A numerical approach to evaluating the transient distribution of a quasi birth-death process

Mandjes, M.; Sollie, B.

DOI
10.1007/s11009-021-09882-6

Publication date
2022

Document Version
Final published version

Published in
Methodology and Computing in Applied Probability

Methodology and Computing in Applied Probability (2022) 24:1693–1715
https://doi.org/10.1007/s11009-021-09882-6

A Numerical Approach for Evaluating the Time-Dependent Distribution of a Quasi Birth-Death Process

Michel Mandjes[1,2,3] · **Birgit Sollie**[4]

Received: 29 May 2020 / Revised: 25 June 2021 / Accepted: 30 June 2021 /
Published online: 15 July 2021

Abstract
This paper considers a continuous-time quasi birth-death (QBD) process, which informally can be seen as a birth-death process of which the parameters are modulated by an external continuous-time Markov chain. The aim is to numerically approximate the time-dependent distribution of the resulting bivariate Markov process in an accurate and efficient way. An approach based on the Erlangization principle is proposed and formally justified. Its performance is investigated and compared with two existing approaches: one based on numerical evaluation of the matrix exponential underlying the QBD process, and one based on the uniformization technique. It is shown that in many settings the approach based on Erlangization is faster than the other approaches, while still being highly accurate. In the last part of the paper, we demonstrate the use of the developed technique in the context of the evaluation of the likelihood pertaining to a time series, which can then be optimized over its parameters to obtain the maximum likelihood estimator. More specifically, through a series of examples with simulated and real-life data, we show how it can be deployed in model selection problems that involve the choice between a QBD and its non-modulated counterpart.

Keywords Quasi birth-death processes · Time-dependent probabilities · Erlang distribution · Maximum likelihood estimation

Mathematics Subject Classification (2010) 60J22 · 92D25 · 62Fxx

1 Introduction

Birth-death (BD) processes are continuous-time Markov processes where transitions can only increase or decrease the state by one—usually referred to as births and deaths, respectively. These well-known processes are widely used and have applications in many areas such as biology, epidemiology and operations research. In some real life systems, however,

✉ Birgit Sollie
birgit.corporaal@gmail.com

Extended author information available on the last page of the article.

it is likely that there is a higher variability in the birth- and/or the death rates than modelled by a conventional BD process. Observe for example the data in Fig. 1, displaying the annual counts of the female population of the whooping crane (see Stratton (2020) for the original data, and Davison et al. (2020) for the female counts). There are some fluctuations visible in the evolution of the population size, which could be indicative of a higher variability in some, or all, model parameters. One wonders whether specific generalizations of the BD process could be more suitable for this data. The major aim of this paper is to develop methodologies that can be used to rigorously compare the fit of a conventional BD process with more general alternatives.

An example of a more general alternative to the conventional BD process is the *quasi birth-death* (QBD) process. The population process, called the level process, in a QBD process is given by a BD process of which the transition rates are modulated by a continuous-time Markov chain, called the phase process. This means that the transition rates of the QBD process switch between multiple distinct values at the jump times of the phase process. Together, the level and the phase process form a bivariate Markov process. In an even more general QBD process, the number of states of the phase process can depend on the current value of the level process. This leads to a so-called level-dependent QBD process, which is the process that we consider in this paper. Over the years, various properties of level-dependent QBD processes have been studied. We refer to e.g. (Bright and Taylor 1995) for calculations concerning the equilibrium distribution, Ramaswami and Taylor (1996) for the computation of certain matrices that play an important role in the QBD context, and Mandjes and Taylor (2016) for a characterization of the process' running maximum.

In the above whooping crane example, one would like to statistically compare the scenario of the data stemming from a conventional BD process with that of the data stemming from the more general QBD process. In order to do so, a prerequisite is that we have a methodology to compute, for both models, the likelihood of our dataset. This, in turn, requires techniques for the evaluation of the time-dependent probabilities corresponding to BD and QBD processes. In this paper we investigate different approaches to compute the time-dependent probabilities of the joint Markov process of level and phase in the level-dependent QBD process. In particular, we propose, justify and test an approach based on the so-called *Erlangization* principle, which we compare with existing alternatives. Then we point out through a series of experiments, including the whooping crane example, how such

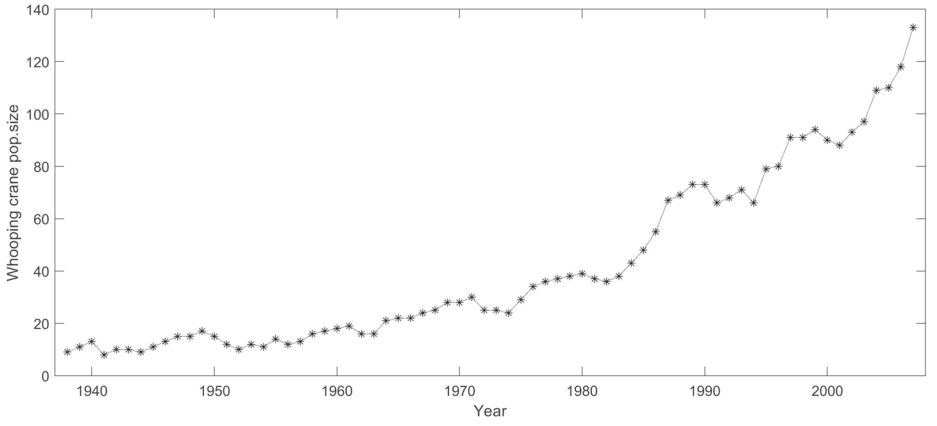

Fig. 1 Yearly population count of female whooping cranes arriving in Texas each autumn

techniques can be used in determining whether a BD process or a QBD process yields the better fit.

In order to numerically evaluate probabilities pertaining to BD and QBD processes, various methods have been developed. For all practical purposes, it is natural to let the underlying Markov chain live on a finite state space. A commonly applied approach to compute the time-dependent distribution boils down to computing the matrix exponential of the transition rate matrix, say Q, of the corresponding Markov chain (of which the states, in the QBD case, encode all level/phase combinations). More precisely, the (i, j)-th entry of e^{Qt} provides us with the probability of being in state j at time t given that the initial state was i, where in the QBD context, i and j correspond to specific phase/level combinations. It is known, however, that the computation of matrix exponentials may involve various numerical complications. We refer to e.g. the survey (Moler and Van Loan 2003), where about all approaches available at that time, it is stated that 'none are completely satisfactory'. We remark that since the publication of Moler and Van Loan (2003) substantial progress has been made in order to resolve the numerical issues: various novel, more sophisticated approaches are being developed (Al-Mohy and Higham 2009). Alternatively, one could solve the linear system of differential equations resulting from the Kolmogorov equations. As argued in e.g. Reibman and Trivedi (1988), this method has various intrinsic problems as well. Most notably, if the underlying system is large, the Q matrix is ill-conditioned, or the differential equations are stiff, the evaluation can be slow and/or inaccurate.

Owing to the special structure of the transition rate matrix (i.e., the Q-matrix having non-negative off-diagonal entries, row sums equal to 0), another approach is possible. In the *uniformization* technique the continuous-time Markov chain is converted to a discrete-time Markov chain (say with transition matrix P) of which the jump times correspond to a Poisson process with a constant rate (say σ). Here P and σ are chosen in such a way that the newly defined process and the original continuous-time Markov chain are statistically identical, i.e. all distributional properties are equivalent. The distribution of the continuous-time Markov chain at time t can thus be obtained by weighing matrices P^k by the probabilities that the Poisson process has jumped k times in $[0, t]$, and summing these over k ($k = 0, 1, \ldots$). This method performs well in many cases, but it has disadvantages as well. Evidently, in numerical computations the above summation has to be truncated at some finite threshold, where the issue is to choose this threshold high enough to make sure that the error made is negligible. In addition, to compute all k-step transition matrices P^k, the corresponding matrix multiplications need to be executed, which may make the procedure prohibitively slow. Uniformization was introduced in the 1950s in Jensen (1953); see also Grassmann (1991), Gross and Miller (1984), and Melamed and Yadin (1984) for other seminal contributions; an extensive discussion on its pros and cons can be found in van Dijk et al. (2018).

In this paper we discuss an alternative approach, based on the Erlangization principle, which has previously been explored (in other contexts) in e.g. Asmussen et al. (2002), Ramaswami et al. (2008), and Mandjes and Taylor (2016). It uses the fact that, although the computation of the distribution of the state of the Markov chain at a deterministic time is challenging, its counterpart at an exponentially distributed epoch just requires solving a system of linear equations. A second observation is that the sum of k independent exponentially distributed random variables with mean t/k—corresponding to an Erlang distribution with scale parameter k and shape parameter k/t—converges to the deterministic number t as k grows large. Combining these two properties, the idea is to evaluate the transition probabilities at an exponentially distributed epoch with parameter k/t, and to raise the resulting matrix to the power k. It is tempting to believe that our deterministic-time transition

probabilities are accurately approximated by this procedure as long as k is chosen large enough. This approach has the inherent advantage that the number of matrix multiplications is limited: if k is a power of two, it suffices to square the exponential-time transition matrix $\log_2 k$ times. Importantly, we can prove the theoretical correctness of the approach, in that we show that it becomes increasingly precise as $k \to \infty$, with an argumentation that relies on large-deviations theory. By means of a series of numerical examples, we also show that this approach is in many settings computationally faster than the approaches based on the matrix exponential and uniformization, without compromising the accuracy.

Going back to the whooping crane data from Fig. 1, an interesting question remains if a QBD process indeed provides a better fit to the data than a conventional BD process, as one might suspect from the graph. In the last section of this paper we investigate this type of model selection problems, both with simulated and real-life data. By the techniques discussed in this paper, we can compute the likelihood pertaining to a time series, thus enabling the evaluation of maximum likelihood estimates. In this respect, note that all three approaches (i.e., matrix exponential, uniformization, Erlangization) can be applied in the QBD as well as the BD setting. As the class of QBD processes contains the class of BD processes, evidently the former by definition leads to a better fit, but this comes at the price of additional parameters. To 'fairly' compare the two models, taking into account the corresponding numbers of parameters, we perform the model selection relying on the celebrated Akaike information criterion (AIC).

The remainder of this paper is organized as follows. The level-dependent QBD process and its corresponding time-dependent distribution are defined in Section 2. Section 3 shows how the transition probabilities at an exponentially distributed epoch can be computed by solving a system of linear equations. The findings of Section 3 are then used in Section 4 to motivate the Erlangization approach; in addition the theoretical correctness of this approach is established. Section 5 experimentally investigates the performance of the three approaches discussed above. Section 6 discusses the model selection problem of choosing between BD processes and QBD processes, using examples with simulated as well as real-life data, with all likelihood computations relying on Erlangization. We conclude the paper, in Section 7, with a brief discussion.

2 Model and Preliminaries

In this section we introduce the class of QBD processes that will be considered in this paper. Next, we define the object of our study, viz. the time-dependent distribution of the corresponding bivariate Markov process, and briefly discuss established approaches to numerically evaluate it.

2.1 Model

A QBD process is a bivariate process comprising *levels* and *phases*. The level process, in the sequel denoted by $\{M_t\}_{t\geqslant 0}$, attains values in $\{0, 1, \ldots, C\}$ for some $C \in \mathbb{N}$. The phase process is denoted by $\{X_t\}_{t\geqslant 0}$; when the level M_t equals m, the phase X_t attains values in $\{1, \ldots, d_m\}$, for some $d_m \in \mathbb{N}$. In many applications the number of phases is uniform in the level, or, more concretely, $d_m = d \in \mathbb{N}$ for all $m \in \{0, \ldots, C\}$. The birth-death nature of the process is reflected by the fact that at any transition the level can increase or decrease by at most 1.

We provide a more precise description of the model $\{M_t, X_t\}_{t\geqslant 0}$ by formally defining the corresponding transition rates.

- In the first place, $Q^{(m)}$, for $m \in \{0, 1, \ldots, C\}$, is a transition rate matrix of dimension $d_m \times d_m$ that corresponds to a continuous-time Markov chain living on the state space $\{1, \ldots, d_m\}$. Its elements are denoted by $q_{ij}^{(m)}$; they are non-negative for $i \neq j$ and in addition the row sums are zero. Whenever $M_t = m$, a jump from phase i to phase j *that leaves the level unchanged* occurs with rate $q_{ij}^{(m)}$, for $i \neq j$. In addition, we define the total rate out of phase i (while the level remains at m),

$$q_i^{(m)} := -q_{ii}^{(m)} = \sum_{j\neq i} q_{ij}^{(m)};$$

  here the sum on the right hand side should be understood to be over all $j \in \{1, \ldots, d_m\}$ such that $j \neq i$.
- In the second place, there are transitions in which the level goes up by 1, while at the same time the phase potentially changes as well. For $m \in \{0, 1, \ldots, C-1\}$, the matrix $\Lambda^{(m)}$ has dimension $d_m \times d_{m+1}$. Its (i, j)-th element contains the rate $\lambda_{ij}^{(m)} \geqslant 0$ at which the level increases by 1 while simultaneously the phase jumps from i to j; note that $i = j$ is allowed (under the proviso that $i \leqslant \min\{d_m, d_{m+1}\}$). Throughout this paper we write

$$\lambda_i^{(m)} := \sum_{j=1}^{d_{m+1}} \lambda_{ij}^{(m)},$$

  to denote the total rate corresponding to an increase in level from phase i, with $i \in \{1, \ldots, d_m\}$.
- Finally, there are transitions in which the level goes down by 1, again potentially simultaneously with a phase change. The (i, j)-th element of the matrix $\mathcal{M}^{(m)}$, which has dimension $d_m \times d_{m-1}$ for $m \in \{1, 2, \ldots, C\}$, contains the rate $\mu_{ij}^{(m)} \geqslant 0$ at which the level decreases by 1 while the phase jumps from i to j; again, $i = j$ is allowed (if $i \leqslant \min\{d_{m-1}, d_m\}$). We compactly write for the total rate of a decrease in level from phase i, with $i \in \{1, \ldots, d_m\}$,

$$\mu_i^{(m)} := \sum_{j=1}^{d_{m-1}} \mu_{ij}^{(m)}.$$

In this work we assume that the matrices $Q^{(m)}$, $\Lambda^{(m)}$, and $\mathcal{M}^{(m)}$ are such that the joint Markov process $\{M_t, X_t\}_{t\geq 0}$ is irreducible, implying that, with positive probability any level/phase pair can be reached from any other level/phase pair in any amount of time. The number of states of this process is $D := \sum_{m=0}^{C} d_m$. We let Q be the $D \times D$ transition rate matrix of $\{M_t, X_t\}_{t\geq 0}$, that is,

$$Q := \begin{pmatrix} \bar{Q}^{(0)} & \Lambda^{(0)} & 0 & \cdots & 0 & 0 \\ \mathcal{M}^{(1)} & \bar{Q}^{(1)} & \Lambda^{(1)} & \cdots & 0 & 0 \\ 0 & \mathcal{M}^{(2)} & \bar{Q}^{(2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & \bar{Q}^{(C-1)} & \Lambda^{(C-1)} \\ 0 & 0 & 0 & \cdots & \mathcal{M}^{(C)} & \bar{Q}^{(C)} \end{pmatrix},$$

where $\bar{Q}^{(m)}$ is defined as $Q^{(m)}$ with the diagonal entries adapted such that the row sums of Q are zero. More precisely, the definition of $\bar{Q}^{(m)}$ entails that the diagonal of Q consists of entries of the form $-\sigma_i^{(m)}$, where (for $m \in \{0, 1, \ldots, C\}$ and $i \in \{1, \ldots, d_m\}$)

$$\sigma_i^{(m)} := q_i^{(m)} + \lambda_i^{(m)} 1_{\{m<C\}} + \mu_i^{(m)} 1_{\{m>0\}}. \tag{1}$$

These rates $\sigma_i^{(m)}$ are to be interpreted as the 'total flux' when the level is m and the phase is i. For later reference we define the largest entry among these fluxes by

$$\sigma := \max_{m\in\{0,1,\ldots,C\}} \left(\max_{i\in\{1,\ldots,d_m\}} \sigma_i^{(m)} \right). \tag{2}$$

We finally introduce the $D \times D$ matrix P_t that describes the process' time-dependent distribution. It contains probabilities of the type

$$p_{ij}(m, m'; t) := \mathbb{P}(M_t = m', X_t = j \mid M_0 = m, X_0 = i), \tag{3}$$

with the states ordered in the same way as is done in Q. The remainder of this section is devoted to describing two often used methods to numerically evaluate P_t, with which we compare our method in Section 5.

2.2 Time-Dependent Probabilities: Matrix Exponential

It is commonly known that P_t, as given in Eq. 3, can be expressed as a matrix exponential, i.e., $P_t = e^{Qt}$. As argued extensively in (Moler and Van Loan 2003), the numerical evaluation of such matrix exponentials is a delicate issue. In numerical computing environments various types of algorithms have been implemented. MATLAB's implementation $\texttt{expm}(\cdot)$ is based on the algorithm developed in Higham (2005), and is claimed to be highly accurate; see also the further refinements in Al-Mohy and Higham (2009).

Approximation 1 (Matrix exponential) *P_t is approximated by*

$$P_t^{(m)} := \texttt{expm}(Qt), \tag{4}$$

based on MATLAB*'s implementation* $\texttt{expm}(\cdot)$.

2.3 Time-Dependent Probabilities: Uniformization

An alternative existing approach to obtain time-dependent probabilities relies on uniformization. The main idea is to convert the continuous-time Markov chain to a discrete-time Markov chain of which the jump times follow a Poisson process with a constant rate. For the QBD process we let this uniform rate be σ, as defined in Eq. 2. Define, with self-evident notation,

$$\mathscr{P}_{(m,i),(m',j)} := \begin{cases} \sigma^{-1} Q_{(m,i),(m',j)} & \text{if } (m,i) \neq (m',j), \\ 1 - \sigma^{-1} \sum_{(m',j)\neq(m,i)} Q_{(m,i),(m',j)} & \text{if } (m,i) = (m',j), \end{cases}$$

or, equivalently, $Q = \sigma\mathscr{P} - \sigma I$. Observe that by definition of σ all these entries are in $[0, 1]$; in fact, $\mathscr{P}$ is a transition probability matrix of a discrete-time Markov chain. Sampling the number of jumps in $(0, t]$ of this discrete-time Markov chain according to a Poisson distribution with parameter σt, we find that

$$P_t = e^{Qt} = e^{(\sigma\mathscr{P}-\sigma I)t} = \sum_{k=0}^{\infty} e^{-\sigma t} \frac{(\sigma t)^k}{k!} \mathscr{P}^k,$$

The following approximation is based on this representation.

Approximation 2 (Uniformization) *For a given $\ell \in \mathbb{N}$, P_t is approximated by*

$$P_t^{(\mathrm{u},\ell)} := \sum_{k=0}^{\ell} e^{-\sigma t} \frac{(\sigma t)^k}{k!} \mathscr{P}^k. \tag{5}$$

A question is: how to select a value of ℓ to make sure that the error made is below some allowable threshold $\delta > 0$? While in practical situations one typically relies on pragmatic criteria to determine ℓ, a formally justified, but potentially somewhat conservative, approach is the following. Realize that, trivially, as $\ell \to \infty$,

$$0 \leqslant p_{ij}(m, m'; t) - p_{ij}^{(\mathrm{u},\ell)}(m, m'; t) \leqslant \mathbb{P}(\mathrm{Pois}(\sigma t) \geqslant \ell + 1) \to 0,$$

where $\mathrm{Pois}(\sigma t)$ denotes a Poisson random variable with mean σt. This bound entails that one could use for example the Chernoff bound to find the ℓ for which $\mathbb{P}(\mathrm{Pois}(\sigma t) \geqslant \ell+1) < \delta$:

$$\begin{aligned} \mathbb{P}(\mathrm{Pois}(\sigma t) \geqslant \ell + 1) &\leqslant \inf_{\theta>0} e^{-\theta(\ell+1)}\, \mathbb{E}\, e^{\theta\, \mathrm{Pois}(\sigma t)} \\ &= \inf_{\theta>0} e^{-\theta(\ell+1)}\, e^{\sigma t(e^{\theta}-1)} = \left(\frac{\sigma t}{\ell+1}\right)^{\ell+1} e^{\ell+1-\sigma t}; \end{aligned} \tag{6}$$

equating the right-hand side to δ yields an ℓ with the desired property.

Note that an important advantage of uniformization is its implementational simplicity: the matrix $\mathscr{P}$ is trivially computed from Q, and it is straightforward to evaluate its powers. The main disadvantage of uniformization is that *many* matrix multiplications are needed, as the approximation uses *all* matrices $\mathscr{P}^k$ for $k = \{0, 1, \ldots, \ell\}$; particularly when σ is relatively large, implying that ℓ has to be chosen large as well, the procedure may become rather time consuming. To remedy this disadvantage of uniformization, we pursue an alternative approach, based on the concept of *Erlangization*. This approach combines two ideas: (i) if the time horizon is exponentially distributed rather than deterministic, then the corresponding transition probability follows simply by solving a linear system of equations, and (ii) one can approximate a deterministic number by a sum of a large number of independent exponentially distributed random variables with an appropriately chosen parameter. Section 3 first elaborates on property (i). Then, in Section 4, it is pointed out how, based on these two properties, P_t can be efficiently and accurately approximated. In Section 5 we numerically compare the performance of Erlangization with the matrix exponential approach (4) and uniformization (5).

3 Time-Dependent Probabilities at Exponential Epochs

The main goal of this section is to show that the evaluation of the distribution of $\{M_t, X_t\}$ at an exponentially distributed epoch essentially reduces to solving a linear system of equations. Let T_η be an exponentially distributed random variable with mean η^{-1} (with $\eta > 0$), independent of $\{M_t, X_t\}_{t\geqslant 0}$. We define

$$\pi_{ij}(m, m'; \eta) := \mathbb{P}(M_{T_\eta} = m', X_{T_\eta} = j \mid M_0 = m, X_0 = i).$$

We now point out how to compute these probabilities $\pi_{ij}(m, m'; \eta)$, with $m, m' \in \{0, 1, \ldots, C\}$, $i \in \{1, \ldots, d_m\}$, and $j \in \{1, \ldots, d_{m'}\}$. Recall the definition of $\sigma_i^{(m)}$ in Eq. 1.

The standard 'Markovian reasoning' yields

$$\pi_{ij}(m,m';\eta) = \sum_{i'=1,i'\neq i}^{d_m} \frac{q_{ii'}^{(m)}}{\sigma_i^{(m)}+\eta}\,\pi_{i'j}(m,m';\eta) + \sum_{i'=1}^{d_{m+1}} \frac{\lambda_{ii'}^{(m)}}{\sigma_i^{(m)}+\eta}\,\pi_{i'j}(m+1,m';\eta)1_{\{m<C\}} + \sum_{i'=1}^{d_{m-1}} \frac{\mu_{ii'}^{(m)}}{\sigma_i^{(m)}+\eta}\,\pi_{i'j}(m-1,m';\eta)\,1_{\{m>0\}} + \frac{\eta}{\sigma_i^{(m)}+\eta}\,1_{\{m=m',i=j\}}.$$

Multiplying both sides of the equation with $\sigma_i^{(m)}+\eta$ results in

$$(\sigma_i^{(m)}+\eta)\,\pi_{ij}(m,m';\eta) = \sum_{i'=1,i'\neq i}^{d_m} q_{ii'}^{(m)}\,\pi_{i'j}(m,m';\eta) + \sum_{i'=1}^{d_{m+1}} \lambda_{ii'}^{(m)}\,\pi_{i'j}(m+1,m';\eta)1_{\{m<C\}} + \sum_{i'=1}^{d_{m-1}} \mu_{ii'}^{(m)}\,\pi_{i'j}(m-1,m';\eta)\,1_{\{m>0\}} + \eta\,1_{\{m=m',i=j\}}. \tag{7}$$

The sum of the coefficients on the right equals $\sigma_i^{(m)}+\eta$, making this system of linear equations strictly diagonally dominant, and therefore non-singular (Horn and Johnson 2013, Thm 6.1.10). As a consequence, the system can be numerically solved in $\pi_{ij}(m,m';\eta)$ through various efficient evaluation techniques, such as the iterative Jacobi and Gauss-Seidel methods (Atkinson 1989, Section VIII.6).

The above linear system can be written in a compact matrix form. Define the $d_m \times d_{m'}$ matrix $\Pi_\eta(m,m')$ as the matrix whose (i,j)-th entry is $\pi_{ij}(m,m';\eta)$. In addition, let $\Sigma^{(m)} := \mathrm{diag}\{\sigma_1^{(m)},\ldots,\sigma_{d_m}^{(m)}\}$ and $\check{Q}^{(m)} := \mathrm{diag}\{q_1^{(m)},\ldots q_{d_m}^{(m)}\}$; the matrix $I^{(m)}$ is an identity matrix of dimension d_m. We thus obtain

$$(\Sigma^{(m)}+\eta I^{(m)})\Pi_\eta(m,m') = (Q^{(m)}+\check{Q}^{(m)})\Pi_\eta(m,m') + \Lambda^{(m)}\Pi_\eta(m+1,m')1_{\{m<C\}} + \mathcal{M}^{(m)}\Pi_\eta(m-1,m')1_{\{m>0\}} + \eta I^{(m)}1_{\{m=m'\}}.$$

We define Π_η as a $D\times D$ matrix, which is a block matrix of which the components are the matrices $\Pi_\eta(m,m')$:

$$\Pi_\eta := \begin{pmatrix} \Pi_\eta(0,0) & \Pi_\eta(0,1) & \Pi_\eta(0,2) & \cdots & \Pi_\eta(0,C) \\ \Pi_\eta(1,0) & \Pi_\eta(1,1) & \Pi_\eta(1,2) & \cdots & \Pi_\eta(1,C) \\ \Pi_\eta(2,0) & \Pi_\eta(2,1) & \Pi_\eta(2,2) & \cdots & \Pi_\eta(2,C) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Pi_\eta(C,0) & \Pi_\eta(C,1) & \Pi_\eta(C,2) & \cdots & \Pi_\eta(C,C) \end{pmatrix}. \tag{8}$$

Observe that in the linear equations (7) both the 'destination level' (namely, m') and the 'destination phase' are constant. This means that we can write the equations in Eq. 7 corresponding to a given phase j and level m', as a system of the form $A\boldsymbol{x}_{jm'} = \boldsymbol{b}_{jm'}$, where $\boldsymbol{b}_{jm'}$ is a known vector of dimension D, $\boldsymbol{x}_{jm'}$ is an unknown vector of dimension D, and A is known matrix of dimension $D\times D$. Importantly, the matrix A *does not depend on* j *and* m', and can be checked to equal $Q-\eta I^{(D)}$. As a consequence, with $\boldsymbol{x}_{jm'} = A^{-1}\boldsymbol{b}_{jm'}$, we have to compute (for this specific (j,m')-pair, that is) the matrix A^{-1} just once. In case the linear system is solved in the conventional way, this takes time $O(D^3)$. This means that, due to the elementary structure of $\boldsymbol{b}_{jm'}$ (containing one entry with value $-\eta$ and zeroes elsewhere), the computational effort of evaluating the full matrix Π_η is $O(D^3)$.

The above reasoning can be compactly rephrased differently as follows. It is readily verified from Eq. 7 that Π_η can be rewritten as $-\eta(Q-\eta I^{(D)})^{-1}$, and computing the inverse $(Q-\eta I^{(D)})^{-1}$ requires $O(D^3)$ time. This matrix Π_η will appear in the approximation of P_t based on Erlangization, introduced in the next section.

4 Erlangization

In this section, we discuss the approach based on Erlangization to approximate P_t. We first introduce the approximation and then provide the theoretical correctness of this approach. Let $S_{\ell,t}$ be an Erlang-distributed random variable with rate parameter ℓ/t and shape parameter ℓ. Let $P_t^{(\mathrm{e},\ell)}$ be a $D \times D$ matrix with entries

$$p_{ij}^{(\mathrm{e},\ell)}(m,m';t) := \mathbb{P}(M_{S_{\ell,t}} = m', X_{S_{\ell,t}} = j \mid M_0 = m, X_0 = i).$$

It is clear that $P_t^{(\mathrm{e},\ell)} = (\Pi_{\ell/t})^\ell$, with Π_η as defined in Eq. 8, owing to the fact that an Erlang random variable with rate parameter μ and shape parameter k can be written as the sum of k independent and identically distributed exponential random variables with rate μ. We propose the following approximation.

Approximation 3 (Erlangization) *For a given $\ell \in \mathbb{N}$, P_t is approximated by,*

$$P_t^{(\mathrm{e},\ell)} = (\Pi_{\ell/t})^\ell. \tag{9}$$

As we will argue below, $P_t^{(\mathrm{e},\ell)}$ converges to P_t as $\ell \to \infty$. The above idea is usually referred to as 'Erlangization': the time $t \geqslant 0$ is approximated by the Erlang time $S_{\ell,t}$. This distribution has mean t and variance t^2/ℓ, so that the corresponding coefficient of variation converges to 0 as $\ell \to \infty$.

Our goal is to assess how much $p_{ij}(m,m';t)$ differs from $p_{ij}^{(\mathrm{e},\ell)}(m,m';t)$. The resulting bounds are then used to show that this difference vanishes as ℓ grows large. We start by establishing an upper bound. For any given $\delta \in (0,t)$,

$$\begin{aligned} p_{ij}^{(\mathrm{e},\ell)}(m,m';t) &= \mathbb{P}(M_{S_{\ell,t}} = m', X_{S_{\ell,t}} = j \,\big|\, |S_{\ell,t}-t| \leqslant \delta, M_0 = m, X_0 = i)\mathbb{P}(|S_{\ell,t}-t| \leqslant \delta) \\ &\quad + \mathbb{P}(M_{S_{\ell,t}} = m', X_{S_{\ell,t}} = j \,\big|\, |S_{\ell,t}-t| > \delta, M_0 = m, X_0 = i)\mathbb{P}(|S_{\ell,t}-t| > \delta) \\ &\leqslant \mathbb{P}(M_{S_{\ell,t}} = m', X_{S_{\ell,t}} = j \,|\, |S_{\ell,t}-t| \leqslant \delta, M_0 = m, X_0 = i) + \mathbb{P}(|S_{\ell,t}-t| > \delta). \end{aligned}$$

Note that $\mathbb{P}(M_{S_{\ell,t}} = m', X_{S_{\ell,t}} = j \,|\, |S_{\ell,t}-t| \leqslant \delta, M_0 = m, X_0 = i)$ is equal to the transition probability $p_{ij}(m,m';S_{\ell,t})$ additionally imposing the condition that $|S_{\ell,t}-t| \leqslant \delta$. The difference between this probability and $p_{ij}(m,m';t)$ can thus be at most δ times the maximum slope of $p_{ij}(m,m';s)$ for s in $[t-\delta, t+\delta]$. Hence

$$p_{ij}^{(\mathrm{e},\ell)}(m,m';t) \leqslant p_{ij}(m,m';t) + \delta \left(\sup_{s\in[t-\delta,t+\delta]} \left| \frac{\mathrm{d}}{\mathrm{d}s} p_{ij}(m,m';s) \right| \right) + \mathbb{P}(|S_{\ell,t}-t| > \delta).$$

Recall that Q is the transition rate matrix of the D-dimensional continuous-time Markov process $\{M_t, X_t\}_{t\geqslant 0}$ and $\sigma := \max_{m,i} \sigma_i^{(m)}$. Then, using the Kolmogorov equations in combination with the triangle inequality, uniformly in $s \geqslant 0$,

$$\left| \frac{\mathrm{d}}{\mathrm{d}s} p_{ij}(m,m';s) \right| \leqslant \sum_{m'',j'} p_{ij'}(m,m'';s) \left| Q_{(m'',j'),(m',j)} \right| \leqslant \sum_{m'',j'} p_{ij'}(m,m'';s)\, \sigma = \sigma.$$

We proceed by finding an upper bound on $\mathbb{P}(|S_{\ell,t} - t| > \delta)$. Noting that $S_{\ell,t}$ can be written as ℓ^{-1} times an Erlang random variable $\bar{S}_{\ell,t}$ with rate parameter $1/t$ and shape parameter ℓ,

$$\mathbb{P}(|S_{\ell,t} - t| > \delta) = \mathbb{P}(|\ell^{-1}\bar{S}_{\ell,t} - t| > \delta) = \mathbb{P}(\ell^{-1}\bar{S}_{\ell,t} - t < -\delta) + \mathbb{P}(\ell^{-1}\bar{S}_{\ell,t} - t > \delta). \quad (10)$$

We can majorize both probabilities on the right-hand side by using the Chernoff bound. Starting with $\mathbb{P}(\ell^{-1}\bar{S}_{\ell,t} - t > \delta)$, we have

$$\mathbb{P}(\ell^{-1}\bar{S}_{\ell,t} - t > \delta) = \mathbb{P}(\bar{S}_{\ell,t} > \ell(\delta + t)) \leqslant \inf_{\theta>0} e^{-\theta\,\ell(\delta+t)}\mathbb{E}\,e^{\theta\bar{S}_{\ell,t}}.$$

Using the moment generating function of the Erlang distribution, we find that

$$e^{-\theta\,\ell(\delta+t)}\mathbb{E}\,e^{\theta\bar{S}_{\ell,t}} = \left(\frac{e^{-\theta(\delta+t)}}{1-t\theta}\right)^{\ell},$$

implying that

$$\mathbb{P}(\ell^{-1}\bar{S}_{\ell,t} - t > \delta) \leqslant \inf_{\theta>0}\left(\frac{e^{-\theta(\delta+t)}}{1-t\theta}\right)^{\ell} = \left(\inf_{\theta>0}\frac{e^{-\theta(\delta+t)}}{1-t\theta}\right)^{\ell} = e^{-\ell\delta/t}\left(1+\frac{\delta}{t}\right)^{\ell}.$$

In a similar way we can majorize $\mathbb{P}(\ell^{-1}\bar{S}_{\ell,t} - t < -\delta)$:

$$\mathbb{P}(\ell^{-1}\bar{S}_{\ell,t} - t < -\delta) \leqslant e^{\ell\delta/t}\left(1-\frac{\delta}{t}\right)^{\ell}.$$

Combining these upper bounds with equation (10), we conclude

$$\mathbb{P}(|S_{\ell,t} - t| > \delta) \leqslant e^{\ell\delta/t}\left(1-\frac{\delta}{t}\right)^{\ell} + e^{-\ell\delta/t}\left(1+\frac{\delta}{t}\right)^{\ell} =: \Psi_{\ell,t}(\delta). \quad (11)$$

We thus find, uniformly in $\delta \in (0, t)$,

$$p_{ij}^{(\mathrm{e},\ell)}(m, m'; t) \leqslant p_{ij}(m, m'; t) + \delta \cdot \sigma + \Psi_{\ell,t}(\delta).$$

Now take $\delta = \ell^{-\alpha}$ with $\alpha > 0$. Using elementary Taylor expansions, it can be shown that $\Psi_{\ell,t}(\delta)$ behaves as $\exp(-\ell^{1-2\alpha}/t^2)$, which converges to 0 as $\ell \to \infty$ for all $\alpha < 1/2$. To see this, first note that

$$e^{\ell\delta/t}\left(1-\frac{\delta}{t}\right)^{\ell} = \exp\left(\frac{\ell}{t}\delta + \ell\log\left(1-\frac{\delta}{t}\right)\right). \quad (12)$$

Now consider the exponential in the right-hand side of Eq. 12. Plugging in $\delta = \ell^{-\alpha}$ and using Taylor expansions, one indeed obtains

$$\frac{1}{t}\ell^{1-\alpha} + \ell\log\left(1-\frac{1}{t}\ell^{-\alpha}\right) = -\frac{1}{t^2}\ell^{1-2\alpha} + o(\ell^{1-3\alpha}).$$

A similar analysis can be performed for the other term in the definition of $\Psi_{\ell,t}(\delta)$. We conclude that, for all $\alpha < 1/2$, $\Psi_{\ell,t}(\ell^{-\alpha})$ converges to 0 as $\ell \to \infty$. Upon combining the above, and picking $\alpha = \frac{1}{3}$, the desired upper bound follows:

$$\limsup_{\ell\to\infty} p_{ij}^{(\mathrm{e},\ell)}(m, m'; t) \leqslant \limsup_{\ell\to\infty} p_{ij}(m, m'; t) + \ell^{-1/3}\cdot\sigma + \Psi_{\ell,t}(\ell^{-1/3}) = p_{ij}(m, m'; t).$$

We proceed by deriving a lower bound, which is established using elements that resemble those used in the upper bound. It is based on the inequality

$$\begin{aligned} p_{ij}^{(\mathrm{e},\ell)}(m,m';t) &\geqslant \mathbb{P}(M_{S_{\ell,t}} = m', X_{S_{\ell,t}} = j \mid M_0 = m, X_0 \\ &= i, |S_{\ell,t} - t| \leqslant \delta) \cdot \mathbb{P}(|S_{\ell,t} - t| \leqslant \delta) \\ &\geqslant \big(p_{ij}(m,m';t) - \delta \cdot \sigma\big) \cdot \big(1 - \mathbb{P}(|S_{\ell,t} - t| > \delta)\big) \\ &\geqslant p_{ij}(m,m';t) - \delta \cdot \sigma - \Psi_{\ell,t}(\delta). \end{aligned}$$

Pick again $\delta = \ell^{-1/3}$, so as to obtain

$$\liminf_{\ell\to\infty} p_{ij}^{(\mathrm{e},\ell)}(m,m';t) \geqslant p_{ij}(m,m';t).$$

The following theorem summarizes the above findings, thus justifying the use of the Erlangization procedure.

Theorem 1 *For any $\ell \in \mathbb{N}$, $t > 0$, and $\delta \in (0,t)$, with σ defined as in* Eq. 2 *and $\Psi_{\ell,t}(\delta)$ defined as in* Eq. 11*,*

$$\left| p_{ij}^{(\mathrm{e},\ell)}(m,m';t) - p_{ij}(m,m';t) \right| \leqslant \delta \cdot \sigma + \Psi_{\ell,t}(\delta). \tag{13}$$

In addition, for any $t > 0$,

$$\lim_{\ell\to\infty} p_{ij}^{(\mathrm{e},\ell)}(m,m';t) = p_{ij}(m,m';t).$$

Note that the advantage of Erlangization is that the number of matrix multiplications is low, in comparison with uniformization. More precisely, picking ℓ a power of two, one just needs to square $\Pi_{\ell/t}$ only $\log_2 \ell$ times. The disadvantage is that the computation of the matrix $\Pi_{\ell/t}$ requires the solution of a linear system of dimension D, as argued in Section 3.

In addition, we note that the maximum diagonal element (in absolute terms) σ appears in the error bound of Theorem 1. As a consequence, the upper bound in Eq. 13 tends to be rather generous for some (m,m') and (i,j) pairs when the diagonal elements are relatively non-uniform.

5 Performance Analysis of Erlangization

In this section we examine the performance of the Erlangization approximation of P_t, as given by Eq. 9. We compare it with the matrix exponential approach given by Eq. 4 as well as uniformization (5). We study the accuracy (i.e., error) and efficiency (i.e., computational time) of the Erlangization approximation. In the sequel we refer to the Erlangization approach by 'E', to the matrix exponential approach by 'M', and to the uniformization approach by 'U'.

In our performance analysis we focus on three QBD processes that are effectively the modulated counterparts of frequently used BD processes. In all three settings the modulating process (also referred to as environmental process) is of dimension 2, irrespectively of the level $m \in \{0,1,\ldots,C\}$. In other words, we have that $d_m = d = 2$, so that

$$Q^{(m)} = \begin{pmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{pmatrix}$$

In addition, we let $\lambda_{ij}^{(m)} = 0$ for $i \neq j$, which (informally) means that an increase in level cannot occur at the same time as a phase jump. The three settings are parameterized by a function $f(m, C)$, in the sense that

$$\lambda_i^{(m)} = \lambda_{ii}^{(m)} := f(m, C)\,\lambda_i,$$

for a known positive function $f(m, C)$ and parameter $\lambda_i \geqslant 0$. Similarly, we let $\mu_{ij}^{(m)} = 0$ for $i \neq j$, and define

$$\mu_i^{(m)} = \mu_{ii}^{(m)} := g(m, C)\,\mu_i,$$

for a known positive function $g(m, C)$ and parameter $\mu_i \geqslant 0$. Hence, there are at most six parameters in these models: $q_1, q_2, \lambda_1, \lambda_2, \mu_1$, and μ_2. We proceed by detailing the dynamics underlying the three models.

Experiment 1 (Infinite-server queue) *Here we consider a system, which can also been seen as a population process, in which individuals arrive according to some arrival process and are served in parallel, in the literature also known as an infinite-server queue* (Kleinrock 1975; Kulkarni 1995). *The special feature is that the Poissonian arrival rate as well as the exponential service rate depend on the state of the modulating process, so that the system at hand is a Markov-modulated infinite-server queue* (Anderson et al. 2016; Blom et al. 2016; 2017). *This concretely means that* $f(m, C) = 1$ *and* $g(m, C) = m$ *(the latter reflecting that the individuals are served in parallel), so that* $\Lambda^{(m)} = diag\{\lambda_1, \lambda_2\}$ *and* $\mathcal{M}^{(m)} = diag\{m\,\mu_1, m\,\mu_2\}$. *We impose a truncation at level* C.

Experiment 2 (Linear birth-death process) *In this setting we consider the stochastic version of the classical Malthusian growth model, also known as the linear birth-death model* (Davison et al. 2020; Karlin and Taylor 1975)*: the rate upward as well as the rate downward is proportional to the number of individuals present. This concretely means that* $f(m, C) = m$ *and* $g(m, C) = m$. *The rates of moving upward and downward are modulated, which entails that in this case* $\Lambda^{(m)} = diag\{m\,\lambda_1, m\,\lambda_2\}$ *and* $\mathcal{M}^{(m)} = diag\{m\,\mu_1, m\,\mu_2\}$. *We again impose a truncation at* C.

Experiment 3 (SIS-type model) *The SIR model is a so-called compartmental model used to describe epidemic growth, that keeps track of the number of susceptible individuals, the number of infectious individuals, and the number of recovered individuals; see e.g. the textbook treatments in* (Allen 2003; Andersson and Britton 2000; Daley and Gani 1999). *In a related variant, the SIS model, recovered individuals eventually become susceptible again. In this experiment we consider a model of the latter type, which, in the non-modulated context, has the following dynamics. There are* C *individuals, to be divided into infected and healthy. Let* M_t *be the number of healthy individuals. When* $M_t = m$, *an arbitrary healthy person becomes infected with rate* $\lambda(C - m)$*; as a result the rate from* m *to* $m + 1$ *is* $\lambda\, m(C - m)$. *Every infected person becomes healthy again independently of the state of all other individuals; as a result, the rate from* m *to* $m - 1$ *is* $m\,\mu$. *If we add modulation, then the* λ *and* μ *become dependent on the environmental process. We thus get that in this model* $f(m, C) = m(C - m)$ *and* $g(m, C) = m$, *so that the upward rates become* $\Lambda^{(m)} = diag\{m(C - m)\lambda_1, m(C - m)\lambda_2\}$, *whereas the downward rates are given by* $\mathcal{M}^{(m)} = diag\{m\,\mu_1, m\,\mu_2\}$.

We start, in Section 5.1, with an extensive analysis of Experiment 1, the infinite-server queue. In particular we study the impact of the parameters ℓ and C on the accuracy (i.e.,

error) and efficiency (i.e., computational time) of the Erlangization approximation, and compare these with the other two approaches. In Section 5.2 we consider Experiments 2 and 3.

Importantly, whenever presenting computational times, we report the time it takes to evaluate the entire matrix $P_t^{(\mathrm{e},\ell)}$ ($P_t^{(\mathrm{m})}$ and $P_t^{(\mathrm{u},\ell)}$ likewise), providing us with $p_{ij}^{(\mathrm{e},\ell)}(m, m'; t)$ for all $i, j \in \{1, 2\}$ and $m, m' = 0, \ldots, C$. Furthermore, we use MATLAB's implementation `timeit(·)` to evaluate the computational times. It is noted that, so as to obtain a reliable estimate, the function `timeit(·)` calls the specified function multiple times, measures the time required each time, and finally outputs the median of all these values.

5.1 Analysis of Experiment 1

We consider Experiment 1 with the parameter values $q_1 = 0.015$, $q_2 = 0.045$, $\lambda_1 = 2$, $\lambda_2 = 9$ and $\mu_1 = \mu_2 = 0.3$, and we let $C = 60$. Observe that in this instance the phase process modulates the arrival rate, but does not affect the service rate. We compute the transition probability $p_{ij}(m, m'; t)$, as defined in Eq. 3, using the three approaches that we discussed. As a representative illustration, we took $i = j = 1$, $m = 6$ and $t = 1$ for varying m', leading to the output that is presented in Table 1. Since the three approaches resulted in almost identical outcomes, Table 1 shows the outcomes for the matrix exponential approach and the absolute differences with the other two approaches. The last row displays the computational time (in seconds) corresponding to the approximation of P_t, which shows that Erlangization performs well compared with the alternative approaches. The values of ℓ for both Erlangization and uniformization are determined by increasing ℓ until the percentage difference between subsequent outcomes of $p_{11}(6, m'; t)$ was below $\varepsilon = 10^{-3}$ for all $m' = 0, \ldots, 15$. For the Erlangization approach, ℓ was doubled each step, and for the uniformization approach, ℓ was increased by one at a time. This resulted in $\ell = 8192 = 2^{13}$ and $\ell = 174$, respectively. The results in Table 1 indicate that, for these values of ℓ, the accuracy of the three approaches is similar.

Evidently, computational times increase in C. To compare the computational times of the Erlangization approach and the uniformization approach as fairly as possible, we apply the following procedure to determine the required values of ℓ. We use $P_t^{(\mathrm{m})}$ in Eq. 4 as benchmark, since the sophisticated implementation `expm(·)` that MATLAB is using is claimed to perform highly accurate and has been tested intensively. Then, for both Erlangization and uniformization, we increase ℓ until the percentage difference between the outcome of $p_{11}(6, 6; t)$ and the one in $P_t^{(\mathrm{m})}$ is below a chosen tolerance $\varepsilon > 0$. Table 2 shows, for various values of ε, the obtained values of ℓ (which is, by construction, a power of two for Erlangization). Evidently, a smaller error ε can be achieved by increasing ℓ. Importantly, the $\log_2 \ell$ values for Erlangization are considerably lower than the ℓ values for uniformization, which is indicative of Erlangization being the more efficient approach.

To investigate the impact of C on the computational time, we increase C from 50 to 500 in steps of 50, compute for each C the values of ℓ with $\varepsilon = 10^{-3}$ (in the way discussed above, that is), and then use these values of ℓ to evaluate the computational times. Table 3 shows the obtained values of ℓ for the increasing values of C, and Fig. 2 shows the corresponding computational times in seconds.

From Table 3 we observe that for Erlangization the value of ℓ is not influenced by C, but that for uniformization the value of ℓ increases roughly linearly in C. Furthermore, Fig. 2

Table 1 *Infinite-server queue*: $p_{ij}(m, m'; t)$ (and absolute differences) with $i = j = 1$, $m = 6$, $t = 1$ and $C = 60$

m'	M	|M - E| $\ell = 8192$	|M - U| $\ell = 174$
0	5.3057 e-05	2.5257 e-08	6.7763 e-20
1	1.0016 e-03	1.7085 e-07	2.8189 e-18
2	8.1536 e-03	4.3531 e-08	5.2042 e-18
3	3.7419 e-02	2.5787 e-06	7.6328 e-17
4	1.0645 e-01	5.1373 e-06	1.1102 e-16
5	1.9546 e-01	2.2531 e-06	2.2204 e-16
6	2.3703 e-01	1.1333 e-05	1.1102 e-16
7	1.9643 e-01	2.8468 e-06	1.6653 e-16
8	1.1861 e-01	4.6012 e-06	5.5511 e-17
9	5.5265 e-02	3.8581 e-06	1.3878 e-17
10	2.0774 e-02	1.1024 e-06	1.7347 e-17
11	6.5164 e-03	1.7150 e-07	7.8063 e-18
12	1.7528 e-03	3.1073 e-07	2.1684 e-18
13	4.1473 e-04	1.6793 e-07	5.4210 e-19
14	8.9208 e-05	6.3282 e-08	2.7105 e-20
15	1.8552 e-05	1.9978 e-08	6.7763 e-21
CPU times	2.39 e-02	2.16 e-02	1.95 e-01

CPU times corresponding to the approximation of P_t (for M, E and U respectively). Parameter values: $q_1 = 0.015$, $q_2 = 0.045$, $\lambda_1 = 2$, $\lambda_2 = 9$ and $\mu_1 = \mu_2 = 0.3$

Table 2 *Infinite-server queue*: The values of ℓ (and, for Erlangization, in addition $\log_2 \ell$ between brackets) for decreasing values of ε, with $C = 60$ and using $P_t^{(\mathrm{m})}$ as benchmark

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
E	8 (3)	64 (6)	512 (9)	4096 (12)	65536 (16)
U	33	39	44	48	51

Table 3 *Infinite-server queue*: The values of ℓ for the increasing values of C, with $\varepsilon = 10^{-3}$ and using $P_t^{(\mathrm{m})}$ as benchmark

C	50	100	150	200	250	300	350	400	450	500
E	512	512	512	512	512	512	512	512	512	512
U	40	59	77	95	113	130	148	165	182	199

Observe that $512 = 2^9$, so that Erlangization requires just 9 matrix multiplications

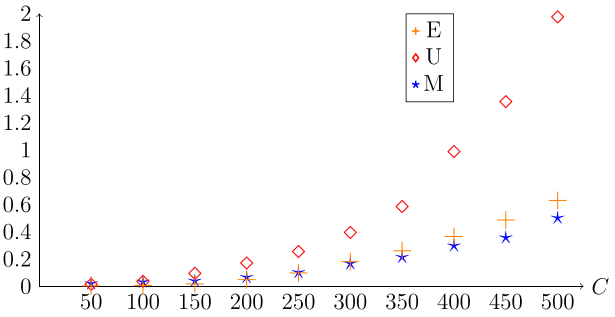

Fig. 2 *Infinite-server queue*: Computational times (in seconds) corresponding to the approximation of P_t with $t = 1$, for the three different methods. Values of ℓ as displayed in Table 3

reveals that the computational times for the matrix exponential method and Erlangization approach are essentially of the same order. For small values of C, the Erlangization method is (slightly) faster, whereas for higher values of C the matrix exponential method is (slightly) faster. The computational cost of the uniformization method, however, is significantly higher. The latter observation is in line with what we expected: as seen in Table 3, uniformization typically needs a relatively large number of matrix multiplications.

To systematically assess the impact of C on the computational time, which we denote by T, we fit the curve $T(C) = \alpha C^{\beta}$. This we do by applying least squares to $T(C) - \alpha C^{\beta}$, i.e., we determine

$$\min_{\alpha,\beta} \sum_{C\in\{50,\ldots,500\}} (T(C) - \alpha C^{\beta})^2.$$

We find that, as a function of C, the CPU time of both the matrix exponential method and Erlangization is superquadratic but subcubic ($\beta = 2.20$ and $\beta = 2.57$, respectively), whereas the CPU time of uniformization is essentially cubic ($\beta = 3.15$). Evidently, these β values serve as an indication only, because they are based on ten observations only.

5.2 Other Experiments

To explore if other settings yield similar results, we investigate the two other experiments as well. We consider Experiment 2 with parameter values $q_1 = 0.3$, $q_2 = 0.9$, $\lambda_1 = \lambda_2 = 0.19$, $\mu_1 = 0.16$, $\mu_2 = 0.08$ (i.e., the phase process does not affect the birth rate) and $C = 300$, and we consider Experiment 3 with parameter values $q_1 = 0.1$, $q_2 = 0.4$, $\lambda_1 = 0.0035$, $\lambda_2 = 0.01$, $\mu_1 = \mu_2 = 0.3$ (i.e., the phase process does not affect the recovery rate) and $C = 100$. We briefly present the results, focusing on the differences with the results of Experiment 1.

First, as in the previous section, we compute the values of ℓ for decreasing values of ε. As the counterparts to the results in Table 2 for Experiment 1, Tables 4 and 5 show the

Table 4 *Linear birth-death process*: The values of ℓ (and, for Erlangization, in addition $\log_2 \ell$ between brackets) for decreasing values of ε, with $C = 300$ and using $P_t^{(\mathrm{m})}$ as benchmark

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
E	8 (3)	128 (7)	1024 (10)	8192 (13)	65536 (16)
U	117	129	137	144	151

Table 5 *SIS-type model*: The values of ℓ (and, for Erlangization, in addition $\log_2 \ell$ between brackets) for decreasing values of ε, with $C = 100$ and using $P_t^{(\mathrm{m})}$ as benchmark

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
E	8 (3)	64 (6)	1024 (10)	8192 (13)	65536 (16)
U	50	58	64	68	73

obtained values of ℓ for Experiment 2 and Experiment 3, respectively. We see that the values of ℓ for Erlangization are similar across the three experiments. The $\log_2 \ell$ values differ at most by one, which corresponds to only one additional matrix multiplication. The results for uniformization, however, are drastically different across the three experiments. This effect could be explained by the fact that the maximum diagonal entry σ, that plays a crucial role in the uniformization approximation (5), depends highly on the functions $f(m, C)$ an $g(m, C)$ chosen.

Next, we examine again the impact of C on the computational time. As we did in Experiment 1, we increase C from 50 to 500 in steps of 50, compute for each C the values of ℓ with $\varepsilon = 10^{-3}$, and use these values of ℓ to evaluate the computational times. Table 6 and 7 show the obtained values of ℓ for the values of C that we considered. Comparing with Experiment 1, we need a slightly higher ℓ in Experiment 2 to obtain the same error $\varepsilon = 10^{-3}$. In Experiment 3, however, the ℓ should be increased considerably, but recall that for the Erlangization approach this only requires a few additional matrix multiplications.

Figure 3 shows for each specific approximation the computational times corresponding to the three experiments. The main conclusions from this figure are:

- observing each of the graphs individually, we see that for each of the three computational methods the three experiments roughly take the same amount of computational time (with the SIS-type model taking somewhat longer than the other two models under the uniformization approach, as will be explained in Remark 1 below);
- comparing the three graphs, we see that for uniformization the computational times are substantially higher, while the other two methods require roughly the same computational cost.

When fitting the curve $T = \alpha C^{\beta}$, we observe from Table 8 that the β-values obtained for the linear birth-death process and the SIS-type model align with those found for the infinite-server queue, in the sense that the matrix exponential method and Erlangization yield a β between 2 and 3, whereas uniformization yields a β larger than 3.

Remark 1 The fact that uniformization is slow for the SIS-type model can be understood as follows. The number of terms needed in Eq. 5*, which in turn determines the number of matrix multiplications to be performed, increases in σ, where we recall that σ denotes*

Table 6 *Linear birth-death process*: The values of ℓ for the increasing values of C, with $\varepsilon = 10^{-3}$ and using $P_t^{(\mathrm{m})}$ as benchmark

C	50	100	150	200	250	300	350	400	450	500
E	1024	1024	1024	1024	1024	1024	1024	1024	1024	1024
U	31	54	75	96	117	137	157	177	197	216

Table 7 *SIS-type model*: The values of ℓ for the increasing values of C, with $\varepsilon = 10^{-3}$ and using $P_t^{(\mathrm{m})}$ as benchmark

C	50	100	150	200	250	300	350	400	450	500
E	1024	1024	1024	1024	2048	4096	4096	8192	8192	8192
U	29	64	110	168	240	323	419	527	648	782

the (absolute value of) the largest diagonal entry of Q. For the infinite-server model and the linear birth-death model, this largest entry is of the order C. For the SIS-type model, however, recalling that $f(m, C) = m(C - m)$, the largest entry is of the order C^2. As a consequence, the number of terms in Eq. 5 *is relatively large, leading to a relatively long computational time.*

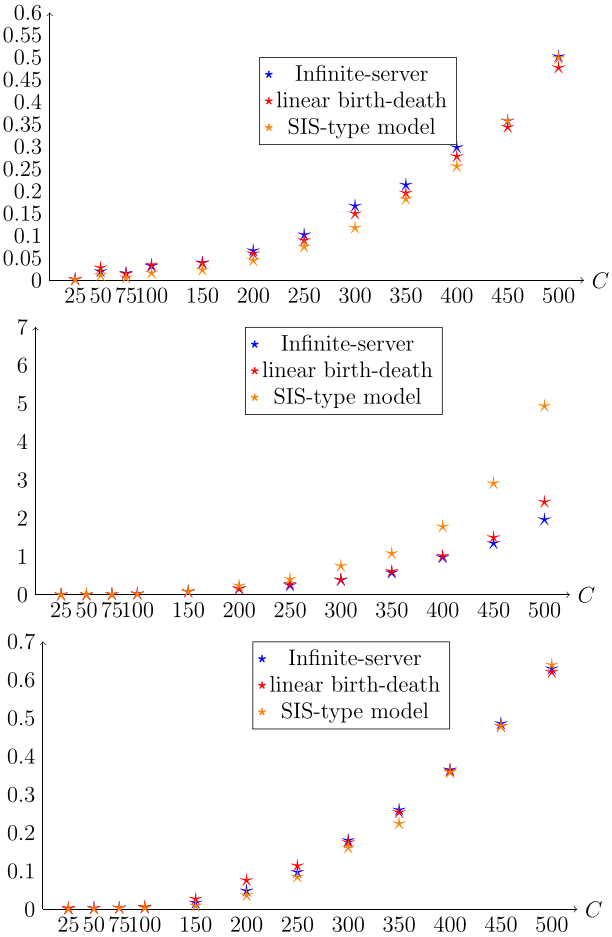

Fig. 3 Computational times (in seconds) corresponding to the approximation of P_t with $t = 1$, for the three experiments and the three different methods; from the top to bottom panel, matrix exponential method, uniformization method and Erlangization method, with values of ℓ as in Tables 3, 6 and 7

Table 8 β values for the different experiments and different approaches

Experiment	E	U	M
Infinite-server	2.57	3.15	2.20
linear birth-death	2.45	3.67	2.29
SIS-type model	2.83	4.20	2.79

6 Model Selection

We started our paper with a motivating example: can we statistically distinguish whether data stems from a QBD or from its non-modulated counterpart? We argued that to answer this question, we need machinery to evaluate the likelihood corresponding to a given time series. Now that we have at our disposal techniques to evaluate probabilities of the type (3), we return to our model selection problem of distinguishing between QBD processes and conventional (non-modulated, that is) BD processes. In this section we do so, using both simulated data and real-life data.

We wish to distinguish between the following four scenarios:

1. No modulation on neither the birth rate λ nor the death rate μ, i.e., $\theta = (\lambda, \mu)$
2. Modulation on the birth rate λ only ($\mu_1 = \mu_2$), i.e., $\theta = (q_1, q_2, \lambda_1, \lambda_2, \mu)$
3. Modulation on the death rate μ only ($\lambda_1 = \lambda_2$), i.e., $\theta = (q_1, q_2, \lambda, \mu_1, \mu_2)$
4. Modulation on both the birth rate λ and the death rate μ, i.e., $\theta = (q_1, q_2, \lambda_1, \lambda_2, \mu_1, \mu_2)$

We start by considering the setting of Experiment 1 with simulated data, and then use the model of Experiment 2 to analyze the whooping crane data featured in the introduction. We investigate which of these scenarios provides the best fit for the data, using the commonly used Akaike information criterion. This criterion includes a penalty that equals twice the number of estimated parameters (i.e., two times 2, 5, 5, and 6 in the above four scenarios), thus preventing overfitting from happening.

In all experiments below there is a time interval $\Delta > 0$ so that the observations correspond to measurements performed at times $0, \Delta, 2\Delta, \ldots, n\Delta$ for some $n \in \mathbb{N}$. We call these observations $m_0, \ldots, m_n$. With θ the vector of parameters, the likelihood is

$$\mathcal{L}(\theta \mid m_0, \ldots, m_n) = \mathbb{P}_\theta(M_0 = m_0, \ldots, M_{n\Delta} = m_n). \tag{14}$$

Regarding scenarios 2, 3, and 4, note that the modulating process is not observed. However, with $\boldsymbol{x} = (x_0, \ldots, x_n) \in \{1, 2\}^{n+1}$, we can rewrite Eq. 14 as

$$\begin{aligned} &\sum_{\boldsymbol{x} \in \{1,2\}^{n+1}} \mathbb{P}_\theta(M_0 = m_0, X_0 = x_0, \ldots, M_{n\Delta} = m_n, X_{n\Delta} = x_n) \\ &= \sum_{\boldsymbol{x} \in \{1,2\}^{n+1}} \prod_{i=1}^{n} p_{x_{i-1}, x_i}(m_{i-1}, m_i; \Delta), \end{aligned} \tag{15}$$

where it is noted that the probabilities in the last expression are of the type (3), and can be evaluated with the techniques discussed in this paper. Importantly, there is no need to enumerate all paths $\boldsymbol{x} \in \{1, 2\}^{n+1}$. Instead we can evaluate Eq. 15 efficiently by, abbreviating $p_{x_{i-1}, x_i}(m[i]) \equiv p_{x_{i-1}, x_i}(m_{i-1}, m_i; \Delta)$, evaluating the matrix product

$$\boldsymbol{\alpha} \begin{pmatrix} p_{11}(m[1]) & p_{12}(m[1]) \\ p_{21}(m[1]) & p_{22}(m[1]) \end{pmatrix} \begin{pmatrix} p_{11}(m[2]) & p_{12}(m[2]) \\ p_{21}(m[2]) & p_{22}(m[2]) \end{pmatrix} \cdots \begin{pmatrix} p_{11}(m[n]) & p_{12}(m[n]) \\ p_{21}(m[n]) & p_{22}(m[n]) \end{pmatrix} \mathbf{1}, \tag{16}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ is the distribution of X_0 and $\mathbf{1}$ is an all-ones vector. Note that the matrices in Eq. 16 appear as blocks in the matrix P_Δ. Maximization of the likelihood gives us the maximum likelihood estimate $\hat{\theta}$ for θ. As we will discuss below, this likelihood can be used in model selection problems. In the experiments below, all calculations involving probabilities of the type $p_{x_{i-1},x_i}(m[i])$ have been performed by the Erlangization approach.

6.1 Simulated Data

We consider the setting of Experiment 1. We simulate data ($n = 2000$) with parameter values $q_1 = 0.015, q_2 = 0.045, \lambda_1 = 0.2, \lambda_2 = 0.9, \mu = 0.03, \Delta = 1$ and $C = 50$. This means that the true model for this data is an infinite-server queue with modulation on λ only. Based on this simulated data, we perform the model selection based on the Akaike information criterion, i.e., using $\text{AIC} = 2N - 2\log L(\hat{\theta})$, with N the dimension of the parameter vector θ.

From Table 9 we observe that the AIC value is smallest for scenario 2, which agrees with the ground truth of the simulated data (i.e., it succeeds in finding the scenario with modulation on the parameter λ only). Interestingly, the number of observations has impact on the conclusions drawn. To illustrate this, see Table 10 showing the results using the first 101 data points of the dataset only (i.e., $n = 100$ instead of $n = 2000$). The AIC value is now minimized by scenario 1, the scenario without modulation, indicating that the dataset is too short to detect the modulation.

6.2 Whooping Crane Population

We proceed by considering the linear birth-death setting of Experiment 2 in relation to the four scenarios mentioned above. We use the whooping crane data (Davison et al. 2020; Stratton 2020), as displayed in Fig. 1, of annual counts of the female population of the whooping crane $n = 69$. From Fig. 1 we could suspect that a model with modulation could lead to a better fit than a model without modulation. We (conservatively) set $C = 200$. The outcomes of the model selection procedure are shown in Table 11. As it turns out, the AIC value is smallest for scenario 1, i.e., the setting corresponding with no modulation. This is in line with the results that one would obtain using the matrix exponential approach. More

Table 9 *Experiment 1, simulated data*: parameter estimates, loglikelihood value and AIC for the four different scenarios ($n = 2000$), with $\ell = 1024$ and $C = 50$

parameter	scenario			
	1.	2.	3.	4.
$\hat{q}_1$	n/a	0.0120	0.0953	0.0122
$\hat{q}_2$	n/a	0.0456	0.0357	0.0462
$\hat{\lambda}_1$ (or λ)	0.3373	0.2093	0.3374	0.2097
$\hat{\lambda}_2$	n/a	0.8904	n/a	0.8790
$\hat{\mu}_1$ (or μ)	0.0302	0.0312	0.0175	0.0314
$\hat{\mu}_2$	n/a	n/a	0.0361	0.0299
$\log L(\hat{\theta})$	−2370.1	−2306.5	−2368.2	−2306.4
AIC	4744.1	4622.9	4746.4	4624.8

True parameter values: $q_1 = 0.015, q_2 = 0.045, \lambda_1 = 0.2, \lambda_2 = 0.9, \mu = 0.03$ with $\Delta = 1$

Table 10 *Experiment 1, simulated data*: parameter estimates, loglikelihood value and AIC for the four different scenarios ($n = 100$), with $\ell = 1024$ and $C = 50$

parameter	scenario			
	1.	2.	3.	4.
$\hat{q}_1$	n/a	0.5265	0.5484	2.4 e-07
$\hat{q}_2$	n/a	0.5548	0.5715	0.7045
$\hat{\lambda}_1$ (or λ)	0.2351	0.2343	0.2351	0.2351
$\hat{\lambda}_2$	n/a	0.2360	n/a	0.5651
$\hat{\mu}_1$ (or μ)	0.0281	0.0281	0.0280	0.0281
$\hat{\mu}_2$	n/a	n/a	0.0282	0.0769
$\log L(\hat{\theta})$	-104.10	-104.10	-104.10	-104.10
AIC	212.20	218.20	218.20	220.20

True parameter values: $q_1 = 0.015$, $q_2 = 0.045$, $\lambda_1 = 0.2$, $\lambda_2 = 0.9$, $\mu = 0.03$ with $\Delta = 1$

specifically, all values of the loglikelihood and AIC coincide up to high level of precision. In line with the experiments performed in the previous section, the computational effort of both approaches is roughly similar. One should bear in mind, though, that the number of observations in this dataset is low, making the detection of modulation (involving 5 or 6 parameters) difficult. Additional literature on parameter estimation for linear birth-death models can be found in e.g. Chen and Hyrien (2011), Crawford et al. (2012), Crawford and Suchard (2012), Davison et al. (2020), and Xu et al. (2015).

7 Concluding Remarks

We have examined various approaches to compute the time-dependent distribution of QBD processes, with emphasis on the Erlangization approach. This approach has provable asymptotic correctness properties, and is, in terms of computational time, typically relatively fast. The latter property pays off in particular in settings where many time-dependent probabilities have to be evaluated. In this context, one could think of instances in which a function

Table 11 *Whooping crane data*: parameter estimates using Erlangization, loglikelihood value and AIC for the four different scenarios ($n = 69$), with $\ell = 256$ and $C = 200$

parameter	scenario			
	1.	2.	3.	4.
$\hat{q}_1$	n/a	0.9496	0.7931	0.9479
$\hat{q}_2$	n/a	0.1941	0.5123	0.1575
$\hat{\lambda}_1$	0.1928	0.1592	0.1789	0.1200
$\hat{\lambda}_2$	n/a	0.1964	n/a	0.1891
$\hat{\mu}_1$	0.1492	0.1465	0.0971	7.9 e-07
$\hat{\mu}_2$	n/a	n/a	0.1604	0.1576
$\log L(\hat{\theta})$	-179.67	-179.66	-179.58	-179.41
AIC	363.34	369.33	369.17	370.82

of the time-dependent probabilities is to be optimized over a set of model parameters, e.g. when performing maximum likelihood estimation.

Our study was motivated by model selection problems, in which one wishes to distinguish between models with and without modulation, i.e., between QBD processes and their BD counterparts. Through a series of experiments, with simulated as well as real-life data, we have shown how the techniques for computing time-dependent distributions can play a role in this context.

Our Erlangization approach gives rise to various directions for further research. For the class of QBD processes, the method's first step (solving the system of linear equations that yield the probabilities at exponential epochs) can exploit the convenient underlying structure, thus allowing an efficient numerical algorithm. We anticipate, however, that Erlangization has the potential to be applied more widely. One could think of multi-type population models, where various types of individuals are considered, which can in turn interact with each other. Another interesting extension concerns the multivariate model in which a population of individuals lives on a network and can move between its nodes. In this respect we refer to our recent paper (de Gunst et al. 2021), approximating time-dependent probabilities in such a network, relying on saddlepoint approximations. The crucial simplification made in de Gunst et al. (2021) is that a discrete-time model is considered, as opposed to the continuous-time model featuring in the present paper. It would therefore be interesting to explore whether an Erlangization-based approach could be developed for the continuous-time setting of such a network population process.

Acknowledgements MM was supported by the NWO Gravitation program NETWORKS, grant 024002003. We thank M. de Gunst (Vrije Universiteit, Amsterdam) and S. Hautphenne (University of Melbourne) for their helpful comments and suggestions.

Data Availability The simulated datasets generated in the context of the present study can be obtained from the corresponding author upon request.

References

Al-Mohy A, Higham N (2009) A new scaling and squaring algorithm for the matrix exponential. SIAM J Matrix Anal Appl 31:970–989

Allen L (2003) An introduction to stochastic processes with applications to biology. Prentice-Hall, Upper Saddle River

Anderson D, Blom J, Mandjes M, Thorsdottir H, de Turck K (2016) A functional central limit theorem for a Markov-modulated infinite-server queue. Methodol Comput Appl Probab 18:153–168

Andersson H, Britton T (2000) Stochastic epidemic models and their statistical analysis. Lecture Notes in Statistics, vol 151. Springer, New York

Asmussen S, Avram F, Usabel M (2002) The Erlang approximation of finite time ruin probabilities. ASTIN Bulletin 32:267–281

Atkinson K (1989) An introduction to numerical analysis, 2nd edn. Wiley, Chichester

Blom J, de Turck K, Mandjes M (2016) Functional central limit theorems for Markov-modulated infinite-server systems. Mathematical Methods of Operations Research 83:351–372
Blom J, de Turck K, Mandjes M (2017) Refined large deviations asymptotics for Markov-modulated infinite-server systems. Eur J Oper Res 259:1036–1044
Bright L, Taylor P (1995) Calculating the equilibrium distribution in level dependent Quasi-Birth-and-Death processes. Stoch Model 11:497–526
Chen R, Hyrien O (2011) Quasi-and pseudo-maximum likelihood estimators for discretely observed continuous-time Markov branching processes. J Stat Plan Inference 141:2209–2227
Crawford F, Minin V, Suchard M (2012) Estimation for general birth-death processes. J Am Stat Assoc 109:730–747
Crawford F, Suchard M (2012) Transition probabilities for general birth-death processes with applications in ecology, genetics, and evolution. J Math Biol 65:553–580
Daley D, Gani J (1999) Epidemic modelling: an introduction. Cambridge studies in mathematical biology, vol 15. Cambridge University Press, Cambridge
Davison A, Hautphenne S, Kraus A (2020) Parameter estimation for discretely observed linear birth-and-death processes. Biometrics. Published online. https://doi.org/10.1111/biom.13282
de Gunst M, Hautphenne S, Mandjes M, Sollie B (2021) Parameter estimation for multivariate population processes: A saddlepoint approach. Stoch Model 37:168–196
Grassmann W (1991) Finding transient solutions in Markovian event systems through randomization. Numerical Solution of Markov Chains 8:37–61
Gross D, Miller D (1984) The randomization technique as a modeling tool and solution procedure for transient Markov processes. Oper Res 32:343–361
Higham N (2005) The scaling and squaring method for the matrix exponential revisited. SIAM J Matrix Anal Appl 26:1179–1193
Horn RA, Johnson CR (2013) Matrix analysis, 2nd edn. Cambridge University Press, Cambridge
Jensen A (1953) Markoff chains as an aid in the study of Markoff processes. Scand Actuar J: 87–91
Karlin S, Taylor H (1975) A first course in stochastic processes. Academic Press, New York
Kleinrock L (1975) Queueing systems, volume 1: Theory. Wiley, Chichester
Kulkarni V (1995) Modeling and analysis of stochastic systems, 1st edn. Chapman & Hall, London
Mandjes M, Taylor P (2016) The running maximum of a level-dependent quasi birth-death process. Probability in the Engineering and Informational Sciences 30:212–223
Melamed B, Yadin M (1984) Randomization procedures in the computation of cumulative-time distributions over discrete state Markov processes. Oper Res 32:926–944
Moler C, Van Loan C (2003) Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev 45:3–49
Ramaswami V, Taylor P (1996) Some properties of the rate matrices in level dependent Quasi-Birth-and-Death processes with a countable number of phases. Stoch Model 12:143–164
Ramaswami V, Woolford D, Stanford D (2008) The Erlangization method for Markovian fluid flows. Ann Oper Res 160:215–225
Reibman A, Trivedi K (1988) Numerical transient analysis of Markov models. Comput Oper Res 15:19–36
Stratton DA (2020) Case studies in ecology and evolution. Book in progress. University of Vermont. http://www.uvm.edu/dstratto/bcor102/
van Dijk N, van Brummelen S, Boucherie R (2018) Uniformization: basics, extensions and applications. Perform Eval 118:8–32
Xu J, Guttorp P, Kato-Maeda M, Minin VN (2015) Likelihood-based inference for discretely observed birth-death-shift processes, with applications to evolution of mobile genetic elements. Biometrics 71:1009–1021

Affiliations

Michel Mandjes[1,2,3] · Birgit Sollie[4]

[1] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

[2] Eurandom, Eindhoven University of Technology, Eindhoven, The Netherlands

[3] Amsterdam Business School, Faculty of Economics and Business, University of Amsterdam, Amsterdam, The Netherlands

[4] Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands