Parameter-wise predictions and sensitivity analysis for random walk models in the life sciences

Yihan Liu[1], David J. Warne[1], and Matthew J. Simpson[1,2]

[1]School of Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, Australia.

[2]ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems, QUT, Brisbane, Australia.

Abstract

Sensitivity analysis characterises input–output relationships for mathematical models, and has been widely applied to deterministic models across many applications in the life sciences. In contrast, sensitivity analysis for stochastic models has received less attention, with most previous work focusing on well-mixed, non-spatial problems. For explicit spatio-temporal stochastic models, such as random walk models (RWMs), sensitivity analysis has received far less attention. Here we present a new type of sensitivity analysis, called *parameter-wise prediction*, for two types of biologically-motivated and computationally expensive RWMs. To overcome the limitations of directly analysing stochastic simulations, we employ continuum-limit partial differential equation (PDE) descriptions as surrogate models, and we link these efficient surrogate descriptions to the RWMs using a range of biophysically-motivated *measurement error models*. Our approach is likelihood-based, which means that we also consider likelihood-based parameter estimation and identifiability analysis along with parameter sensitivity. The new approach is presented for two important classes of lattice-based RWM including a classical model where crowding effects are neglected, and an exclusion process model that explicitly incorporates crowding. Our workflow illustrates how different process models can be combined with different measurement error models to reveal how each parameter impacts the outcome of the expensive stochastic simulation. Open-access software to replicate all results is available on GitHub.

1 Introduction

Sensitivity analysis aims to examine how the output of a mathematical model relates to, or is influenced by the inputs. Model inputs can include model parameter values and initial conditions [1], while outputs typically refer to the solution of the mathematical model, the result of a single realisation of a stochastic model, or some average result obtained from an ensemble of stochastic simulations [2]. Many studies focus on deterministic models as they are widely used and easily interpretable in the sense that the model output is free from randomness and uncertainty [3–7]. Sensitivity analysis for stochastic models has been explored to a lesser extent, primarily for stochastic reaction networks (e.g. chemical reaction networks, birth–death models, gene expression networks) [8–12], where individuals within the population are assumed to be well-mixed and model outputs (e.g. the density of chemical species) are translationally invariant.

Lattice-based random walk models (RWMs), that describe the motion of individual agents on a lattice, are frequently used to interpret cell biology experiments involving cell migration and proliferation [13–19]. Here we focus on two types of RWMs, each motivated by a different kind of cell biology experiment. The first RWM is inspired by scratch assay data [20], using Fluorescence Ubiquitin Cell Cycle Indicator (FUCCI) labelling [21], illustrated in Figure 1(a), where cells in different phases of the cell cycle fluoresce red or green. A schematic of the cell cycle is illustrated in Figure 1(b). These experiments provide information about how cell position relates to cell cycle status [21], and are often performed under low-density conditions where cells are free to progress through the cell cycle without any contact inhibition [20, 22]. The second RWM is motivated by a circular barrier assay, illustrated in Figure 1(c), which involves growing high-density, uniform monolayers of cells within a circular barrier to mimic a high cell density *in vivo* environment [23, 24]. The circular barrier assay involves placing cells within the barrier and allowing time for cells to attach the cell culture plate. The barrier is then lifted, non-viable cells are washed away, and the subsequent longer-term radial expansion of population is imaged. Schematics of the circular barrier assay are illustrated in Figure 1(f)-(g).

Parameter estimation and model predictions are important if we wish to gain mechanistic insight or to make predictions about future experimental scenarios. Estimation and prediction can be made using likelihood-based [13, 25–28] or Bayesian methods [17, 18, 29–33]. For RWMs in the theoretical biology literature, it has become customary to focus on likelihood-free Bayesian approaches [17, 18, 29, 30, 33], but more recently we have focused on computationally efficient likelihood-based approaches that we will follow here [13].

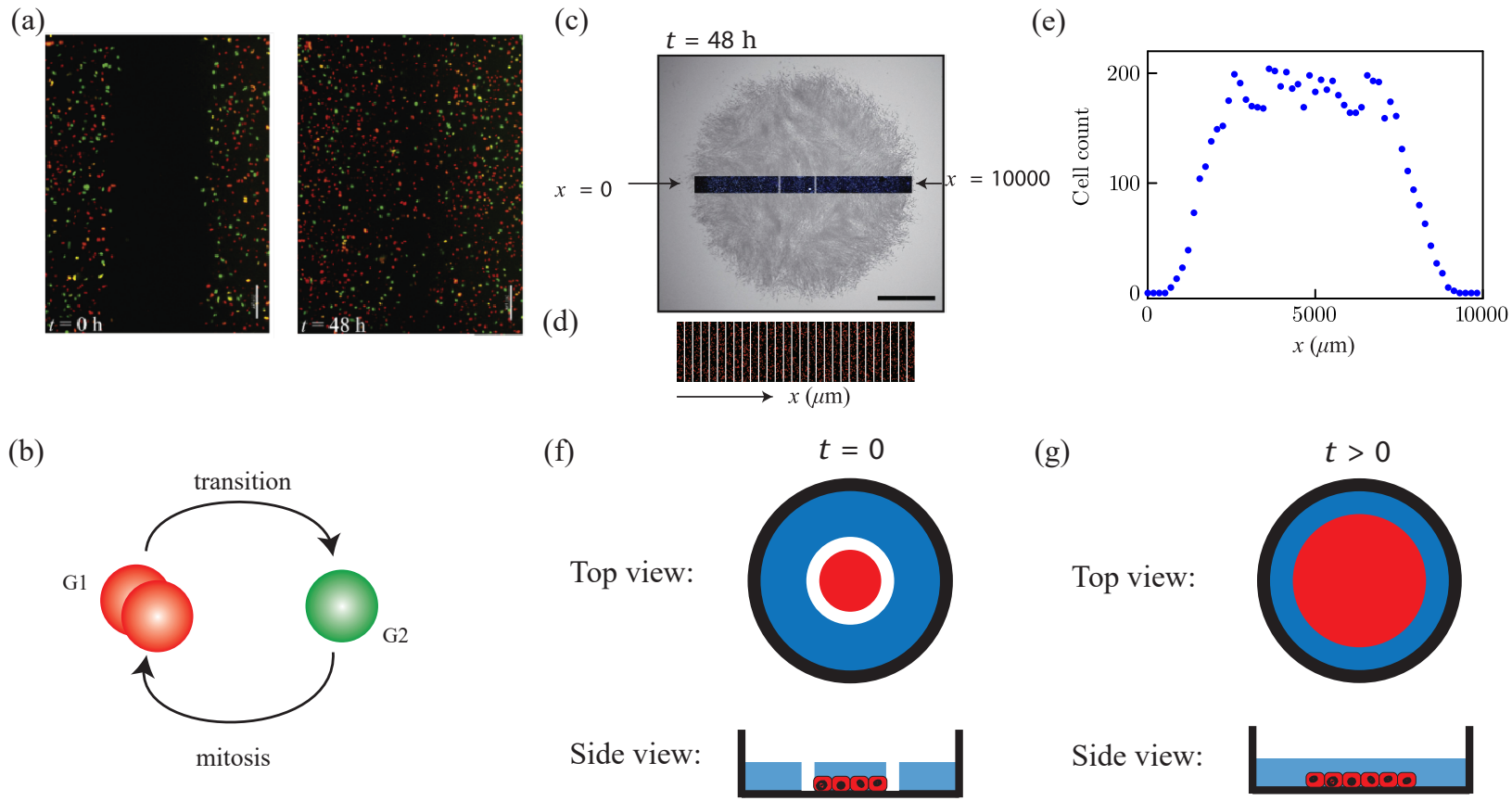

Figure 1: Experimental motivation. (a) Images of a scratch assay using melanoma cells with FUCCI at 0 h and 48 h. Scale bar corresponds to 200 μm [20]. (b) Cell cycle schematic indicating FUCCI fluorescence. (c) Barrier assay at 48 h with 2 mm scale bar [23]. (d) Transect through the centre of the spreading population in (c). (e) Count data collected from (d). (f)–(g) Schematic of the circular barrier assay. Black ring (open rectangle) represents the well, the blue region denotes the cell culture medium, the red regions indicate the expanding cell monolayer, and the white ring (bars) in (e) indicate the barrier. Images are reused with permission.

In this work, we follow a likelihood-based approach by using maximum likelihood estimation (MLE) to estimate best-fit model parameters for a range of RWMs. To ensure these estimates are biologically meaningful and interpretable, we consider parameter identifiability analysis to determine whether the data contain sufficient information to yield precise estimates [3, 4, 13, 34–36]. In particular, we use the *profile likelihood* to assess the practical identifiability [35, 37–40] by partitioning the parameter vector θ into two parts, $\theta = (\psi, \omega)^\top$, where ψ represents the *interest* parameter(s) and ω represents the *nuisance* parameter(s) [37]. This targeting property of the profile likelihood allows us to assess parameter identifiability for complex models with many parameters by systematically isolating and analysing each parameter while optimising out the nuisance parameters. While predicting model outcomes across the full parameter space is routinely used to understand how uncertainties in parameter estimates propagate into uncertainties in model predictions [13, 41–43], this standard approach does not provide information about the role of individual parameters as in the context of sensitivity analysis that we focus on.

For RWMs in which agents migrate in physical space, sensitivity analysis has predominantly been conducted using methods originally designed for deterministic models, where the only source of uncertainty affecting the output is the variation in the input parameters and initial conditions [44]. Output uncertainties for RWMs arise from input variation (e.g. parameter values, initial condition) and from inherent stochasticity. To reduce the impact of inherent model stochasticity, model outputs are often averaged over many identically prepared realisations which can dramatically increase the computational cost of the analysis [45–48]. Of the few studies that perform sensitivity analyses for RWMs, most employ the partial rank correlation coefficient (PRCC) [45–48]. The PRCC quantifies the influence of each input parameter through a numerical index, called the *sensitivity index* [44, 49]. A positive (negative) sensitivity index indicates a positive (negative) correlation between the parameter and the model output. Sensitivity analysis using PRCC is typically combined with Latin hypercube sampling (LHS) [50]. To perform LHS-PRCC, a distribution is specified for each input parameter, and LHS is used to generate random parameter combinations that cover the parameter space. For each random parameter combination, the model is simulated to produce the model output(s). All input and outputs are *rank-transformed*. To isolate the impact of a target input, two linear regressions are performed: (i) the target input is regressed on the remaining inputs, and (ii) the output is regressed on all inputs except the target input. The sensitivity index is calculated as the correlation between the residuals from the two regressions. While LHS-PRCC can isolate the effect of parameters one at a time, randomly drawn parameter samples can lead to working with insufficient or implausible parameter

combinations [44] which means that LHS-PRCC can be computationally inefficient. Therefore, addressing the limitations of working with random parameter samples by developing computationally efficient approaches for sensitivity analysis remains a significant challenge in the field of RWMs.

One approach to reduce the computational cost of sensitivity analysis is to use efficient surrogate models to approximate the output of a RWM [51–54]. This can include working with continuum-limit descriptions [13, 41, 51, 55, 56]. In addition to improving computational efficiency via working with surrogates, we also avoid working with randomly chosen samples parameters as in LHS-PRCC. To this end, Simpson and Maclaren [3, 57] proposed an efficient method for sensitivity analysis based on the targeting property of the profile likelihood. Parameter-wise prediction intervals provide both qualitative and quantitative insight into how uncertainty in individual parameters, or combinations of parameters, influences variability in model predictions. Importantly, the parameter combinations used to construct parameter-wise prediction intervals are not random combinations. Instead, parameter combinations are based on optimising a log-likelihood function which ensures that we focus on relevant parameter combinations that lead to model solutions that are close to the observed data. The parameter-wise prediction approach has been applied to various continuum models, including both identifiable [3, 4, 57] and non-identifiable models [34], but it has not yet been employed in the context of more complicated and computationally expensive stochastic models, such as RWMs. In this work, we conduct the first sensitivity analysis for RWMs using parameter-wise prediction.

2 Methods

2.1 Motivating experiments

We use two types of lattice-based RWMs motivated by two different kinds of cell biology experiments presented schematically in Figure 1. The first experiment is a scratch assay performed under low cell density conditions, as illustrated in Figure 1(a) [20]. Scratch assays involve growing uniform monolayers of cells on tissue culture plates before part of the monolayer is scratched away. The resulting temporal recolonisation of the scratched region is imaged [58, 59]. The scratch assay of interest is conducted using FUCCI labelling where cells in G1 phase fluoresce red, and cells in G2 phase fluoresce green [20]. Cells in G1 phase progress through the cell cycle into G2 phase, and cells in G2 phase undergo mitosis,

producing two daughter cells that are both in G1 phase, see Figure 1(b). These experiments are performed under low density conditions to minimise the impact of contact inhibition so it is natural to model this experiment using a non-interacting RWM [19]. The RWM is non-interacting in the sense that multiple agents can reside on the same site, and agent movement is unaffected by the location of other agents. Since the experiments involve two subpopulations, each of which corresponds to cells in different phases of the cell cycle, the non-interacting RWM will also involve two subpopulations that represent cells in the two different phases of the cell cycle as identified in the experiments.

The second experiment is a circular barrier assay conducted under high cell density conditions, describing the spatial expansion of a circular monolayer of cells, as illustrated schematically in Figure 1(f)–(g). In this experiment, circular barriers are placed at the centre of each well in a 24-well tissue culture plate. Cells are seeded uniformly within circular barriers, and some time is allowed for cells to attach to the culture plate before the barriers are removed to initiate the experiment at $t = 0$. The initial cell monolayer is briefly washed to remove non-viable and unattached cells, and the subsequent outward radial expansion of the cell population is imaged at later times, as shown in Figure 1(c). As the circular barrier assays are conducted under high cell density conditions, it is important to account for crowding effects when modelling these experiments so we use an interacting RWM. The RWM is interacting in the sense that each lattice site can be occupied by, at most, a single agent, and agent movement is affected by the location of other agents on the lattice [19].

2.2 Mathematical models

Sections 2.2.1–2.2.2 present details of the non-interacting and interacting RWMs, respectively. Both models involve a 2D square lattice with lattice spacing Δ, and individual events are simulated using a random sequential update method [60]. In all simulations, each agent represents an individual biological cell. The lattice has dimensions $I \times J$, consisting of I columns and J rows, with reflecting boundary conditions applied along all boundaries. Each lattice site is indexed by (i, j), and is associated with a unique Cartesian coordinate (x_i, y_j), where $x_i = (i-1)\Delta$ for $i = 1, 2, 3, \ldots, I$, and $y_j = (j-1)\Delta$ for $j = 1, 2, 3, \ldots, J$. All RWMs are discrete-time models, and the evolution of the models involve time steps of constant duration, τ. Time steps are indexed by k so that $t_k = k\tau$ for $k = 1, 2, 3, \ldots, K$. To keep our simulation framework general, we work with dimensionless simulations by setting $\Delta = \tau = 1$. These simulations can be re-scaled using appropriate length and time scales to accommodate cells of different sizes, motility rates, and cell cycle progression rates [61].

2.2.1 Non-interacting random walk model

Let $R(t)$ and $G(t)$ denote the total number of red and green agents that represent cells in G1 and G2 phase at time t, respectively, so that the total number of agents is $T(t) = R(t)+G(t)$. We initialise the RWM by placing $R(0)$ red agents, and $G(0)$ green agents on the lattice. Since the RWM is non-interacting, each site can be occupied by multiple agents. To evolve the model from time t to time $t+\tau$, we consider four types of stochastic events: (i) motility of red agents; (ii) motility of green agents; (iii) red agents progressing through the cell cycle and transitioning into green agents (red-to-green transition); and, (iv) green agents undergoing mitosis to produce two daughter agents that will be at the beginning of their cell cycle, belonging to the G1 subpopulation and accordingly coloured red.

To simulate agent motility, $R(t)$ red agents and $G(t)$ green agents are selected independently at random, one at a time, with replacement. Each selected agent is given an opportunity to move. Experimental evidence suggests motility of cells in G1 phase can be indistinguishable from cells in G2 phase [62]. We incorporate this into our model by allowing each selected agent to moveve with probability $P \in [0, 1]$, regardless of their cell cycle status. The target site for a potential motility event is chosen as follows: the probability that a motile agent at site (i, j) steps to $(i, j \pm 1)$ is $1/4$, while the probability that it steps to $(i \pm 1, j)$ is $(1 \pm \rho)/4$, where $\rho \in [-1, 1]$ is a bias parameter that influences directional bias in the horizontal direction. Again, the bias parameter is taken to be identical for both subpopulations. Setting $\rho = 0$ corresponds to unbiased motility. The bias is considered only in the horizontal direction, as the macroscopic density of both subpopulations is independent of vertical location due to the geometry of the experiment in Figure 1(a). While the scratch assay in Figure 1(a) does not provide any obvious visual evidence of migration bias, it is often thought that bias mechanisms (e.g. chemotaxis) can play important roles under certain circumstances [63, 64]. For completeness, we include the bias term in our model so it is clear that our approach is applicable to both biased and unbiased cell migration. To simulate the red-to-green transitions, $R(t)$ red agents are selected independently at random, one at a time, with replacement. Each selected red agent progresses through the cell cycle, transitioning into a green agent with probability $\lambda_1 \in [0, 1]$. To simulate proliferation, $G(t)$ green agents are selected independently at random, one at a time, with replacement. Each selected green agent undergoes mitosis to produce two red agents with probability $\lambda_2 \in [0, 1]$. When proliferation takes place, both red daughter agents are placed on the same lattice site as the original parent green agent.

Although cells are free to move in any direction in the experiment in Figure 1(a), the macro-

scopic cell density is independent of the vertical position and varies with horizontal position. This geometric simplification is due to the fact that scratches are made within uniform monolayers. Experimental images within this setting can be quantified using count data obtained by superimposing a series of uniformly spaced columns across the image and counting the number of red and green cells separately in each column [41]. This approach summarises the experimental outcome as a series of count data as a function of the horizontal location of each column, as in Figure 1(c)–(e). For simplicity, it is common to report count data at a single time point at the end of the experiment, and we model this approach by using our models to simulate count data at a single time point only.

We will now explain how to generate count data from the non-interacting RWM so that it mimics the geometry and count data collected from the experiment in Figure 1(a). The experiment is designed to allow the population of cells to migrate inwards into an initially-vacant region, where the moving fronts eventually meet and the scratched region becomes occupied, as in Figure 1(a). In contrast, our RWM is designed such that agents tend to move outwards into empty space, resulting in a persistent spreading population with leading edges that progress outward without meeting. This outward migration design provides more information, particularly for simulations over longer time periods [65]. Despite these differences in movement direction, the important geometric features of our RWM match the experiment. All simulations are initialised such that the expected occupancy of lattice sites within each column is identical. Combined with reflecting boundary conditions, this ensures that the average density of any lattice site remains independent of vertical position throughout the simulation as in the experiment. The outcome of any RWM simulation provides the number of agents at each lattice site for different subpopulations as a function of time. We define $C_s^*(i,j,k) \in \{0,1,2,3,\ldots\}$ as the unbounded integer number of agents from the sth subpopulation at site (i,j) after k time steps, where $s = 1,2,\ldots,S$. For our application, we have $S = 2$. With this notation, the total number of agents from subpopulation s in the ith column after k time steps is

$$C_s^{\mathrm{o}}(i,k) = \sum_{j=1}^{J} C_s^*(i,j,k). \tag{1}$$

While the RWM provides a high-fidelity means of generating noisy data that mimic those collected from cell biology experiments, it can be computationally expensive to generate large numbers of forward simulations to obtain detailed information about the impact of different parameter choices. To address this limitation, we use a computationally efficient

continuum-limit surrogate model that is given by

$$\frac{\partial c_1}{\partial t} = D\frac{\partial^2 c_1}{\partial x^2} - v\frac{\partial c_1}{\partial x} - k_1 c_1 + 2k_2 c_2, \tag{2}$$

$$\frac{\partial c_2}{\partial t} = D\frac{\partial^2 c_2}{\partial x^2} - v\frac{\partial c_2}{\partial x} + k_1 c_1 - k_2 c_2, \tag{3}$$

where

$$D = \lim_{\Delta,\tau\to 0}\frac{P\Delta^2}{4\tau}, \quad v = \lim_{\Delta,\tau\to 0}\frac{P\rho\Delta}{2\tau}, \quad k_1 = \lim_{\tau\to 0}\frac{\lambda_1}{\tau}, \qquad k_2 = \lim_{\tau\to 0}\frac{\lambda_2}{\tau}. \tag{4}$$

Here, $c_1(x,t) \geq 0$ and $c_2(x,t) \geq 0$ are the dimensionless densities of red and green agents at location x and time t, respectively. For the geometry and initial condition that we consider, Equations (2)–(3) have an exact solution for which the details are provided in Section 1 of the Supplementary Material.

An implicit assumption in deriving the surrogate continuum-limit model is that we are working with a sufficiently large lattice so that the discrete count data can be approximated by a smooth continuous function of position. Under this idealised scenario, it is possible to precisely relate the count data generated from the RWM to the solution of the corresponding continuum-limit PDE model in the following way,

$$\begin{aligned} c_s(x_i,t) &= \lim_{J\to\infty}\left[\frac{\sum_{j=1}^{J} C_s^*(i,j,k)}{J}\right], \\ &= \lim_{J\to\infty}\left[\frac{C_s^{\mathrm{o}}(i,k)}{J}\right]. \end{aligned} \tag{5}$$

This idealised relationship is confirmed in Section 3 of the Supplementary Material. In practice however, we must deal with the case that J is finite, as experiments always involve a relatively small field of view, corresponding to small, finite J. These practical conditions mean that count data typically exhibits large fluctuations, such as the experimental count data in Figure 1(e). Consequently, the relationship between the observed count data, $C_s^{\mathrm{o}}(i,k)$, and the solution of the continuum-limit model, $c_s(x,t)$, is unclear. We will address this question by introducing physically-motivated *measurement error models* for different kinds of RWMs in Section 2.4 [4]. We note that measurement error models are also referred to as error models or observation models in the literature. In this work we use the term measurement error model.

2.2.2 Interacting random walk model

Simulations of circular barrier assays are initialised by uniformly placing $T(0)$ agents on lattice sites within a circular region of radius r_0. Each site within this circular region has

an occupancy probability $c_0 \in [0, 1]$ and since we are dealing with an interacting RWM each site can be occupied by, at most, a single agent. To evolve the stochastic algorithm from time t to time $t + \tau$, we consider two types of stochastic events: (i) unbiased agent motility; and, (ii) agent proliferation. Motility events are simulated in a similar way to those in the non-interacting RWM, except that here we have a single population, $S = 1$, and we set $\rho = 0$, since our experimental images indicate the outward spreading population maintains a circular shape. To simulate proliferation, $T(t)$ agents are selected independently at random, one at a time, with replacement, and each is given an opportunity to proliferate. When an agent is selected, it attempts to proliferate with probability $\lambda \in [0, 1]$. Upon proliferation, the agent deposits a daughter agent at one of its four nearest neighbour sites with equal probability. The key difference between the non-interacting and interacting RWM is that here we account for crowding effects by aborting potential motility or proliferation events that would place any agent on an occupied site. This means that the interacting RWM is closely related to an exclusion process [60].

The outcome of the circular barrier assay in Figure 1(c) is quantified by placing a narrow rectangular transect across the spreading monolayer [23]. Count data are collected by superimposing a series of uniformly spaced columns across the transect, as illustrated in Figure 1(d). We summarise the experimental data as a series of cell counts as a function of horizontal position, as shown in Figure 1(e). The interacting RWM can be used to generate count data that closely resembles the experimental data. To generate count data we place a rectangular transect across the centre of the lattice. The transect has dimensions $I \times H$, with the lower horizontal boundary at $j = j_1$ and the upper horizontal boundary at $j = j_1 + H$. We let $C^*(i, j, k)$ represent the occupancy of site (i, j) after k time steps. If site (i, j) is occupied we have $C^*(i, j, k) = 1$, otherwise $C^*(i, j, k) = 0$. The count data at the ith column after k time steps within the transect is

$$C^{\mathrm{o}}(i, k) = \sum_{j=j_1}^{j_1+H} C^*(i, j, k). \tag{6}$$

so that $C^{\mathrm{o}}(i, k) \in \{0, 1, 2, 3, \ldots, H\}$. In summary, a key difference between the two types of column-based count data that we consider in this work is that count data from the non-interacting RWM are unbounded $C_s^{\mathrm{o}}(i, k) \in \{0, 1, 2, 3, \ldots\}$, while count data from the interacting RWM are bounded $C^{\mathrm{o}}(i, k) \in \{0, 1, 2, 3, \ldots, H\}$.

The continuum-limit description of the interacting RWM is a 2D generalisation of the well-known Fisher–Kolmogorov equation [61, 66], given by

$$\frac{\partial c}{\partial t} = D\nabla^2 c + \Lambda c\,(1 - c)\,, \tag{7}$$

where

$$D = \lim_{\Delta,\tau\to 0} \frac{P\Delta^2}{4\tau} \quad \text{and} \quad \Lambda = \lim_{\tau\to 0} \frac{\lambda}{\tau}. \tag{8}$$

Since both the experiment and simulation maintain radial symmetry, we write Equation (7) as

$$\frac{\partial c}{\partial t} = \frac{D}{r}\frac{\partial}{\partial r}\left(r\frac{\partial c}{\partial r}\right) + \Lambda c\,(1-c)\,, \tag{9}$$

where $c(r,t) \geq 0$ represents the dimensionless cell density at radius r and time t. Since Equation (9) cannot be solved analytically, we obtain numerical solutions as outlined in Section 2 of the Supplementary Material. For the simulations we consider, the computational time required to solve Equation (9) is approximately three orders of magnitude faster than simulating the RWM. Therefore, despite working with a numerical approach, the surrogate PDE model is very attractive for the purposes of estimation, identifiability and sensitivity analysis over working with the stochastic model. To link the solution of the continuum-limit model in Equation (9) with count data from the interacting RWM, we make the assumption that the agent density is constant within each rectangular discretisation along the transect. This will be a reasonable approximation provided that H is sufficiently small compared to the radius of the spreading population.

2.3 Data

In this study we use synthetic data generated by the RWMs described in Section 2.2 to produce count data with properties analogous to those of relevant experiments. The work is presented as two different cases: (i) Case 1 uses data generated from a non-interacting RWM described in Section 2.2.1; and, (ii) Case 2 uses data generated from an interacting RWM described in Section 2.2.2. For Case 2, we explore results based on data collected at the end of two identically prepared simulations: Case 2a uses early-time data after 100 time steps and Case 2b uses late-time count data after 600 time steps. As we demonstrate, this distinction turns has an important impact on parameter identifiability as we will detail in Section 3.2. In all cases our data are stored in a vector denoted y^{o} noting that in the applied statistics literature it is customary not to use bold face notation for vectors.

In reality, generating count data from high-density experiments is very time-consuming, and this can be addressed by subsampling the data. We take the same approach by working with subsampled count data from our RWMs to ensure that our methods do not rely on working with impractical amounts of cell counts. To this end, we collect N data points for

each subpopulation from the full count data. Consequently, y^{o} is a vector of length $2N$ for Case 1, whereas y^{o} is a vector of length N for Case 2.

In Case 1, we simulate the non-interacting RWM on a 200×20 lattice with $(P, \rho, \lambda_1, \lambda_2)^\top = (1, 0.5, 0.02, 0.03)^\top$. Initially, red and green agents occupy sites within columns $81 \leq i \leq 120$, with relatively low initial densities of 0.5 and 0.2, respectively, as in Figure 2(a). Full count data at $t = 0$ for the red and green subpopulations are shown in Figure 2(c)–(d). Data are then collected after $k = 100$ time steps, producing the snapshots in Figure 2(b), where we can see that the total population has spread, drifted in the positive x-direction, and the total number of agents grows over the simulation. Full count data for both the red and green subpopulations after $k = 100$ time steps are presented in Figure 2(e)–(f), respectively. We subsample $N = 12$ equally-spaced data points from the full count data in Figure 2(e)–(f) for both subpopulations, given in Figure 2(g)–(h). We denote the sampled data as $C_1^{\mathrm{o}}(n, k)$ and $C_2^{\mathrm{o}}(n, k)$, where $n = 1, 2, 3, \ldots, N$ and $k = K = 100$. The relationship between the data index n and the lattice column index i is $i = (n - 1)18$.

In Case 2, we simulate the interacting RWM on a 200×200 lattice with $(P, \lambda, c_0)^\top = (1, 0.01, 0.8)^\top$. Agents are initially placed within a circular region of radius $r_0 = 20\Delta$. Within this region, each site is occupied by a single agent with probability 0.8 to reflect the relatively high density design of the barrier assay. A snapshot of the initial condition is shown in Figure 3(a). The early-time snapshot after $k = 100$ time steps, and a late-time snapshot after $k = 600$ time steps are shown in Figure 3(b)–(c), respectively. A green rectangular transect is superimposed on each snapshot, with detailed views of the corresponding transects shown in Figure 3(d)–(f), respectively. The corresponding full count data from the rectangular transect after $k = 0, 100$ and 600 time steps are shown in Figure 3(g)–(i), respectively. We subsample $N = 20$ equally-spaced data points from the full count data after $k = 0, 100$ and 600 time steps. The subsampled data is presented in Figure 3(j)–(l), respectively. In Case 2, we denote the sampled data as $C^{\mathrm{o}}(n, k)$, with $k = 100$ for Case 2a and $k = 600$ for Case 2b. The relationship between the data index n and the column index i in Case 2 is given by $i = (n - 1)10$.

In this work Cases 2a and 2b illustrate that estimation, identifiability, and sensitivity analysis can be performed across varying experimental timescales. As we will demonstrate, a key feature of Case 2 is that the identifiability of various simulation parameters depends upon the simulation timescale. For completeness, we also repeat a similar analysis using late-time data for Case 1 in Section 4 of the Supplementary Material. Further exploration involving different choices of parameter values and simulation durations can be explored by adapting

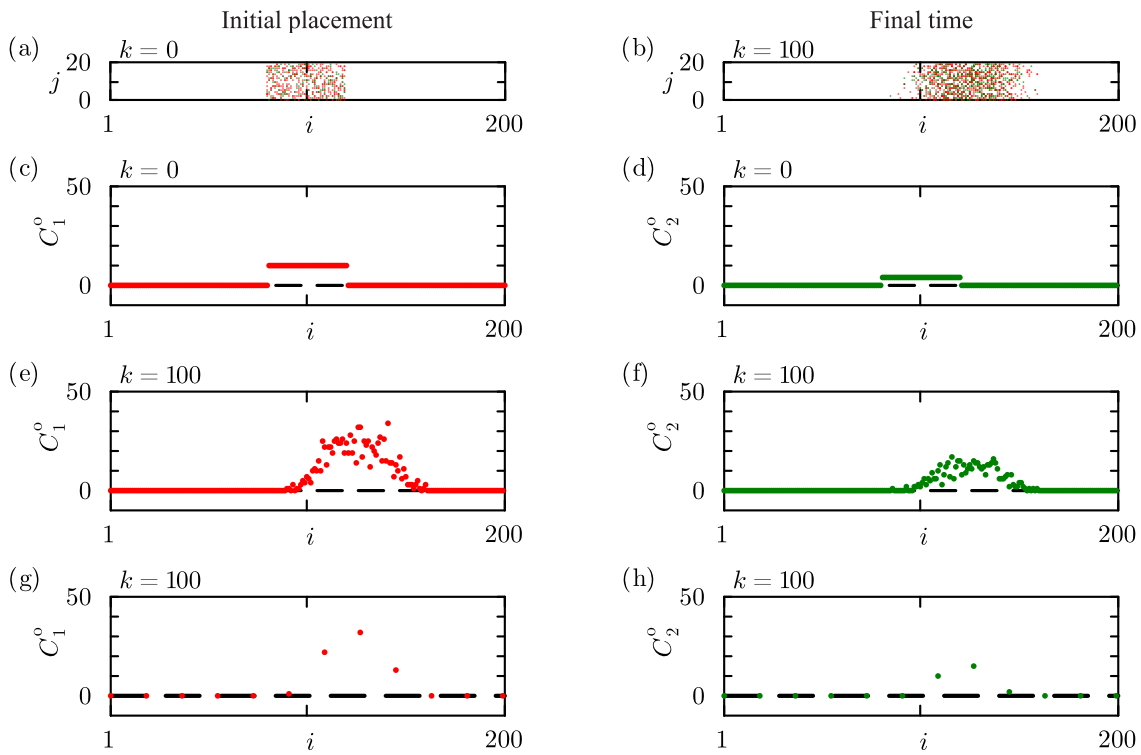

Figure 2: Snapshots of agent distributions and count data from the non-interacting RWM with $(P, \rho, \lambda_1, \lambda_2)^\top = (1, 0.5, 0.02, 0.03)^\top$. (a)–(b) Snapshots of agent distribution. (c)–(f) Full count data. (g)–(h) Subsampled count data obtained from data in (e)–(f), respectively. Horizontal dashed lines in (e)–(h) show the $C_s^{\mathrm{o}} = 0$ lower bound for the count data. Results correspond to $I = 200$ and $J = 20$.

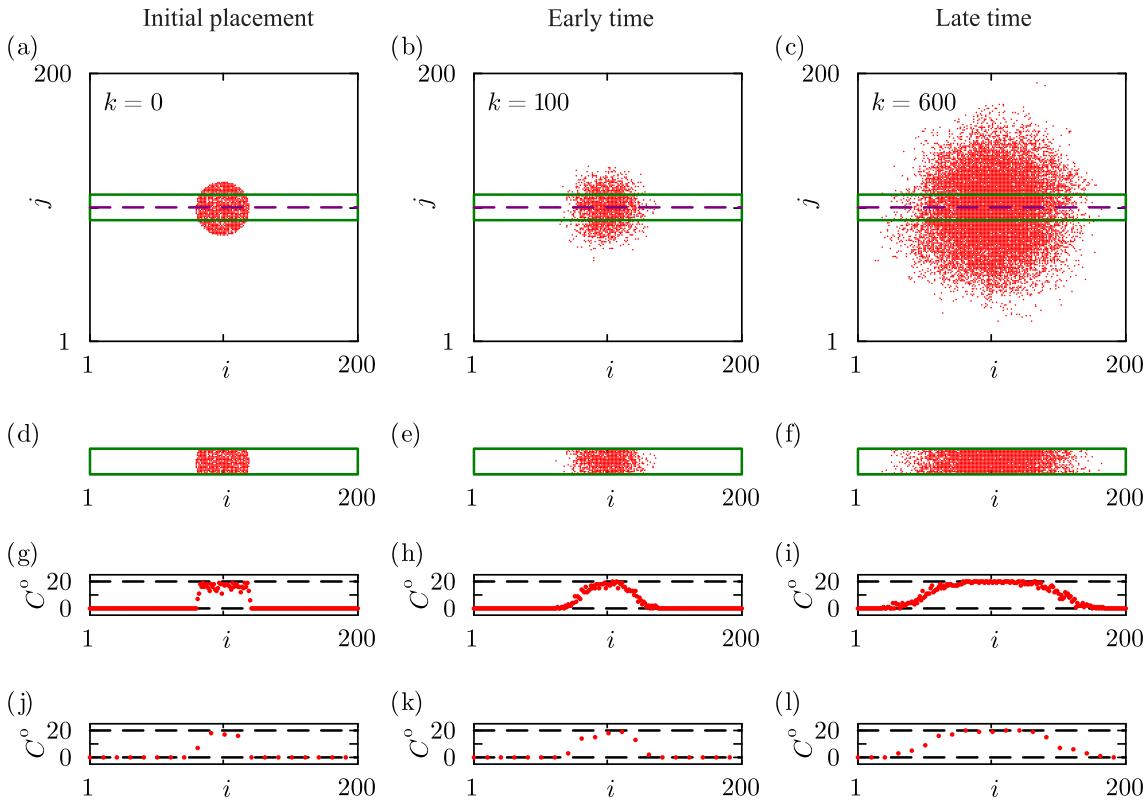

Figure 3: Snapshots of agent distributions and count data from the interacting RWM with $(P, \lambda, c_0)^\top = (1, 0.01, 0.8)^\top$. (a)–(c) Snapshots of agent distribution superimposed with a green rectangular transect of height $H = 20$ and purple line showing the centre of the spreading population where $j = 100$. (d)–(f) Agent distributions within the transects in (a)–(c), respectively. (g)–(i) Count data from the snapshots in (d)–(f), respectively. (j)–(l) Subsampled count data obtained from data in (g)–(i), respectively. Horizontal dashed lines in (g)–(l) show the $C^{\mathrm{o}} = 0$ lower bound and $C^{\mathrm{o}} = 20$ upper bound for the count data. Results correspond to $I = J = 200$ and $H = 20$.

open-access software on GitHub.

2.4 Likelihood function

In this section, we describe how two physically-motivated measurement error models that probabilistically relate the count data summarised in Figure 2–3 to appropriate solutions of the relevant surrogate PDE models. For Case 1, we use a Poisson-based measurement error model, whereas for Case 2 we use a binomial measurement error model. These different choices of measurement error models accommodate differences in the properties of the count data, which we now describe.

For the red subpopulation in Case 1, we interpret the integer count data $C_1^{\mathrm{o}}(n,k) \in \{0,1,2,3,\ldots\}$ at a given spatial location after k time steps as representing a finite number of independent samples from an underlying stochastic process (i.e. the RWM or a real biological experiment). These data are considered to be samples from a distribution in which the expected total number of agents in the column corresponding to the nth data point is related to the solution of the continuum-limit model, given by $J \times c_1(x_n, t_k)$. Here, x_n denotes the horizontal Cartesian coordinate associated with the nth data point and t_k is time after k time steps. This relationship, together with the non-negative, unbounded integer count data motivates the use of a Poisson measurement error model to link the observed count data with the continuum-limit solution. While other options are possible, we use a Poisson-based measurement error model for simplicity as it does not introduce any additional parameters to be estimated [13]. Following a similar argument we can link count data from the green subpopulation to the solution of Equation (2)–(3) using a Poisson measurement error model. Invoking a standard independence assumption, we can evaluate the probability density for each subpopulation and take the product of these two probability densities, which yields a Poisson-based likelihood for the non-interacting RWM in Case 1 at a single spatial location x_n:

$$\begin{aligned}\mathcal{L}(\theta \mid y^{\mathrm{o}}(n)) \propto & \left[Jc_1(x_n,t_k)\right]^{C_1^{\mathrm{o}}(n,k)} \exp\left[-Jc_1(x_n,t_k)\right] \\ & \times \left[Jc_2(x_n,t_k)\right]^{C_2^{\mathrm{o}}(n,k)} \exp\left[-Jc_2(x_n,t_k)\right].\end{aligned} \tag{10}$$

Invoking a standard independence assumption we can consider count data at all spatial locations to give a log-likelihood function as follows,

$$\ell(\theta \mid y^{\mathrm{o}}) \propto \sum_{s=1}^{S}\sum_{n=1}^{N} \log\left(\left[Jc_s(x_n,t_k)\right]^{C_s^{\mathrm{o}}(n,k)} \exp\left[-Jc_s(x_n,t_k)\right]\right), \tag{11}$$

where $S = 2$ in for our RWM.

Count data for Case 2 consists of agent counts within each column of the transect, which implicitly determine the number of vacant lattice sites in each column. Thus, the data take the form $C^{\mathrm{o}}(n,k)$ and $E^{\mathrm{o}}(n,k) = H - C^{\mathrm{o}}(n,k)$ for $n = 1, 2, 3, \ldots, N$. Interpreting $C^{\mathrm{o}}(n,k)$ at a given spatial location as a finite number of independent samples from an underlying stochastic process (i.e. the RWM or a real biological experiment), these samples approximate a continuous measurement of agent occupancy within that column. These data are effectively drawn from a distribution in which the expected occupancy fraction is given by the solution of the continuum-limit model, $H \times c(r,t)$. However, while $C^{\mathrm{o}}(n,k)$ represents count data along the centreline of the spreading population, $H \times c(r,t)$ denotes the counts along the radius. By symmetry, we have $r = |x - x_0|$, where x_0 denotes the x-coordinate of the centre of the spreading population. These observations, together with the fact that the count data are bounded non-negative integers, we make use of a binomial measurement error model to link the observed count data with the solution of the PDE model, as the binomial distribution is defined over non-negative integers with a finite upper bound [13]. Here, we adopt a binomial measurement error model for Case 2 at a single spatial location x_n leading to a likelihood function of the form

$$\mathcal{L}(\theta \mid y^{\mathrm{o}}(n,k)) \propto c(x_n,t_k)^{C^{\mathrm{o}}(n,k)} \left[1 - c(x_n,t_k)\right]^{E^{\mathrm{o}}(n,k)}, \tag{12}$$

giving rise to the log-likelihood function for data across multiple spatial locations as

$$\ell(\theta \mid y^{\mathrm{o}}) \propto \sum_{n=1}^{N} \log \left[c(x_n,t_k)^{C^{\mathrm{o}}(n,k)} \left[1 - c(x_n,t_k)\right]^{E^{\mathrm{o}}(n,k)} \right]. \tag{13}$$

In this work, we set the proportionality constants in Equation (11) and Equation (13) to unity [41].

2.5 Likelihood based estimation and identifiability

In all cases we aim to estimate parameters in terms of rates within the stochastic model. For example, we will estimate $P^{\star} = P/\tau$, $\rho^{\star} = \rho/\tau$, $k_1^{\star} = k_1/\tau$ and $k_2^{\star} = k_2/\tau$ for Case 1. For a set of count data y^{o} together with a relevant process and measurement error model, we have access to a log-likelihood function, $\ell(\theta \,|\, y^{\mathrm{o}})$. The parameters that best match the data, $\hat{\theta}$, correspond to the choice of θ that maximises $\ell(\theta \,|\, y^{\mathrm{o}})$, giving rise to the maximum likelihood estimate (MLE):

$$\hat{\theta} = \arg\max_{\theta} \left[\ell(\theta \,|\, y^{\mathrm{o}})\right]. \tag{14}$$

Throughout this work we use the Nelder–Mead algorithm with simple bound constraints [67] to compute $\hat{\theta}$ [68]. In general, our numerical optimisation calculations for identifiable parameters are robust to the choice of initial estimates for the iterative solver, and the algorithm converges using the default stopping criteria.

To compare results against asymptotic thresholds we work with a normalised log-likelihood function

$$\bar{\ell}(\theta \mid y^{\mathrm{o}}) = \ell(\theta \mid y^{\mathrm{o}}) - \ell(\hat{\theta} \mid y^{\mathrm{o}}), \tag{15}$$

so that $\bar{\ell}(\hat{\theta} \mid y^{\mathrm{o}}) = 0$. Given the normalised log-likelihood function, we use the profile likelihood approach to quantify the precision of parameter estimates by examining the curvature of the log-likelihood function [35, 37–40]. We partition the full parameter set θ into interest parameters, ψ, and nuisance parameters, ω, such that $\theta = (\psi, \omega)^\top$. In this work, we focus exclusively on univariate profile log-likelihood functions, where the interest parameter is always a single parameter. For a given dataset y^{o}, the profile likelihood for the interest parameter ψ under the partition $\theta = (\psi, \omega)^\top$ is defined as:

$$\bar{\ell}_p(\psi \mid y^{\mathrm{o}}) = \max_{\omega|\psi} \left[\bar{\ell}(\psi, \omega \mid y^{\mathrm{o}})\right], \tag{16}$$

which implicitly defines a function $\omega^\dagger(\psi)$, representing the optimal values of ω for each value of ψ. Similar to the MLE, we calculate all profile likelihoods using the Nelder-–Mead algorithm using the same bound constraints and default stopping criteria. As a specific example, consider the non-interacting RWM in Case 1 with $\theta = (P^\star, \rho^\star, \lambda_1^\star, \lambda_2^\star)^\top$ where we can compute four univariate profile likelihood functions by choosing the interest and nuisance parameters as follows: (i) $\psi = P^\star$ with $\omega = (\rho^\star, \lambda_1^\star, \lambda_2^\star)^\top$; (ii) $\psi = \rho^\star$ with $\omega = (P^\star, \lambda_1^\star, \lambda_2^\star)^\top$; (iii) $\psi = \lambda_1^\star$ with $\omega = (P^\star, \rho^\star, \lambda_2^\star)^\top$; and, (iv) $\psi = \lambda_2^\star$ with $\omega = (P^\star, \rho^\star, \lambda_1^\star)^\top$. For all profile likelihood calculations, we evaluate $\bar{\ell}_p$ over a uniformly-discretised interval containing the MLE. For example, for $\psi = P^\star$ we consider the interval $P^{\star -} < \hat{P}^\star < P^{\star +}$, and compute $\bar{\ell}_p$ across a uniform discretisation of that interval. The resulting univariate profile likelihood function can be visualised by simply plotting $\bar{\ell}_p(\psi \mid y^{\mathrm{o}})$. The curvature of $\bar{\ell}_p$ provides information about the practical identifiability of the interest parameter [37]. A flat profile likelihood indicates that the parameter is not identifiable [34], whereas the degree of curvature of $\bar{\ell}_p$ indicates inferential precision [35, 37–40]. We determine likelihood-based confidence intervals by identifying intervals where $\bar{\ell}_p \geq -\Delta_{q,n}/2$, where $\Delta_{q,n}$ denotes the qth quantile of the χ^2 distribution with n degrees of freedom, which we take to be the relevant number of free parameters [69, 70]. For example, identifying the interval where $\bar{\ell}_p \geq -\Delta_{0.95,1}/2 \approx -1.9207$ allows us to compute the asymptotic 95% confidence interval [4, 37] for a univariate profile likelihood [41].

2.6 Parameter-wise prediction

To propagate uncertainty in a single interest parameter ψ into the uncertainty in the model solution, we define an interval for ψ such that $\psi \in \{\psi \mid \bar{\ell}_p(\psi \mid y^{\mathrm{o}}) \geq -\Delta_{0.95,1}/2 \approx -1.9207\}$ [57]. We take a uniform discretisation of that interval, with M mesh points, and solve the model at each mesh point with $\theta_m = (\psi_m, \omega_m^\dagger(\psi))$ for $m = 1, 2, 3, \ldots, M$. For each of these M model solutions, we construct a prediction interval for data realisations [4] as follows. Each solution of the surrogate model can be interpreted as the mean of the relevant noise model, and we account for the width of the noise model distribution by computing the 5% and 95% quantiles of the measurement error model at each spatial location. The union of these M intervals at each spatial location yields a parameter-wise prediction interval for data realisations. In this work, we use $M = 100$ and the procedure can be repeated for all parameters in θ.

As a concrete example, consider the non-interacting RWM with $\theta = (P^\star, \rho^\star, \lambda_1^\star, \lambda_2^\star)^\top$. To construct parameter-wise prediction interval for $\psi = P$ with $\omega = (\rho^\star, \lambda_1^\star, \lambda_2^\star)^\top$, we first define the interval $(P^\star_{\min}, P^\star_{\max})$. As we will illustrate, for identifiable problems the bounds of this interval are values of $P^\star$ for which $\bar{\ell}_p(P^\star \mid y^{\mathrm{o}}) = -\Delta_{0.95,1}/2 \approx -1.9207$. These bounds can be determined using a root finding method (e.g. we use the bisection algorithm). For non-identifiable parameters, a sufficiently flat univariate profile likelihood function can fail to intersect one or both asymptotic thresholds. In such cases, user-defined bounds can be implemented, as we will illustrate in Section 3.2. Once the interval $(P^\star_{\min}, P^\star_{\max})$ is identified, it is uniformly-discretised into M mesh points, giving $P^\star_m$ for $m = 1, 2, 3, \ldots, M$. For each $P^\star_m$, we compute the corresponding nuisance parameters $\omega_m^\dagger(P^\star_m) = (\rho^\star, \lambda_1^\star, \lambda_2^\star)_m^\top$ and solve the surrogate PDE for the M parameter combinations to give $c_s^{(m)}(x, t)$ for $m = 1, 2, 3, \ldots, M$ and $s = 1, 2, \ldots, S$. Each PDE solution is uniformly-discretised over space, giving $c_s^{(m)}(x_i, t)$, for $i = 1, 2, 3, \ldots, 200$. For each of these discretised solutions, we define a prediction interval for the data realisation as $c_{s,0.05}^{(m)}(x_i, t) < c_s^{(m)}(x_i, t) < c_{s,0.95}^{(m)}(x_i, t)$, where the lower and upper bounds correspond to the 5% and 95% quantiles of the Poisson-based measurement error model, respectively [41]. After calculating these intervals at each spatial location for the M solutions, we take the union of the M solutions at each spatial location to give the parameter-wise prediction interval at the ith spatial location

$$\left[\min_{m=1,2,3,\ldots,M} \left(c_{s,0.05}^{(m)}(x_i, t)\right), \max_{m=1,2,3,\ldots,M} \left(c_{s,0.95}^{(m)}(x_i, t)\right)\right].$$

A similar procedure is applied to the interacting RWM in Case 2 to construct various parameter-wise prediction intervals. This computational approach provides a complete rep-

resentation of the parameter-wise prediction intervals, which can be interpreted as a form of tolerance interval [71]. These intervals account for both individual parameter uncertainty and data realisation uncertainty in an efficient, objective manner that avoids random sampling such as in a more standard LHS-PRCC approach [41].

3 Results and Discussion

In this section, we present the results for the estimation, identifiability analysis, and sensitivity analysis using the data described in Section 2.3 for both the non-interacting and interacting RWMs. For the non-interacting RWM in Case 1, we present one set of results corresponding to a simulation at early time, the result corresponding to a late-time simulation is presented in Section 4 of the Supplementary Material document. For the interacting RWM in Case 2, we present two results: Case 2a corresponds to an early-time simulation, and Case 2b to a late-time simulation. The results are presented in Figure 4–6, where each row corresponds to a different parameter, and each column (from left to right) shows the univariate profile likelihood, the parameter-wise prediction interval, and the difference between the confidence set and the solution of the continuum-limit evaluated at the MLE, respectively.

3.1 Case 1

Using the noisy count data collected from the non-interacting RWM at early-time, presented in Figure 2(g)–(h), we use the Poisson-based measurement error model and associated log-likelihood function to estimate the MLE, $\hat{\theta} = (\hat{P}^\star, \hat{\rho}^\star, \hat{\lambda}_1^\star, \hat{\lambda}_2^\star)^\top = (0.7548, 0.5986, 0.0199, 0.0391)^\top$. Compared with the true values, $\theta = (P^\star, \rho^\star, \lambda_1^\star, \lambda_2^\star)^\top = (1, 0.5, 0.02, 0.03)^\top$, these point estimates appear quite reasonable but they do not provide any indication of their identifiability. We explore the practical identifiability of each individual parameter constructing various univariate profile likelihoods shown in column (a) of Figure 4. Each profile likelihood is relatively narrow with a single peak around the MLE which indicates that all parameters are practically identifiable. For each parameter, we define a 95% asymptotic confidence interval as the region where $\bar{\ell}_p \geq -\Delta_{0.95,1}/2 \approx -1.9207$ [69]. For example, to construct the univariate profile likelihood for the interest parameter $P^\star$, as shown in the first row of column (a) in Figure 4, we evaluate $\bar{\ell}_p$ over a specified interval $P^{\star -} < \hat{P}^\star < P^{\star +}$, where we choose $P^{\star -} = 0$ and $P^{\star +} = 2$. The 95% confidence interval for $P^\star$ is computed using the bisection algorithm,

giving $[P^\star_{\min}, P^\star_{\max}] = [0.4386, 1.5379]$. The 95% confidence intervals for the remaining parameters in Figure 4 are: (i) $\rho^\star = 0.5986$ [0.2961, 1.1925]; (ii) $\lambda_1^\star = 0.0199$ [0.0147, 0.0268], and (iii) $\lambda_2^\star = 0.0391$ [0.0284, 0.0556]. The true value for each parameter lies within the corresponding 95% confidence interval, indicating that the estimates are reasonably precise. The relatively narrow shape of each univariate profile provides further evidence that all parameters are well identified by the count data we have considered.

Parameter-wise sensitivity analysis is performed by constructing parameter-wise prediction intervals, as shown in columns (b)–(c) of Figure 4. Each sub-figure in these columns illustrates how the results at early-time simulation depend on each of the parameters, with results in column (b) detailing the impact of a single parameter on C_1^{o} and results in column (c) showing the impact on C_2^{o}. In column (d) of Figure 4 we superimpose plots of the difference between the respective confidence sets for model realisations and the surrogate PDE solution evaluated at the MLE using the same visualisation approach as in [4]. Here $C^{\mathrm{o}}_{s,0.9}$ is the confidence set bounded by 5% and 95% quantiles of the relevant distribution associated with the measurement error model. These plots effectively show the width of the parameter-wise prediction interval for each parameter as a function of position. This analysis provides detailed insight into the width of the parameter-wise prediction intervals across spatial locations and enables a comparison between the parameter sensitivity for C_1^{o} and C_2^{o}.

All parameter-wise prediction intervals in Figure 4 are wider in the middle regions of the population (i.e. $100 \lessapprox i \lessapprox 150$), indicating larger sensitivity in the predicted count data in the bulk of the population, and a smaller sensitivity in the tail regions of the population (i.e. $i \lessapprox 100$ and $i \gtrapprox 150$). This pattern is consistent with our intuitive expectations, as the higher density middle regions can involve larger stochastic variations compared to the low or zero agent density in the tail region characterized by minimal variations. The parameter-wise prediction interval for C_1^{o} associated with $\lambda_2^\star$, shown in the last row of column (b) in Figure 4 is noticeably wider than the prediction intervals for the other parameters, thus suggesting that $\lambda_2^\star$ has a stronger influence on C_1^{o} than the other parameters. Additionally, $\lambda_1^\star$ appears to affect both C_1^{o} and C_2^{o} similarly, while $\lambda_2^\star$ has a larger impact on C_1^{o} and a reduced impact on C_2^{o}.

All results in Figure 4 are based on early-time ($k = 100$) count data, and the associated profile likelihoods indicate that all four parameters are identifiable. To assess whether these main findings are sensitive to the timing of data collection, we conduct an similar analysis using count data collected at a later time point, as reported in Section 4 of the Supplementary

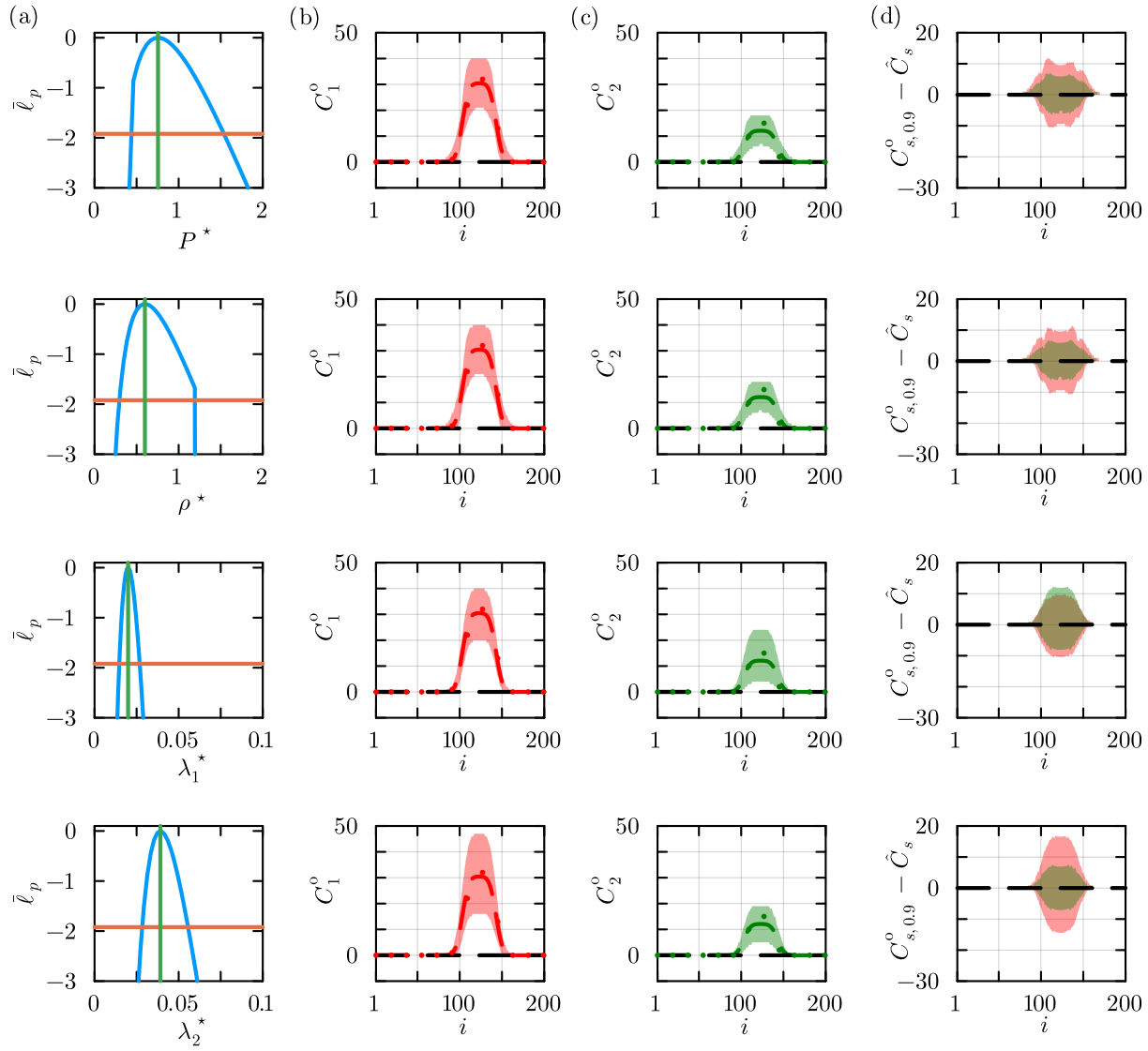

Figure 4: Estimation, identifiability and sensitivity results for the non-interacting RWM. (a) Univariate profiles for $\theta = (P^\star, \rho^\star, \lambda_1^\star, \lambda_2^\star)^\top$ indicate the MLE (vertical green line) and the 95% asymptotic threshold (horizontal orange line). The MLE is $\hat{\theta} = (0.7548, 0.5986, 0.0199, 0.0391)^\top$. (b)–(c) Profile-wise prediction intervals for C_1 (shaded red) and C_2 (shaded green) for each parameter superimposed with count data (green dots), and the solution of the PDE evaluated at the MLE for C_1 (dashed red) and C_2 (dashed green). (d) Difference between confidence set and the solution of the mathematical model evaluated at the MLE for C_1 (shaded red) and C_2 (shaded green).

Material. In brief, we find that the conclusions remain consistent regardless of whether we consider early or later-time data. Interestingly, this is different to the case for the interacting RWM in Case 2 as we will now explore and discuss.

3.2 Case 2

Using noisy count data from the interacting RWM at early-time, presented in Figure 3(k) for Case 2a, we aim to estimate: (i) the motility rate $P^\star$; (ii) the proliferation rate $\lambda^\star$; and (iii) the initial density c_0. In Case 2, we include the initial condition c_0 as an unknown parameter since the initial number of cells in a barrier assay is not always known precisely [23]. Using the log-likelihood function based on the binomial measurement error model, the MLE is $\hat{\theta} = (\hat{P}^\star, \hat{\lambda}^\star, \hat{c}_0)^\top = (0.6716, 0.0300, 0.2709)^\top$. These estimates differ substantially from the true values, $\theta = (P^\star, \lambda^\star, c_0)^\top = (1, 0.01, 0.8)^\top$. The large discrepancies arising from the use of noisy data suggest are a potential indication of challenges with parameter identifiability that we will now explore.

We explore the practical identifiability of each parameter for Case 2a by constructing three univariate profile likelihoods shown in Figure 5, column (a). The profiles likelihood for both $P^\star$ and c_0 exhibit a single peak around the MLE, but both profiles are very wide indicating that these parameters are not well-identified by this data, with 95% confidence intervals of $P^\star = 0.6716$ [0.3718, 1.3822] and $c_0 = 0.2709$ [0.0263, 0.8820], respectively. The profile likelihood for $\lambda^\star$ indicates that $\lambda^\star$ is one-sided identifiable, with the 95% asymptotic confidence interval bounded below at $\lambda^\star = 0.0097$. We construct the profile likelihood for $\lambda^\star$ using a conservative upper bound of 0.04. Values of $\lambda^\star > 0.04$ are not considered, as simulations with $\lambda^\star = 0.04$ show a clear mismatch with the data, indicating that larger values are not biologically relevant. The poorly identified initial density c_0 suggests that the data collected after $k = 100$ time steps are insufficient to accurately identify c_0. Since the data are collected relatively early in the simulation (i.e. the number of time steps is small compared to the proliferation time scale $1/\lambda$), the simulation does not involve significant agent proliferation, which explains why $\lambda^\star$ is poorly identified. These results highlight that parameter non-identifiability can arise even in the relatively simple setting of Case 2a, in contrast to Case 1, where all parameters are identifiable.

Sensitivity analysis for Case 2a is performed by constructing parameter-wise prediction intervals shown in Figure 5, column (b). We use $\lambda^\star = 0.04$ as the upper bound when constructing the parameter-wise prediction intervals for the non-identifiable parameter $\lambda^\star$, as 0.04 is the

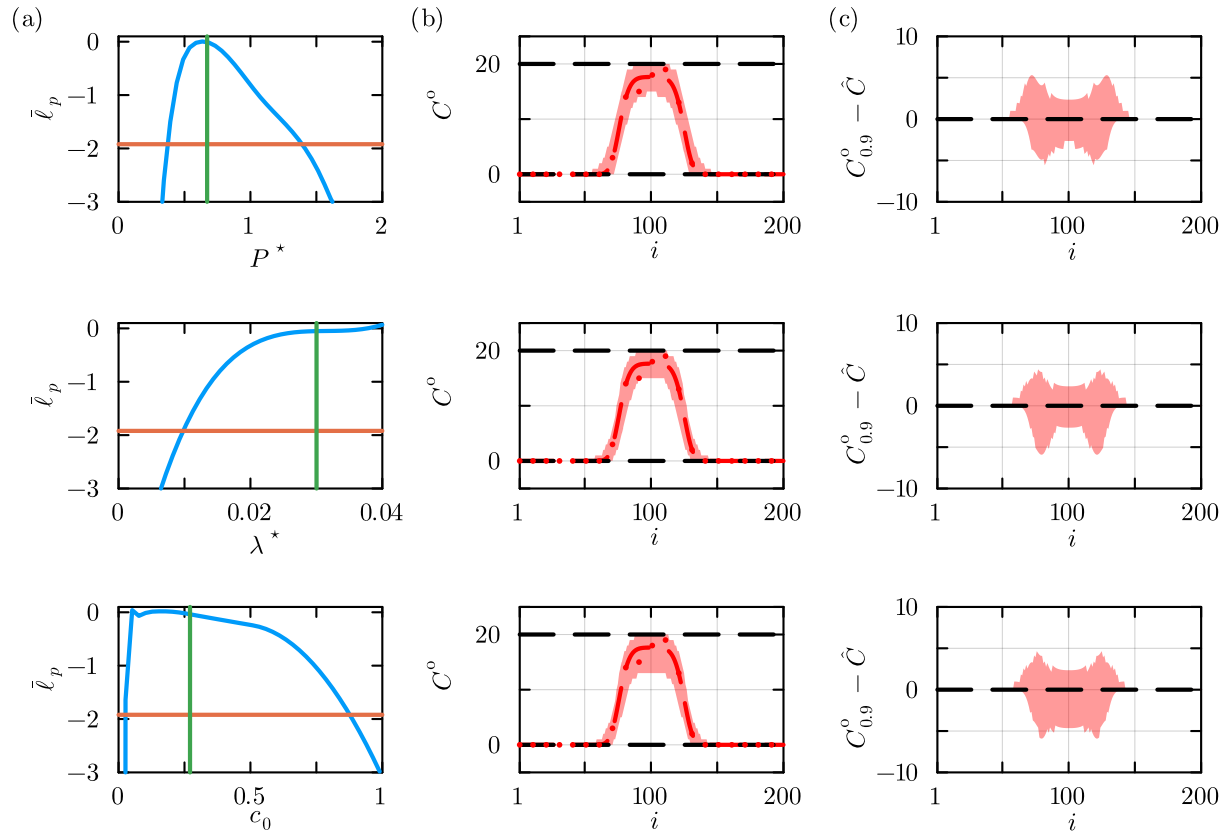

Figure 5: Estimation, identifiability and sensitivity results for the non-interacting RWM using early-time data. (a) Univariate profiles for $\theta = (P^\star, \lambda^\star, c_0)^\top$ indicate the MLE (vertical green line) and the 95% asymptotic threshold (orange horizontal line). The MLE is $\hat{\theta} = (0.6716, 0.0300, 0.2709)^\top$. (b) Profile-wise prediction intervals for C (shaded red) for each parameter superimposed with count data (red dots) and the solution of the PDE evaluated at the MLE. (c) Difference between confidence set and the solution of the mathematical model evaluated at the MLE.

maximum plausible $\lambda^\star$ value considered in the model. Each sub-figure in Figure 5, column (b), illustrates how the early-time simulation results depend on each parameter: $P^\star$, $\lambda^\star$, or c_0, respectively. We find no obvious differences in the parameter-wise prediction intervals for these three parameters, indicating that all three parameters have a similar impact on the model outcome. The differences between the respective confidence sets and the PDE solution evaluated at the MLE for each parameter in column (b) of Figure 5 shows that the prediction intervals in Case 2a are relatively narrow in the middle region of the spreading population (i.e. $80 \lessapprox i \lessapprox 120$) and wider in the transition regions (i.e. $70 \lessapprox i \lessapprox 80$ and $120 \lessapprox i \lessapprox 130$). This pattern is in contrast to Case 1, where the middle region of the population exhibited larger variability. The difference are because the crowding mechanism of RWM in Case 2, which ensures count data are bounded above by $H = 20$. In the high-density middle region, fluctuations are suppressed as the agent count approaches H, thereby reducing variability. In contrast, the non-interacting RWM in Case 1 has no such upper bound, allowing increasing fluctuations in high-density regions. The relatively wider prediction intervals in the transition regions of Case 2a are explained by the fact that fluctuations in count data in these regions do not reach the upper or lower bounds, leading to greater variability compared to the middle region. Both observations are also supported by the data shown in Figure 3(e) and (h).

In Case 2b, we perform a similar analysis using data from late-time simulation ($k = 600$), as presented in Figure 3(l). Using the same binomial measurement error model, the MLE for Case 2b is given by $\hat{\theta} = (\hat{P}^\star, \hat{\lambda}^\star, \hat{c}_0)^\top = (1.0886, 0.0095, 0.5950)^\top$. The estimates of $P^\star$ and $\lambda^\star$ are close to the expected true values ($P^\star = 1$, $\lambda^\star = 0.01$), while the estimate for c_0 differs more substantially from expected true value ($c_0 = 0.8$). The univariate profile likelihoods presented in the first and second rows of Figure 6, column (a), show that $P^\star$ and $\lambda^\star$ are well identified by this data, with relatively narrow profiles and 95% confidence intervals of $P^\star = 1.0886$ [0.8531, 1.4201] and $\lambda^\star = 0.0095$ [0.0077, 0.0119]. The flat profile likelihood for c_0 indicates that this parameter is not well-identified by this data. This finding differs from Case 2a, where $\lambda^\star$ is non-identifiable and c_0 is identifiable. This difference in identifiability reflects the timing of data collection. In Case 2b, the data are collected at a late-time simulation ($k = 600$), by which time the system has evolved for a sufficient duration of time that the influence of the initial condition has diminished. Interestingly, the extended simulation time allows for agent proliferation to have a more pronounced impact which means that $\lambda^\star$ is well-identified by this late-time data.

Sensitivity analysis for Case 2b involves constructing various parameter-wise prediction intervals shown in Figure 6, column (b). For the non-identifiable parameter c_0, the prediction

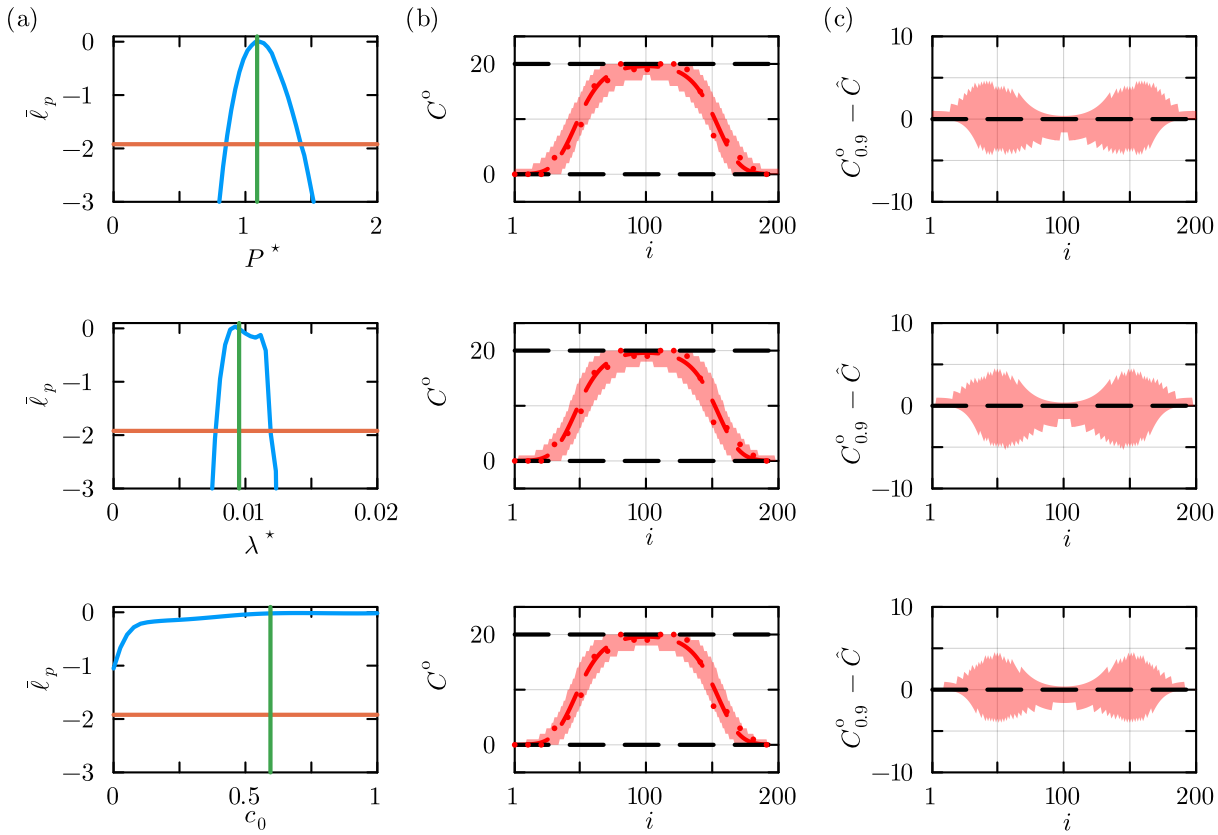

Figure 6: Estimation, identifiability and sensitivity results for the non-interacting RWM using late-time data. (a) Univariate profiles for $\theta = (P^\star, \lambda^\star, c_0)^\top$ indicate the MLE (vertical green line) and the 95% asymptotic threshold (orange horizontal line). The MLE is $\hat{\theta} = (1.0886, 0.0095, 0.5950)^\top$. (b) Profile-wise prediction intervals for C (shaded red) for each parameter superimposed with count data (red dots) and the solution of the PDE evaluated at the MLE. (c) Difference between confidence set and the solution of the mathematical model evaluated at the MLE.

interval is constructed over the range of all physically–relevant values, $c_0 \in [0, 1]$. As with Case 2a, we see no obvious differences in the parameter-wise prediction intervals for these three parameters. Plotting the differences between the boundary of the confidence sets and the continuum model solution evaluated at the MLE, given in Figure 6 column (c), we see that the width of the prediction interval in the middle region of the population is even narrower than in Case 2a. This occurs because the data in Case 2b are collected at a later simulation time, during which the agent density in the middle region continues to increase towards the carrying capacity density, as shown in Figure 3(f) and (i). At these higher densities, the fluctuations are further suppressed, leading to reduced variability. In fact, agent counts in the middle region of the population reaches the upper bound $H = 20$, which means that there are no fluctuations above this value owing to crowding effects in the RWM. This explains why the width of the prediction interval above the MLE becomes zero in the centre of the middle region of the spreading population.

4 Conclusion and Future Work

In this work we present a novel and efficient approach for sensitivity analysis that can be applied to computationally expensive RWMs. A key element of our approach is to link sparse noisy count data to the solution of an efficient surrogate PDE model using various forms of physically-motivated measurement error models. These tools place our analysis within a likelihood-based framework that can be used for for parameter estimation, parameter identifiability analysis and sensitivity analysis via the targeting property of the profile likelihood. Count data from a classical non-interacting RWM are described using a Poisson-based measurement error model, while count data from an interacting exclusion process-based RWM is captured using a binomial measurement error model. Within this general framework we illustrate how to obtain point estimates using maximum likelihood estimation, we analyse the identifiability of those estimates using profile likelihood, and we use parameter-wise predictions as a form of sensitivity analysis. Across our two case studies, we illustrate how the approach applies to both identifiable and poorly-identified problems, and we illustrate how the approach can be implemented using exact or numerical solutions of the relevant surrogate PDE.

Our approach makes use of the targeting property of the profile likelihood to focus on parameter estimates that are targeted towards those regions in the parameter space where the solution of the surrogate model are close to the data. This is a significant departure from random sampling-based approaches, such the LHS-PRCC approach, that relies on random parameter samples to explore parameter sensitivity. A key limitation of working with random parameter samples is that those samples are not necessarily guided by the aim of keeping the model solution close to the target data. While all parameter-wise prediction analysis implemented in this work focuses on univariate parameter-wise prediction intervals, it is possible to use the same approach to generate multivariate prediction intervals (e.g. bivariate profile-wise predictions [72]) where the interest parameters are taken to be a small group of model parameters instead of a single parameter.

Our methodology for sensitivity analysis is demonstrated via two biologically-motivated case studies. Each case study involves a particular choice of process model in the form of a computationally expensive lattice-based RWM and a physically-motivated measurement error model that is guided by the properties of the data generated by the RWM. While we focus on making particular choices of biologically-motivated process and measurement error models, it is important to point out that this framework can be applied quite generally by varying either the process model, the measurement error model or both the process

and measurement error models. This means that the same framework can be applied to many extensions of the current case studies. For example, in Case 1 we assume that both subpopulations of agents share the same migration parameters P and ρ. This assumption can be easily relaxed to allow each subpopulation to have a unique parameters so that $\theta = (P_1, P_2, \rho_1, \rho_2, \lambda_1, \lambda_2)^\top$, where P_i and ρ_i represent the motility probabilities and the bias parameter for the ith subpopulation, respectively. Under this extension the relevant surrogate PDE model is given by

$$\frac{\partial c_1}{\partial t} = D_1 \frac{\partial^2 c_1}{\partial x^2} - v_1 \frac{\partial c_1}{\partial x} - k_1 c_1 + 2k_2 c_2, \tag{17}$$

$$\frac{\partial c_2}{\partial t} = D_2 \frac{\partial^2 c_2}{\partial x^2} - v_2 \frac{\partial c_2}{\partial x} + k_1 c_1 - k_2 c_2, \tag{18}$$

with

$$D_1 = \lim_{\Delta,\tau\to 0} \frac{P_1 \Delta^2}{4\tau}, \quad D_2 = \lim_{\Delta,\tau\to 0} \frac{P_2 \Delta^2}{4\tau}, \quad v_1 = \lim_{\Delta,\tau\to 0} \frac{P_1 \rho_1 \Delta}{2\tau}, \quad \text{and} \quad v_2 = \lim_{\Delta,\tau\to 0} \frac{P_2 \rho_2 \Delta}{2\tau}, \tag{19}$$

where k_1, and k_2 are defined previously in Equation (4). For this extension the surrogate model is best solved numerically rather than generalising the transform approach outlined in the Supplementary Material document. The rest of the workflow for estimation, identifiability and parameter-wise sensitivity analysis remains unchanged.

In this work we use a Poisson-based measurement error model for data generated by the non-interacting RWM because the Poisson distribution is defined over unbounded, non-negative integers. We note that other distributions, such as the negative binomial distribution, also share these properties and could serve as alternative measurement error models [73]. Similarly, we use a binomial measurement error model for data generated by the interacting RWM since the binomial distribution is defined over non-negative integers with a finite upper bound. The truncated Poisson distribution has similar features and so could be used as an alternative measurement error model [74].

All results in our case studies are presented for data generated at a single time point only. While many experiments focus only on data at a single time point, it is conceptually and computationally straightforward to generalise our approach to apply to problems where data is collected at K time points for S subpopulations [41]. For example, for a Poisson-based log-likelihood function with count data collected at K time points across S subpopulations can be written as

$$\ell(\theta \mid y^{\text{o}}) \propto \sum_{s=1}^{S} \sum_{k=1}^{K} \sum_{n=1}^{N} \log \left(\left[J c_s(x_n, t_k) \right]^{C_s^{\text{o}}(n,k)} \exp\left[-J c_s(x_n, t_k) \right] \right), \tag{20}$$

where the data vector y^{o} has length $S \times K \times N$. With this modest generalisation of the log-likelihood function all other steps for estimation, identifiability analysis and parameter-wise sensitivity analysis remain unchanged. A similar generalisation can be made for the log-likelihood function based on the binomial measurement error model with data collected at K time points across S subpopulations, which can be expressed as

$$\ell(\theta \mid y^{\mathrm{o}}) \propto \sum_{s=1}^{S} \sum_{k=1}^{K} \sum_{n=1}^{N} \log \left[c_s(x_n, t_k)^{C_s^{\mathrm{o}}(n,k)} \left(1 - c_s(x_n, t_k)\right)^{E^{\mathrm{o}}(n,k)} \right]. \tag{21}$$

Both Equation (20) and Equation (21) can be further explored in the context of more complex models.

A key feature of our work is to use approximate continuum-limit PDE model as a computationally-efficient surrogate. While in many cases the continuum-limit models are well–established in the literature, and known to lead to accurate approximations [15, 19], it is not strictly necessary to use this approach. Other methods for deriving approximate surrogate models from data, such as working with Gaussian processes [75, 76] or equation learning methods can also be used [77, 78], and we leave these approaches for future exploration.

References

[1] Razavi S, Jakeman A, Saltelli A, Prieur C, Iooss B, Borgonovo E, et al. The future of sensitivity analysis: an essential discipline for systems modeling and policy support. Environmental Modelling & Software. 2021;137:104954.

[2] Erban R, Chapman SJ. Stochastic modelling of reaction–diffusion processes: algorithms for bimolecular reactions. Physical Biology. 2009;6(4):046001.

[3] Simpson MJ, Maclaren OJ. Profile-wise analysis: A profile likelihood-based workflow for identifiability analysis, estimation, and prediction with mechanistic mathematical models. PLoS Computational Biology. 2023;19(9):e1011515.

[4] Murphy RJ, Maclaren OJ, Simpson MJ. Implementing measurement error models with mechanistic mathematical models in a likelihood-based framework for estimation, identifiability analysis and prediction in the life sciences. Journal of the Royal Society Interface. 2024;21(210):20230402.

[5] Arriola L, Hyman JM. Sensitivity analysis for uncertainty quantification in mathematical models. Springer; 2009.

[6] Dela A, Shtylla B, de Pillis L. Multi-method global sensitivity analysis of mathematical models. Journal of Theoretical Biology. 2022;546:111159.

[7] Chitnis N, Hyman JM, Cushing JM. Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model. Bulletin of Mathematical Biology. 2008;70:1272-96.

[8] Gupta A, Khammash M. An efficient and unbiased method for sensitivity analysis of stochastic reaction networks. Journal of The Royal Society Interface. 2014;11(101):20140979.

[9] Gunawan R, Cao Y, Petzold L, Doyle FJ. Sensitivity analysis of discrete stochastic systems. Biophysical Journal. 2005;88(4):2530-40.

[10] Rathinam M, Sheppard PW, Khammash M. Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks. The Journal of Chemical Physics. 2010;132(3).

[11] Sheppard PW, Rathinam M, Khammash M. A pathwise derivative approach to the computation of parameter sensitivities in discrete stochastic chemical systems. The Journal of Chemical Physics. 2012;136(3).

[12] Bergman DR, Karikomi MK, Yu M, Nie Q, MacLean AL. Modeling the effects of EMT-immune dynamics on carcinoma disease progression. Communications Biology. 2021;4(1):983.

[13] Liu Y, Warne DJ, Simpson MJ. Likelihood-based inference, identifiability, and prediction using count data from lattice-based random walk models. Physical Review E. 2024;110(4):044405.

[14] Khain E, Katakowski M, Hopkins S, Szalad A, Zheng X, Jiang F, et al. Collective behavior of brain tumor cells: The role of hypoxia. Physical Review E. 2011;83(3):031920.

[15] Codling EA, Plank MJ, Benhamou S. Random walk models in biology. Journal of the Royal Society Interface. 2008;5(25):813-34.

[16] Callaghan T, Khain E, Sander LM, Ziff RM. A stochastic model for wound healing. Journal of Statistical Physics. 2006;122(5):909-24.

[17] Ross RJH, Baker RE, Parker A, Ford MJ, Mort RL, Yates CA. Using approximate Bayesian computation to quantify cell–cell adhesion parameters in a cell migratory process. NPJ Systems Biology and Applications. 2017;3(1):9.

[18] Jones PJ, Sim A, Taylor HB, Bugeon L, Dallman MJ, Pereira B, et al. Inference of random walk models to describe leukocyte migration. Physical Biology. 2015;12(6):066001.

[19] Plank MJ, Simpson MJ, Baker RE. Random walk models in the life sciences: Including births, deaths and local interactions. Journal of the Royal Society Interface. 2025;22(222):20240422.

[20] Vittadello ST, McCue SW, Gunasingh G, Haass NK, Simpson MJ. Mathematical models for cell migration with real-time cell cycle dynamics. Biophysical Journal. 2018;114(5):1241-53.

[21] Sakaue-Sawano A, Kurokawa H, Morimura T, Hanyu A, Hama H, Osawa H, et al. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. Cell. 2008;132(3):487-98.

[22] Shirmanova MV, Gorbachev DA, Sarkisyan KS, Parnes AP, Gavrina AI, Polozova AV, et al. FUCCI-Red: A single-color cell cycle indicator for fluorescence lifetime imaging. Cellular and Molecular Life Sciences. 2021;78:3467-76.

[23] Haridas P, Penington CJ, McGovern JA, McElwain DLS, Simpson MJ. Quantifying rates of cell migration and cell proliferation in co-culture barrier assays reveals how skin and melanoma cells interact during melanoma spreading and invasion. Journal of Theoretical Biology. 2017;423:13-25.

[24] Das AM, Eggermont AM, Ten Hagen TL. A ring barrier–based migration assay to assess cell migration in vitro. Nature Protocols. 2015;10(6):904-15.

[25] Callaham JL, Loiseau JC, Rigas G, Brunton SL. Nonlinear stochastic modelling with Langevin regression. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2021;477(2250):20210092.

[26] Cassidy T. A continuation technique for maximum likelihood estimators in biological models. Bulletin of Mathematical Biology. 2023;85(10):90.

[27] Handcock MS, Jones JH. Likelihood-based inference for stochastic models of sexual network formation. Theoretical Population Biology. 2004;65(4):413-22.

[28] Hartig F, Calabrese JM, Reineking B, Wiegand T, Huth A. Statistical inference for stochastic simulation models–theory and application. Ecology Letters. 2011;14(8):816-27.

[29] Liepe J, Taylor H, Barnes CP, Huvet M, Bugeon L, Thorne T, et al. Calibrating spatio-temporal models of leukocyte dynamics against *In vivo* live-imaging data using approximate Bayesian computation. Integrative Biology. 2012;4(3):335-45.

[30] Vo BN, Drovandi CC, Pettitt AN, Pettet GJ. Melanoma cell colony expansion parameters revealed by approximate Bayesian computation. PLoS Computational Biology. 2015;11(12):1-22.

[31] Klinke DJ. An empirical Bayesian approach for model-based inference of cellular signaling networks. BMC Bioinformatics. 2009;10:1-18.

[32] Ellam L, Girolami M, Pavliotis G, Wilson A. Stochastic modelling of urban structure. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2018;474(2213):20170700.

[33] Browning AP, McCue SW, Binny RN, Plank MJ, Shah ET, Simpson MJ. Inferring parameters for a lattice-free model of cell migration and proliferation using experimental data. Journal of Theoretical Biology. 2018;437:251-60.

[34] Simpson MJ, Maclaren OJ. Making predictions using poorly identified mathematical models. Bulletin of Mathematical Biology. 2024;86(7):80.

[35] Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009;25(15):1923-9.

[36] Simpson MJ, Baker RE. Parameter identifiability, parameter estimation and model prediction for differential equation models. To apear in SIAM Review. 2025.

[37] Pawitan Y. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford University Press; 2001.

[38] Raue A, Kreutz C, Theis FJ, Timmer J. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2013;371(1984):20110544.

[39] Villaverde AF, Tsiantis N, Banga JR. Full observability and estimation of unknown inputs, states and parameters of nonlinear biological models. Journal of the Royal Society Interface. 2019;16(156):20190043.

[40] Villaverde AF, Pathirana D, Fröhlich F, Hasenauer J, Banga JR. A protocol for dynamic model calibration. Briefings in Bioinformatics. 2021;23(1):bbab387.

[41] Simpson MJ, Murphy RJ, Maclaren OJ. Modelling count data with partial differential equation models in biology. Journal of Theoretical Biology. 2024;580:111732.

[42] Bouman JA, Hauser A, Grimm SL, Wohlfender M, Bhatt S, Semenova E, et al. Bayesian workflow for time-varying transmission in stratified compartmental infectious disease transmission models. PLoS Computational Biology. 2024;20(4):e1011575.

[43] Piironen J, Vehtari A. Comparison of Bayesian predictive methods for model selection. Statistics and Computing. 2017;27:711-35.

[44] Marino S, Hogue IB, Ray CJ, Kirschner DE. A methodology for performing global uncertainty and sensitivity analysis in systems biology. Journal of Theoretical Biology. 2008;254(1):178-96.

[45] Byun JH, Jung IH. Mathematical modeling of the receptor-mediated endocytosis process of targeted therapeutic agents in drug delivery systems. Applied Mathematical Modelling. 2020;79:300-13.

[46] Linderman JJ, Riggs T, Pande M, Miller M, Marino S, Kirschner DE. Characterizing the dynamics of CD4+ T cell priming within a lymph node. Journal of Immunology. 2010;184(6):2873-85.

[47] Johnson SC, Frattolin J, Edgar LT, Jafarnejad M, Moore Jr JE. Lymph node swelling combined with temporary effector T cell retention aids T cell response in a model of adaptive immunity. Journal of the Royal Society Interface. 2021;18(184):20210464.

[48] Segovia-Juarez JL, Ganguli S, Kirschner D. Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent-based model. Journal of theoretical biology. 2004;231(3):357-76.

[49] Pianosi F, Beven K, Freer J, Hall JW, Rougier J, Stephenson DB, et al. Sensitivity analysis of environmental models: A systematic review with practical workflow. Environmental Modelling & Software. 2016;79:214-32.

[50] McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics. 2000;42(1):55-61.

[51] Bergman DR, Norton KA, Jain HV, Jackson T. Connecting agent-based models with high-dimensional parameter spaces to multidimensional data using SMoRe ParS: A surrogate modeling approach. Bulletin of Mathematical Biology. 2024;86(1):11.

[52] Alizadeh R, Allen JK, Mistree F. Managing computational complexity using surrogate models: A critical review. Research in Engineering Design. 2020;31(3):275-98.

[53] Gherman IM, Abdallah ZS, Pang W, Gorochowski TE, Grierson CS, Marucci L. Bridging the gap between mechanistic biological models and machine learning surrogates. PLoS Computational Biology. 2023;19(4):e1010988.

[54] Fonseca LL, Böttcher L, Mehrad B, Laubenbacher RC. Optimal control of agent-based models via surrogate modeling. PLOS Computational Biology. 2025;21(1):e1012138.

[55] Chaplain MA, Lorenzi T, Macfarlane FR. Bridging the gap between individual-based and continuum models of growing cell populations. Journal of Mathematical Biology. 2020;80(1):343-71.

[56] Almeida L, Audebert C, Leschiera E, Lorenzi T. Discrete and continuum models for the coevolutionary dynamics between CD8+ cytotoxic T lymphocytes and tumour cells. Mathematical Medicine and Biology: A Journal of the IMA. 2023;40(2):141-74.

[57] Simpson MJ, Walker SA, Studerus EN, McCue SW, Murphy RJ, Maclaren OJ. Profile likelihood-based parameter and predictive interval analysis guides model choice for ecological population dynamics. Mathematical Biosciences. 2023;355:108950.

[58] Grada A, Otero-Vinas M, Prieto-Castrillo F, Obagi Z, Falanga V. Research techniques made simple: Analysis of collective cell migration using the wound healing assay. Journal of Investigative Dermatology. 2017;137(2):e11-6.

[59] Liang CC, Park AY, Guan JL. *In vitro* scratch assay: A convenient and inexpensive method for analysis of cell migration *in vitro*. Nature Protocols. 2007;2(2):329-33.

[60] Chowdhury D, Schadschneider A, Nishinari K. Physics of transport and traffic phenomena in biology: From molecular motors and cells to organisms. Physics of Life Reviews. 2005;2(4):318-52.

[61] Simpson MJ, Landman KA, Hughes BD. Cell invasion with proliferation mechanisms motivated by time-lapse data. Physica A: Statistical Mechanics and its Applications. 2010;389(18):3779-90.

[62] Vittadello ST, McCue SW, Gunasingh G, Haass NK, Simpson MJ. Examining go-or-grow using fluorescent cell-cycle indicators and cell-cycle-inhibiting drugs. Biophysical Journal. 2020;118(6):1243-7.

[63] Pettet GJ, Byrne HM, McElwain DLS, Norbury J. A model of wound-healing angiogenesis in soft tissue. Mathematical Biosciences. 1996;136(1):35-63.

[64] Owen MR, Alarcón T, Maini PK, Byrne HM. Angiogenesis and vascular remodelling in normal and cancerous tissues. Journal of Mathematical Biology. 2009;58:689-721.

[65] Kramer N, Walzl A, Unger C, Rosner M, Krupitza G, Hengstschläger M, et al. *In vitro* cell migration and invasion assays. Mutation Research/Reviews in Mutation Research. 2013;752(1):10-24.

[66] Murray JD. Mathematical biology: I. An introduction. Springer; 2002.

[67] Nelder JA, Mead R. A simplex method for function minimization. The Computer Journal. 1965;7(4):308-13.

[68] Johnson SG. The NLopt module for Julia; 2018. Available from: `https://github.com/JuliaOpt/NLopt.jl`.

[69] Royston P. Profile likelihood for estimation and confidence intervals. The Stata Journal: Promoting Communications on Statistics and Stata. 2007;7(3):376-87.

[70] Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics. 1938;9(1):60-2.

[71] Vardeman SB. What about the other intervals? The American Statistician. 1992;46(3):193-7.

[72] Simpson MJ. Data-informed model reduction for inference and prediction from non-identifiable models. Journal of Theoretical Biology. 2025;611:112155.

[73] Davis RA, Wu R. A negative binomial model for time series of counts. Biometrika. 2009;96(3):735-49.

[74] Li XJ, Sun Y, Tian GL, Liang J, Shi J. Mean regression model for the zero-truncated Poisson distribution and its generalization. Computational Statistics & Data Analysis. 2023;179:107650.

[75] Gramacy RB, Lee HKH. Bayesian treed Gaussian process models with an application to computer modeling. Journal of the American Statistical Association. 2008;103(483):1119-30.

[76] Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. Journal of Mathematical Psychology. 2018;85:1-16.

[77] Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences. 2016;113(15):3932-7.

[78] Rudy SH, Brunton SL, Proctor JL, Kutz JN. Data-driven discovery of partial differential equations. Science Advances. 2017;3(4):e1602614.

Supplementary Material: Parameter-wise predictions and sensitivity analysis for random walk models in the life sciences

Yihan Liu[1], David J. Warne[1], and Matthew J. Simpson[1,2]

[1]School of Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, Australia.

[2]ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems, QUT, Brisbane, Australia.

Contents

S1 Exact solution of surrogate PDE for Case 1 **3**

S2 Numerical solution of surrogate PDE for Case 2 **5**

S3 Model validation **6**

S3.1 Case 1 validation 6

S3.2 Case 2 validation 8

S4 Additional Results: Case 1 over a longer timescale **11**

S1 Exact solution of surrogate PDE for Case 1

Here, we implement a transformation method to obtain analytical solutions for systems of linear partial differential equations (PDEs) with coupled source terms relevant to Case 1. This method, based on diagonalisation, can be applied to systems with n subpopulations [1] such as

$$\frac{\partial c(x,t)}{\partial t} - D\frac{\partial^2 c(x,t)}{\partial x^2} - v\frac{\partial c(x,t)}{\partial x} = Kc(x,t), \tag{S1}$$

where $c(x,t)$ is a vector of length n, K is an $n \times n$ constant matrix, and D and v are scalar constants. As we will illustrate, a key assumption in this approach is that the matrix K is diagonalisable.

The first step in the solution procedure is to determine an $n \times n$ constant matrix S, where each column of S corresponds to an eigenvector of K [1]. We then define a second vector of length n, given by $b(x,t) = S^{-1}c(x,t)$. Consequently, we can write $c(x,t) = Sb(x,t)$, and Equation (S1) can be expressed as:

$$S\frac{\partial b(x,t)}{\partial t} - SD\frac{\partial^2 b(x,t)}{\partial x^2} - Sv\frac{\partial b(x,t)}{\partial x} = KSb(x,t). \tag{S2}$$

Left-multiplying Equation (S2) on both sides by S^{-1} yields:

$$\frac{\partial b(x,t)}{\partial t} - D\frac{\partial^2 b(x,t)}{\partial x^2} - v\frac{\partial b(x,t)}{\partial x} = S^{-1}KSb(x,t), \tag{S3}$$

where $S^{-1}KS$ is a constant diagonal matrix [1]. Equation (S3) thus decouples into n independent PDEs, each of which can be solved analytically for certain choices of boundary conditions and initial conditions. Once $b(x,t)$ is determined, the solution is given by $c(x,t) = Sb(x,t)$.

Now consider the system of PDEs, Equations(2)–(3) in the main document, with $c(x,t) = (c_1(x,t), c_2(x,t))^\top$,

$$\begin{bmatrix} \dfrac{\partial c_1(x,t)}{\partial t} \\ \dfrac{\partial c_2(x,t)}{\partial t} \end{bmatrix} - D \begin{bmatrix} \dfrac{\partial^2 c_1(x,t)}{\partial x^2} \\ \dfrac{\partial^2 c_2(x,t)}{\partial x^2} \end{bmatrix} + v \begin{bmatrix} \dfrac{\partial c_1(x,t)}{\partial x} \\ \dfrac{\partial c_2(x,t)}{\partial x} \end{bmatrix} = \begin{bmatrix} -k_1 & 2k_2 \\ k_1 & -k_2 \end{bmatrix} \begin{bmatrix} c_1(x,t) \\ c_2(x,t) \end{bmatrix}. \tag{S4}$$

We consider this PDE model on an infinite domain $-\infty < x < \infty$, with the initial condition for $c_1(x,t)$ given by $c_1(x,0) = \alpha > 0$ for $|x - x_0| \leq h$, and $c_1(x,0) = 0$ for $|x - x_0| > h$. Similarly, the initial condition for $c_2(x,t)$ is given by $c_2(x,0) = \beta > 0$ for $|x - x_0| \leq h$, and $c_2(x,0) = 0$ for $|x - x_0| > h$. This initial condition specifies that both subpopulations vanish

outside an interval of width $2h$ centred at $x = x_0$. Within this interval, the density of each subpopulation takes on a positive, constant value. For our approach, we require $k_1 > 0$ and $k_2 > 0$. The transformation and inverse of the transformation matrix are given by

$$S = \begin{bmatrix} \dfrac{k_2 - k_1 - \Delta}{2k_1} & \dfrac{k_2 - k_1 + \Delta}{2k_1} \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad S^{-1} = \begin{bmatrix} \dfrac{-k_1}{\Delta} & \dfrac{k_2 - k_1 + \Delta}{2\Delta} \\ \dfrac{k_1}{\Delta} & \dfrac{k_1 - k_2 + \Delta}{2\Delta} \end{bmatrix}, \tag{S5}$$

where $\Delta = \sqrt{{k_1}^2 + 6k_1k_2 + {k_2}^2}$. With these ingredients, we can write Equation (S4) in terms of the uncoupled variables $b(x,t) = (b_1(x,t), b_2(x,t))^\top$ as,

$$\begin{bmatrix} \dfrac{\partial b_1(x,t)}{\partial t} \\ \dfrac{\partial b_2(x,t)}{\partial t} \end{bmatrix} - D \begin{bmatrix} \dfrac{\partial^2 b_1(x,t)}{\partial x^2} \\ \dfrac{\partial^2 b_2(x,t)}{\partial x^2} \end{bmatrix} + v \begin{bmatrix} \dfrac{\partial b_1(x,t)}{\partial x} \\ \dfrac{\partial b_2(x,t)}{\partial x} \end{bmatrix} = \begin{bmatrix} \dfrac{-k_1 - k_2 - \Delta}{2} & 0 \\ 0 & \dfrac{-k_1 - k_2 + \Delta}{2} \end{bmatrix} \begin{bmatrix} b_1(x,t) \\ b_2(x,t) \end{bmatrix}, \tag{S6}$$

where we see that the square matrix associated with the source terms is now diagonal, which means that the two PDEs are uncoupled. Using $b(x,0) = S^{-1}c(x,0)$, the initial condition for $b_1(x,t)$ is given by $b_1(x,0) = -\alpha k_1/\Delta + \beta(k_2 - k_1 + \Delta)/(2\Delta)$ for $|x - x_0| < h$, and $b_1(x,0) = 0$ for $|x - x_0| > h$. Similarly, the initial condition for $b_2(x,t)$ is given by $b_2(x,0) = \alpha k_1/\Delta + \beta(k_1 - k_2 + \Delta)/(2\Delta)$ for $|x - x_0| < h$, and $b_2(x,0) = 0$ for $|x - x_0| > h$. With these initial conditions, each decoupled PDE can be solved using a standard integral transform approach [2] to give

$$\begin{aligned} b_1(x,t) &= \left[-\alpha \frac{k_1}{2\Delta} + \beta \frac{(k_2 - k_1 + \Delta)}{4\Delta}\right] \left[\operatorname{erf}\left(\frac{h - (x - x_0 - vt)}{2\sqrt{Dt}}\right) + \operatorname{erf}\left(\frac{h + (x - x_0 - vt)}{2\sqrt{Dt}}\right)\right] \\ &\quad \exp\left[\frac{-t(k_1 + k_2 + \Delta)}{2}\right], \\ b_2(x,t) &= \left[\alpha \frac{k_1}{2\Delta} + \beta \frac{(k_1 - k_2 + \Delta)}{4\Delta}\right] \left[\operatorname{erf}\left(\frac{h - (x - x_0 - vt)}{2\sqrt{Dt}}\right) + \operatorname{erf}\left(\frac{h + (x - x_0 - vt)}{2\sqrt{Dt}}\right)\right] \\ &\quad \exp\left[\frac{-t(k_1 + k_2 - \Delta)}{2}\right]. \end{aligned} \tag{S7}$$

Using the relation $c(x,t) = Sb(x,t)$ the exact solution to the coupled system of PDEs (Equations (2)–(3) in the main document) is

$$\begin{aligned} c_1(x,t) &= \frac{(k_2 - k_1 - \Delta)}{2k_1} b_1(x,t) + \frac{(k_2 - k_1 + \Delta)}{2k_1} b_2(x,t), \\ c_2(x,t) &= b_1(x,t) + b_2(x,t). \end{aligned} \tag{S8}$$

These solutions are strictly valid on the infinite domain, $-\infty < x < \infty$. In practice, we will apply these solutions on a finite domain, $0 \leq x \leq L$. In the usual way [3], choosing x_0 and L to be sufficiently large that the density profiles for $c_1(x,t)$ and $c_2(x,t)$ remain close to zero near $x = 0$ and $x = L$ allows us to use an exact infinite domain solution to approximate the solution of a closely-related finite domain problem [2].

S2 Numerical solution of surrogate PDE for Case 2

Here we solve the continuum-limit PDE for the interacting random walk model (Equation (9) in the main document) using a numerical method-of-lines approach. To implement this approach we spatially discretise the PDE model using standard finite difference approximations for the spatial terms and then solve the resulting system of coupled ordinary differential equations (ODEs) in time using the DifferentialEquations.jl package in Julia to take advantage of automatic time stepping and temporal truncation error control [4].

We solve Equation (9) on $0 < r < R$, with the initial condition given by $c(r,0) = c_0 > 0$ for $r < r_0$ and $c(r,0) = 0$ for $r > r_0$. This initial condition specifies that the population vanishes outside a circle of radius r_0 centred at the origin. Within this circle, the population density assumes a positive, constant value c_0. We discretise the spatial terms on a uniform grid with grid spacing $\delta > 0$, such that $c(r_n,t) = c_n$ for $n = 1,2,3,\ldots,N$, where $r_n = (n-1)\delta$. At $r = 0$ and $r = R$, corresponding to mesh points r_1 and r_N, respectively, we impose zero-flux boundary conditions to give

$$\frac{\mathrm{d}c_1}{\mathrm{d}t} = \frac{D}{\delta^2}\left(c_2 - c_1\right) + \Lambda c_1\left(1 - c_1\right), \tag{S9}$$

$$\frac{\mathrm{d}c_n}{\mathrm{d}t} = \frac{D}{\delta^2}\left(c_{n+1} - 2c_n + c_{n-1}\right) + \frac{D}{2r_n\delta}\left(c_{n+1} - c_{n-1}\right) + \Lambda c_n\left(1 - c_n\right), \quad \text{for } n = 2,3,4,\ldots,N-1, \tag{S10}$$

$$\frac{\mathrm{d}c_N}{\mathrm{d}t} = \frac{D}{\delta^2}\left(c_{N-1} - c_N\right) + \Lambda c_N\left(1 - c_N\right). \tag{S11}$$

We solve this system of ODEs using Heun's method within the DifferentialEquations.jl package in Julia [4]. All results presented in this work correspond to $\delta = 0.5$. To ensure our results are grid independent, we considered a number of test cases and checked that numerical results with $\delta = 0.5$ were indistinguishable from results with $\delta = 0.25$.

S3 Model validation

In this work, we employ two computationally efficient continuum-limit approximations of two types of RWMs. For the non-interacting RWM with two subpopulations, we use a system of PDEs (Equations (2)–(3) in the main document). For the interacting RWM with a single population, we use a two-dimensional (2D) generalisation of the well-known Fisher–Kolmogorov equation in radial coordinates (Equation (9) in the main document). Here, we present evidence demonstrating that the solutions of the surrogate PDE models gives an accurate approximation of appropriately averaged data from the relevant RWM.

S3.1 Case 1 validation

To validate the surrogate PDE for case 1 (Equation (2)–(3) in the main document), we compare exact solutions of the PDEs given by Equation (S8) with count data obtained from simulations of the noninteracting RWM. Since the solution of the PDE models gives us non-dimensional densities, $c_1(x,t)$ and $c_2(x,t)$, and the RWM simulations give count data, compare the count data with the quantities $J \times c_1(x,t)$ and $J \times c_2(x,t)$ to give a direct comparison, noting that the PDE solution does not capture any fluctuations. To accurately compare noise-free PDE solutions with the noisy count data, we deliberately reduce stochastic variability in data from the non-interacting RWM by averaging the count data over 100 identically prepared realisations of the discrete model. Our comparison involves identical parameters, initial conditions, boundary conditions, and simulation time as in Case 1. In summary, the parameters of the simulation are $(\alpha, \beta, h, x_0, D, v, k_1, k_2)^\top = (0.5, 0.2, 19.5, 99.5, 0.25, 0.25, 0.02, 0.03)^\top$ and the comparison is made at $t = 100$ on a large but truncated domain.

We present snapshots of a single realisation of the RWM simulation at the initial placement and after $k = 100$ time steps in Figure S1(a)–(b). The comparison in Figure S1(c)–(d) confirms that the same initial condition is used in both the PDE model and the RWM. The comparison after 100 time steps for both subpopulations is shown in Figure S1(e)–(f) which demonstrates that the solution of the surrogate PDE model accurately describes the output of the RWM when care is taken to miminise stochastic fluctuations.

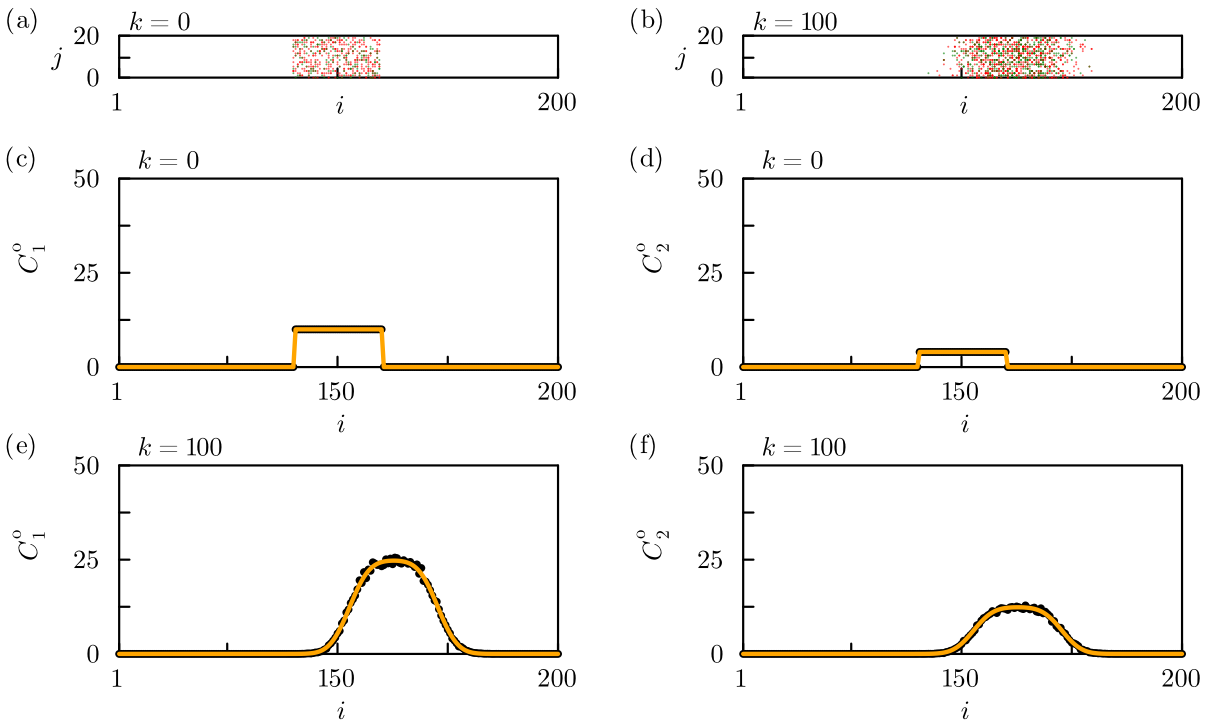

Figure S1: Validation of the surrogate PDE model for Case 1. he PDE solution is evaluated using $(\alpha, \beta, h, x_0, D, v, k_1, k_2)^\top = (0.5, 0.2, 19.5, 99.5, 0.25, 0.25, 0.02, 0.03)^\top$. (a)–(b) Snapshots of the agent distribution. (c)–(f) Averaged count data from the non-interacting RWM over 100 realisations (black dots), superimposed with the corresponding PDE solution (solid orange line). (c) and (e) Comparison for subpopulation 1. (d) and (f) Comparison for subpopulation 2.

S3.2 Case 2 validation

To validate the PDE for Case 2 (Equation (9) in the main document), we compare the solution of the PDE with count data obtained from a simulation using the interacting RWM. The PDE model is solved numerically as detailed in Section S2. The RWM is simulated using the same parameters, initial conditions, boundary conditions, and simulation time as in Case 2b of the main document. In summary we specify $(r_0, D, \Lambda, c_0)^\top = (20, 0.25, 0.01, 0.8)^\top$ and compare the solutions at $t = 600$ for $0 < r < 100$. Under these conditions we will now explain how we check that the solution of the surrogate PDE model provides a good prediction of the computationally expensive RWM.

We validate the PDE model by comparing the radial and column count data. For the radial count data, we consider the count data as a function of radius, r. The solution of the PDE, $c(r, t)$, represents the population density at radius r and time t. In contrast, the RWM simulation can be summarised by the variable $C^*(i, j, k)$, which gives the number of agents at site (i, j) after k time steps in a single realisation. To convert the simulation result into a radial density profile, we superimpose Q equally-spaced concentric circles on the lattice. Each circle, shown in Figure S2(b) with $Q = 20$, is centred at $(i, j) = (100, 100)$, and the qth circle has radius $r_q = 100q/Q$ for $q = 1, 2, 3, \ldots, Q$. For $q \geq 2$, the density at radius r is approximated by counting the number of agents within the annulus $r_{q-1} < r < r_q$. For the smallest concentric circle, $q = 1$, the density is approximated by counting agents in the circular region $r < r_1$. This process yields Q equally-spaced estimates of the agent counts as a function of r. To reduce the impact of stochastic fluctuations, we average these counts over 100 identically prepared realisations of the discrete model, noting that averaged count data can lead to non-integer values. We compare this averaged count data with the solution of the surrogate PDE (Equation (9) in the main document) in terms of the quantity $(2\pi r\Delta) \times c(r, t)$, where $c(r, t)$ is the solution of the surrogate PDE model. The factor $2\pi r\Delta$ is the area of the annulus of thickness Δ, centred at radial position r. In our simulations, we have $\Delta = 100/Q$.

For the column count data, we consider the count data as a function of the horizontal position, x. The count data are obtained from the RWM simulation, consistent with the approach described in Section 2.3 of the main document, except that the data are averaged over 100 identically prepared realisations. To ensure consistency with the units of the RWM simulation, the PDE solution is rescaled by $H \times c(r, t)$, which represents the agent count along the radius. As in Section 2.3, we take $H = 20$. By symmetry, we have $r = |x - x_0|$, where x_0 denotes the x-coordinate of the centre of the spreading population. The scaled

PDE solution can therefore be expressed as a function of horizontal position, $C^{\mathrm{o}}(x,t)$.

Our discrete-continuum comparison is performed for the radial count data, $C^{\mathrm{o}}(r,t)$, and column count data, $C^{\mathrm{o}}(x,t)$, summarised in Figure S2. The initial distribution of agents is given in Figure S2(a), and the resulting distribution of agents after $k = 600$ time steps is given in Figure S2(b) where we see the impact of combined migration and proliferation as the population spreads away from the initial region containing the agents. Note that the impact of proliferation is clear as the number of agents on the lattice after $k = 600$ time steps has clearly increased, and the density of agents towards the centre of the population, near $(i,j) = (100,100)$ where $r = 0$, is high. In this region the population forms an almost confluent monolayer, which is close to the maximum packing density in the interacting RWM. The distribution of radial count data, $C^{\mathrm{o}}(r,t)$, as a function of radial position obtained from the discrete model is superimposed on count estimates using the surrogate PDE model plotted in Figure S2(c)–(d) for $t = 0$ and $t = 600$, respectively. The distribution of column count data, $C^{\mathrm{o}}(x,t)$, as a function of horizontal position obtained from the discrete model is superimposed on count estimates using the surrogate PDE model plotted in Figure S2(e)–(f) for $t = 0$ and $t = 600$, respectively. Comparing the discrete count data with the distribution of count data from the surrogate PDE solution confirms that the solution of the surrogate PDE provides a good approximation of RWM in terms of agent counts where care has been taken to construct stochastic count data with minimal fluctuations.

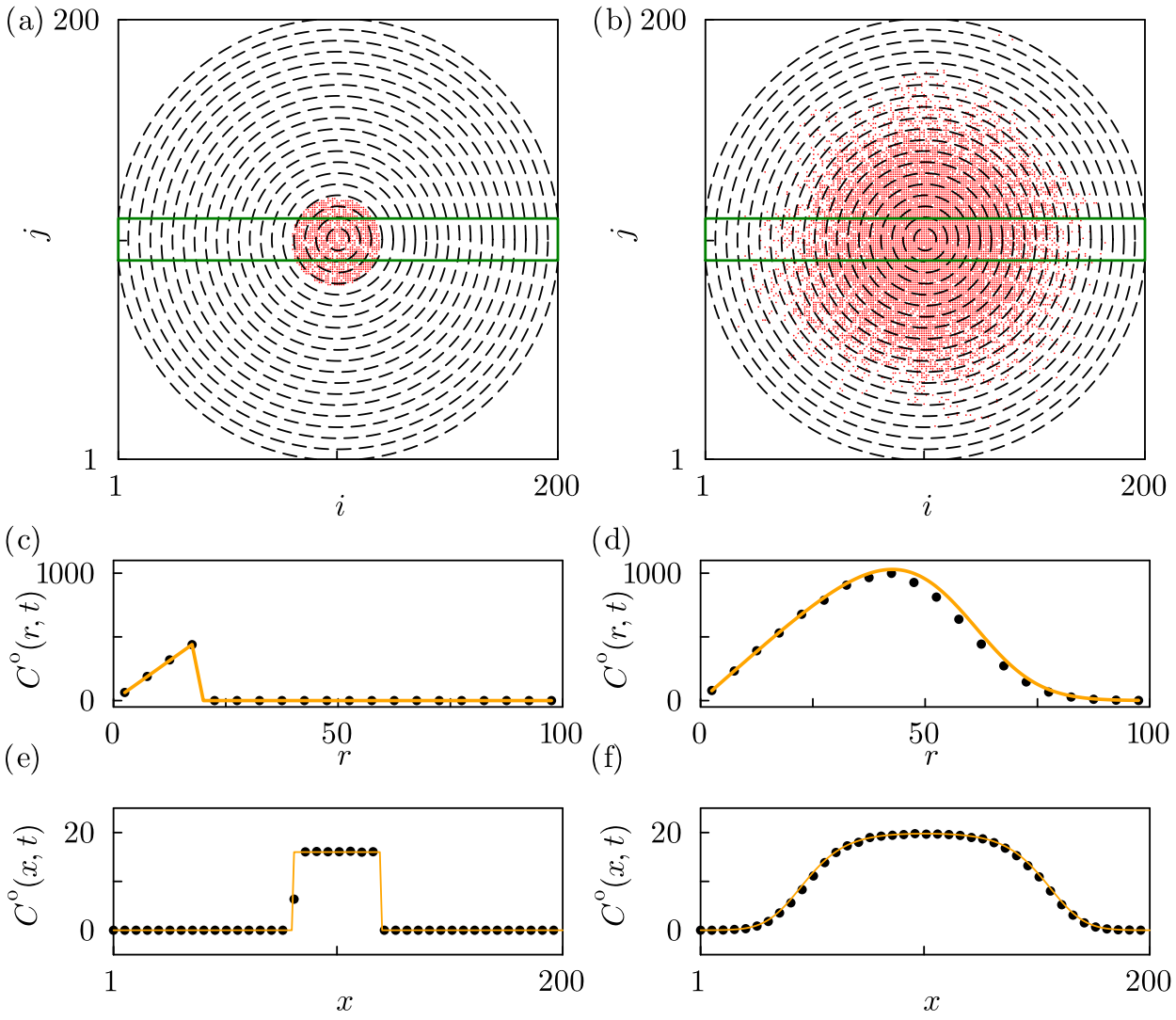

Figure S2: Validation of the surrogate PDE model for Case 2. The PDE is solved using $(r_0, D, \Lambda, c_0)^\top = (20, 0.25, 0.01, 0.8)^\top$. (a)–(b) Snapshots of the agent distribution at $k = 0$ and $k = 600$, respectively. Each snapshot is superimposed with 20 concentric circles (black dashed lines) and a rectangular box (green line), which are used to construct the radial and column count data from the stochastic model, respectively. (c)–(f) Discrete simulation results (black dots) superimposed with the solution of the surrogate PDE model (solid orange line). (c)–(d) Radial count data comparisons at $t = 0$ and $t = 600$, respectively. (e)–(f) Column count data comparisons at $t = 0$ and $t = 600$, respectively.

S4 Additional Results: Case 1 over a longer timescale

Results for estimation, identifiability and sensitivity analysis for Case 1 in the main document are based on data captured after a relatively short simulation, $k = 100$. Here we repeat the exact same analysis for data collected after a longer duration of time, $k = 200$. The data for this additional case is summarised in Figure S3 where these simulations are precisely the same as used in the main document except that data are collected after $k = 200$ time steps instead of $k = 100$ time steps.

New results in terms of parameter estimation, parameter identifiability and sensitivity analysis via profile-wise predictions are summarised in Figure S4 for the new data collected after $k = 200$ time steps. Comparing these additional results with those in the main document (Figure 4) for the shorter time scale simulation reveal that there is no striking difference in terms of parameter estimation, parameter identifiability and sensitivity analysis in this case.

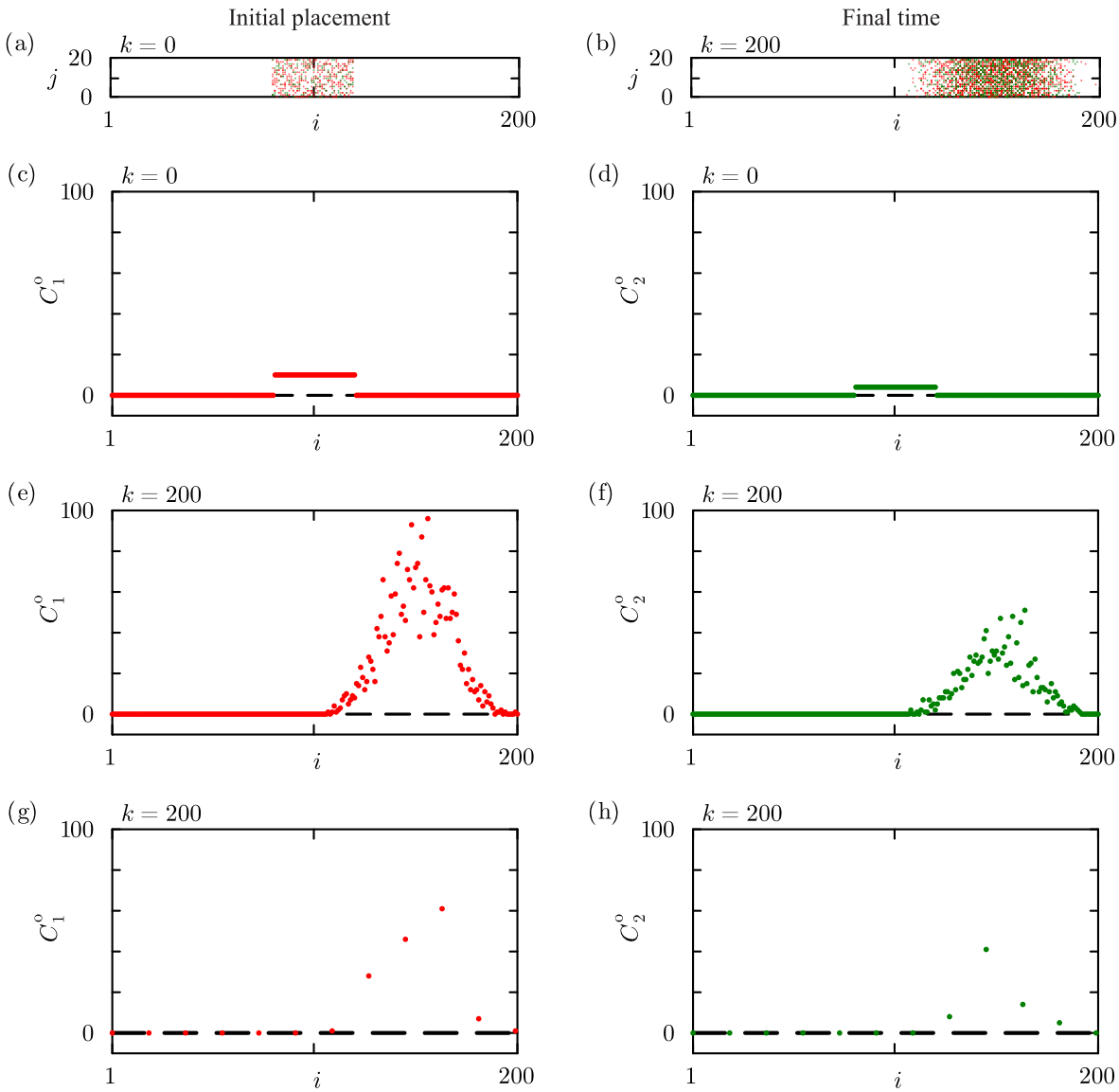

Figure S3: Snapshots of agent distributions and count data from the non-interacting RWM with $(P, \rho, \lambda_1, \lambda_2)^\top = (1, 0.5, 0.02, 0.03)^\top$. (a)–(b) Snapshots of agent distribution. (c)–(f) Full count data. (g)–(h) Subsampled count data obtained from data in (e)–(f), respectively. Horizontal dashed lines in (e)–(h) show the $C_s^{\mathrm{o}} = 0$ lower bound for the count data. Results correspond to $I = 200$ and $J = 20$.

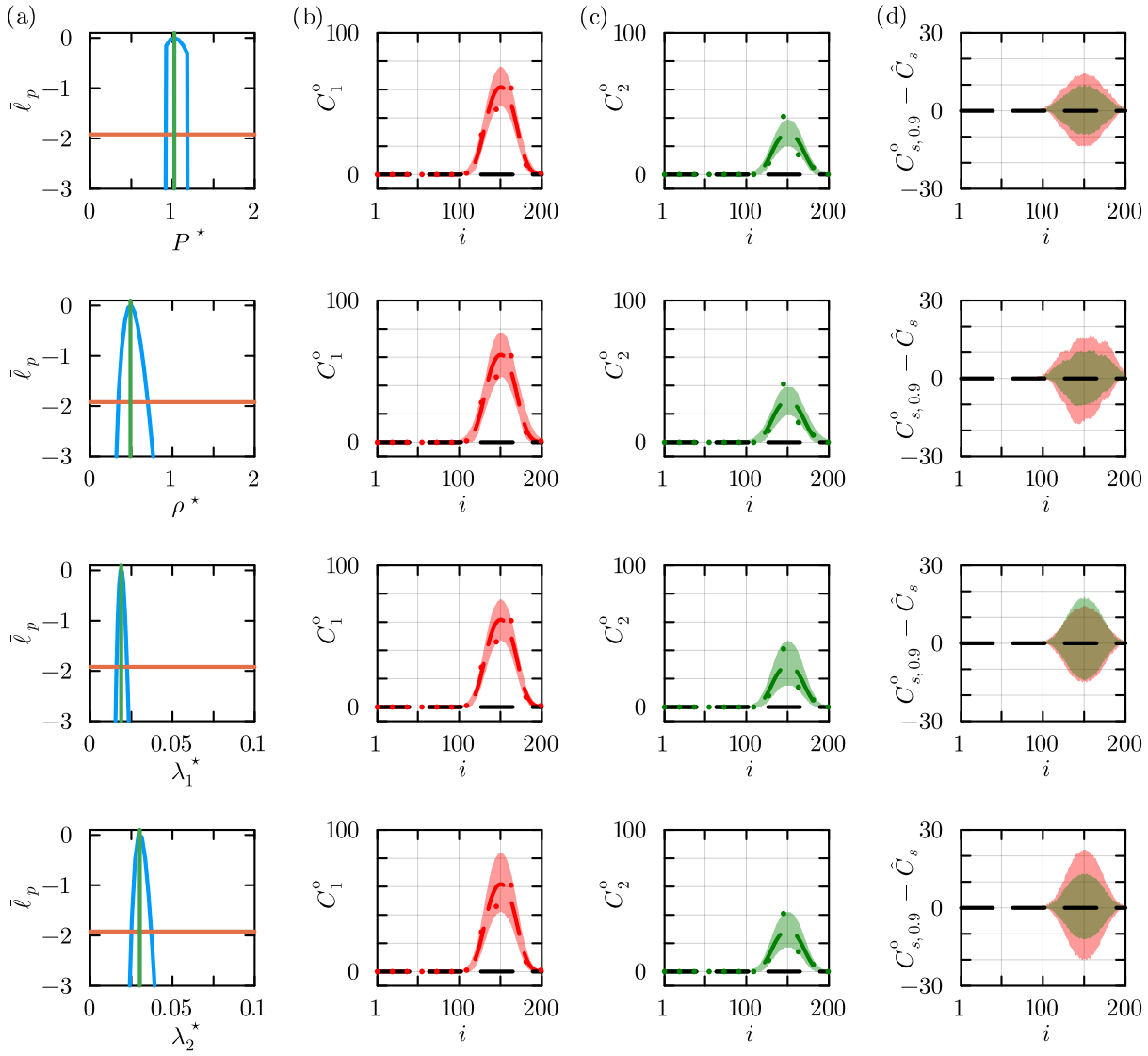

Figure S4: Estimation, identifiability and sensitivity results for the non-interacting RWM. (a) Univariate profiles for $\theta = (P^\star, \rho^\star, \lambda_1^\star, \lambda_2^\star)^\top$ indicate the MLE (vertical green line) and the 95% asymptotic threshold (horizontal orange line). The MLE is $\hat{\theta} = (1.0227, 0.4887, 0.0188, 0.0302)^\top$. (b)–(c) Profile-wise prediction intervals for C_1 (shaded red) and C_2 (shaded green) for each parameter superimposed with count data (green dots), and the solution of the PDE evaluated at the MLE for C_1 (dashed red) and C_2 (dashed green). (d) Difference between confidence set and the solution of the mathematical model evaluated at the MLE for C_1 (shaded red) and C_2 (shaded green).

References

1. Clement TP. Generalized solution to multispecies transport equations coupled with a first-order reaction network. Water Resources Research. 2001; 37:109–7. DOI: 10.1029/2000WR900239
2. Crank J. The Mathematics of Diffusion. Clarendon Press, 1964
3. Simpson MJ and McCue SW. Fisher–KPP-type models of biological invasion: open source computational tools and practical examples. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 2024; 480:20240186. DOI: 10.1098/rspa.2024.0186. Available from: https://doi.org/10.1098/rspa.2024.0186
4. Rackauckas C and Nie Q. DifferentialEquations.jl – A performant and feature-rich ecosystem for solving differential equations in Julia. Journal of Open Research Software. 2017; 5:15. DOI: 10.5334/jors.151