studied. I also thank Associate Professor Michael Plank and Dr Rachelle Binny for their helpful comments, feedback and guidance. I also thank them for their hospitality and willingness to host me in Christchurch as we coauthored the paper. I also appreciate the assistance of Ms Esha Shah and Ms Parvathi Haridas for performing the experiments that feature in our work. Finally, I appreciate the advice of Mr David Warne, for helping me to understand and interpret certain computational and statistical aspects of the work.

I thank and acknowledge the financial support of the Australian Research Council and additional funding provided by the School of Mathematical Sciences, Institute of Health and Biomedical Innovation and my principal supervisor. I am thankful that I was able to present my work at the AMSI Winter School 2016 and ANZIAM 2017. Lastly, I am grateful of the opportunity to travel to Christchurch to work with our collaborators, Associate Professor Michael Plank and Dr Rachelle Binny.

I would like to thank all the fellow Research Masters and PhD students around me, for their support, encouragement, advice and friendship. And to my friends and family, Mum, Dad, Millie, Raiha, Asad, Emma, Jonathan and Brad, for always showing me what's possible.

Abstract

Cell proliferation assays are routinely used to study collective cell behaviour, and can be interpreted with mathematical models. Often, it is assumed that cells proliferate logistically, and the classical mean-field logistic equation is employed to study how interactions between cells cause the growth rate to reduce as the population approaches confluence. Rarely is this assumption tested, particularly by applying experimental data. In this thesis, we apply a stochastic, on-lattice, individual based model to explore the experimental duration required to gain insight into a crowding mechanism that can best describe experimental data. Next, we demonstrate a technique to calibrate an off-lattice individual based model to three independent experimental data sets. Our model is able to both describe and predict the evolution of the population and spatial structure in a cell proliferation assay.

Contents

Chapter 1

Introduction

1.1 Overview

The collective behaviour of cell populations is routinely investigated using two-dimensional *in vitro* cell biology experiments (Liang et al., 2007). Understanding the behaviour of cell populations is relevant to many normal and pathological processes, such as tissue regeneration and malignant spreading, respectively. One of the most common *in vitro* experimental methods employed to understand these processes is called a *cell proliferation assay* (Bosco et al., 2015; Bourseguin et al., 2016). Cell proliferation assays involve a monolayer of cells, placed uniformly at relatively low density, on a two-dimensional substrate. Individual cells proliferate and move, and the assay is observed as the density of the monolayer increases (Tremel et al., 2009). Comparing the behaviour of the cells with and without application of a putative drug plays an important role in drug design (Bosco et al., 2015; Bourseguin et al., 2016). In this thesis, the experimental results involve cell proliferation assays conducted using 3T3 fibroblast cells and PC-3 prostate cancer cells (Kaighn et al., 1979).

Typically, data collected from a cell proliferation assay consists of a series of images, collected over time, that comprise of a field-of-view that represents the centre of each proliferation assay (Figure 1.1). This data may be summarised as a time-series of the approximate number of cells or approximate cell density. Spatial structure at any point in time can be characterised with measures of the numbers of pairs separated by distance. Commonly, the cell behaviour is summarised by a proliferation rate, which can then be used to compare a target and control assay (Johnston et al., 2015). A typical initial density of a cell proliferation assay is about 10% of carrying capacity, and a typical cell doubling time is approximately

15 hours (Jin et al., 2016b; Treloar et al., 2014). The classical logistic equation (Pearl, 1927; Edelstein-Keshet, 1988; Murray, 2002) is commonly applied to describe the growth of a cell population (Jin et al., 2016a; Cai et al., 2007; Maini et al., 2004b; Sherratt and Murray, 1990), and is given by

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \lambda N(t)\left(1 - \frac{N(t)}{N_{\max}}\right), \tag{1.1}$$

which has the exact solution

$$N(t) = \frac{N_{\max}}{1 + A\exp(-\lambda t)}, \tag{1.2}$$

where $N(t)$ is the number of cells, λ is a proliferation rate and $N_{\max}$ is the carrying capacity population. A is chosen to ensure the population at $t = 0$ is $N(0)$. The solution to Equation (1.2) is shown in Figure 1.2(a) for parameters corresponding to a typical cell doubling time of 15 hours, and a typical initial density of 10%. For $N(t) \ll N_{\max}$, Equation (1.2) implies $N(t) \sim N(0)\exp(\lambda t)$, and for small t, $N(t) \sim N(0)(1 + \lambda t + \mathcal{O}(t^2))$. This behaviour is observed in Figure 1.2(b), where the solution to the logistic equation 1.2 is presented for a typical experimental duration of 24 hours. These qualitative observations demonstrate that, for a typical initial cell density, the growth of a population of cells is initially approximately exponential. This means that a standard experimental duration of 24 hours is long enough to calculate the cell proliferation rate, but crowding effects that inhibit the net proliferation rate are not clearly observable. Additional cost is involved in extending the duration of these experiments. Exploration into the experimental duration required to gain information about these crowding effects is discussed in Chapter 2 of this thesis.

Applying a mathematical model to understand a cell proliferation assay provides quantitative insight into the mechanisms involved (Maini et al., 2004b; Sengers et al., 2007). A calibrated mathematical model can be used to form experimental hypotheses and guide experimental design (Stoll et al., 2003). A standard approach to modelling a cell proliferation assay is to use a mean-field model, which is equivalent to assuming that individuals interact in proportion to the mean population density (Tremel et al., 2009; Sengers et al., 2007; Maini et al., 2004b; Sarapata and de Pillis, 2014; Sherratt and Murray, 1990), neglecting spatial structure. Additionally, most previous studies that interpret cell biology assays using continuum mathematical models make the implicit assumption that cells proliferate logistically (Cai et al., 2007; Maini et al., 2004b,a; Sheardown and Cheng, 1996). While the classical logistic equation (Equation (1.1)) is commonly used to

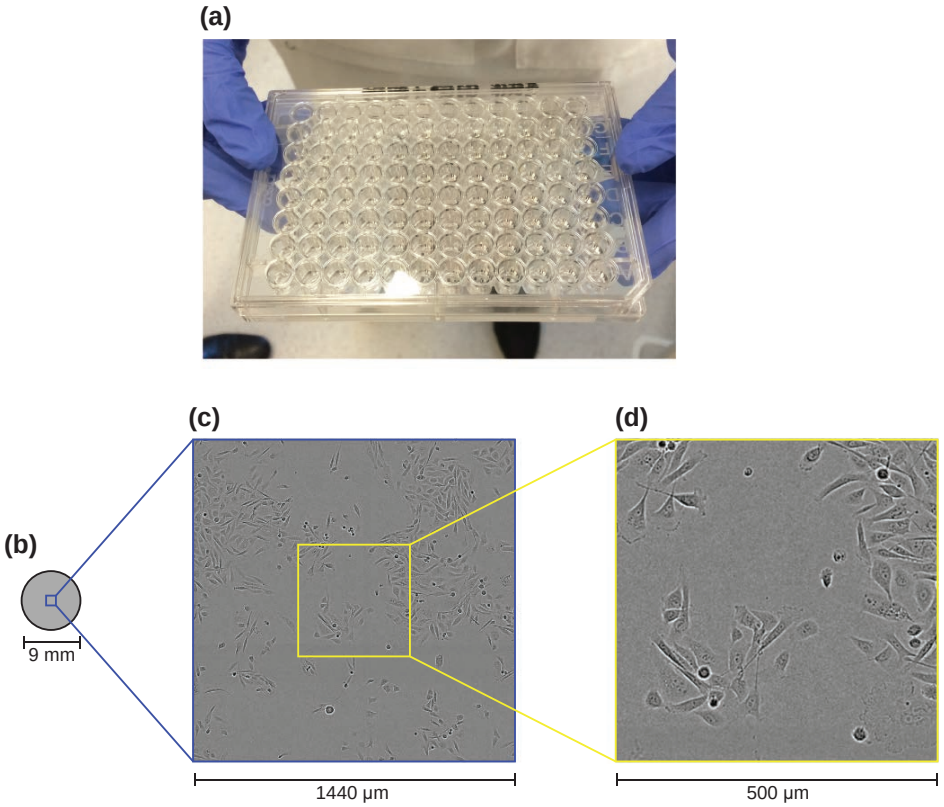

Figure 1.1: (a) An experimentalist holding a 96-well plate, used to conduct cell proliferation assays. Each well has diameter 9 mm. (b) A simple schematic of a single well in (a). (c) Example image taken from the centre 1440 $\times$ 1440 μm of a well, during a cell proliferation assay with PC-3 prostate cancer cells. (d) Enlargement of the centre 500 $\times$ 500 μm region of (c), as indicated. Individual cells are visible.

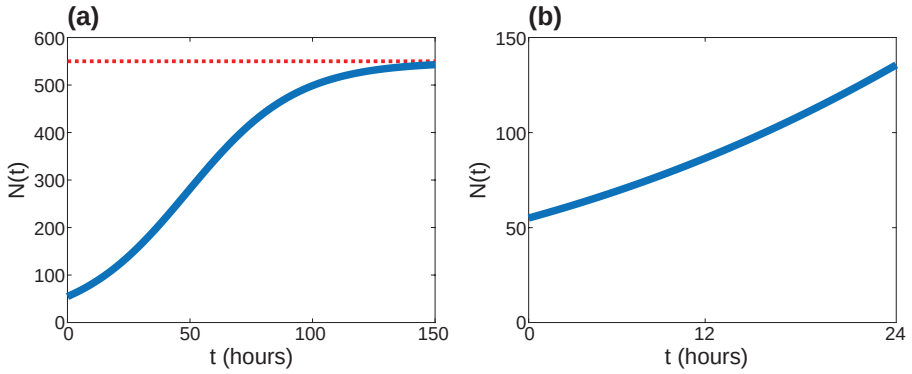

Figure 1.2: Solution to the classical logistic equation (1.2) with $N_{\max} = 550$, $N(0) = 55$ and $\lambda = 0.045$, corresponding to a doubling time of approximately 15 hours, and a initial density of 10% of carrying capacity. (a) The solution to late time where the population is observed to asymptote to the carrying capacity. (b) The solution over a typical experimental duration of 24 hours, when the population is observed to grow exponentially.

model populations (Pearl, 1927; Edelstein-Keshet, 1988; Murray, 2002), this choice is often made without a careful examination of its validity (Treloar et al., 2014). There is awareness that biological populations do not always grow according to the classical logistic equation (Gerlee, 2013; Zwietering et al., 1990). Recent work of Sarapata and de Pillis (2014) explores a range of sigmoid growth models for different types of tumours, including those of the bladder, breast, liver, lung, and melanoma tumours. Sarapata and de Pillis (2014) show that the classical logistic model does not always provide the best match to observed data, and they test a range of other sigmoidal growth models for each different kind of tumour. A key difference between these sigmoidal growth models is the way in which they characterise crowding effects, which, as mentioned, are not clearly observable in experimental data from a typical experimental duration of 24 hours. In our work, we implement a generalised logistic growth model (Tsoularis and Wallace, 2002; Jin et al., 2016a), and explore our ability to distinguish between sigmoidal growth models as the experimental duration is effectively increased.

Mean-field models, such as ordinary and partial differential equations, do not consider the spatial structure of a population. By definition, these models assume that cells interact according to a mean population density (Maini et al., 2004b; Sarapata and de Pillis, 2014; Sherratt and Murray, 1990). Baker and Simpson (2010) saw that, in a simulation-based study, some initial random spatial arrangements of cells can lead to clustering at later times. In this work, we see spatial

structure present in the experimental data at early time, which disappears as the population approaches confluence. An alternative to continuum mathematical models are stochastic discrete models, such as individual based models (IBMs). More recently, increased computation power has meant that these models have been used to directly model cell-level behaviour (Johnston et al., 2014; Binny et al., 2016a). IBMs are attractive for modelling biological phenomena because they can be used to represent properties of individual agents, such as cells, in the system of interest. For example, an IBM of a cell proliferation assay may describe the proliferation and movement rates of individual agents, which may depend on interactions between neighbouring agents (Jin et al., 2016a; Binny et al., 2016b). In addition, more complex mechanisms, such as directional movement bias, can be studied with an IBM. Because IBMs do not assume that interactions between individuals occur randomly, as in mean-field models, they are able to model populations without first making simplifying assumptions about spatial structure. Typical IBMs restrict the position of agents to be on a lattice, restricting both the position, and direction of movement, of agents (Codling et al., 2008). In contrast, lattice-free IBMs are more realistic as they allow agents to move in continuous space, in any direction. However, the extra freedom associated with lattice-free models come at the cost of higher computational requirements (Plank and Simpson, 2012). This thesis will consider both approaches to implementing an IBM.

In this thesis, we connect stochastic mathematical models to experimental and *in silico* data. Taking a Bayesian approach, we assume that cell proliferation assays are stochastic processes, and the model parameters are random variables. There is an awareness that, as stochastic models account for variation between experimental replicates, inferences can be drawn independent of the variability present in the experimental data (Ramin and Arhonditsis, 2013). This is evident in our results in Chapter 3, where we calibrate a mathematical model to multiple sets of experimental data. We use approximate Bayesian computation (Tanaka et al., 2006; Sunnaker et al., 2013; Liepe et al., 2014) to update our knowledge of the unknown model parameters as data is applied, to produce posterior distributions for each parameter, from which a point estimate can be calculated. We explore how uncertainty in the recovered parameters is decreased as the experimental duration is effectively increased. We also perform inference on the parameters in a lattice-free IBM using experimental data. We are unaware of any existing work calibrating a lattice-free IBM to experimental data. The point estimates for the parameters that produce the best fit vary between experimental replicates, however posterior distributions that describe the parameters agree between the data sets. As well as

suggesting the model is valid, the agreement of these posterior distributions implies that a combined posterior distribution may be formed as an attempt to remove variation between experimental replicates. This allows us to validate the model further by comparing a model prediction to experimental data, for an additional, independent, data set.

1.2 Structure of this thesis

This thesis comprises of two main chapters, each forming a submission to an academic journal. Chapter 2 has been accepted for publication in the *Bulletin of Mathematical Biology*. Chapter 3 is under consideration in the *Journal of Theoretical Biology*. The works are presented in this thesis as the original submission, with minor typographical changes for notational consistency. In this section, we provide an overview of the contents of each chapter.

In Chapter 2, we explore the optimal duration of a cell proliferation assay. To do this, we present a suite of *in silico* cell proliferation assays where the cells proliferate according to a generalised logistic growth model. Using approximate Bayesian computation, we then demonstrate that data from a standard cell proliferation assay, with duration 24 hours, cannot reliably distinguish between growth models. We then effectively increase the duration of the *in silico* experiments and quantify the decrease in uncertainty in the recovered parameters in the crowding mechanism.

In Chapter 3, we use data from a series of cell proliferation assays to calibrate a spatially continuous (lattice-free) IBM of collective cell migration and proliferation. We repeat this for three independent experimental data sets. As the posterior parameter estimates from each data set are similar, we combine them to form a combined posterior distribution for each unknown parameter. We then use this combined posterior distribution to confirm the predictive power of the calibrated IBM by accurately forecasting the evolution of an additional, independent, experimental replicate.

In Chapter 4, we summarise our work and the main findings of this thesis. We also provide a summary of possible future extensions to the work presented in this thesis.

1.3 Statement of joint authorship

In this section, we outline the contribution of the Masters student and the co-authors of each paper.

Chapter 2: Exploring the optimal duration of a cell proliferation assay

This chapter is derived from a paper titled "A Bayesian computational approach to explore the optimal duration of a cell proliferation assay", to appear in the *Bulletin of Mathematical Biology* (accepted June 2017) (Browning et al., 2017). Co-authors for this paper are listed with their contributions below.

- **Browning, AP** implemented the methodology, performed all data analysis, contributed to the writing of the manuscript and produced all figures and supplementary material.
- McCue, SW oversaw the research, provided technical assistance, helped interpret the results and critically reviewed the manuscript.
- Simpson MJ initiated the concept for the manuscript, oversaw and directed the research, contributed to the writing of the manuscript, critically reviewed and revised the manuscript, and is listed as the corresponding author.

Chapter 3: Inferring parameters for a lattice-free model using experimental data

This chapter is derived from a paper titled "Inferring parameters for a lattice-free model of cell migration and proliferation using experimental data" under consideration in the *Journal of Theoretical Biology* (submitted June 2017). Co-authors for this paper are listed with their contributions below.

- **Browning, AP** extracted and processed experimental data, implemented the mathematical methodology, performed all data analysis, contributed to the writing of the manuscript and produced all figures and supplementary material.
- McCue, SW oversaw the research, provided technical assistance, helped interpret the results and critically reviewed the manuscript.
- Binny, RN helped in implementing the methodology, helped interpret the results and critically reviewed the manuscript.

- Plank, MJ helped interpret the results and critically reviewed the manuscript.
- Shah, ET performed the experiments.
- Simpson, MJ initiated the concept for the manuscript, oversaw and directed the research, contributed to the writing of the manuscript, critically reviewed and revised the manuscript, and is listed as the corresponding author.

Chapter 2

Exploring the optimal duration of a cell proliferation assay

A paper to appear in the Bulletin of Mathematical Biology as

A Bayesian computational approach to explore the optimal duration of a cell proliferation assay.

Alexander P Browning, Scott W McCue, Matthew J Simpson

Abstract

Cell proliferation assays are routinely used to explore how a low density monolayer of cells grows with time. For a typical cell line with a doubling time of 12 hours (or longer), a standard cell proliferation assay conducted over 24 hours provides excellent information about the low-density exponential growth rate, but limited information about crowding effects that occur at higher densities. To explore how we can best detect and quantify crowding effects, we present a suite of *in silico* proliferation assays where cells proliferate according to a generalised logistic growth model. Using approximate Bayesian computation we show that data from a standard cell proliferation assay cannot reliably distinguish between classical logistic growth and more general non-logistic growth models. We then explore, and quantify, the trade-off between increasing the duration of the experiment and the associated decrease in uncertainty in the crowding mechanism.

2.1 Introduction

Two-dimensional *in vitro* cell biology experiments play an invaluable role in improving our understanding of the collective behaviour of cell populations (Liang et al., 2007). Understanding collective cell behaviour is relevant to a number of normal and pathological processes, such as tissue regeneration and malignant spreading, respectively. One of the most common *in vitro* cell biology experiments is called a *proliferation assay* (Bosco et al., 2015; Bourseguin et al., 2016). Cell proliferation assays are initiated by uniformly placing a monolayer of cells, at low density, on a two-dimensional substrate. Individual cells in the population undergo both movement and proliferation events, and the assay is observed as the density of the monolayer of cells increases. Comparing cell proliferation assays with and without a putative drug plays an important role in drug design (Bosco et al., 2015; Bourseguin et al., 2016).

One approach to interpret a cell proliferation assay is to use a mathematical model. This approach can provide quantitative insight into the mechanisms involved (Maini et al., 2004b; Sengers et al., 2007). For example, it is possible to estimate the proliferation rate of cells by calibrating a mathematical model to data from a cell proliferation assay. Results can then be used to compare a target and control assay (Johnston et al., 2015). Typically, most previous studies that interpret cell biology assays using continuum mathematical models make the assumption that cells proliferate logistically (Cai et al., 2007; Dale et al., 1994; Doran et al., 2009; Jin et al., 2016b; Maini et al., 2004b,a; O'Dea et al., 2012; Savla et al., 2004; Sengers et al., 2007; Sheardown and Cheng, 1996; Sherratt and Murray, 1990). The classical logistic equation is given by

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = \lambda C(t)(1 - C(t)), \tag{2.1}$$

where $C(t)$ is the scaled cell density, such that $C(t) = 1$ represents the carrying capacity density, t is time and λ is the cell proliferation rate. For example, by calibrating the solution of Equation (2.1) to data from a cell biology assay, Treloar et al. (2014) showed that the proliferation rate of 3T3 fibroblast cells is approximately 0.048 /hour. However, while the classical logistic model is routinely used to study biological population dynamics (Pearl, 1927; Edelstein-Keshet, 1988; Murray, 2002), this choice is often made without a careful examination of whether the classical logistic model is valid (Treloar et al., 2014).

In the literature, there is an awareness that biological populations do not always grow according to the classical logistic equation (Gerlee, 2013; Zwietering et al., 1990). For example, West and coworkers investigate the growth of cell populations from a wide range of animal models and find that the growth is not logistic; instead, they find that a more general model provides a better match to the experimental data (West et al., 2001). Likewise, Laird (1964) examines tumour growth data and shows that the Gompertz growth law matches the data better than the classical logistic model. Similar observations have also been made more recently for different types of tumour growth by Sarapata and de Pillis (2014).

Therefore, it is not always clear that the classical logistic model ought to be used to describe cell proliferation assays. The classical logistic model, and its generalisations (Tsoularis and Wallace, 2002), all lead to similar growth dynamics during the early phase of the experiment when the density is small. The key differences between these models occur at larger densities as the cell population grows towards the carrying capacity density. The question of whether cells in a proliferation assay grow logistically, or by some other mechanism, is obscured by the fact that most cell proliferation assays are conducted for a relatively short period of time. To illustrate this, we note that a typical cell proliferation rate of $\lambda = 0.048$ /hour (Treloar et al., 2014) corresponds to a doubling time of approximately 14 hours. Given that a typical initial cell density in a cell proliferation assay is approximately $C(0) \approx 0.1$, and the typical time scale of a cell proliferation assay is no more than 24 hours, the cell density will grow to be no more than 0.4, Figure 2.1(a)-(d). Indeed, the evolution of the cell density data in Figure 2.1(d) shows that the cell density grows approximately linearly over the standard experimental duration of 24 hours. This linear increase is consistent with the early time behaviour of the exponential growth phase, but provides less information about later time behaviour where crowding effects play a role. Therefore, standard experimental durations are inappropriate for the purposes of examining how cells grow at high densities. The focus of the current work is to explore how we can determine the optimal duration of a cell proliferation assay so that it can be used to reliably distinguish between classical logistic and generalised logistic growth models. In summary, this study is the first time that an individual based model has been used to explore the duration of a cell proliferation assay, in order to reliably distinguish between different types of growth models. This chapter is organised as follows. We first present a suite of results from a stochastic *in silico* cell proliferation assay. The benefit of working with an *in silico* assay is that it can be used to describe the evolution of a cell proliferation assay corresponding to a known, but

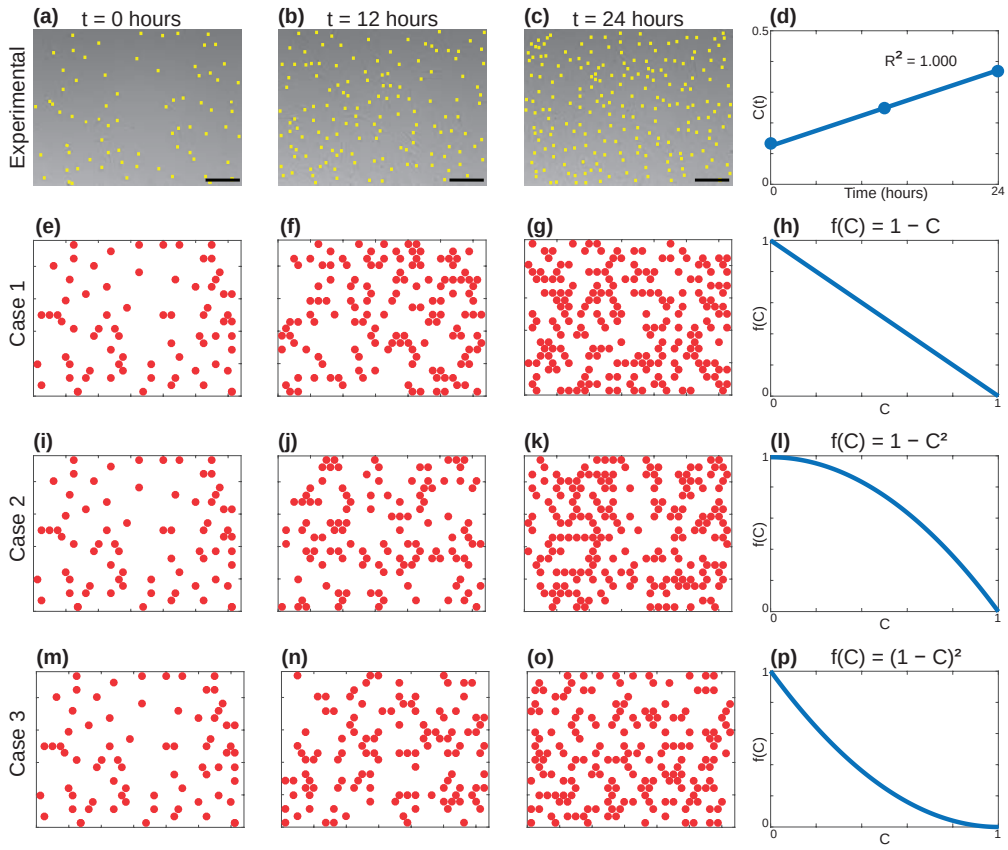

Figure 2.1: Snapshots of experimental and *in silico* cell proliferation assays. (a)-(c) show experimental images obtained from a typical cell proliferation assay performed using 3T3 fibroblast cells, shown at 12 hour intervals, as indicated. The position of each cell is identified with a yellow marker. Each scale bar corresponds to 100 μm. (d) shows the scaled cell density from this experiment at each time point, and a least squares linear fit. (e),(i) and (m) show the initial distribution of agents in the *in silico* assays. The location of agents in (e),(i) and (m) are identical, and chosen so that the location of each cell in (a) is mapped to a hexagonal lattice, and placed on the nearest vacant lattice site. Simulation results are shown for: Case 1 in (e)-(g); Case 2 in (i)-(k); and, Case 3 in (m)-(o). The crowding function used in each Case is shown (h),(l) and (p) for Cases 1, 2 and 3 respectively. All simulations correspond to parameters $I = 25$, $J = 22$, $P_m = 0.579$, $P_p = 0.002$, $\tau = 0.0417$ hour and $\Delta = 25$ μm, which are introduced in Section 2.2.1. These discrete parameters correspond to $\lambda = 0.048$ /hour and $D = 2200$ μm^2/hour. Images in (a)-(c) are reprinted with permission from the Bulletin of Mathematical Biology (2013) Simpson et al. Experimental and modelling investigation of monolayer development with clustering. 75, 871-889.

general, proliferation mechanism,

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = \lambda C(t) f(C), \tag{2.2}$$

where $f(C) \in [0, 1]$ is a crowding function of our choice (Jin et al., 2016a). The crowding function is a smooth decreasing function that satisfies $f(0) = 1$ and $f(1) = 0$. In general, we could study any choice of $f(C)$ that satisfies these conditions. However, for the purposes of this study we restrict our attention to the family of crowding functions given by

$$f(C) = (1 - C^{\alpha})^{\beta}, \tag{2.3}$$

where α and β are positive constants (Tsoularis and Wallace, 2002). This choice of $f(C)$ is still general and we note that different choices of α and β correspond to well-known biological growth models such as the classical logistic growth model, the Gompertz growth model, and the von Bertalanaffy growth model (Tsoularis and Wallace, 2002). Our choice of $f(C)$ is partly motivated by the recent work of Sarapata and de Pillis (2014), who explore a range of sigmoid growth models for different types of tumours, including bladder, breast, liver, lung, and melanoma tumours. Sarapata and de Pillis (2014) show that the classical logistic growth model does not always provide the best match to observed data, and they test a range of other sigmoid growth models for each different kind of tumour. The different forms of sigmoid growth models that Sarapata and de Pillis (2014) explore are encompassed in our choice of crowding function, Equation (2.3), simply by making different choices of the constants α and β.

In this work we focus on three particular choices of $f(C)$:

Case 1: $\alpha = 1$ and $\beta = 1$. Here, $f(C)$ is a linear function that corresponds to the classical logistic equation 2.1. See Figure 2.1(h);

Case 2: $\alpha = 2$ and $\beta = 1$. Here, $f(C)$ is a non-linear, concave-down function. See Figure Figure 2.1(l); and,

Case 3: $\alpha = 1$ and $\beta = 2$. Here, $f(C)$ is a non-linear, concave-up function. See Figure Figure 2.1(p).

Setting $\alpha = 1$ and $\beta = 1$ recovers the classical logistic equation 2.1, whereas other choices of α and β lead to different, general logistic growth models. Typical *in silico* experiments showing snapshots of the growing populations are given in Figure 2.1(e)-(g) for Case 1, Figure 2.1(i)-(k) for Case 2 and Figure 2.1(m)-(o) for Case 3. After we have generated typical *in silico* results for these different choices

of $f(C)$, we then examine our ability to distinguish between data corresponding to different choices of $f(C)$ using approximate Bayesian computation (ABC) (Liepe et al., 2014; Sunnaker et al., 2013; Tanaka et al., 2006; Collis et al., 2017) to estimate the parameters α and β. This procedure clearly shows that the duration of a standard cell proliferation assay is too short to reliably recover the values of α and β. Therefore, to provide quantitative insight into the benefit of performing the experiment for a longer duration, we quantify the decrease in our uncertainty of the parameters and the increase in information as we effectively run the experiment for longer periods of time.

2.2 Methods

2.2.1 Discrete mathematical model

We use a lattice-based random walk model to describe a cell proliferation assay (Liggett, 1999). Throughout the work, we will refer to a realisation of the stochastic model as either an *in silico* experiment, or a simulation. In the model cells are treated as equally-sized discs, and this is a typical assumption (Deroulers et al., 2009; Vo et al., 2015) that is supported by experimental measurements (Simpson et al., 2013). We use a hexagonal lattice, with no more than one agent per site. The lattice spacing, Δ, is chosen to be equal to the mean cell diameter (Jin et al., 2016a). This means we have a circular packing of agents, which corresponds to the maximum carrying capacity for a population of uniformly sized discs. The relationship between the scaled density, $C(t)$, and the number of agents, $N(t)$, is

$$C(t) = \frac{N(t)}{N_{\max}}, \tag{2.4}$$

so that $C(t) = 1$ corresponds to the carrying capacity of $N_{\max}$ agents, which is the number of lattice sites. Motivated by the experimental images of the cell proliferation assay in Figure 2.1(a)-(c), that is conducted with 3T3 fibroblast cells, we set $\Delta = 25$ μm to be the mean cell diameter (Simpson et al., 2013). As the images in Figure 2.1(a)-(c) show a fixed field of view that is much smaller than the spatial extent of the uniformly distributed cells in the experiment, we apply zero net flux boundary conditions (Johnston et al., 2015).

Each lattice site, indexed (i, j) where $i, j \in \mathbb{Z}^+$, has position

$$(x, y) = \begin{cases} \left(i\Delta, j\Delta\sqrt{3}/2\right) & \text{if } j \text{ is even,} \\ \left((i + 1/2)\Delta, j\Delta\sqrt{3}/2\right) & \text{if } j \text{ is odd,} \end{cases}$$

such that $1 \leq i \leq I$ and $1 \leq j \leq J$. To match a typical physical domain, such as the experiment in Figure 2.1(a)-(c) where the field of view is 625 μm $\times$ 480 μm and the cell diameter is $\Delta = 25$ μm, we set $I = 25$ and $J = 22$. When this domain is packed to confluence, the field of view can hold no more than $N_{\max} = 550$ agents.

In any single realisation of the discrete model, the occupancy of site $\mathbf{s}$ is denoted $C_{\mathbf{s}}$, with $C_{\mathbf{s}} = 1$ if the site is occupied, and $C_{\mathbf{s}} = 0$ if vacant. We report results from the model by summing the total number of agents at time t, which we denote $N(t)$. Each site $\mathbf{s}$ is associated with a unique index (i, j). We denote the set of nearest neighbour sites surrounding site $\mathbf{s}$ as $\mathcal{N}\{\mathbf{s}\}$, and the size of $\mathcal{N}\{\mathbf{s}\}$ is $|\mathcal{N}\{\mathbf{s}\}|$. For a typical lattice site, not on any boundary, $\mathcal{N}\{\mathbf{s}\}$ corresponds to the usual six nearest neighbour sites and $|\mathcal{N}\{\mathbf{s}\}| = 6$. However, for any lattice site on a boundary, we adjust $\mathcal{N}\{\mathbf{s}\}$ and $|\mathcal{N}\{\mathbf{s}\}|$ as appropriate to enforce no-flux boundary conditions.

To initiate simulations of a cell proliferation assay, we randomly select a lattice site and place an agent on that site, provided the site is vacant. We repeat this process until $N(0) = 55$ agents have been randomly placed. This corresponds to each simulation starting with $C(0) = 0.1$, which is typical of the initial density, such as in Figure 2.1(a). The following algorithm is used to simulate the way in which cells migrate and proliferate during the experiment. At any time, t, there are $N(t)$ agents on the lattice. In each discrete time step, of duration τ, we allow motility and proliferation events to occur in the following two sequential steps.

First, $N(t)$ agents are selected independently at random, one at a time with replacement, and given the opportunity to move with probability $P_m \in [0, 1]$. A motile agent attempts to move to one of the six nearest neighbour sites, selected at random. To simulate crowding effects, potential motility events are aborted if an agent attempts to move to an occupied site or attempts to move outside the domain.

Second, another $N(t)$ agents are selected independently, at random, one at a time with replacement, and given the opportunity to proliferate with probability $P_p \in [0, 1]$. To assess how crowding affects the ability of a cell to proliferate, we follow the approach of (Jin et al., 2016a) and assume that an agent at site $\mathbf{s}$ senses the occupancy of the six nearest neighbour sites, and can detect a measure of the

average occupancy of those sites,

$$\bar{C}_{\mathbf{s}} = \frac{1}{|\mathcal{N}(\mathbf{s})|} \sum_{\mathbf{s}' \in \mathcal{N}\{\mathbf{s}\}} C_{\mathbf{s}'}. \tag{2.5}$$

This means that $\bar{C}_{\mathbf{s}} \in [0, 1]$ is a measure of the local crowdedness in $\mathcal{N}(\mathbf{s})$. We use $\bar{C}_{\mathbf{s}}$ to determine whether a potential proliferation event succeeds by introducing a *crowding function*, $f(C) \in [0, 1]$ with $f(0) = 1$ and $f(1) = 0$. To incorporate crowding effects we sample a random number, $R \sim U(0, 1)$. If $R < f(\bar{C}_{\mathbf{s}})$, a daughter agent is placed at a randomly chosen, vacant, nearest neighbouring site, whereas if $R > f(\bar{C}_{\mathbf{s}})$, the potential proliferation event is aborted. After the $N(t)$ potential proliferation events have been attempted, $N(t + \tau)$ is updated.

These two steps are repeated until the desired end time, T, is reached.

As previously demonstrated (Jin et al., 2016a), the continuum limit description of this discrete model gives rise to

$$\frac{\partial C(x, y, t)}{\partial t} = D \left(\frac{\partial^2 C(x, y, t)}{\partial x^2} + \frac{\partial^2 C(x, y, t)}{\partial y^2} \right) + \lambda C(x, y, t) f(C), \tag{2.6}$$

where,

$$\lambda = \lim_{\Delta, \tau \to 0} \frac{P_p}{\tau}, \tag{2.7}$$

$$D = \lim_{\Delta, \tau \to 0} \frac{P_m \Delta^2}{4\tau}. \tag{2.8}$$

Here, λ is the proliferation rate, and the motility of agents is characterised by a diffusivity, D. Since the agents are initially distributed uniformly we have $\partial C(x, y, t)/\partial x \approx \partial C(x, y, t)/\partial y \approx 0$. This means that the partial differential equation simplifies to an ordinary differential equation,

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = \lambda C(t) f(C), \tag{2.9}$$

which is a generalised logistic growth model.

In this study, we only ever vary the parameters in the crowding function, α and β. All other parameters are fixed, and chosen to represent a typical cell population. As previously stated, we set $N(0) = 55$, $I = 25$ and $J = 22$, to accommodate the typical geometry and initial condition of a cell proliferation assay with a population of cells whose mean diameter is $\Delta = 25$ μm (Simpson et al., 2013). To describe the rate at which cells move, we set $P_m = 0.579$ and $\tau = 0.0417$ hours. This corresponds to $D = 2200$ $\mu\mathrm{m}^2$/hour, which is a typical value of the cell diffusivity

for a mesenchymal cell line (Simpson et al., 2014). To describe the rate at which cells proliferate, we set $P_p = 0.002$ and $\tau = 0.0417$ hours. This corresponds to $\lambda = 0.048$ /hour, which is a typical value of the cell proliferation rate (Treloar et al., 2014). This proliferation rate is consistent with the experimental data in Figure 2.1(d).

Using these parameter estimates, we show the evolution of $C(t)$ for a single realisation of the discrete model, for each choice of crowding function, in Figure 2.2(a)-(b), for $T = 24$ and 96 hours, respectively. Results in Figure 2.2(a)-(b) show some stochastic fluctuations, as expected. To approximate the expected behaviour, we perform 20 identically prepared realisations of the discrete model and show the mean density profile, $\hat{C}(t)$, in Figure 2.2(c)-(d), for $T = 24$ and 96 hours, respectively. Comparing the single realisations with the mean behaviour confirms that there are minimal fluctuations, at this scale. Furthermore, we see minimal differences in the overall behaviour of the model when we consider a profile of a single realisation and the mean from an ensemble of 20 identically prepared realisations.

2.2.2 Parameter estimation using ABC rejection

Using a Bayesian framework, we consider the crowding function parameters $\boldsymbol{\theta} = (\alpha, \beta)$ as random variables, and the uncertainty in the $\boldsymbol{\theta}$ is updated using observed data (Gelman et al., 2004; Sunnaker et al., 2013; Tanaka et al., 2006; Collis et al., 2017). Under this assumption, we note that the cell density profile, $C(t)$, is also a random variable. In this section we refer to the variables using vector notation to keep the description of the inference algorithm as succinct as possible. However, in the main text we refer to the variables using ordered pairs, (α, β), so that our results are presented as clearly as possible.

To begin with, we perform three *in silico* experiments with fixed, known parameter values, which we refer to as the target parameters, $\boldsymbol{\theta}_*$, corresponding to each Case considered. We take care to ensure that the three *in silico* experiments lead to typical $C(t)$ data, as we demonstrate in Figure 2.2. The data from these experiments is treated as *observed* data, denoted $\mathbf{X}_{\text{obs}}$. Then, we use an ABC approach to explore, and quantify, how well the target values of $\boldsymbol{\theta}$ can be estimated using the observed data. In particular, we are interested in the effect of varying the duration over which the observation data is collected, T.

In the absence of any experimental observations, information about $\boldsymbol{\theta}$ is characterised by a specified prior distribution (Gelman et al., 2004; Sunnaker et al.,

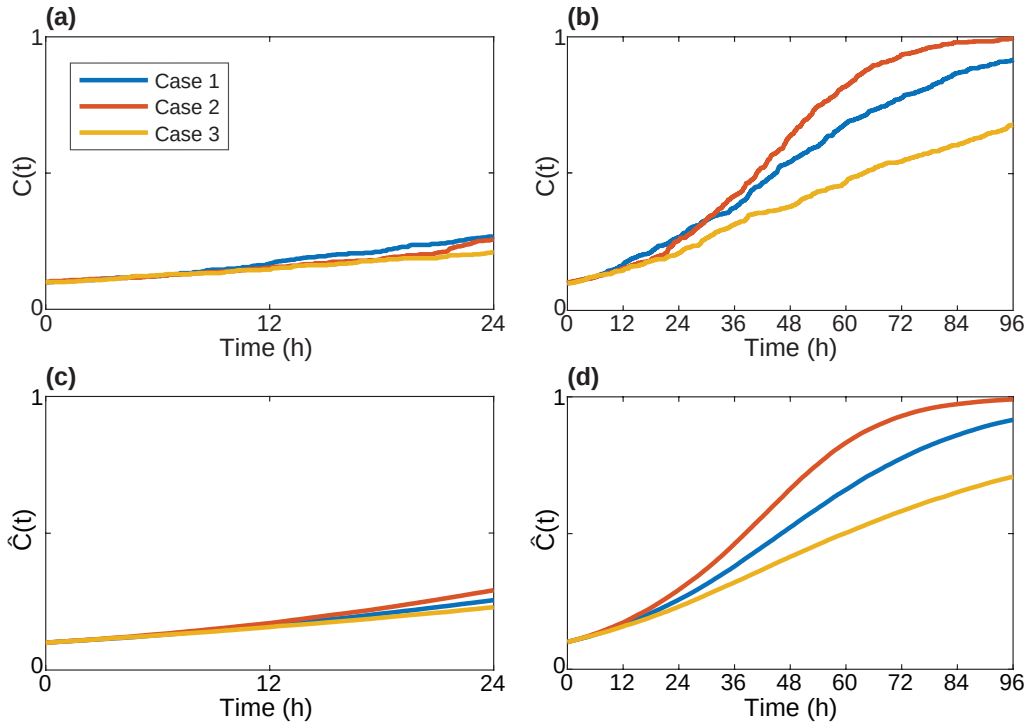

Figure 2.2: The role of the crowding function and timing for a typical proliferation assay. Results show the evolution of: (a) $C(t)$, for $T = 24$ hours; (b) $C(t)$, for $T = 96$ hours; (c) $\hat{C}(t)$, for $T = 24$ hours; and, (d) $\hat{C}(t)$, for $T = 96$ hours. All simulations correspond to $N(0) = 55$, $I = 25$, $J = 22$, $P_m = 0.579$, $P_p = 0.002$, $\tau = 0.0417$ hours and $\Delta = 25$ μm. These discrete parameters correspond to $\lambda = 0.048$ /hour and $D = 2200$ μm^2/hour. The carrying capacity is $N_{\max} = 550$.

2013). For our choices of α and β, we set the prior to be

$$\pi(\boldsymbol{\theta}) = \frac{1}{9}, \quad \boldsymbol{\theta} \in (0,3) \times (0,3), \tag{2.10}$$

which is a uniform distribution across $(\alpha, \beta) \in (0,3) \times (0,3)$.

We summarise data, $\mathbf{X}$, with a lower-dimensional summary statistic, S. Under a Bayesian framework, the information from the prior is updated by the likelihood of the observations, $p(S_{\text{obs}}|\boldsymbol{\theta})$, to produce posterior distributions, $p(\boldsymbol{\theta}|S_{\text{obs}})$, of $\boldsymbol{\theta}$. In this study, we use the most fundamental ABC algorithm, known as ABC rejection (Liepe et al., 2014; Tanaka et al., 2006; Sunnaker et al., 2013). Our aim is to quantify the trade off between the duration of the experiment, T, and the reduction in uncertainty of the value of $\boldsymbol{\theta}$ as well as the information gain.

In this work, we choose S to be the number of agents observed at equally spaced intervals of 24 hours. Let $N_{\text{obs}}(t)$ and $N_{\text{sim}}(t)$ denote the number of agents present in the observed data and a simulated cell proliferation assay at time t, respectively. We choose a discrepancy measure, $\rho(S_{\text{obs}}, S_{\text{sim}})$, to be the cumulative sum of the square difference between $N_{\text{sim}}(t)$ and $N_{\text{obs}}(t)$ at each 24 hour interval, up to the duration of the experiment, T, such that

$$\rho(S_{\text{obs}}, S_{\text{sim}}) = \sum_{i=1}^{T/24} \left[N_{\text{sim}}(24i) - N_{\text{obs}}(24i)\right]^2. \tag{2.11}$$

With these definitions, the ABC rejection algorithm is given by Algorithm 2.1.

Algorithm 2.1 ABC rejection sampling

1: Set $P_m = 0.579$, $P_p = 0.002$, $\Delta = 25$ μm, $\tau = 0.0417$ hours, $N(0) = 55$.
2: Draw $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta})$.
3: Simulate cell proliferation assay with $\boldsymbol{\theta}_i$.
4: Record $S_{\text{sim}_i} = \{N_{\text{sim}}(24j)\}$, $j = 1, 2, 3, 4$.
5: Compute $\epsilon_i = \rho(S_{\text{obs}}, S_{\text{sim}_i})$, given in Equation (2.11).
6: Repeat steps 2-5 until 10^6 samples $\{\theta_i, \epsilon_i\}_{i=1}^{10^6}$ are simulated.
7: Retain a small proportion, $u = 0.01$, with the smallest discrepancy, ϵ_i, as posterior samples.

To present and perform calculations with posterior distributions, we use a kernel density estimate with grid spacing 0.01 to form an approximate continuous posterior distribution from the samples. We do this using the `ksdensity` function in the MATLAB Statistics Toolbox (Mathworks, 2017). All ABC posterior results presented in the main part of the chapter correspond to retaining the 10,000 simulations out of 1,000,000 simulations with the smallest discrepancy, giving $u = 0.01$.

To confirm that our results are insensitive to this choice of u we also present equivalent results with $u = 0.02$ in Appendix 2.A.

Kullback-Leibler divergence

To quantitatively compare posterior distributions, we calculate the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951; Burnham and Anderson, 2002), $D_{KL}(p\|\pi)$, for each posterior distribution. The KL divergence is a measure of the information gain in moving from the prior, $\pi(\boldsymbol{\theta})$, to the posterior, $p(\boldsymbol{\theta}|S_{\text{obs}})$, in Bayesian inference, and is defined as

$$D_{KL}(p(\boldsymbol{\theta}|S_{\text{obs}})\|\pi(\boldsymbol{\theta})) = \iint_{\Theta} p(\boldsymbol{\theta}|S_{\text{obs}}) \log\left(\frac{p(\boldsymbol{\theta}|S_{\text{obs}})}{\pi(\boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{\theta}, \tag{2.12}$$

where $\Theta = (0,3)\times(0,3)$ is the prior support. To calculate $D_{KL}(p(\boldsymbol{\theta}|S_{\text{obs}})\|\pi(\boldsymbol{\theta}))$ we use quadrature to estimate the integral in Equation (2.12), taking care to ensure that the result is independent of the discretisation. Note that D_{KL} is a measure of the amount of information gained when moving from the prior distribution to the posterior distribution.

Other measures

We also make use of several other measures to help quantify various properties of the posterior densities. For each Case we always know, in advance, the target parameter values, $\boldsymbol{\theta}_*$, and we also estimate the mode, $\boldsymbol{\theta}_m$, using the kernel density estimate. Note that the mode is the value of $\boldsymbol{\theta}$ corresponding to the maximum posterior density,

$$\boldsymbol{\theta}_m = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|S_{\text{obs}}). \tag{2.13}$$

It is useful to report the posterior density at the target, $p(\boldsymbol{\theta}_*|S_{\text{obs}})$, for various values of T. It is also instructive to report the posterior density at the mode, $p(\boldsymbol{\theta}_m|S_{\text{obs}})$, for various values of T. Another useful measure is the Euclidean distance between the target and the mode, given by

$$d = ||\boldsymbol{\theta}_* - \boldsymbol{\theta}_m||_2, \tag{2.14}$$

where $||\cdot||_2$ denotes the Euclidean norm.

2.3 Results and discussion

Results from a typical cell proliferation assay are shown in Figure 2.1(a)-(c). The cell density profile, shown in Figure 2.1(d), increases approximately linearly with time. This indicates that the experimental duration is not long enough for us to observe crowding effects, which occur at higher densities, and cause the net growth rate to reduce so that cell density profile, $C(t)$, becomes concave down at later times. Therefore, by using typical experimental data, it is unclear whether the growth process follows a classical logistic model, or some other more general growth model.

To provide further insight into the limitations of this standard experimental design, we show results from the discrete model in Figure 2.2(a) for a standard experimental duration of $T = 24$ hours, for three different crowding functions. These results show several interesting features: (i) the cell density profile for each Case appears to increase linearly with time, which is similar to the experimental results in Figure 2.1(d); (ii) it is difficult to distinguish between the three different profiles, despite each profile corresponding to a different crowding function; and (iii) comparing the cell density profiles of a single realisation in Figure 2.2(a) to the expected behaviour in Figure 2.2(c) confirms that the expected cell density profiles for each Case are similar for the first 24 hours.

To examine when crowding effects begin to significantly influence the cell density profile, we perform simulations over longer durations of time. In particular, we examine $T \leq 96$ hours. Results for a single realisation in Figure 2.2(b) show that the cell density profiles for each Case are indistinguishable during the first 24 hours. However the profile for each Case does become increasingly distinguishable at times greater than 24 hours. For example, each Case is clearly discernible by 72 hours. Comparing the cell density profiles of a single realisation in Figure 2.2(b) to the expected behaviour in Figure 2.2(d) confirms that each Case is only distinguishable at times greater than 24 hours. These observations motivate several questions that we will explore. The two main questions we focus on are: (i) what experimental duration is required to reliably distinguish between Cases 1, 2 and 3; and, (ii) can we quantify the trade off between allowing the experiment to run for a sufficiently long period of time to distinguish between the Cases, while still minimising the duration of the experiment.

To quantify the increase in information we can obtain by running the experiment for longer durations of time, we attempt to recover the parameters in the crowding function for each Case using ABC to produce a posterior distribution for α and β, which we refer to as the ordered pair (α, β). To achieve this aim, we

produce *in silico* observed data, using a target parameter set for each Case: Case 1 corresponds to $(\alpha, \beta) = (1, 1)$; Case 2 corresponds to $(\alpha, \beta) = (2, 1)$; and Case 3 corresponds to $(\alpha, \beta) = (1, 2)$. All other parameters in the simulations are held fixed at the values given previously.

The data we use to perform inference takes the form of the size of the population, $N(t)$, recorded at equally spaced intervals, each of duration 24 hours. In particular, we examine the effect of varying the total duration of the experiment, T. This means that if we consider an experimental design with $T = 24$ hours, then we record $N(24)$ only. In contrast, if we consider an experimental design with $T = 72$ hours, we record $N(24), N(48)$ and $N(72)$. Overall, we examine four durations, $T = 24, 48, 72$ and 96 hours.

Results in Figure 2.3(a)-(d) show the bivariate posterior distributions of α and β for Case 1, with $T = 24, 48, 72$ and 96 hours, respectively. Recall that the target parameters for Case 1 are $(\alpha, \beta) = (1, 1)$. The results indicate that the choice of prior, $\pi(\boldsymbol{\theta})$, on the domain $(0, 3) \times (0, 3)$, is reasonable because the posterior distribution has full support within this region. The distribution in Figure 2.3(a) shows there are many parameter combinations that are likely to match the observed data, with $T = 24$ hours. This observation is consistent with the results in Figure 2.2(a) where we observe that setting $T = 24$ hours is insufficient to distinguish between the three Cases. Comparing the posterior distributions in Figure 2.3(a)-(d), we see that increasing T leads to a narrowing of the posterior distribution, and the mode of the distribution moves toward the target parameter combination. For this Case, we see the largest benefit when increasing T from 48 to 72 hours. For example, for $T = 48$ hours, the mode of the distribution is (1.82, 2.16), which means that each parameter estimate is almost double each target value. In contrast, the mode of the distribution at $T = 72$ hours is (1.06, 0.95), so each parameter is able to be estimated within 6% of the target.

To quantify the properties in the posterior distributions, Figure 2.3(a)-(d), there are many features that we may consider. Figure 2.3(e) compares the posterior density at the target parameter values and the maximum posterior density of the distribution, which corresponds to the mode. The maximum posterior density increases with T, confirming that the posterior distribution narrows as the duration of the experiment is increased. Results in Figure 2.3(f) show that d eventually decreases with T, indicating that the mode of the distribution moves towards the target as T increases. Together, these results show that the density at the mode is close to the density at the target, and that both these quantities increase with T. This indicates that the target parameter combination is always as likely as the

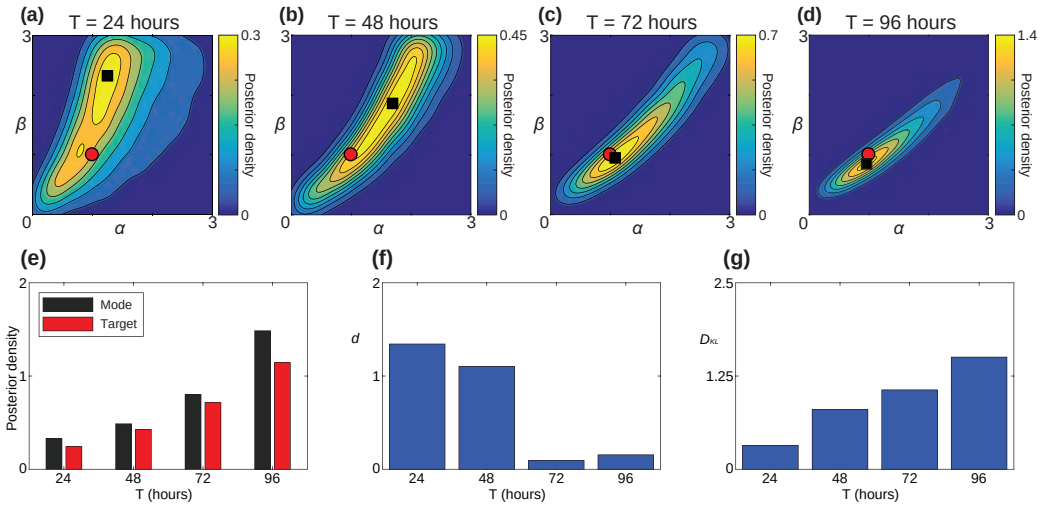

Figure 2.3: Posterior distributions for Case 1: $(\alpha, \beta) = (1, 1)$. (a)-(d) ABC posterior distributions for: (a) $T = 24$ hours; (b) $T = 48$ hours; (c) $T = 72$ hours and (d) $T = 96$ hours. The posterior distributions are approximated using the best 10,000 samples from 1,000,000 prior samples ($u = 0.01$), as measured by ρ, given by Equation (2.11). The red circles show the location of the target parameters used to generate the observed data ($\alpha = 1$, $\beta = 1$). The black squares indicate the mode of the posterior distribution. The modes are (1.32, 2.43), (1.82, 2.16), (1.06, 0.95) and (0.96, 0.86) in (a)-(d), respectively. (e)-(g) Show measures of accuracy and precision. (e) Quantitatively compares the posterior density at the mode and the target parameter values. (f) Shows d, the Euclidean distance between the mode and target parameter values, given by Equation (2.14). (g) Shows D_{KL}, the Kullback-Leibler divergence from the prior, for each posterior distribution, given by Equation (2.12).

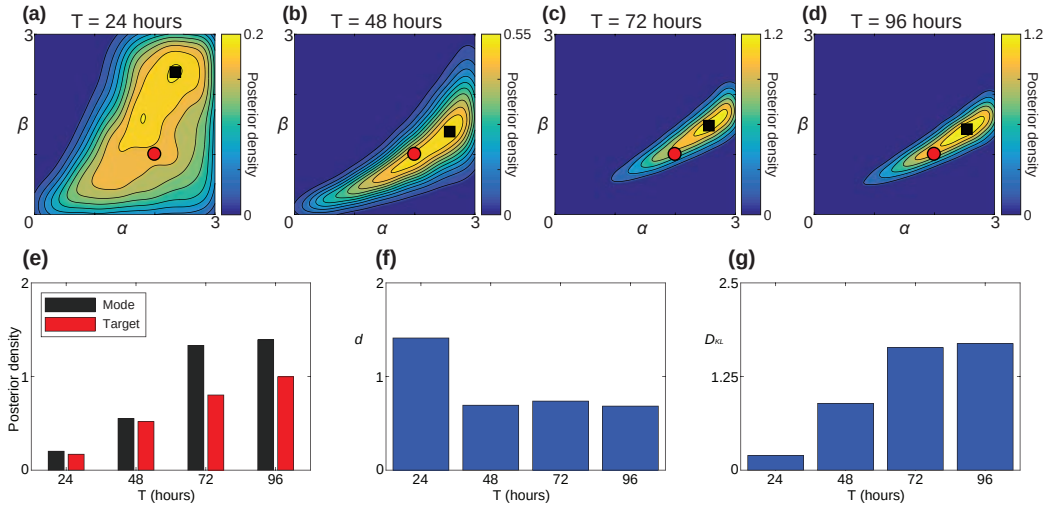

Figure 2.4: Posterior distributions for Case 2: $(\alpha, \beta) = (2, 1)$. (a)-(d) ABC posterior distributions for: (a) $T = 24$ hours; (b) $T = 48$ hours; (c) $T = 72$ hours and (d) $T = 96$ hours. The posterior distributions are approximated using the best 10,000 samples from 1,000,000 prior samples ($u = 0.01$), as measured by ρ, given by Equation (2.11). The red circles show the location of the target parameters used to generate the observed data ($\alpha = 2$, $\beta = 1$). The black squares indicate the mode of the posterior distribution. The modes are (1.89, 1.81), (2.55, 1.38), (2.54, 1.46) and (2.53, 1.41) in (a)-(d), respectively. (e)-(g) Show measures of accuracy and precision. (e) Quantitatively compares the posterior density at the mode and the target parameter values. (f) Shows d, the Euclidean distance between the mode and target parameter values, given by Equation (2.14). (g) Shows D_{KL}, the Kullback-Leibler divergence from the prior, for each posterior distribution, given by Equation (2.12).

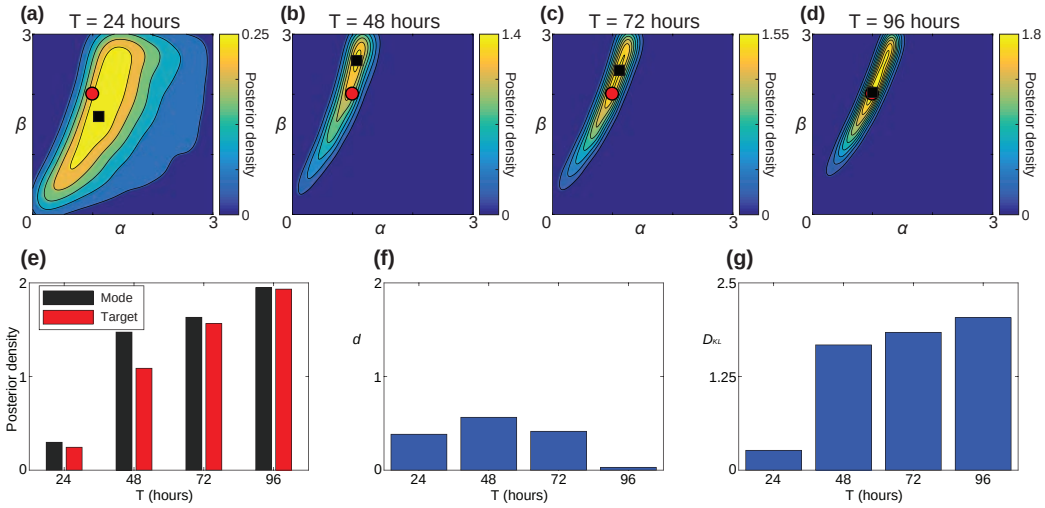

Figure 2.5: Posterior distributions for Case 3: $(\alpha, \beta) = (1, 2)$. (a)-(d) ABC posterior distributions for: (a) $T = 24$ hours; (b) $T = 48$ hours; (c) $T = 72$ hours and (d) $T = 96$ hours. The posterior distributions are approximated using the best 10,000 samples from 1,000,000 prior samples ($u = 0.01$), as measured by ρ, given by Equation (2.11). The red circles show the location of the target parameters used to generate the observed data ($\alpha = 1$, $\beta = 2$). The black squares indicate the mode of the posterior distribution. The modes are (1.13, 1.70), (1.09, 2.57), (1.20, 2.67) and (1.03, 2.11) in (a)-(d), respectively. (e)-(g) Show measures of accuracy and precision. (e) Quantitatively compares the posterior density at the mode and the target parameter values. (f) Shows d, the Euclidean distance between the mode and target parameter values, given by Equation (2.14). (g) Shows D_{KL}, the Kullback-Leibler divergence from the prior, for each posterior distribution, given by Equation (2.12).

mode. Results in Figure 2.3(g) shows how the KL divergence (Equation (2.12)) also increases with T. We see that the largest gain in information for this Case occurs when T is increased from 24 hours ($D_{KL} = 0.33$) to 48 hours ($D_{KL} = 0.84$). The quantitative measures in Figure 2.3(e)-(g) suggest that there is always value in increasing T, however the value of increasing T varies. For example, there is a substantial benefit in extending the experiment from $T = 48$ to 72 hours, whereas the benefit in extending the experiment from $T = 72$ to 96 hours is less pronounced.

Results in Figure 2.4(a)-(d) and Figure 2.5(a)-(d) show the bivariate posterior distributions of α and β for Cases 2 and 3, respectively. Note that all data presented for Cases 2 and 3 is given in the same format as used for the results corresponding to Case 1 in Figure 2.3. As before, we always observe a narrowing of the posterior distribution as T increases. Results in Figure 2.4(e) and Figure 2.5(e) clearly show that the target parameter combination becomes more likely as T is increased. Data for d in Figure 2.4(f) confirms that the distance between the target and the mode is reduced for larger values of T. Data for d in Figure 2.5(f) shows that the distance between the target and the mode increases, at first, when T is increased from 24 to 48 hours. However, the most important feature is that d always decreases eventually for large enough T. Again, as T is increased, D_{KL} increases in both Figure 2.4(g) and Figure 2.5(g).

Overall, the essential trends in Figure 2.4 and Figure 2.5 are consistent with those in Figure 2.3, namely: (i) the standard choice of $T = 24$ hours is insufficient to determine the parameters in the crowding function and hence it is impossible to reliably distinguish between classical logistic growth and more general logistic growth models; and, (ii) as the value of T is increased, our ability to recover the parameters in the crowding function increases. However, certain details differ between the cases. For example, choosing $T = 72$ hours allows us to recover estimates of α and β to an accuracy of at least 6, 46 and 34% in Cases 1, 2 and 3, respectively. Therefore, with this choice of T we are able to recover the parameters for Case 1 relatively accurately. In contrast, if we choose $T = 96$ hours, we recover estimates of α and β to an accuracy of at least 14, 41 and 6% in Cases 1, 2 and 3, respectively. Therefore, with this choice of T we are able to recover the parameters for Cases 1 and 3 relatively accurately, yet Case 2 remains relatively unclear.

2.4 Conclusion

In this work, we implement a random walk model to simulate a cell proliferation assay. In particular, we focus on exploring whether the typical experimental

design of a cell proliferation assay, with $C(0) \approx 0.1$, $\lambda \approx 0.05$ /hour and $T = 24$ hours, is sufficient to make a clear distinction between classical logistic growth and more general logistic growth models. We are motivated to explore this question because many theoretical modelling studies choose to represent cell proliferation with the classical logistic model, yet this assumption is rarely tested using experimental data. Furthermore, there is a growing awareness in the mathematical biology literature that the choice of using a classical logistic model can be inappropriate. For example, Sarapata and de Pillis (2014) show that a range of tumour growth data is more accurately predicted using a generalised logistic model rather than the classical logistic model. Therefore, the question of whether standard designs of cell proliferation assays can make a clear and unambiguous distinction between classical logistic growth and more general logistic growth is important as cell proliferation assays are commonly employed. It is currently unclear whether the standard experimental design is sufficient to distinguish between different sigmoid growth mechanisms. This study discusses the first time that a stochastic individual based model has been used to explore the optimal duration of a cell proliferation assay. In particular, we explore how to choose the duration of the assay to reliably distinguish between different types of growth models.

One of the main conclusions of our study is that the typical experimental design for a cell proliferation assay, with $C(0) \approx 0.1$, $\lambda \approx 0.05$ /hour and $T = 24$ hours, cannot be used to make a distinction between classical logistic growth and more general logistic growth. Further, we use our stochastic modelling and parameter inference tools to explore how the experimental design can be altered so that this distinction can be made with confidence. In particular we explore the option of increasing the duration of the experiment, T. Our parameter inference results show that increasing T always provides more information about the crowding function parameters. However, the trends are subtle, and there is no simple guideline for prescribing the ideal experimental duration that one could implement in practice. Our results show that we can recover the crowding function for the case of classical logistic growth (Case 1: $\alpha = 1$, $\beta = 1$) to within an accuracy of 6% if the experimental duration is increased to $T = 72$ hours. Beyond this duration, we encounter diminishing returns for this Case. For example, further increasing the duration of the experiment to $T = 96$ hours leads to only a small increase in additional information about the crowding function. In other cases where we consider generalised logistic growth (Case 2: $\alpha = 1$, $\beta = 2$), we see that the parameter estimates remain relatively poor, even if the experimental duration is increased to $T = 72$ hours. For Case 2 we recover the parameters to an accuracy of within 33% if $T = 72$ hours, and to within 5% if $T = 96$ hours. Therefore, it is not possible

to make a simple conclusion that cell proliferation assays ought to be conducted until $T = 48$ or $T = 72$ hours since the increase in information with T is subtle. Despite this complication, our results certainly show that the standard choice of $T = 24$ hours is insufficient, and that the experiment ought to be conducted for a long as practically possible.

One aspect of a cell proliferation assay that we have not explored is the dependence of the results on the initial cell density, $C(0)$. All results in this work, both the *in vitro* experimental data in Figure 2.1, and the *in silico* data in Figures 2.2-2.5, deal with initial densities of $C(0) \approx 0.1$, where $C = 1$ corresponds to the maximum carrying capacity of the confluent monolayer. This initial density corresponds to a fairly standard choice of initiating a cell proliferation assay with approximately 20,000 cells placed into the wells of a 24-well tissue culture plate where each well has a diameter of approximately 15 mm. Alternatively, a similar initial density can be obtained by initiating a cell proliferation assay with approximately 10,000 cells placed into the wells of a 96-well tissue culture plate, where each well has a diameter of approximately 9 mm. While it is true that crowding effects in a cell proliferation assay might be more clearly discernable by initiating the experiment with larger numbers of cells, we warn against this for two reasons. First, from a practical point of view, our experience in initiating a two-dimensional *in vitro* cell biology assay with large numbers of cells is problematic as the cells can tend to cluster together, and pile up in the vertical direction instead of spreading as a monolayer (Treloar et al., 2013). Second, established methods for initiating cell proliferation assays with $C(0) \approx 0.1$ are perfectly well suited to observe the low density exponential phase of the growth process, which is important to estimate the intrinsic proliferation rate, λ. For example, the data shown in Figure 2.1(a)-(c) corresponds to a cell proliferation assay initialised with 20,000 cells in a 24-well tissue culture plate, and results in Figure 2.1(d) show $C(t)$ grows linearly over the first 24 hours. This result is consistent with the early part of the growth process where we expect $C(t) \sim C(0)\exp(\lambda t) = C(0)\left[1 + \lambda t + \mathcal{O}(t^2)\right]$. Therefore, we do not suggest that the standard experimental design for a cell proliferation assay ought to be altered by increasing $C(0)$. This is why, throughout this study, we have treated λ and $C(0)$ as known, constant values, in the experimental design.

All of the results presented here have focused on exploring whether we can make a reliable distinction between classical logistic growth and more general logistic growth in a cell proliferation assay. To achieve this we use *in silico* simulations in which the crowding function can be specified. While the discrete simulation algorithm can be used to model a cell proliferation assay with any crowding function, $f(C)$, to illustrate the key points of our study we focus on three particular cases.

Case 1 corresponds to classical logistic growth, while Cases 2 and 3 are examples of more general logistic growth. Of course, the methods outlined in this work apply equally well to any other choice of crowding function. Furthermore, while all crowding functions explored here involve two parameters, α and β, it is possible that other choices of crowding function might contain additional parameters. Under these conditions, the procedures described here to quantitatively measure the potential for parameter recovery as a function of the experimental design apply in exactly the same way regardless of the number of unknown parameters in the crowding function.

Chapter 2: Supplementary material

2.A Posterior distributions

In the main chapter we present a series of ABC posterior distributions that are constructed using the top 10,000 samples from a total of 1,000,000 prior samples, giving $u = 0.01$. In this supplementary material chapter we present equivalent posterior distributions that are constructed using the top 20,000 of the same 1,000,000 samples, giving $u = 0.02$. Results in Figures 2.6-2.8 correspond to the results in Figures 2.3-2.5 in the main part of the chapter. Both a qualitative and quantitative comparison of these two sets of results shows that the results are insensitive to our choice of setting $u = 0.01$.

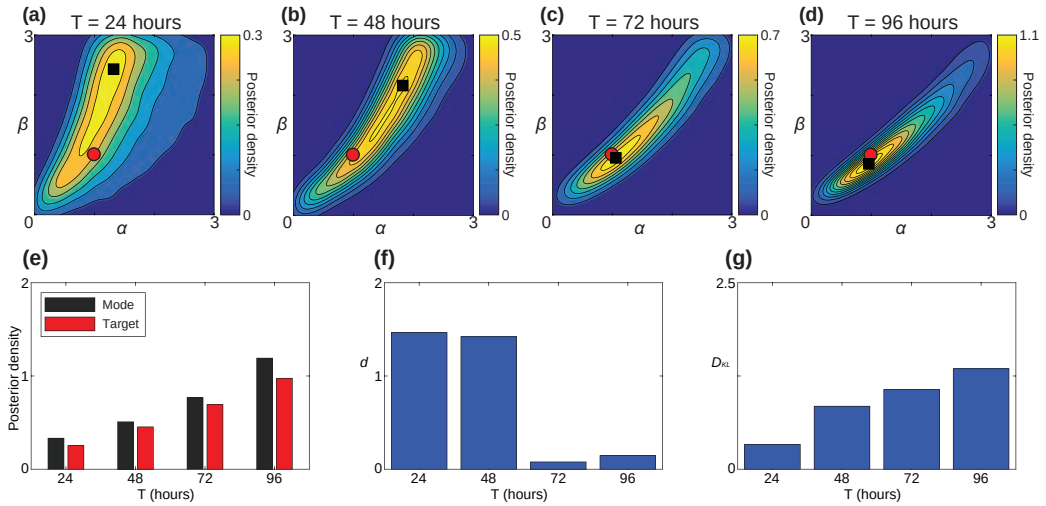

Figure 2.6: Posterior distributions for Case 1: $(\alpha, \beta) = (1, 1)$. (a)-(d) ABC posterior distributions for: (a) $T = 24$ hours; (b) $T = 48$ hours; (c) $T = 72$ hours and (d) $T = 96$ hours. The posterior distribution is approximated using the best 20,000 samples from 1,000,000 prior samples ($u = 0.02$), as measured by ρ, given by Equation (2.11). The red circles shows the location of the target parameters used to generate the observed data ($\alpha = 1$, $\beta = 1$). The black squares indicate the mode of the posterior density. The modes are (1.32, 2.43), (1.82, 2.16), (1.06, 0.95) and (0.96, 0.86) in (a)-(d), respectively. (e)-(g) Show measures of accuracy and precision. (e) Quantitatively compares the posterior density at the mode and the target parameter values. (f) Shows d, the Euclidean distance between the mode and target parameter values, given by Equation (2.14). (g) Shows D_{KL}, the Kullback-Leibler divergence from the prior, for each posterior distribution, given by Equation (2.12).

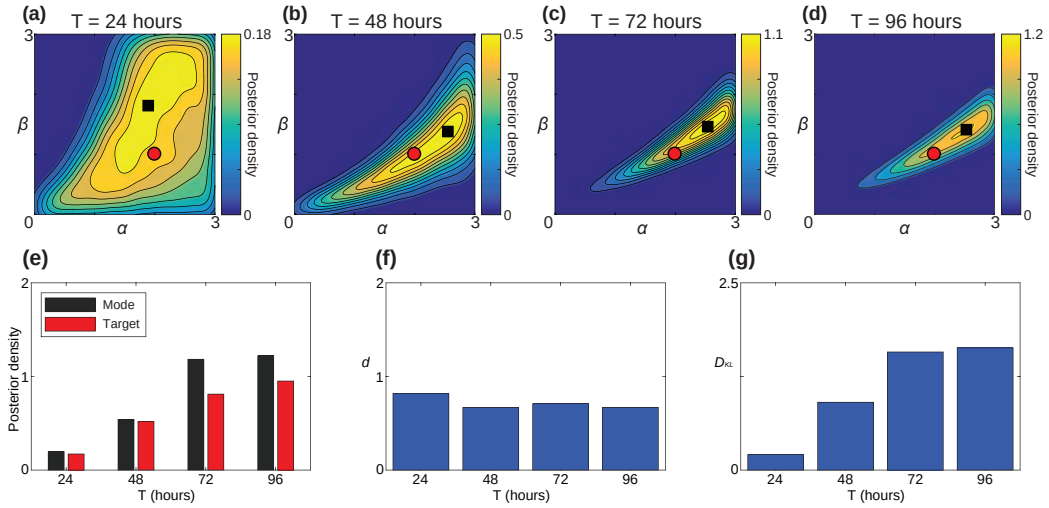

Figure 2.7: Posterior distributions for Case 2: $(\alpha, \beta) = (2, 1)$. (a)-(d) ABC posterior distributions for: (a) $T = 24$ hours; (b) $T = 48$ hours; (c) $T = 72$ hours and (d) $T = 96$ hours. The posterior distribution is approximated using the best 20,000 samples from 1,000,000 prior samples ($u = 0.02$), as measured by ρ, given by Equation (2.11). The red circles shows the location of the target parameters used to generate the observed data ($\alpha = 2$, $\beta = 1$). The black squares indicate the mode of the posterior density. The modes are (1.89, 1.81), (2.55, 1.38), (2.54, 1.46) and (2.53, 1.41) in (a)-(d), respectively. (e)-(g) Show measures of accuracy and precision. (e) Quantitatively compares the posterior density at the mode and the target parameter values. (f) Shows d, the Euclidean distance between the mode and target parameter values, given by Equation (2.14). (g) Shows D_{KL}, the Kullback-Leibler divergence from the prior, for each posterior distribution, given by Equation (2.12).

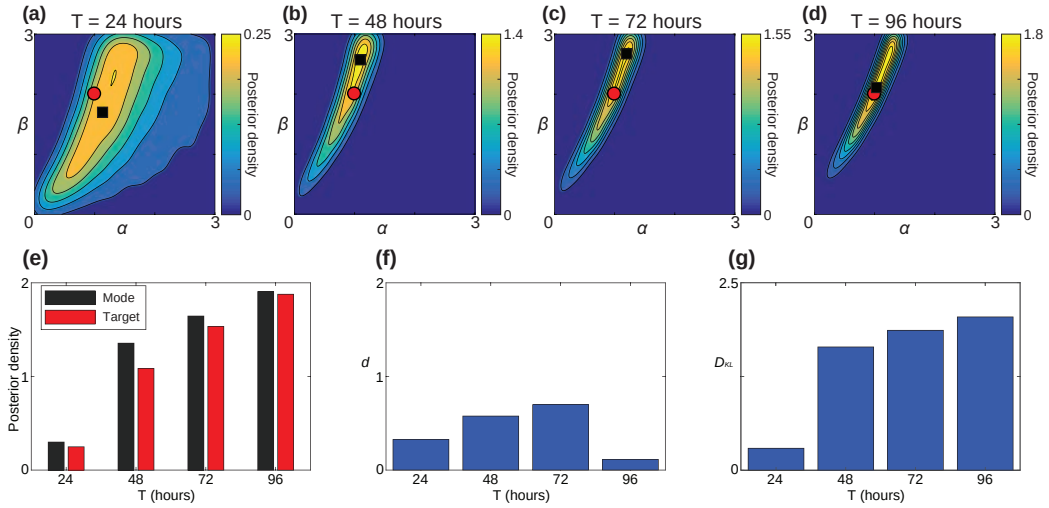

Figure 2.8: Posterior distributions for Case 3: $(\alpha, \beta) = (1, 2)$. (a)-(d) ABC posterior distributions for: (a) $T = 24$ hours; (b) $T = 48$ hours; (c) $T = 72$ hours and (d) $T = 96$ hours. The posterior distribution is approximated using the best 20,000 samples from 1,000,000 prior samples ($u = 0.02$), as measured by ρ, given by Equation (2.11). The red circles shows the location of the target parameters used to generate the observed data ($\alpha = 1$, $\beta = 2$). The black squares indicate the mode of the posterior density. The modes are (1.13, 1.70), (1.09, 2.57), (1.20, 2.67) and (1.03, 2.11) in (a)-(d), respectively. (e)-(g) Show measures of accuracy and precision. (e) Quantitatively compares the posterior density at the mode and the target parameter values. (f) Shows d, the Euclidean distance between the mode and target parameter values, given by Equation (2.14). (g) Shows D_{KL}, the Kullback-Leibler divergence from the prior, for each posterior distribution, given by Equation (2.12).

Chapter 3

Inferring parameters for a lattice-free model using experimental data

A paper under consideration in the Journal of Theoretical Biology as

Inferring parameters for a lattice-free model of cell migration and proliferation using experimental data.

Alexander P Browning, Scott W McCue, Rachelle N Binny, Michael J Plank, Esha T Shah, Matthew J Simpson

Abstract

Collective cell spreading takes place in spatially continuous environments, yet it is often modelled using discrete lattice-based approaches. Here, we use data from a series of cell proliferation assays, with a prostate cancer cell line, to calibrate a spatially continuous individual based model (IBM) of collective cell migration and proliferation. The IBM explicitly accounts for crowding effects by modifying the rate of movement, direction of movement, and the rate of proliferation by accounting for pair-wise interactions. Taking a Bayesian approach we estimate the free parameters in the IBM using rejection sampling on three separate, independent experimental data sets. Since the posterior parameter estimates from each experiment are similar, we combine the estimates. Performing simulations with parameters sampled from the combined distribution allows us to confirm the predictive power of the calibrated IBM by accurately forecasting the evolution of a fourth, experimental data set. Overall, we show how to calibrate a lattice-free IBM to experimental data, and our work highlights the importance of interactions between individuals. Despite great care taken to distribute cells as uniformly as possible experimentally, we find evidence of significant spatial clustering over short distances, suggesting that standard mean-field models could be inappropriate.

3.1 Introduction

One of the most common *in vitro* cell biology experiments is called a *cell proliferation assay* (Bosco et al., 2015; Bourseguin et al., 2016; Browning et al., 2017). These assays are conducted by placing a monolayer of cells, at low density, on a two-dimensional substrate. Individual cells undergo proliferation and movement events, and the assay is monitored over time as the density of cells in the monolayer increases (Tremel et al., 2009). One approach to interpret a cell proliferation assay is to use a mathematical model. Calibrating the solution of a mathematical model to data from a cell proliferation assay can provide quantitative insight into the underlying mechanisms, by, for example, estimating the cell proliferation rate (Tremel et al., 2009; Sengers et al., 2007). A standard approach to modelling a cell proliferation assay is to use a mean-field model, which is equivalent to assuming that individuals within the population interact in proportion to the average population density and that there is no spatial structure, such as clustering, present (Tremel et al., 2009; Sengers et al., 2007; Maini et al., 2004a; Sarapata and de Pillis, 2014; Sherratt and Murray, 1990). More recently, increased computational power has meant that individual based models (IBMs) have been used to directly model the cell-level behaviour (Binny et al., 2016a; Frascoli et al., 2013; Johnston et al., 2014). IBMs are attractive for modelling biological phenomena because they can be used to represent properties of individual agents, such as cells, in the system of interest (Binny et al., 2016a,b; Frascoli et al., 2013; Peirce et al., 2004; Read et al., 2012; Treloar et al., 2013). Typical IBMs use a lattice, meaning that both the position of agents, and the direction of movement, are restricted (Codling et al., 2008). In contrast, lattice-free IBMs are more realistic because they enable agents to move in continuous space, in any direction. However, this extra freedom comes at the cost of higher computational requirements (Plank and Simpson, 2012).

In this chapter we consider a continuous-space, continuous-time IBM (Binny et al., 2016b). This IBM is well-suited to studying experimental data from a cell proliferation assay with PC-3 prostate cancer cells (Kaighn et al., 1979), as shown in Figure 3.1(a)-(d). The key mechanisms in the experiments include cell migration and cell proliferation, and we note that there is no cell death in the experiments on the time scales that we consider. Therefore, agents in the IBM are allowed to undergo both proliferation and movement events. Crowding effects that are often observed in two-dimensional cell biology experiments (Cai et al.,

2007) are explicitly incorporated into the IBM as the rates of proliferation and movement in the model are inhibited in regions of high agent density. In this study we specifically choose to work with the PC-3 cell line because these cells are known to be highly migratory, mesenchymal cells (Kaighn et al., 1979). This means that cell-to-cell adhesion is minimal for this cell line, and cells tend to migrate as individuals. We prefer to work with a continuous-space, lattice-free IBM as this framework gives us the freedom to identically replicate the initial location of all cells in the experimental data when we specify the initial condition in the IBM. In addition, lattice-free IBMs do not restrict the direction of movement like a lattice-based approach.

Taking a Bayesian approach, we assume that cell proliferation assays are stochastic processes, and model parameters are random variables, allowing us to update information about the model parameters using ABC (Collis et al., 2017; Tanaka et al., 2006). For this purpose we perform a large number of IBM simulations using parameters sampled from a prior distribution. Previous work, based on mean-field models, suggests that the proliferation rate and cell diffusivity for PC-3 cells is $\lambda \approx 0.05$ /hour and $D \approx 175$ μm^2/hour, respectively (Johnston et al., 2015). The prior distribution for the IBM parameters are taken to be uniform and to encompass these previous estimates. We generate 10^6 realisations of the IBM using parameters sampled from the prior distribution, and accept 1% of simulations that provide the best match to the experimental data. Our approach to connect the experimental data and the IBM is novel, we are unaware of any previous work that has used ABC to parameterise a lattice-free IBM of a cell proliferation assay.

Applying the ABC algorithm to data from three sets of identically prepared experiments leads to three similar posterior distributions. This result provides confidence that the IBM is a realistic representation of the cell proliferation assays and leads us to produce a combined posterior distribution from which we use the mode to give point estimates of the model parameters. To provide further validation of the IBM, we use the combined posterior distribution and the IBM to make a prediction of the fourth experimental data set. Simulating the IBM with parameters sampled from the combined posterior distribution allows us to predict both the time evolution of the population size, $N(t)$, and the pair correlation within a small neighbourhood of radius 50 μm, $\mathcal{P}(t)$, which provides a measure of spatial structure. These results indicate that the *in silico* predictions are consistent with the experimental observations.

This chapter is organised as follows. Section 3.2.1 and Section 3.2.2 describe the experiments and the IBM, respectively. In Section 3.2.3 we explain how to apply the ABC algorithm to estimate the IBM parameters. In Section 3.3 we

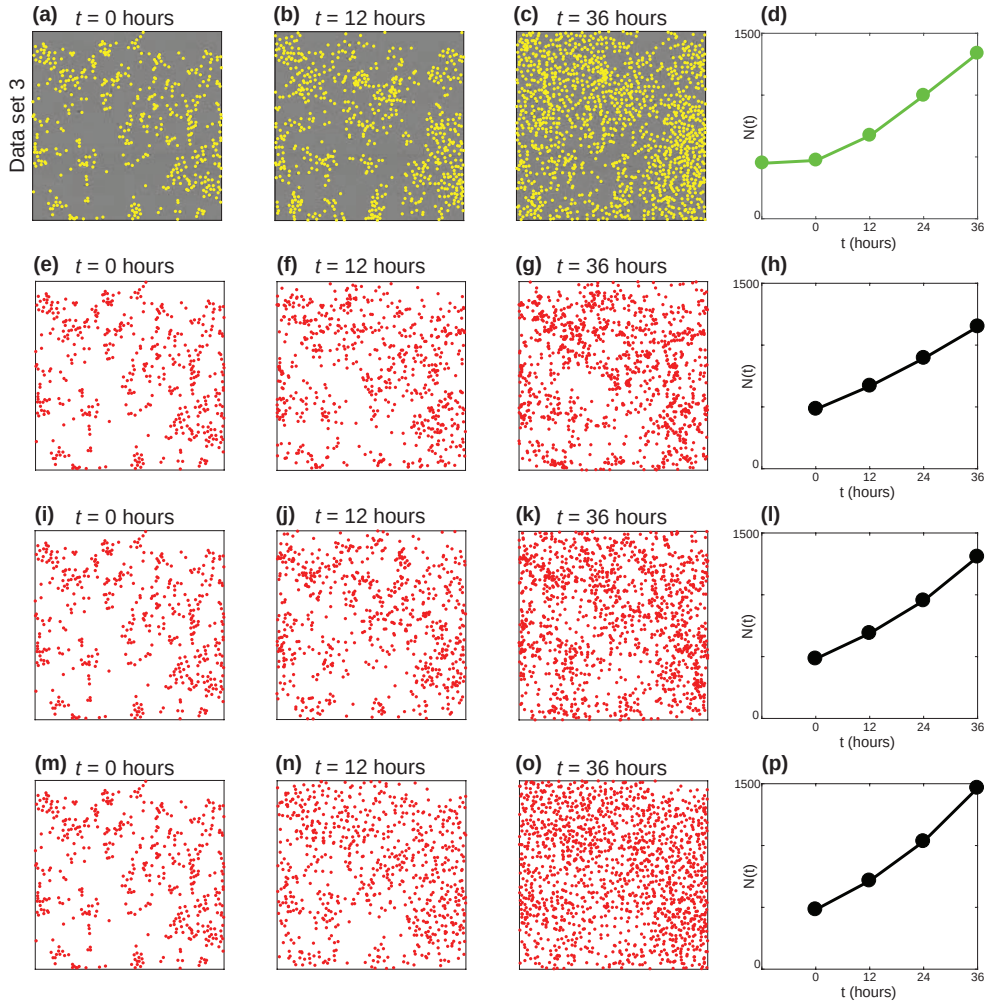

Figure 3.1: (a)-(c) Experimental data set 3 at $t = 0$, 12 and 36 hours. The position of each cell is identified with a yellow marker. The field of view is a square of length 1440 μm. (d) Population size, $N(t)$ for experimental data set 3. (e)-(h) One realisation of the IBM with $\gamma_b = 0$ μm, leading to an overly clustered distribution of agents. (i)-(l) One realisation of the IBM with $\gamma_b = 4$ μm, leading to a distribution of agents with similar clustering to the experimental data. (m)-(p) One realisation of the IBM with $\gamma_b = 20$ μm, leading to an overly segregated distribution of agents. All IBM simulations are initiated using the same distribution of agents as in (a), with $m = 0.56$ /hour, $p = 0.041$ /hour, and $\sigma = 24$ μm.

present the marginal posterior distributions of the IBM parameters using data from the first three sets of experiments. The predictive power of the calibrated IBM is demonstrated by using the combined marginal posterior distributions to predict the fourth experimental data set. The predictive power of the calibrated IBM is compared with the standard mean-field logistic equation (Murray, 2002). While both models can accurately predict $N(t)$, the logistic equation provides no information about the spatial structure in the experimental data. Finally, in Section 3.4, we conclude and summarise opportunities for further research.

3.2 Methods

3.2.1 Experimental methods

We perform a series of proliferation assays using the IncuCyte ZOOM$^{\text{TM}}$ live cell imaging system (Essen BioScience, MI USA) (Jin et al., 2017). All experiments are performed using the PC-3 prostate cancer cell line (Kaighn et al., 1979). These cells, originally purchased from American Type Culture Collection (Manassas, VA, USA), are a gift from Lisa Chopin (April, 2016). The cell line is used according to the National Health and Medical Research Council (NHMRC) National statement on ethical conduct in human research with ethics approval for the QUT Human Research Ethics Committee (QUT HREC 59644, Chopin). Cells are propagated in RPMI 1640 medium (Life Technologies, Australia) with 10% foetal calf serum (Sigma-Aldrich, Australia), 100 U/mL penicillin, and 100 μg/mL streptomycin (Life Technologies), in plastic tissue culture flasks (Corning Life Sciences, Asia Pacific). Cells are cultured in 5% CO_2 and 95% air in a Panasonic incubator (VWR International) at 37 ^{o}C. Cells are regularly screened for *Mycoplasma*.

Approximately 8,000 cells are distributed in the wells of the tissue culture plate as uniformly as possible. After seeding, cells are grown overnight to allow for attachment and some subsequent growth. The plate is placed into the IncuCyte ZOOM$^{\text{TM}}$ apparatus, and images showing a field of view of size 1440 $\times$ 1440 μm are recorded every 12 hours for a total duration of 48 hours. An example of a set of experimental images is shown in Figure 3.1(a)-(c), while images from the other three data sets are provided in Appendix 3.A, Figure 3.8.

Experimental images are recorded at five time points, at intervals of 12 hours, giving $t' = 0, 12, 24, 36$ and 48 hours. Comparing the evolution of $N(t')$ in Figure 3.2(a) shows the number of cells in some experiments do not increase appre-

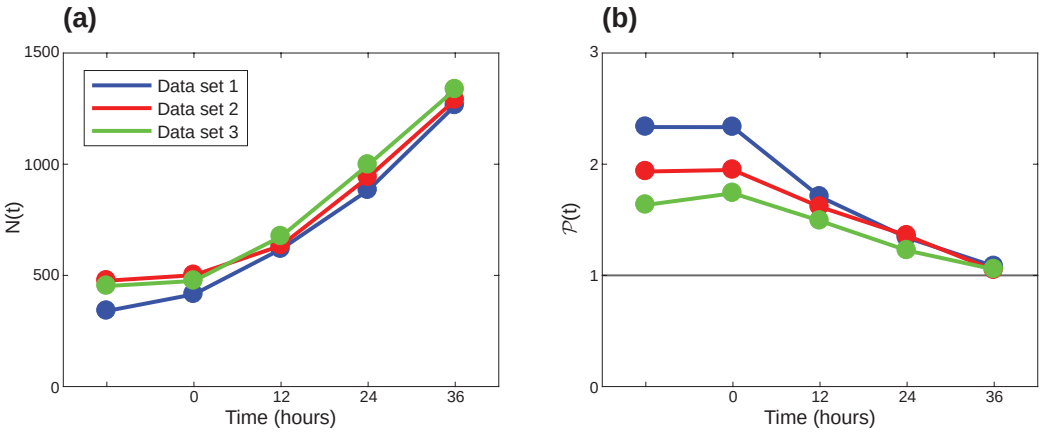

Figure 3.2: Summary statistics for experimental data sets 1, 2 and 3, shown in blue, red and green, respectively. (a) Population size, $N(t)$. (b) Pair correlation, $\mathcal{P}(t)$. Unprocessed experimental data is given in Appendix 3.A.

ciably during the first 12 hours. This suggests that the cells may experience a settling phase, so some time is required for the cells to commence normal proliferation (Tremel et al., 2009; Jin et al., 2017). Therefore, we treat the image at $t' = 12$ hours as the first image after the settling phase, and shift time, $t = t' - 12$ hours. Therefore, excluding the first experimental image at $t' = 0$ hours, we have images recorded at four time points after the settling time, $t = 0, 12, 24$ and 36 hours.

3.2.2 Mathematical model

Individual based model

We consider an IBM describing the proliferation and movement of individual cells (Binny et al., 2016a,b). Since cell death is not observed in the experiments, the IBM does not include agent death. The IBM allows the net proliferation rate and the net movement rate of agents to depend on the spatial arrangement of other agents. To be consistent with previous experimental observations, the IBM incorporates a biased movement mechanism so that agents tend to move away from nearby crowded regions (Cai et al., 2007). We use the IBM to describe the dynamics of a population of agents on a square domain of length $L = 1440$ μm to match the field-of-view of the experimental data (Figure 3.1(a)-(c)). Agents in the model are treated as a series of points which we may interpret as a population of

uniformly-sized discs with diameter $\sigma = 24$ μm (Appendix 3.B). Each agent has location $\mathbf{x}_n = (x_1, x_2)$, for $n = 1, ..., N(t)$. Since the field-of-view of each image is much smaller than the size of the well in the tissue culture plate, we apply periodic boundary conditions (Jin et al., 2017).

Proliferation and movement events occur according to a Poisson process over time (Binny et al., 2016b). The nth agent is associated with neighbourhood-dependent rates, $P_n \geq 0$ and $M_n \geq 0$, of proliferation and movement, respectively. These rates consist of intrinsic components, $p > 0$ and $m > 0$, respectively. Crowding effects are introduced by reducing the intrinsic rates by a contribution from other neighbouring agents. These crowding effects are calculated using a kernel, $w^{(\cdot)}(r)$, that depends on the separation distance, $r \geq 0$, so that

$$P_n = \max\left(0, p - \sum_{i \neq n}^{N(t)} w^{(p)}(r)\right), \tag{3.1}$$

$$M_n = \max\left(0, m - \sum_{i \neq n}^{N(t)} w^{(m)}(r)\right). \tag{3.2}$$

Following Binny et al. (2016b), we specify the kernels to be Gaussian with width corresponding to the cell diameter, σ, giving

$$w^{(p)}(r) = \gamma_p \exp\left(-\frac{r^2}{2\sigma^2}\right), \tag{3.3}$$

$$w^{(m)}(r) = \gamma_m \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{3.4}$$

Here, γ_p is the value of $w^{(p)}(0)$ and γ_m is the value of $w^{(m)}(0)$. These parameters provide a measure of the strength of crowding effects on agent proliferation and movement, respectively. The kernels, $w^{(p)}(r)$ and $w^{(m)}(r)$, ensure that the interactions between pairs of agents separated by more than roughly 2-3 cell diameters lead to a negligible contribution. For computational efficiency, we truncate the Gaussian kernels so that $w^{(p)}(r) = w^{(m)}(r) = 0$, for $r \geq 3\sigma$ (Law et al., 2003).

To reduce the number of unknown parameters in the IBM, we specify γ_p and γ_m by invoking an assumption about the maximum packing density of the population. Here we suppose that the net proliferation and net movement rates reduce to zero when the agents are packed at the maximum possible density, which is a hexagonal packing (Figure 2.3(a)). For interactions felt between the nearest neighbours only

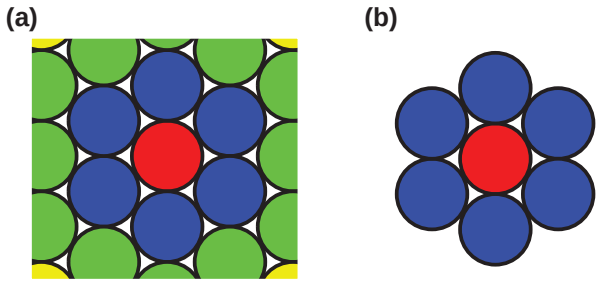

Figure 3.3: (a) Hexagonal packing of uniformly sized discs. The focal agent (red) is surrounding by six nearest neighbouring agents (blue), and twelve next nearest neighbouring agents (green). (b) Hexagonal packing around a focal agent (red) showing the six nearest neighbours only.

(Figure 2.3(b)), we obtain

$$\gamma_p = \frac{p}{6} \exp\left(\frac{1}{2}\right), \tag{3.5}$$

$$\gamma_m = \frac{m}{6} \exp\left(\frac{1}{2}\right), \tag{3.6}$$

which effectively specifies a relationship between γ_p and p, and between γ_m and m. Note that this assumption does not preclude a formation of agents in which some pairs have a separation of less than σ and densities greater than hexagonal packing, which can occur by chance.

When an agent at $\mathbf{x}_n$ proliferates, the location of the daughter agent is selected by sampling from a bivariate normal distribution with mean $\mathbf{x}_n$ and variance σ^2 (Binny et al., 2016b). Since mesenchymal cells in two-dimensional cell culture are known to move with a directional movement bias away from regions of high density (Cai et al., 2007), we allow the model to incorporate a bias so that the preferred direction of movement is in the direction of decreasing agent density. For simplicity, the distance that each agent steps is taken to be a constant, equal to the cell diameter, σ (Plank and Simpson, 2012).

To choose the movement direction, we use a crowding surface, $B(\mathbf{x})$, to measure the local crowdedness at location $\mathbf{x}$, given by

$$B(\mathbf{x}) = \sum_{i=1}^{N(t)} w^{(b)}(\|\mathbf{x} - \mathbf{x}_i\|). \tag{3.7}$$

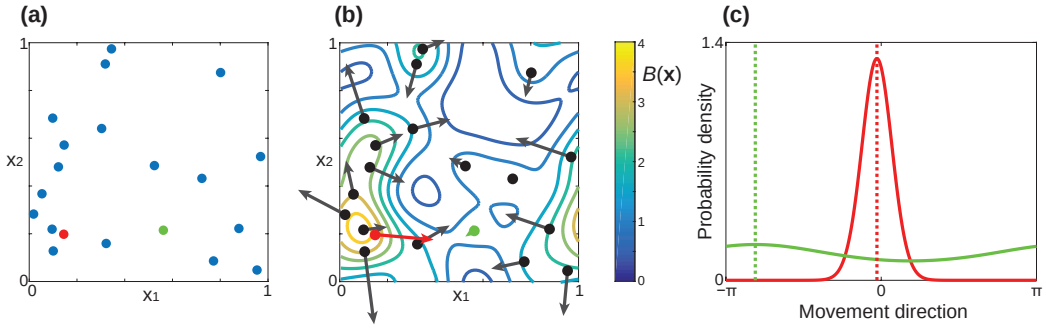

Figure 3.4: (a) Example distribution of agents on a 1×1 periodic domain. (b) Level curves of the corresponding crowding surface, $B(\mathbf{x})$, for this arrangement of agents. The arrows show the preferred direction of movement, $\mathbf{B}_n$. To illustrate how the direction of movement is chosen, (c) shows the probability density of the von Mises distribution for the red and green agents highlighted in (a) and (b). The preferred direction, $\arg(\mathbf{B}_n)$, is shown as dotted vertical lines for both agents. The red agent is in a crowded region so $\|\mathbf{B}_n\|$ is large, meaning that the agent is likely to move in the preferred direction $\arg(\mathbf{B}_n)$. The green agent is in a low density region and $\|\mathbf{B}_n\|$ is small, meaning that the bias is very weak and the agent's direction of movement is almost uniformly distributed. To illustrate the effects of the crowding surface as clearly as possible, we set $\gamma_b = 1$, $\sigma = 0.1$, $L = 1$ in this schematic figure to draw attention to the gradient of the crowding surface.

The crowding surface is the sum of contributions from every agent, given by a bias kernel, $w^{(b)}(r)$. The contributions depend on the distance between $\mathbf{x}$ and the location of the ith agent, $\mathbf{x}_i$, given by $r = \|\mathbf{x} - \mathbf{x}_i\|$. Again, we choose $w^{(b)}$ to be Gaussian, with width equal to the cell diameter, and repulsive strength, $\gamma_b \geq 0$, so that

$$w^{(b)}(r) = \gamma_b \exp\left(-\frac{r^2}{2\sigma^2}\right), \tag{3.8}$$

where γ_b is value of $w^{(b)}(0)$, and has dimensions of length. Note that $B(\mathbf{x})$ is an increasing function of local density, and approaches zero as the local density decreases. A typical crowding surface is shown in Figure 3.4(b) for the arrangement of agents in Figure 3.4(a).

To determine the direction of movement we use the shape of $B(\mathbf{x})$ to specify the bias, or preferred direction, of agent n, $\mathbf{B}_n$, given by

$$\mathbf{B}_n = -\nabla B(\mathbf{x}_n), \tag{3.9}$$

which gives the magnitude and direction of steepest descent. Results in Figure 3.4(b) show $\mathbf{B}_n$ for the arrangement of agents in Figure 3.4(a). To deter-

mine the direction of movement, we consider the magnitude and direction of $\mathbf{B}_n$, and sample the actual movement direction from a von Mises distribution, von Mises$(\arg(\mathbf{B}_n), \|\mathbf{B}_n\|)$ (Binny et al., 2016b; Forbes et al., 2011). Therefore, agents are always most likely to move in the direction of $\mathbf{B}_n$, however as $\|\mathbf{B}_n\| \to 0$, the preferred direction becomes uniformly distributed.

To illustrate how the direction of movement is chosen, we show, in Figure 3.4(b), the bias vector for each agent, $\mathbf{B}_n$. Note that $\mathbf{B}_n$ does not specify the movement step length, and the direction of $\mathbf{B}_n$ does not necessarily specify the actual direction. Rather, $\arg(\mathbf{B}_n)$ specifies the preferred direction. To illustrate this property, we highlight two agents in Figure 3.4(a). The red agent is located on a relatively steep part of the crowding surface, so $\|\mathbf{B}_n\|$ is large. The green agent is located on a relatively flat part of the crowding surface, so $\|\mathbf{B}_n\|$ is close to zero. Figure 3.4(c) shows the von Mises distributions for the red and green agent. Comparing these movement distributions confirms that the crowded red agent is more likely to move in the direction of $\mathbf{B}_n$. The bias is weak for the green agent, so the direction of movement is almost uniformly distributed since $\|\mathbf{B}_n\|$ is smaller.

Under the assumption that proliferation and movement events are exponentially distributed, we follow Baker and Simpson (2010) and apply a modified Gillespie algorithm (Gillespie, 1977) to implement the IBM. To initialise each simulation we specify the initial number and initial location of agents to match to the experimental images at $t = 0$ hours (Appendix 3.A) for experimental data sets 1, 2, 3 and 4. In all simulations we set $\sigma = 24$ μm and $L = 1440$ μm. The remaining three parameters, m, p and γ_b, are varied with the aim of producing posterior distributions using a Bayesian framework.

If $\gamma_m = \gamma_b = 0$, and the variance of the dispersal distribution is large, the IBM corresponds to logistic growth (Binny et al., 2016b; Browning et al., 2017). Under these simplified conditions, a uniformly distributed initial population of agents will grow, at rate p, to eventually reach a uniformly distributed maximum average density of $p/(2\pi\gamma_p\sigma_p^2)$. We do not consider this case here as our initial distribution of cells in the experiments is clustered, and so the logistic growth model is, strictly speaking, not valid (Binny et al., 2016b).

Summary statistics

To match the IBM simulations with the experimental data we use properties that are related to the first two spatial moments (Law et al., 2003). The first spatial moment, the average density, is characterised by the number of agents in the population, $N(t)$. The second spatial moment characterises how agents are spatially

distributed, and is often reported in terms of a pair correlation function (Binny et al., 2016a,b; Law et al., 2003). In this work we consider the pair correlation within a distance of δr, given by

$$\mathcal{P}(t) = \frac{L^2 \sum_{i=1}^{N(t)} \sum_{\substack{j=1 \\ j \neq i}}^{N(t)} \mathbb{I}_{\|\mathbf{x}_i - \mathbf{x}_j\| \leq \delta r}}{N(t)^2 \pi \delta r^2}, \tag{3.10}$$

where $\mathbb{I}$ is an indicator function so that the double sum in Equation (3.10) gives twice the number of distinct pairs within a distance δr, which we set to be 50 μm. Therefore, $\mathcal{P}(t)$ is the ratio of the number of pairs of agents, separated by a distance of less than 50 μm, to the expected number of pairs of agents separated by a distance of less than 50 μm, if the agents were randomly distributed. This means that, $\mathcal{P}(t) = 1$ corresponds to randomly placed agents; $\mathcal{P}(t) > 1$ corresponds to a locally clustered distribution; and, $\mathcal{P}(t) < 1$ corresponds to a locally segregated distribution.

3.2.3 Approximate Bayesian computation

We consider m, p and γ_b as random variables, and the uncertainty in these parameters is updated using observed data (Collis et al., 2017; Tanaka et al., 2006). To keep the description of the inference algorithm succinct, we refer to the unknown parameters as $\mathbf{\Theta} = \langle m, p, \gamma_b \rangle$.

In the absence of any experimental observations, information about $\mathbf{\Theta}$ is characterised by specified prior distributions. The prior distributions are chosen to be uniform on an interval that is wide enough to encompass previous estimates of m and p (Johnston et al., 2015). To characterise the prior for γ_b, we note that this parameter is related to a length scale over which bias interactions are felt. Preliminary results (not shown) use a prior in the interval $0 \leq \gamma_b \leq 20$ μm and suggest that a narrow prior in the interval $0 \leq \gamma_b \leq 10$ μm is appropriate. In summary, our prior distributions are uniform and independent, given by

$$\pi(m) = \mathrm{U}(0, 10) \text{ /hour}, \tag{3.11}$$

$$\pi(p) = \mathrm{U}(0, 0.1) \text{ /hour}, \tag{3.12}$$

$$\pi(\gamma_b) = \mathrm{U}(0, 10) \ \mu\text{m}. \tag{3.13}$$

We always summarise data, $\mathbf{X}$, with a lower-dimensional summary statistic, S. Data and summary statistics from the experimental images are denoted $\mathbf{X}_{\text{obs}}$ and

S_{obs}, respectively. Similarly, data and summary statistics from IBM simulations are denoted $\mathbf{X}_{\text{sim}}$ and S_{sim}, respectively. Information from the prior is updated by the likelihood of the observations, $\pi(S_{\text{obs}}|\boldsymbol{\Theta})$, to produce posterior distributions, $\pi(\boldsymbol{\Theta}|S_{\text{obs}})$. We employ the most fundamental ABC algorithm, known as ABC rejection (Liepe et al., 2014; Tanaka et al., 2006), to sample from the approximate posterior distribution. The approximate posterior distributions are denoted $\pi_u(\boldsymbol{\Theta}|S_{\text{obs}})$.

In this chapter we use a summary statistic that is a combination of $N(t)$ and $\mathcal{P}(t)$ at equally spaced intervals of duration 12 hours. A discrepancy measure, $\rho(S_{\text{obs}}, S_{\text{sim}})$, is used to assess the closeness of S_{obs} and S_{sim},

$$\rho(S_{\text{obs}}, S_{\text{sim}}) = \sum_{j=1}^{3} \left(\frac{[N_{\text{sim}}(12j) - N_{\text{obs}}(12j)]^2}{N_{\text{obs}}(12j)^2} + \frac{[\mathcal{P}_{\text{sim}}(12j) - \mathcal{P}_{\text{obs}}(12j)]^2}{\mathcal{P}_{\text{obs}}(12j)^2} \right). \tag{3.14}$$

Algorithm 3.1 is used to obtain $10^6 u$ samples, $\{\boldsymbol{\Theta}_i\}_{i=1}^{10^6 u}$, from the approximate joint posterior distribution, $\pi_u(\boldsymbol{\Theta}|S_{\text{obs}})$, for each data set. Here, $u \ll 1$ is the accepted proportion of samples.

Algorithm 3.1 ABC rejection sampling algorithm to obtain $10^6 u$ samples from the approximate posterior distribution, $\pi_u(\boldsymbol{\Theta}|S_{\text{obs}})$.

1: Set $\sigma = 24$ μm, $L = 1440$ μm, and set $\mathbf{x}_n$ to match experimental data $\mathbf{X}_{\text{obs}}$ at $t = 0$.
2: Draw parameter samples from the prior $\boldsymbol{\Theta}_i \sim \pi(\boldsymbol{\Theta})$.
3: Simulate cell proliferation assay with $\boldsymbol{\Theta}_i$ and $t \leq 36$ hours.
4: Record summary statistic $S_{\text{sim}_i} = \{N_{\text{sim}}(12j), \mathcal{P}(12j)\}_{j=1}^{3}$, where j is an index that denotes the three observation time points, $t = 12, 24$ and 36 hours.
5: Compute the discrepancy measure $\epsilon_i = \rho(S_{\text{obs}}, S_{\text{sim}_i})$, given in Equation (3.14).
6: Repeat steps 2-5 until 10^6 samples $\{\boldsymbol{\Theta}_i, \epsilon_i\}_{i=1}^{10^6}$ are simulated.
7: Order $\{\boldsymbol{\Theta}_i, \epsilon_i\}_{i=1}^{10^6}$ by ϵ_i such that $\epsilon_1 < \epsilon_2 < ....$
8: Retain the first 1% ($u = 0.01$) of prior samples $\boldsymbol{\Theta}_i$, as posterior samples, $\{\boldsymbol{\Theta}_i\}_{i=1}^{10^6 u}$.

To present and perform calculations with posterior samples, we use a kernel density estimate to form approximate marginal posterior distributions, for each parameter, and each data set using the `ksdensity` function in MATLAB (Mathworks, 2017). This is done by treating the components of the joint posterior samples as samples from each marginal distribution. The `ksdensity` function gives a discrete distribution for each marginal posterior, with grid spacing $\Delta m = 0.01$, $\Delta p = 0.0001$ and $\Delta\gamma_b = 0.01$, for m, p and γ_b, respectively. This discretisation ensures that the marginal posterior densities are approximated using 1000 equally spaced values across the prior support.

Generating and sampling from the combined posterior distribution

The marginal posterior distributions for each parameter are similar for each independent experimental data set. Therefore, we combine the marginal posterior distributions for each independent experimental data set to obtain a combined posterior distribution. If the approximate marginal posterior distribution for m is $\pi_u(m|S^{(k)}_{\text{obs}})$, where $S^{(k)}_{\text{obs}}$ is the summary statistic from the kth experimental data set, then the combined marginal posterior distribution for m is

$$\pi_u(m|\{S^{(k)}_{\text{obs}}\}_{k=1}^3) \propto \prod_{k=1}^{3} \pi_u(m|S^{(k)}_{\text{obs}}). \tag{3.15}$$

Combined marginal posterior distributions for p and γ_b are calculated similarly.

To test the predictive power of the calibrated IBM, we sample parameters from the combined joint posterior distribution by sampling each parameter separately from the corresponding combined marginal posterior distributions. This approach amounts to assuming that m, p and γ_b are independent random variables, and we will make a comment on the validity of this assumption later. For m, we generate a discrete combined posterior distribution, $\pi_u(m|\{S^{(k)}_{\text{obs}}\}_{k=1}^3)$, using the kernel-density estimate for each data set and Equation (3.15). This gives a discrete distribution with bin width $\Delta m = 0.01$, where each bin is denoted by an index, $l = 0, 1, ...$, and has probability density $\pi_u(l\Delta m|\{S^{(k)}_{\text{obs}}\}_{k=1}^3)$. If m is uniformly distributed within each bin, we apply Algorithm 3.2 to obtain 10^4 samples. Repeating this process in a similar way gives 10^4 samples for both p and γ_b.

Algorithm 3.2 Rejection sampling algorithm for sampling from the combined approximate posterior distribution, $\pi_u(m|\{S^{(k)}_{\text{obs}}\}_{k=1}^3)$.

1: Set $\Delta m = 0.01$, $m_{\max} = 10$, which is the upper limit of the prior support.
2: Set maximum density $\nu = \max \pi_u(m|\{S^{(k)}_{\text{obs}}\}_{k=1}^3)$.
3: Sample proposal bin index l_* from $\{0, ..., m_{\max}/\Delta m - 1\}$.
4: Sample $r_1 \sim \text{U}(0, \nu)$.
5: If $r_1 < \pi_u(l_*\Delta m; \{S^{(k)}_{\text{obs}}\}_{k=1}^3)$, accept l_*, else repeat steps 3-5.
6: Sample the location within the chosen bin, $m_i \sim \text{U}(l_*\Delta m, (l_* + 1)\Delta m)$.
7: Repeat steps 3-6 until 10^4 samples, $\{m_i\}_{i=1}^{10^4}$, are obtained.

Predicting experimental data set 4 using the combined posterior distribution

We sample 10^4 parameter sets, $\{\boldsymbol{\Theta}_i\}_{i=1}^{10^4}$, from the combined posterior distribution, $\pi_u(\boldsymbol{\Theta}|\{S^{(k)}_{\text{obs}}\}_{k=1}^3)$. Using these samples, we simulate the IBM initialised with the

actual initial arrangement of cells in data set 4 at $t = 0$. For each parameter combination S_{sim} is recorded at 12 hour intervals, and used to construct distributions of $N(t)$ and $\mathcal{P}(t)$. These distributions are represented as box plots and compared with summary statistics from experimental data set 4.

3.3 Results and discussion

To qualitatively illustrate the importance of spatial structure we show, in rows 2-4 of Figure 3.1, snapshots from the IBM with different choices of parameters. In each case the IBM simulations evolve from the initial condition specified in Figure 3.1(a). Results in the right-most column of Figure 3.1 compare the evolution of $N(t)$ and we see that the parameter combination in the second row underestimates $N(t)$, the parameter combination in the fourth row overestimates $N(t)$, and the parameter combination in the third row produces a reasonable match to the experimental data. A visual comparison of the spatial arrangement of agents in rows 2-4 of Figure 3.1 suggests that these different parameter combinations may lead to different spatial structures. This illustration of how the IBM results vary with the choice of parameters motivates us to use ABC rejection to estimate the joint distribution of the parameters. To do this we will use summary statistics from three identically prepared, independent sets of experiments. The summary statistics for these experiments, $N(t)$ and $\mathcal{P}(t)$, are summarised in Figure 3.2, and tabulated in Appendix 3.B.

The approximate marginal posterior distributions for m, p and γ_b are shown in Figure 3.5(a)-(c), respectively, for experimental data sets 1, 2 and 3. There are several points of interest to note. In each case, the posterior support is well within the interior of the prior support, suggesting that our choice of priors is appropriate. An interesting feature of the marginal posterior distributions for all parameters is that there is significant overlap for each independent experimental data set. There is some variation in the mode between experimental data sets, for each parameter, which is expected under the assumption that cell proliferation assays are stochastic processes.

Since the marginal posterior distributions for each experimental data set overlap, we produce a combined marginal posterior distribution for each parameter using Equation (3.15). The combined marginal posterior distributions are superimposed, and the mode is given by 0.56 /hour, 0.041 /hour and 4.0 μm for m, p and γ_b, respectively. These estimates of p and m give a cell doubling time of $\ln(2)/p \approx 17$ hours, and a cell diffusivity of approximately 320 μm^2/hour, which

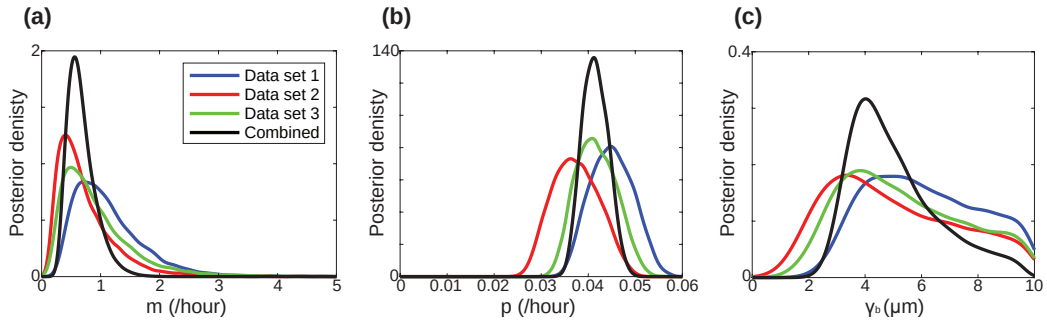

Figure 3.5: (a)-(c) Kernel-density estimates of the approximate marginal posterior distributions for each data set, for parameters m, p and γ_b, respectively, with $u = 0.01$. The combined posterior distribution (black), given by Equation (3.15), is superimposed. The modes of the combined marginal posterior distributions are $m = 0.56$ /hour, $p = 0.041$ /hour and $\gamma_b = 4.0$ μm. All distributions are scaled so that the area under the curve is unity.

are typical values for PC-3 cells at low density (Johnston et al., 2015; Jin et al., 2016b). All results in the main chapter correspond to retaining the top 1% of samples ($u = 0.01$) and additional results (Appendix 3.C) confirm that the results are relatively insensitive to this choice.

To assess the predictive power of the calibrated IBM, we attempt to predict the time evolution of a separate, independently collected data set, experimental data set 4, as shown in Figure 3.6(a)-(d). We use the mode of the combined posterior distribution and the initial arrangement of agents in experimental data set 4 to produce a typical prediction in Figure 3.6(e)-(h). Visual comparison of the experimental data and the IBM prediction suggests that the IBM predicts a similar number of agents, and a similar spatial structure, with some clustering present. To quantify our results, we compare the evolution of $N(t)$ in Figure 3.6(i) which reveals an excellent match. Furthermore, we predict the evolution of $\mathcal{P}(t)$ in Figure 3.6(j) confirming similar trends. The quality of match between the predicted distribution of $N(t)$ and $\mathcal{P}(t)$ supports our assumption that m, p and γ_b can be treated as independent random variables as posited in Section 3.2.3. Nonetheless, the predicted decay in $\mathcal{P}(t)$ is not as rapid as in the experimental data. There are many potential explanations for this, including the choice of summary statistics, and assumption relating p and γ_p, and m and γ_m.

In addition to examining a single, typical realisation of the calibrated model, we now examine a suite of realisations of the calibrated IBM, and compare results with experimental data set 4. The suite of IBM realisations is obtained by sampling from the joint posterior distribution. Results in Figure 3.7(a) compare

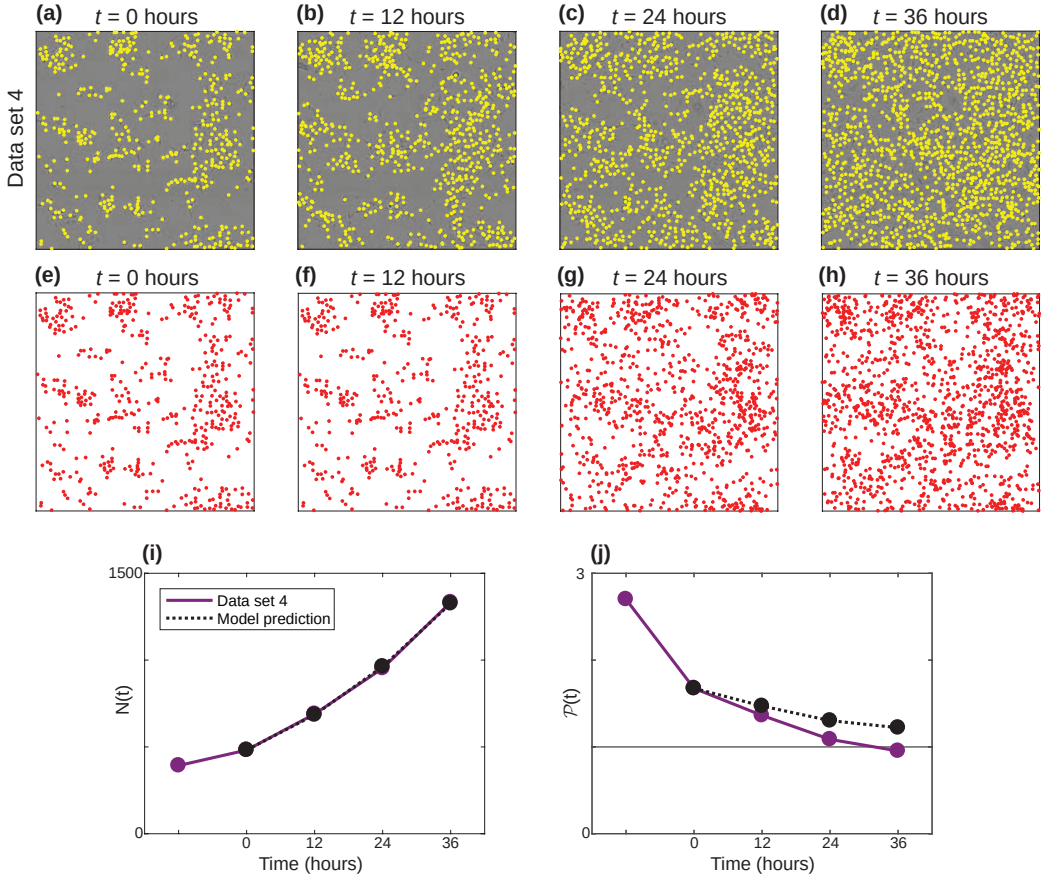

Figure 3.6: (a)-(d) Experimental images for data set 4. The position of each cell is identified with a yellow marker. The field of view is a square of length 1440 μm. (e)-(h) One realisation of the IBM with parameters corresponding to the posterior mode: $m = 0.56$ /hour, $p = 0.041$ /hour and $\gamma_b = 4.0$ μm, with the same initial arrangement of agents as in (a). (i) $N(t)$ for the experimental data (purple) and the IBM prediction (dashed black). (j) $\mathcal{P}(t)$ for the experimental data (purple) and the IBM prediction (dashed black).

$N(t)$ from experimental data set 4 with distributions of $N(t)$ from the suite of IBM simulations, showing an excellent match. The spread of the distributions of $N(t)$ increases with time, which is expected. Results in Figure 3.7(b) compare the evolution of $\mathcal{P}(t)$ from experimental data set 4 with distributions of $\mathcal{P}(t)$ from the suite of IBM simulations, showing the predicted distributions of $\mathcal{P}(t)$ overlap with the experimental data. Overall, the quality of the match between the prediction and the experimental data is high, as the prediction captures both qualitative and quantitative features of the data. To illustrate the importance of considering spatial structure in the IBM, we also calibrate the solution of the classical mean-field logistic equation (Murray, 2002) to experimental data sets 1, 2 and 3. The logistic equation is given by

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \lambda N(t)\left(1 - \frac{N(t)}{N_{\max}}\right), \tag{3.16}$$

where λ is the cell proliferation rate and $N_{\max}$ is the maximum number of agents (Murray, 2002; Jin et al., 2017). Following a similar procedure (Appendix 3.D), we use ABC rejection to form combined posterior distributions of λ and $N_{\max}$. The modes of the combined posterior distributions are $\lambda = 0.036$/hour and $N_{\max} = 4017$. This estimate leads to a doubling time of approximately 19 hours, which is slightly longer than the doubling time predicted using the calibrated IBM. We then examine a suite of solutions to Equation (3.16), where we sample from the joint posterior distribution for λ and $N_{\max}$. The predicted distribution of $N(t)$ is compared with experimental data set 4 in Figure 3.7(c), revealing an excellent match. However, implicit in the logistic equation is the mean-field assumption, which amounts to ignoring spatial structure. Therefore, the logistic equation effectively predicts $\mathcal{P}(t) = 1$ for all $t > 0$, which clearly is unable to match the spatial structure inherent in the experiments, as demonstrated in Figure 3.7(d). Overall, both calibrated models are able to predict the evolution of $N(t)$ over 36 hours. However, the logistic model is unable to describe, or predict, any information relating to spatial structure in the arrangement of cells. The differences in the way that the logistic model and the IBM treat interactions between individuals could explain why the calibration process leads to different estimates of the low density cell proliferation rates, λ and p. These differences affirm that the interactions between individuals at different spatial scales appear to be important for our experimental data.

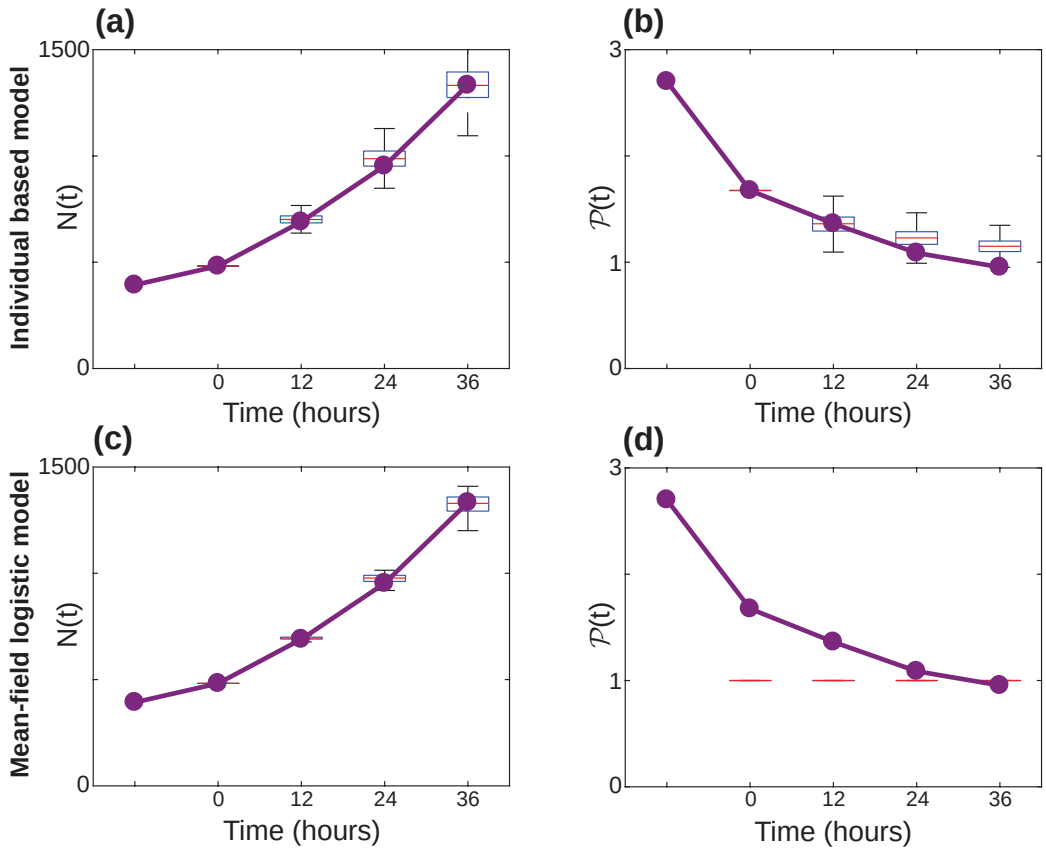

Figure 3.7: (a)-(b) Predictive distributions for $N(t)$ and $\mathcal{P}(t)$, respectively, generated using the IBM. 10^4 parameter samples were taken from the combined posterior distribution, and a model realisation produced for each sample, initiated as in Figure 3.6(a). Box plots show the distribution of $N(t)$ and $\mathcal{P}(t)$ across these realisations in (a) and (b), respectively. (c)-(d) Show the equivalent predictive distributions as box plots, using the same procedure for the mean-field logistic growth model. The procedure and kernel-density estimates of the marginal distributions for the mean-field logistic model are outlined in Appendix 3.D.

3.4 Conclusion

In this work we explore how to connect a spatially continuous IBM of cell migration and cell proliferation to novel data from a cell proliferation assay. Previous work parameterising IBM models of cell migration and cell proliferation to experimental data using ABC have been restricted to lattice-based IBMs (Johnston et al., 2014). This is partly because ABC methods require large numbers of IBM simulations, and lattice-based IBMs are far less computationally expensive than lattice-free IBMs (Plank and Simpson, 2012). We find it is preferable to work with a lattice-free IBM when dealing with experimental data as a lattice-based IBM requires approximations when mapping the distribution of cells from experimental images to a lattice (Johnston et al., 2014, 2016). This mapping can be problematic. For example, if multiple cells in an experimental image are equally close to one lattice site, ad hoc assumptions have to be introduced about how to arrange those cells on the lattice without any overlap. These issues are circumvented using a lattice-free method.

To help overcome the computational cost of using ABC with a lattice-free IBM, we introduce several realistic, simplifying assumptions. The IBM originally presented by Binny et al. (2016b) involves 12 free parameters, which is a relatively large number for standard inference techniques. The model is simplified by noting that our experiments do not involve cell death, and specifying the width of the interaction kernels to be constant, given by the cell diameter. Another simplification is given by assuming that crowding effects reduce the proliferation and movement rates to zero when the agents are packed at the maximum hexagonal packing density. This leads to a simplified model with three free parameters: m, p and γ_b. Using ABC rejection, we arrive at posterior distributions for these parameters for three independent experimental data sets. The marginal posterior distributions for the three parameters are similar, leading us to combine the marginal posterior distributions. The mode of the combined posterior distributions for m and p are consistent with previous parameter estimates (Johnston et al., 2015) and the mode for γ_b is consistent with previous observations that mesenchymal cells in this kind of two-dimensional experiment tend to move away from regions of high cell density (Cai et al., 2007).

In the field of mathematical biology, questions about how much detail to include in a mathematical model, and what kind of mathematical model is preferable for understanding a particular biological process are often settled in an *ad hoc* manner, as discussed by (Maclaren et al., 2015). Our approach in this work is to use a mathematical model that incorporates just the key mechanisms, with an appropri-

ate number of unknown parameters. Other approaches are possible, such as using much more complicated mathematical models that describe additional mechanisms such as: (i) detailed information about the cell cycle in individual cells (Fletcher et al., 2012); (ii) concepts of leader and follower cells (Kabla, 2012); (iii) explicitly coupling cell migration and cell proliferation to the availability of nutrients and growth factors (Tang et al., 2014); or (iv) including mechanical forces between cells (Stichel et al., 2017). However, we do not include these kinds of detailed mechanisms because our experimental data does not suggest that these mechanisms are relevant to our situation. Furthermore, it is not always clear that using a more complicated mathematical model, with additional mechanisms and additional unknown parameters, necessarily leads to improved biological insight. In fact, simply incorporating additional mechanisms and parameters into the mathematical model often leads to a situation where multiple parameter combinations lead to equivalent predictions which limits the usefulness of the mathematical model (Simpson et al., 2006). In this study, our approach is to be guided by experimental data and our ability to infer the parameters in a mathematical model based on realistic amounts of experimental data (Maclaren et al., 2015). In particular we use three experimental data sets to calibrate the IBM, and an additional data set to separately examine the predictive capability of the calibrated IBM. We find that the process of calibrating the IBM leads to well defined posterior distributions of the model parameters, and that the calibrated IBM produces a reasonable match to the experimental data. The process of calibrating the IBM, and then separately testing the predictive capability of the calibrated IBM, provides some confidence that the level of model complexity is appropriate for our purposes.

An interesting feature of our approach is that the ABC marginal posterior distributions for each parameter overlap for each independent experimental data set. This is reassuring as it suggests that the same IBM mechanism matches the three independent experimental data sets using similar parameters. Another approach would be to use ABC to parameterise the IBM by matching all the experimental data sets simultaneously. Although this alternative approach is valid, it does not allow us to examine whether the parameter estimates are consistent across the three independent experiments. Additional confidence in the calibrated IBM is provided by predicting the evolution of a fourth independent experimental data set by performing IBM simulations with parameters sampled from the combined marginal posterior distributions.

An interesting feature of all experimental data at early time, when the cell density is relatively low, is that the pair correlation measure suggests that the cells are clustered at short intervals, and that this clustering becomes less pronounced

with time. This observation is very different to the way that previous theoretical studies have viewed the role of spatial structure. For example, previous simulation-based studies assume that some initial random spatial arrangement of cells can lead to clustering at later times (Baker and Simpson, 2010). In contrast, our experimental data suggests it could be more realistic to consider that the spatial structure is imposed by the initial arrangement of cells. Moreover, since all of our experimental data involves some degree of spatial clustering, our work highlights the importance of using appropriate models to provide a realistic representation of key phenomena. Almost all continuum models of collective behaviour in cell populations take the form of ordinary differential equations and partial differential equations that implicitly invoke a mean-field assumption (Tremel et al., 2009; Sengers et al., 2007; Maini et al., 2004a; Sarapata and de Pillis, 2014; Sherratt and Murray, 1990). Such assumptions ignore the role of spatial structure. While pair-wise models that avoid mean-field assumptions are routine in some fields, such as disease spreading (Sharkey et al., 2006; Sharkey, 2008) and ecology (Law et al., 2003), models that explicitly account for spatial structure are far less common for collective cell behaviour.

Using our parameter estimates, the continuum spatial moment description could be used to interpret experimental data sets with larger numbers of cells (Binny et al., 2016b), such as experimental images showing a wider field-of-view, or experiments initiated with a higher density of cells. Our approach to estimate the parameters in the model is to work with the IBM since this allows us more flexibility in connecting with the experimental data, such as choosing the initial locations of the agents in the IBM to precisely match the initial locations of cells in the experimental images.

There are many ways that our study could be extended. For example, here we choose a summary statistic encoding information about the first two spatial moments. However, other summary statistics may provide different insight, and it could be of interest to explore the effect of this choice. For example, here we describe the spatial structure over a relatively short spatial interval, approximately 2σ. It could be of interest to repeat our analysis with a wider interval, however this would incur additional computational costs. Another approach to extend our work would be to repeat the inference procedure without making any assumptions relating p and γ_p, and m and γ_m. Such an approach would be more computationally expensive and probably require additional experimental data. Therefore, we leave these topics for future consideration.

Chapter 3: Supplementary material

3.A Experimental data

We conduct all analysis using four independently generated experimental data sets. To analyse the data we determine the location of the centre of each cell in each experimental image using ImageJ (Schneider et al., 2012). To provide a visual check on the ImageJ results, we superimpose the experimental images and the estimated cell locations in Figure 3.8. In each image the field of view is a square of length 1440 μm.

Using the (x_1, x_2) coordinate data, we calculate the size of the population, $N(t)$, and the pair correlation locally in the interval $[0, 50]$ μm, $\mathcal{P}(t)$. These summary statistics are given in Table 3.1.

	$N(-12)$	$N(0)$	$N(12)$	$N(24)$	$N(36)$
Data set 1	340	415	618	883	1264
Data set 2	476	501	634	941	1289
Data set 3	452	475	675	997	1336
Data set 4	393	482	690	953	1333
	$\mathcal{P}(-12)$	$\mathcal{P}(0)$	$\mathcal{P}(12)$	$\mathcal{P}(24)$	$\mathcal{P}(36)$
Data set 1	2.3342	2.3332	1.7075	1.3389	1.0817
Data set 2	1.9343	1.9481	1.6158	1.3572	1.0465
Data set 3	1.6335	1.7389	1.4904	1.2239	1.0535
Data set 4	2.7009	1.6751	1.3631	1.0872	0.9539

Table 3.1: Summary statistics. $N(t)$ is the population size at time t. $\mathcal{P}(t)$ is the pair correlation in the interval $[0, 50]$ μm at time t.

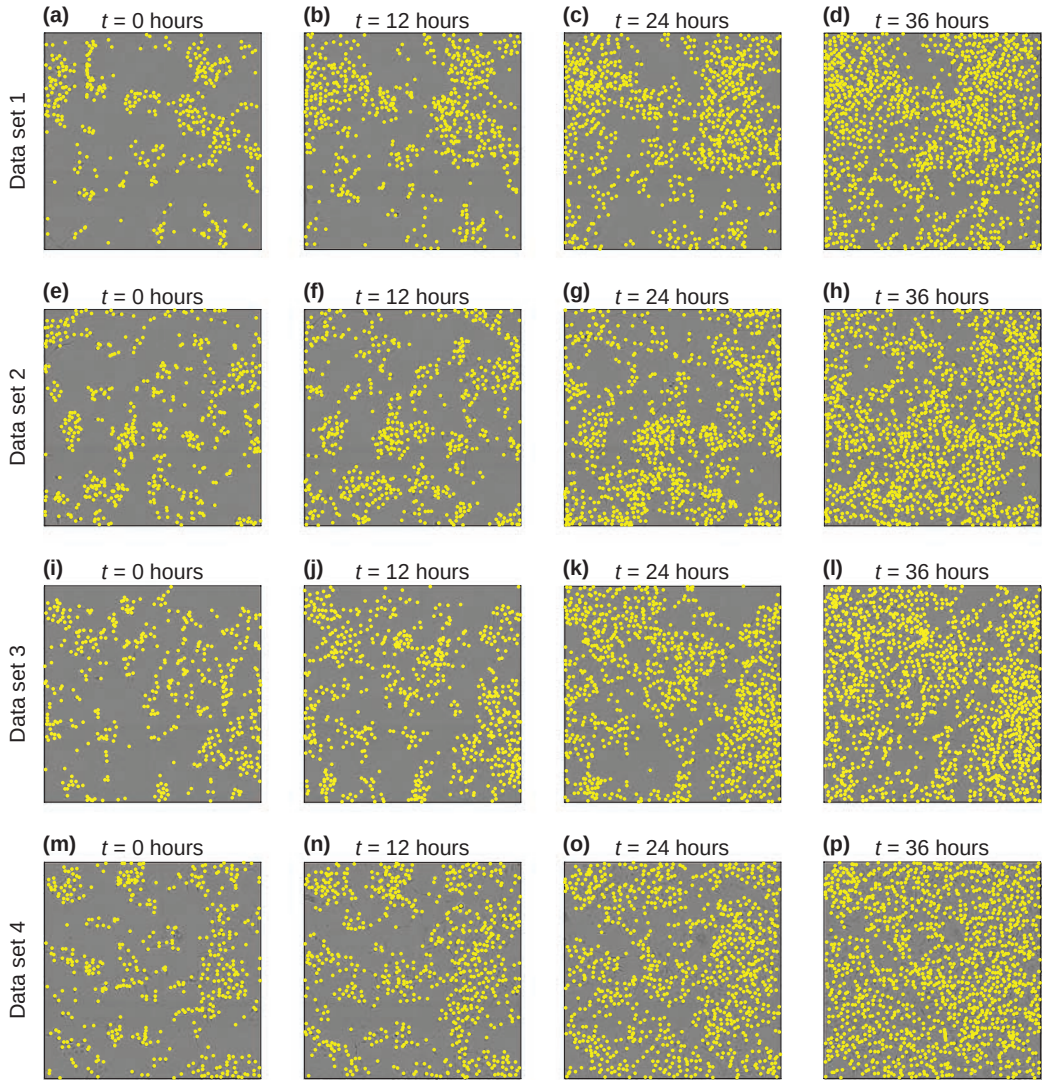

Figure 3.8: Images of each data set at 12 hour intervals. Yellow markers indicate the approximate location of each cell as determined by ImageJ.

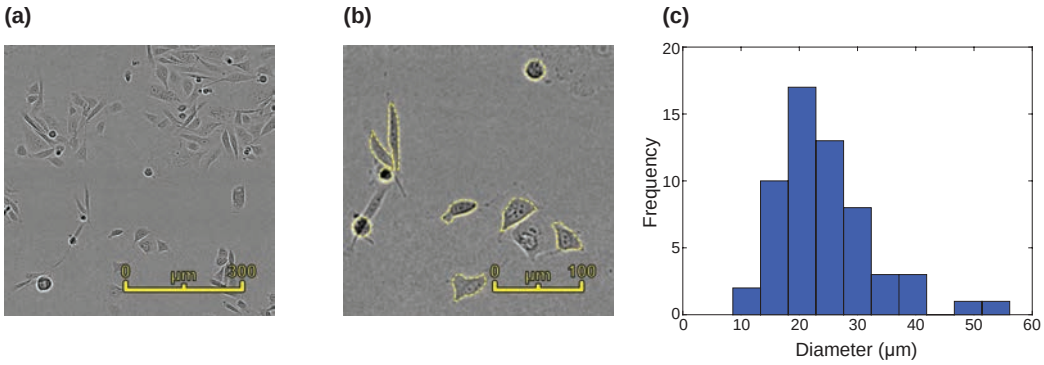

Figure 3.9: (a) A typical experimental image taken from data set 1 at $t = 12$ hours. (b) Shows how the perimeter of randomly chosen cells is traced to determine the area of each cell, and the equivalent cell diameter. (c) Distribution of cell diameter estimates using measurements of 58 cells from data set 1 at $t = 12$. The mean cell diameter is $\sigma = 24$ μm.

3.B Mean cell diameter

Here we present two independent methods to estimate the mean cell diameter. First, we make direct measurements of the area of a reasonable number of randomly chosen cells. Second, we examine the pair correlation function over a range of distances. Both approaches suggest that the mean cell diameter is approximately $\sigma = 24$ μm. This estimate is consistent with previous estimates (Jin et al., 2016b).

We approximate the shape and size of the cells by treating them as a population of uniformly sized discs. We follow the procedure of (Treloar et al., 2014) to approximate the average cell diameter. The area, A, enclosed by the outline of a number of randomly chosen cells is converted into an equivalent diameter, by assuming that, on average, the cell is circular, giving $d = 2\sqrt{A/\pi}$. Results in Figure 3.9(c) show the distribution of cell diameters from 58 randomly chosen cells. On average, this gives the average cell diameter to be $\sigma = 24$ μm.

In addition to directly measuring the area of certain cells, another way to estimate the average cell diameter is to calculate the pair correlation function over a larger range of distances. The pair correlation function, $C(r, t)$, describes the ratio of pairs of cells, separated by distance of r, to the expected number of pairs of cells separated by the same distance if the population were distributed randomly. The pair correlation function for $N(t)$ objects distributed in a square domain of

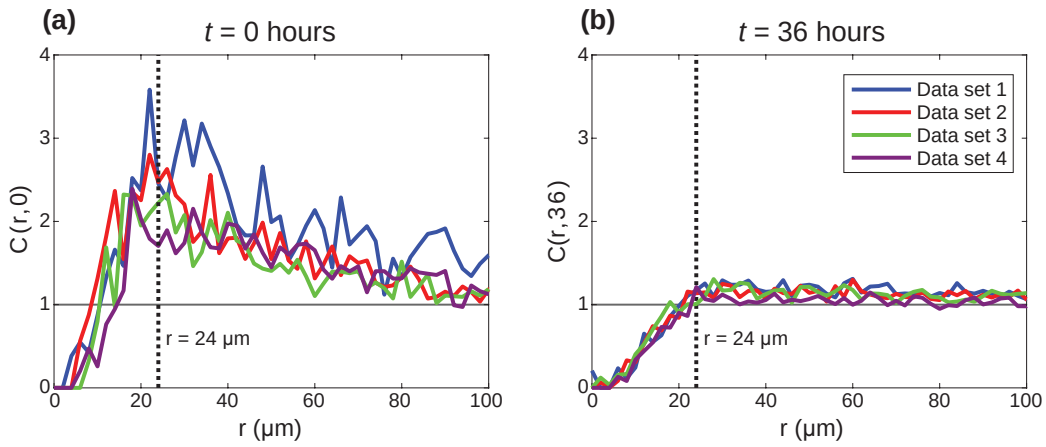

Figure 3.10: Pair correlation function, $C(r,t)$, calculated for small intervals of width $\delta r = 2$ μm at (a) $t = 0$ hours and (b) $t = 36$ hours.

length L is given by (Binny et al., 2016b)

$$C(r,t) = \frac{L^2 \sum_{i=1}^{N(t)} \sum_{\substack{j=1 \\ j\neq i}}^{N(t)} \mathbb{I}_{r \leq \|\mathbf{x}_i - \mathbf{x}_j\| < r+\delta r}}{N(t)^2 \pi \delta r (2r + \delta r)}, \tag{3.17}$$

where $\mathbb{I}$ is an indicator function so that the double sum gives the number of pairs within a sufficiently thin annulus, $[r, r+\delta r)$. If $C(r,t) > 1$, we have a greater number of pairs at distance r than we would expect if the population is distributed randomly, and we refer to this distribution as being clustered. However, if $C < 1$, we have a smaller number of pairs at distance r than we would expect if the population is distributed randomly, and we refer to this distribution as being segregated.

We calculate the pair correlation function with $\delta r = 2$ μm. Results in Figure 3.10(a) show that $C(r,0)$ is a maximum at approximately $r = 24$ μm, meaning that pairs of cells are most likely to be separated by a distance of 24 μm. This is consistent with the diameter of cells being approximately 24 μm. Results in Figure 3.10(b) show that there is a change in the pair correlation function at $r \approx 24$ μm. Cells are approximately randomly distributed at distances $r \geq 24$ μm, and there is a lack of pairs of cells at shorter distances, $r < 24$ μm. Again, this is consistent with our estimate of the cell diameter being approximately 24 μm.

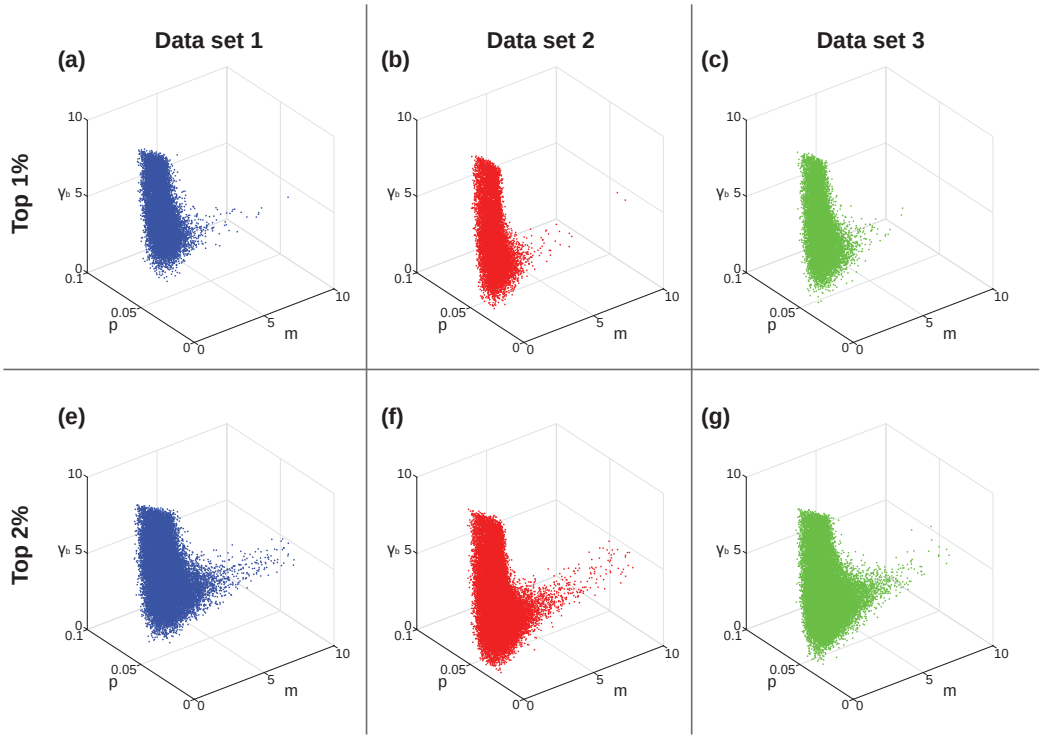

Figure 3.11: (a)-(c) show the accepted proposals, or 10^4 samples from the approximate posterior distribution $\pi_u(\boldsymbol{\Theta})$ with $u = 0.01$, for each data set 1, 2 and 3 in (a)-(c), respectively. (e)-(g) Corresponding accepted proposals for $u = 0.02$.

3.C Posterior distributions

In the main part of the chapter, we show a series of ABC marginal posterior distributions constructed using the top 1% of samples, or $u = 0.01$. In addition to showing the approximate marginal distributions, here we also show the joint posterior distributions in Figure 3.11(a)-(c) for data sets 1-3, respectively. Furthermore, the joint posterior distributions in Figure 3.11(d)-(f), for data set 1-3, respectively, are given for the top 2% samples, or $u = 0.02$. Visually comparing the joint distributions in Figure 3.11(a)-(c) with the joint distributions in Figure 3.11(d)-(f) suggests that the joint distributions have a similar shape, regardless of these choices of u. Furthermore, the mode of the approximate continuous joint posterior distribution is $\langle 0.56, 0.041, 4.0\rangle$ for $u = 0.01$, whereas the mode is $\langle 0.59, 0.040, 3.8\rangle$ for $u = 0.02$. This measure provides a quantitative confirmation that the key features of the posterior distributions are relatively insensitive to these choices of u. Results in Figure 3.12 show the marginal distributions when $u = 0.02$.

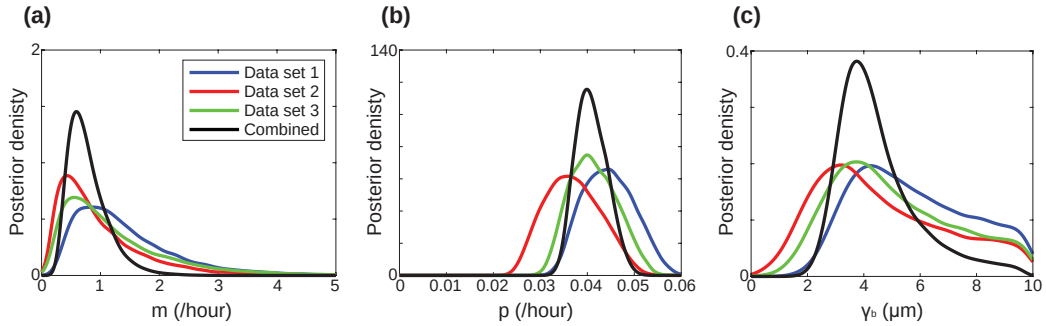

Figure 3.12: (a)-(c) Kernel-density estimates of the approximate marginal posterior distributions for each data set, for parameters m, p and γ_b, respectively, with $u = 0.01$. The combined posterior distribution (black), given by Equation (3.15), is superimposed. The modes of the combined marginal posterior distributions are $m = 0.59$ /hour, $p = 0.040$ /hour and $\gamma_b = 3.8$ μm. All distributions are scaled so that the area under the curve is unity.

3.D Mean-field logistic model

In the main chapter, we compare the predictive performance of the IBM to that of the standard mean-field continuum logistic model. The logistic model is given by

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \lambda N(t)\left(1 - \frac{N(t)}{N_{\max}}\right), \tag{3.18}$$

where $\lambda > 0$ is the proliferation rate, and $N_{\max}$ is the carrying capacity.

We implement an ABC rejection algorithm, similar to Algorithm 3.1 in the main chapter, to estimate λ and $N_{\max}$. However, when we infer the parameters for the logistic model we use a slightly different discrepancy measure, ρ', that only includes $N(t)$, so that

$$\rho'(S_{\text{obs}}, S_{\text{sim}}) = \sum_{i=1}^{3}[N_{\text{sim}}(12i) - N_{\text{obs}}(12i)]^2, \tag{3.19}$$

where $S = \{N(12), N(24), N(36)\}$. To apply ABC rejection we set the prior distributions to be uniform and independent, given by

$$\pi(\lambda) = \mathrm{U}(0, 0.1) \text{ /hour}, \tag{3.20}$$

$$\pi(N_{\max}) = \mathrm{U}(1000, 5000) \text{ agents}. \tag{3.21}$$

Kernel-density estimates of the marginal posterior distributions for each parameter are shown in Figure 3.13, and we note the mode of the combined marginal

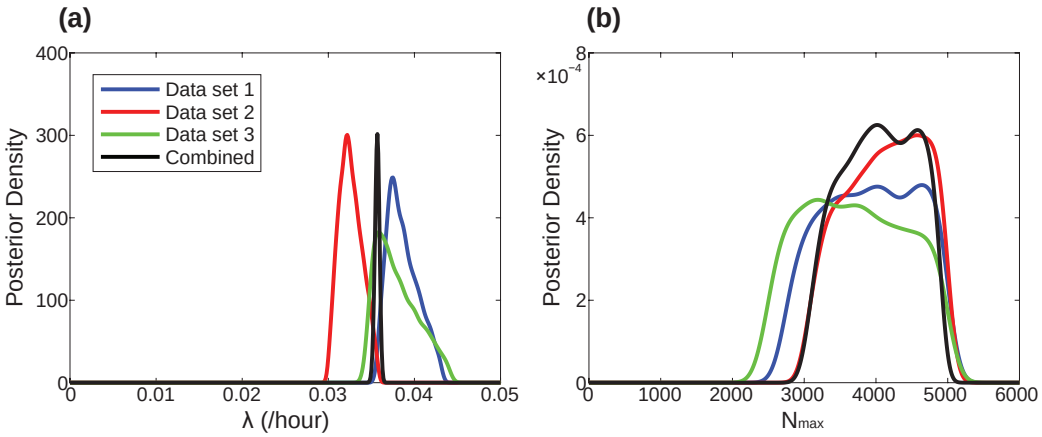

Figure 3.13: Kernel-density estimates of the marginal posterior distributions are shown for each data set, for parameters in the continuum logistic model, Equation (3.18), λ and $N_{\max}$ in (a) and (b) respectively. The combined posterior distribution (black) is the product of the marginal distributions. The combined posterior mode gives a point estimate of $\lambda = 0.0357$ /hour and $N_{\max} = 4017$. All marginal distributions are scaled to an area of unity.

posteriors gives point estimates of $\lambda = 0.0357$ /hour and $N_{\max} = 4017$.

Chapter 4

Conclusion

In this thesis, we apply mathematical modelling techniques to describe and interpret cell proliferation assays. In particular, we consider stochastic individual based models (IBMs) to describe the behaviour of individual cells in proliferation assays, as well as capture experimental variability. We take a Bayesian approach to inference, and treat unknown parameters, and experimental results, as random variables. As the models considered in this thesis have intractable likelihoods, we apply approximate Bayesian computation (ABC) (Tanaka et al., 2006; Sunnaker et al., 2013; Liepe et al., 2014) to produce posterior distributions of the unknown parameters, given experimental and *in silico* data. Previous work calibrating IBMs to experimental data using ABC has been restricted to lattice-based IBMs (Johnston et al., 2014). This is partly due to the fact that ABC based inference techniques require a large number of IBM simulations, and lattice-free models are computationally expensive. We consider both lattice-based and lattice-free models in this thesis. First, we summarise experimental data with only the number of cells in a population to investigate the optimal duration of a cell proliferation assay, using a lattice-based IBM, and *in silico* data. Next, we also summarise experimental data from a cell proliferation assay with a measure of spatial structure, and calibrate a lattice-free IBM that describes the evolution of both the number of cells and the spatial structure.

Cell proliferation assays may be interpreted using mathematical models, which can provide insight into the mechanisms involved (Maini et al., 2004b; Sengers et al., 2007). A common assumption is that a cell population grows according to the classical logistic equation. However, this assumption is rarely tested against experimental data. Recent work of Sarapata and de Pillis (2014) show that, while the classical logistic model is appropriate in some cases, other generalised logistic growth models are better suited to describe different types of tumours. In Chapter 2, we explore whether a typical experimental design for a cell proliferation assay,

with a duration of 24 hours, is enough to distinguish between classical logistic and more generalised logistic growth. Our approach represents the first time that a stochastic IBM has been used to explore the optimal duration of a cell proliferation assay. Our results show that a typical experimental design is not enough to reliably distinguish between classical and generalised logistic growth models. We further explore how the experimental duration can be increased to reduce the uncertainty in the growth mechanism.

The question of the experimental duration required to reliability determine the growth mechanism of a cell proliferation assay has never been explored with an IBM. Our study in Chapter 2 incorporates the inherit stochasticity of experimental data with ABC inference techniques to explore how the uncertainty in the growth mechanism is reduced as the experimental duration is increased. While we restricted our study to a small family of growth mechanisms, with two key parameters, the techniques and measures we present can be trivially extended to any growth mechanism, with any number of unknown parameters. We demonstrate that a standard experimental duration of 24 hours is inadequate to determine the growth mechanism of a cell proliferation assay. As such, while the proliferation rate may still be calculated and used to compare cell proliferation assays from a typical experimental duration, calibrated mathematical models that assume the classical logistic equation should not be used to predict high density behaviour.

Typically, mathematical models that describe cell proliferation assays neglect the importance of spatial structure. Mean-field models, by definition, do not consider spatial structure, and it is common to summarise experimental data by only the population. While pair-wise models that capture spatial structure are routine in some fields, including disease spreading (Sharkey et al., 2006; Sharkey, 2008) and ecology (Law et al., 2003), they are rarely employed to study collective cell behaviour. In this work, we find evidence of clustering at short intervals, in the experimental data, at early time. We summarise spatial structure using a pair correlation measure, which also shows that spatial structure becomes less pronounced with as the cell density increases. In Chapter 3 we present a lattice-free IBM to describe novel experimental data. Our model captures the key mechanisms observed in the experimental data, namely proliferation, motility and a directional movement bias. While more complicated models have been studied (Kabla, 2012; Tang et al., 2014; Stichel et al., 2017), it is not clear whether calibration of these models to experimental data provides improved biological insight (Simpson et al., 2006).

In Chapter 3, we also calibrate a lattice-free IBM to the experimental data. Our approach is novel as lattice-free models are rarely calibrated to experimental

data to describe collective cell behaviour. Despite lattice-free IBMs being more computationally expensive than lattice-based IBMs, we find that it is preferable to work with lattice-free models as they make less simplifying assumptions. For example, lattice-based IBMs require approximations when mapping the distribution of cells in experimental images to a lattice (Johnston et al., 2014, 2016). In addition, lattice-based models restrict the separation of pairs of cells to discrete values. As such, they may be less effective in describing spatial structure than lattice-free models. High performance computing infrastructure, and an appropriate number of unknown parameters, allow us to calibrate three independent experimental data sets. We find that the posterior distributions for each unknown parameter, between experimental data sets, are similar. Observing that the parameters are only weakly correlated, we combine the posterior distributions from each experimental data set, to produce a combined posterior distribution for each parameter. To investigate the predictive power of the calibrated model, we predict distribution estimates for each of summary statistics for an additional experimental data set. We find that the model is able to produce a reasonable match to the additional experimental data set.

There are many ways our study could be extended. In Chapter 2, we assume relatively simple, general logistic models, and explore our ability to distinguish between growth models as the experimental duration is increased. However, we only summarise data by the number of cells, and assume that there is no spatial structure. While this may be true for typical experimental data, we could extend our study to examine this question for higher quality experimental data, such as the experimental data we present in Chapter 3. In addition, we may apply our modelling and inference techniques from Chapter 3 to explore how experimental duration affects our ability to infer the unknown parameters in an IBM.

In Chapter 3, we connect three independent, relatively low density, experimental data sets to a lattice-free IBM. However, we only summarise experimental data by the number of cells, and a relatively simple pair correlation measure. Additional summary statistics that could be employed include the pair correlation at more distances, and the density of triples. This study could be extended by testing the predictive capability of the IBM against an additional summary statistic. Also, these additional summary statistics could be applied in the calibration process. To increase our confidence in the recovered parameters, and the predictive power of the model, we could also repeat our methods with more experimental data sets, including those with larger initial number of cells. We could relax some of the simplifying assumptions, and explore how increasing the number of unknown parameters affects both the parameter estimates, and the predictive power of the

calibrated IBM. Finally, the techniques we employ to calibrate, and test, our IBM can be applied to any model, with any number of parameters.

Connecting mathematical models to experimental data can provide insight into the mechanisms involved in cell proliferation assays. A common assumption is that cells proliferate according to the classical logistic equation, however this assumption is rarely tested by applying experimental data. In this thesis, we describe a technique for determining the optimal duration of a cell proliferation assay in order to determine whether cells proliferate according to classical logistic growth or another growth mechanism. We are able to demonstrate how a typical experimental design, with a duration of 24 hours, provides only very limited insight into the mechanisms that describe the growth of a cell population. We also calibrate a mathematical model that can describe both population growth and spatial structure observed in new experimental data. We test the validity of the calibrated model by accurately predicting the growth of a cell population in an additional independent experimental data set. This thesis highlights the unique ability of mathematical modelling tools to interpret the mechanisms involved in collective cell behaviour.

Bibliography

Baker RE, Simpson MJ (2010) Correcting mean-field approximations for birth-death-movement processes. *Physical Review E* **82**(4), 041905. 4, 49, 60

Binny RN, Haridas P, James A, Law R, Simpson MJ, Plank MJ (2016a) Spatial structure arising from neighbour-dependent bias in collective cell movement. *PeerJ* **4**, e1689. 5, 41, 45, 50

Binny RN, James A, Plank MJ (2016b) Collective cell behaviour with neighbour-dependent proliferation, death and directional bias. *Bulletin of Mathematical Biology* **78**(11), 2277–2301. 5, 41, 45, 46, 47, 49, 50, 58, 60, 64

Bosco DB, Kenworthy R, Zorio DAR, Sang QXA (2015) Human mesenchymal stem cells are resistant to paclitaxel by adopting a non-proliferative fibroblastic state. *PLoS One* **10**(6), e0128511. 1, 13, 41

Bourseguin J, Bonet C, Renaud E, Pandiani C, Boncompagni M, Giuliano S, Pawlikowska P, Karmous-Benailly H, Ballotti R, Rosselli F, Bertolotto C (2016) FANCD2 functions as a critical factor downstream of MiTF to maintain the proliferation and survival of melanoma cells. *Scientific Reports* **6**, 36539. 1, 13, 41

Browning AP, McCue SW, Simpson MJ (2017) A Bayesian computational approach to explore the optimal duration of a cell proliferation assay. *Bulletin of Mathematical Biology* **79**. 7, 41, 49

Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, Berlin. 23

Cai AQ, Landman KA, Hughes BD (2007) Multi-scale modeling of a wound-healing cell migration assay. *Journal of Theoretical Biology* **245**(3), 576–594. 2, 13, 41, 45, 47, 58

Codling EA, Plank MJ, Benhamou S (2008) Random walk models in biology. *Journal of The Royal Society Interface* **5**(25), 813–834. 5, 41

Collis J, Connor AJ, Paczkowski M, Kannan P, Pitt-Francis J, Byrne HM, Hubbard ME (2017) Bayesian calibration, validation and uncertainty quantification for predictive modelling of tumour growth: A tutorial. *Bulletin of Mathematical Biology* **79**(4), 1–36. 17, 20, 42, 50

Dale PD, Sherratt JA, Maini PK (1994) The speed of corneal epithelial wound healing. *Applied Mathematics Letters* **7**(2), 11–14. 13

Deroulers C, Aubert M, Badoual M, Grammaticos B (2009) Modeling tumor cell migration: From microscopic to macroscopic models. *Physical Review E* **79**(3), 031917. 17

Doran MR, Mills RJ, Parker AJ, Landman KA, Cooper-White JJ (2009) A cell migration device that maintains a defined surface with no cellular damage during wound edge generation. *Lab on a Chip* **9**(16), 2364–2369. 13

Edelstein-Keshet L (1988) *Mathematical Models in Biology*. Random House, New York. 2, 4, 13

Fletcher AG, Breward CJ, Chapman SJ (2012) Mathematical modeling of monoclonal conversion in the colonic crypt. *Journal of Theoretical Biology* **300**, 118 – 133. 59

Forbes C, Evans M, Hastings N, Peacock B (2011) *Statistical Distributions*. John Wiley & Sons, New Jersey, 4 edition. 49

Frascoli F, Hughes BD, Zaman MH, Landman KA (2013) A computational model for collective cellular motion in three dimensions: General framework and case study for cell pair dynamics. *PLoS ONE* **8**(3), e59249. 41

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*. CRC Press, Florida, 2 edition. 20

Gerlee P (2013) The model muddle: In search of tumor growth laws. *Cancer Research* **73**(8), 2407. 4, 14

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **81**(25), 2340–2361. 49

Jin W, Penington CJ, McCue SW, Simpson MJ (2016a) Stochastic simulation tools and continuum models for describing two-dimensional collective cell spreading

with universal growth functions. *Physical Biology* **13**(5), 056003. 2, 4, 5, 16, 17, 18, 19

Jin W, Shah ET, Penington CJ, McCue SW, Chopin LK, Simpson MJ (2016b) Reproducibility of scratch assays is affected by the initial degree of confluence: Experiments, modelling and model selection. *Journal of Theoretical Biology* **390**, 136–145. 2, 13, 54, 63

Jin W, Shah ET, Penington CJ, McCue SW, Maini PK, Simpson MJ (2017) Logistic proliferation of cells in scratch assays is delayed. *Bulletin of Mathematical Biology* **79**(5), 1028–1050. 44, 45, 46, 56

Johnston ST, Ross JV, Binder BJ, Sean McElwain DL, Haridas P, Simpson MJ (2016) Quantifying the effect of experimental design choices for in vitro scratch assays. *Journal of Theoretical Biology* **400**, 19–31. 58, 71

Johnston ST, Shah ET, Chopin LK, McElwain DLS, Simpson MJ (2015) Estimating cell diffusivity and cell proliferation rate by interpreting IncuCyte ZOOM™ assay data using the Fisher-Kolmogorov model. *BMC Systems Biology* **9**(1), 38. 1, 13, 17, 42, 50, 54, 58

Johnston ST, Simpson MJ, McElwain SDL, Binder BJ, Ross JV (2014) Interpreting scratch assays using pair density dynamics and approximate Bayesian computation. *Open Biology* **4**(9). 5, 41, 58, 69, 71

Kabla AJ (2012) Collective cell migration: leadership, invasion and segregation. *Journal of The Royal Society Interface* **9**(77), 3268. 59, 70

Kaighn ME, Narayan KS, Ohnuki Y, Lechner JF, Jones LW (1979) Establishment and characterization of a human prostatic carcinoma cell line (PC-3). *Investigative Urology* **17**(1), 16–23. 1, 41, 42, 44

Kullback S, Leibler RA (1951) On information and sufficiency. *The Annals of Mathematical Statistics* **22**(1), 79–86. 23

Laird AK (1964) Dynamics of tumour growth. *British Journal of Cancer* **18**(3), 490–502. 14

Law R, Murrell DJ, Dieckmann U (2003) Population growth in space and time: Spatial logistic equations. *Ecology* **84**(1), 252–262. 46, 49, 50, 60, 70

Liang CC, Park AY, Guan JL (2007) In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro. *Nature Protocols* **2**(2), 329–333. 1, 13

Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MPH (2014) A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nature Protocols* **9**(2), 439–456. 5, 17, 22, 51, 69

Liggett TM (1999) *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*. Springer, Berlin. 17

Maclaren OJ, Fletcher AG, M BH, Maini PK (2015) Models, measurement and inference in epithelial tissue dynamics. *ArXiv e-prints* 1506.05052. 58, 59

Maini PK, McElwain DLS, Leavesley D (2004a) Travelling waves in a wound healing assay. *Applied Mathematics Letters* **17**(5), 575–580. 2, 13, 41, 60

Maini PK, McElwain DLS, Leavesley DI (2004b) Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal mesothelial cells. *Tissue Engineering* **10**(3-4), 475–482. 2, 4, 13, 69

Mathworks (2017) Kernel smoothing function estimate for univariate and bivariate data. http://www.mathworks.com/help/stats/ksdensity.html. Accessed: May 2017. 22, 51

Murray JD (2002) *Mathematical Biology*. Springer, Berlin. 2, 4, 13, 44, 56

O'Dea RD, Byrne HM, Waters SL (2012) *Continuum Modelling of In Vitro Tissue Engineering: A Review*, volume 10. Springer, Berlin, 229–266. 13

Pearl R (1927) The growth of populations. *Quarterly Review of Biology* **2**(4), 532–548. 2, 4, 13

Peirce S, van Gieson E, Skalak T (2004) Multicellular simulation predicts microvascular patterning and in silico tissue assembly. *FASEB J* **18**(6), 731–733. 41

Plank MJ, Simpson MJ (2012) Models of collective cell behaviour with crowding effects: comparing lattice-based and lattice-free approaches. *Journal of The Royal Society Interface* **9**(76), 2983–2996. 5, 41, 47, 58

Ramin M, Arhonditsis GB (2013) Bayesian calibration of mathematical models: Optimization of model structure and examination of the role of process error covariance. *Ecological Informatics* **18**, 107–116. 5

Read M, Andrews PS, Timmis J, Kumar V (2012) Techniques for grounding agent-based simulations in the real domain: a case study in experimental autoimmune

encephalomyelitis. *Mathematical and Computer Modelling of Dynamical Systems* **18**(1), 67–86. 41

Sarapata EA, de Pillis LG (2014) A comparison and catalog of intrinsic tumor growth models. *Bulletin of Mathematical Biology* **76**(8), 2010–2024. 2, 4, 14, 16, 30, 41, 60, 69

Savla U, Olson LE, Waters CM (2004) Mathematical modeling of airway epithelial wound closure during cyclic mechanical strain. *Journal of Applied Physiology* **96**(2), 566. 13

Schneider C, Rasband W, Eliceiri K (2012) NIH image to ImageJ: 25 years of image analysis. *Nature Methods* **9**, 671–675. 61

Sengers BG, Please CP, Oreffo ROC (2007) Experimental characterization and computational modelling of two-dimensional cell spreading for skeletal regeneration. *Journal of The Royal Society Interface* **4**(17), 1107. 2, 13, 41, 60, 69

Sharkey KJ (2008) Deterministic epidemiological models at the individual level. *Journal of Mathematical Biology* **57**(3), 311–331. 60, 70

Sharkey KJ, Fernandez C, Morgan KL, Peeler E, Thrush M, Turnbull JF, Bowers RG (2006) Pair-level approximations to the spatio-temporal dynamics of epidemics on asymmetric contact networks. *Journal of Mathematical Biology* **53**(1), 61–85. 60, 70

Sheardown H, Cheng YL (1996) Mechanisms of corneal epithelial wound healing. *Chemical Engineering Science* **51**(19), 4517–4529. 2, 13

Sherratt JA, Murray JD (1990) Models of epidermal wound healing. *Proceedings of the Royal Society B: Biological Sciences* **241**(1300), 29. 2, 4, 13, 41, 60

Simpson MJ, Landman KA, Hughes BD, F Newgreen D (2006) Looking inside an invasion wave of cells using continuum models: Proliferation is the key. *Journal of Theoretical Biology* **243**(3), 343–360. 59, 70

Simpson MJ, Sharp JA, Baker RE (2014) Distinguishing between mean-field, moment dynamics and stochastic descriptions of birth–death–movement processes. *Physica A: Statistical Mechanics and its Applications* **395**, 236–246. 20

Simpson MJ, Treloar KK, Binder BJ, Haridas P, Manton KJ, Leavesley DI, McElwain DLS, Baker RE (2013) Quantifying the roles of cell motility and cell proliferation in a circular barrier assay. *Journal of The Royal Society Interface* **10**(82). 17, 19

Stichel D, Middleton AM, Müller BF, Depner S, Klingmüller U, Breuhahn K, Matthäus F (2017) An individual-based model for collective cancer cell migration explains speed dynamics and phenotype variability in response to growth factors. *npj Systems Biology and Applications* **3**(1), 5. 59, 70

Stoll BR, Migliorini C, Kadambi A, Munn LL, Jain RK (2003) A mathematical model of the contribution of endothelial progenitor cells to angiogenesis in tumors: implications for antiangiogenic therapy. *Blood* **102**(7), 2555. 2

Sunnaker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C (2013) Approximate Bayesian computation. *PLoS Computational Biology* **9**(1), e1002803. 5, 17, 20, 22, 69

Tanaka MM, Francis AR, Luciani F, Sisson SA (2006) Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics* **173**(3), 1511–1520. 5, 17, 20, 22, 42, 50, 51, 69

Tang L, van de Ven AL, Guo D, Andasari V, Cristini V, Li KC, Zhou X (2014) Computational modeling of 3D tumor growth and angiogenesis for chemotherapy evaluation. *PLoS ONE* **9**(1), e83962. 59, 70

Treloar KK, Simpson MJ, Haridas P, Manton KJ, Leavesley DI, McElwain DLS, Baker RE (2013) Multiple types of data are required to identify the mechanisms influencing the spatial expansion of melanoma cell colonies. *BMC Systems Biology* **7**(1), 137. 31, 41

Treloar KK, Simpson MJ, Sean McElwain DL, Baker RE (2014) Are in vitro estimates of cell diffusivity and cell proliferation rate sensitive to assay geometry? *Journal of Theoretical Biology* **356**, 71–84. 2, 4, 13, 14, 20, 63

Tremel A, Cai A, Tirtaatmadja N, Hughes BD, Stevens GW, Landman KA, O'Connor AJ (2009) Cell migration and proliferation during monolayer formation and wound healing. *Chemical Engineering Science* **64**(2), 247–253. 1, 2, 41, 45, 60

Tsoularis A, Wallace J (2002) Analysis of logistic growth models. *Mathematical Biosciences* **179**(1), 21–55. 4, 14, 16

Vo BN, Drovandi CC, Pettit AN, Simpson MJ (2015) Quantifying uncertainty in parameter estimates for stochastic models of collective cell spreading using approximate Bayesian computation. *Mathematical Biosciences* **263**, 133–142. 17

West GB, Brown JH, Enquist BJ (2001) A general model for ontogenetic growth. *Nature* **413**(6856), 628–631. 14

Zwietering MH, Jongenburger I, Rombouts FM, van't Riet K (1990) Modeling of the bacterial growth curve. *Applied and Environmental Microbiology* **56**(6), 1875–1881. 4, 14