International Journal of Stochastic Analysis

# Applications of Stochastic Processes in Biology and Medicine

Guest Editors: Charles J. Mode, Rick Durrett, Fima Klebaner, and Peter Olofsson

# Applications of Stochastic Processes in Biology and Medicine

International Journal of Stochastic Analysis

# Applications of Stochastic Processes in Biology and Medicine

Guest Editors: Charles J. Mode, Rick Durrett, Fima Klebaner, and Peter Olofsson

Copyright © 2013 Hindawi Publishing Corporation. All rights reserved.

This is a special issue published in "International Journal of Stochastic Analysis." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

# Editorial Board


Lev Abolnikov, USA
Ravi P. Agarwal, Saudi Arabia
Ana M. Aguilera, Spain
E. J. Allen, USA
Bradford D. Allen, USA
Vo V. Anh, Australia
Jürgen Becker, Germany
Michel Benaim, Chile
Fred Espen Benth, Norway
Zdzisław Brzeźniak, UK
Manuel O. Cáceres, Argentina
René Carmona, USA
Mou Hsiung Chang, USA
Pao Liu Chow, USA
Ana Bela Cruzeiro, Portugal
Hans Daduna, Germany
Lokenath Debnath, USA
Jewgeni Dshalalow, USA
Kambiz Farahmand, UK
Hélyette Geman, France
Stephen Gilmore, UK
Qi-Ming He, Canada
Onesimo H. Lerma, Mexico
Bei L. Hu, USA
Yaozhong Hu, USA
Yuri G. Kondratiev, Germany
A. M. Krasnosel'skii, Russia
Ho Woo Lee, Korea
Nikolai N. Leonenko, UK
Wenbo V. Li, USA
Deli Li, Canada
Shoumei Li, China
Josefa Linares-Perez, Spain
R. S. Liptser, Israel
Kai Liu, UK
M. Jesus Lopez-Herrero, Spain
Jin Ma, USA
A. A. Martynyuk, Ukraine
B. Maslowski, Czech Republic
S. Mohammed, USA
Hideo Nagai, Japan
Marcel F. Neuts, USA
Donal O'Regan, Ireland
Enzo Orsingher, Italy
Goran Peskir, UK
Georg C. Pflug, Austria
Huyên Pham, France
Yongcheng Qi, USA
John M. Rassias, Greece
Andrew Rosalsky, USA
B. Schmalfuss, Germany
Henri Schurz, USA
Bruno Sericola, France
Karl Sigman, USA
Anatoly V. Skorohod, USA
H. M. Srivastava, Canada
Lukasz Stettner, Poland
Jordan Stoyanov, UK
Agnès Sulem, France
Denis Talay, France
Ciprian A. Tudor, France
S. A. Utev, UK
Aihua Xia, Australia
Hong K. Xu, Taiwan
Jiongmin Yong, USA
Qing Zhang, USA
Tusheng Zhang, UK
Huaizhong Zhao, UK

# Contents

**Applications of Stochastic Processes in Biology and Medicine**, Charles J. Mode, Rick Durrett, Fima Klebaner, and Peter Olofsson
Volume 2013, Article ID 790625, 2 pages

**Risk of Infectious Disease Outbreaks by Imported Cases with Application to the European Football Championship 2012**, Attila Dénes, Péter Kevei, Hiroshi Nishiura, and Gergely Röst
Volume 2013, Article ID 576381, 9 pages

**Modeling Neutral Evolution Using an Infinite-Allele Markov Branching Process**, Xiaowei Wu and Marek Kimmel
Volume 2013, Article ID 963831, 10 pages

**Simulating the Emergence of Mutations and Their Subsequent Evolution in an Age-Structured Stochastic Self-Regulating Process with Two Sexes**, Charles J. Mode, Candace K. Sleeman, and Towfique Raj
Volume 2013, Article ID 826321, 23 pages

**A Decomposable Branching Process in a Markovian Environment**, Vladimir Vatutin, Elena Dyakonova, Peter Jagers, and Serik Sagitov
Volume 2012, Article ID 694285, 24 pages

**Birth and Death Processes with Neutral Mutations**, Nicolas Champagnat, Amaury Lambert, and Mathieu Richard
Volume 2012, Article ID 569081, 20 pages

**Stochastic Methodology for the Study of an Epidemic Decay Phase, Based on a Branching Model**, Sophie Pénisson and Christine Jacob
Volume 2012, Article ID 598701, 32 pages

Hindawi Publishing Corporation
International Journal of Stochastic Analysis
Volume 2013, Article ID 790625, 2 pages
http://dx.doi.org/10.1155/2013/790625

*Editorial*

# Applications of Stochastic Processes in Biology and Medicine

**Charles J. Mode,[1] Rick Durrett,[2] Fima Klebaner,[3] and Peter Olofsson[4]**

[1] *Department of Mathematics, Drexel University, Philadelphia, PA 19104, USA*
[2] *Department of Mathematics, Duke University, Durham, NC, USA*
[3] *School of Mathematical Sciences, Monash University, Melbourne, VIC, Australia*
[4] *Department of Mathematics, Trinity University, San Antonio, TX, USA*

Correspondence should be addressed to Charles J. Mode; cjmode@comcast.net

Received 27 February 2013; Accepted 27 February 2013

Copyright © 2013 Charles J. Mode et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Even though the title of this special edition suggests that a wide class of stochastic processes may be applicable in biology and medicine, all the six papers that were accepted for publication in this edition were based, either directly or indirectly, on a class of stochastic processes called branching processes. Within all the literature on probability, statistics and stochastic processes, papers, and books on branching processes occupy a relatively small but significant niche, even though their history spans and period of more than one hundred years. In particular, the papers that constitute this edition may be classified into three types; namely, models of epidemics of infectious diseases, genetics and evolution.

In a paper by N. Champagnat et al. on birth and death process with neutral mutation, classes of branching processes are used to study mutations characterized in terms of splitting trees. It is interesting to note that mutations were not described in terms of nucleotide substitutions that are widely used in the classification of human haplotypes based on mitochondrial DNA. Like the analysis used in papers and books on branching processes, limit theorems played a basic role in the paper and the time scale of evolution was evolutionary time rather than generations.

In a paper by A. Denes et al. Galton-Watson multitype branching processes in discrete time with immigration were used to model risks of infectious disease by imported cases with applications to the European Football Championships of 2012. Among the many tools used by these authors in their analysis of the model was probability generating functions.

In the paper of S. C. Penisson and Jacob, the decay phase of the model for of an infectious disease caused by some infectious agent was approximated by a branching process. Limit theorems again played a basic role in the analysis of the model. There is also an interesting application of the model to BSE, mad cow disease, epidemic in Great Britain.

A paper by V. Vatutin et al. is devoted to decomposable branching processes for the two-type case in, among other things, random environments. In particular, it is assumed that the population has two types. One type lives on one island with variable environments, and the other type lives on another island with a constant environment. Moreover, it is assumed that migration occurs only from a random environment to a constant environment. The main focus of the paper is to apply limit theorems to study probabilities of extinction in the critical and subcritical cases.

In a paper by C. J. Mode et al. the focus of attention is to formulate and partially analyze a model of the emergence of mutations and their subsequent evolution in an age-structured self-regulating stochastic process with two sexes. From the genetic point of view, only one autosomal locus with two alleles is considered. The working paradigm of the paper differs from that of other papers in the edition in that it is based on computer intensive methods, and, in particular, Monte Carlo simulation methods. The process is also self-regulating in the sense that the subpopulation of each genotype will attain a population size determined by the carrying capacity of the environment for that genotype. Another feature of this model is that, in contrast to most models of classical branching processes, a set of nonlinear deterministic equations is embedded in the stochastic process and the analysis of this deterministic system is part of the analysis of the model as a whole in computer simulation experiments. Interestingly, at some points in the parameter

space of the process, the embedded deterministic model will exhibit chaotic behavior, but examples with chaos were not considered in the paper.

Lastly in a paper by X. Wu and M. Kimmel, neutral evolution is modeled as an infinite allele Markov branching processes. The main focus of attention in this paper is the proportion of alleles having a given number of copies at some point in time, which is referred to as the frequency spectrum by the authors. A formula for the variance of the frequency spectrum is also derived by the authors that is useful in interval estimation and tests of hypotheses for parameter values. It is also shown that the mean of frequency spectrum has an explicit form expressed in terms of a hyper-geometric function. A formula for the asymptotic rate of convergence is also derived. A technique used in the derivation of the formulas involves probability generating functions. Simulations are also used to illustrate some of the results that were derived from the birth and death process.

*Charles J. Mode*
*Rick Durrett*
*Fima Klebaner*
*Peter Olofsson*

Hindawi Publishing Corporation
International Journal of Stochastic Analysis
Volume 2013, Article ID 576381, 9 pages
http://dx.doi.org/10.1155/2013/576381

*Research Article*

# Risk of Infectious Disease Outbreaks by Imported Cases with Application to the European Football Championship 2012

**Attila Dénes,[1] Péter Kevei,[2] Hiroshi Nishiura,[3,4] and Gergely Röst[1]**

[1] *Bolyai Institute, University of Szeged, Aradi vértanúk tere 1, Szeged 6720, Hungary*
[2] *MTA-SZTE Analysis and Stochastics Research Group, Bolyai Institute, Aradi vértanúk tere 1, Szeged 6720, Hungary*
[3] *School of Public Health, The University of Hong Kong, 100 Cyberport Road, Hong Kong*
[4] *PRESTO, Japan Science and Technology Agency, Saitama, Japan*

Correspondence should be addressed to Attila Dénes; denesa@math.u-szeged.hu

Received 28 September 2012; Accepted 27 November 2012

Academic Editor: Charles J. Mode

Copyright © 2013 Attila Dénes et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The European Centre for Disease Prevention and Control called the attention in March 2012 to the risk of measles in Ukraine among visitors to the 2012 UEFA European Football Championship. Large populations of supporters travelled to various locations in Poland and Ukraine, depending on the schedule of Euro 2012 and the outcome of the games, possibly carrying the disease from one location to another. In the present study, we propose a novel two-phase multitype branching process model with immigration to describe the risk of a major epidemic in connection with large-scale sports-related mass gathering events. By analytic means, we calculate the expected number and the variance of imported cases and the probability of a major epidemic caused by the imported cases in their home country. Applying our model to the case study of Euro 2012 we demonstrate that the results of the football games can be highly influential to the risk of measles outbreaks in the home countries of supporters. To prevent imported epidemics, it should be emphasized that vaccinating travellers would most efficiently reduce the risk of epidemic, while requiring the minimum doses of vaccines as compared to other vaccination strategies. Our theoretical framework can be applied to other future sport tournaments too.

## 1. Introduction

The European Centre for Disease Prevention and Control reported a measles outbreak in Ukraine with more than 11,000 cases from the beginning of 2012 until the end of June 2012 [1, 2]. The 2012 UEFA European Championship (Euro 2012) took place in Ukraine and Poland between 8 June and 1 July 2012, attracting several hundreds of thousands of football fans to these countries [3]. Susceptible visitors not only had a high risk of being infected, but also geographically propagating the epidemic to other countries.

We introduce a discrete time Markov chain model, which is an adaptation of a multitype Galton-Watson process with immigration to give a mathematical model for the evolution of the epidemic. Thus, we calculate the risk of epidemics connected to sports-related mass gathering events. Our model consists of two parts, the first one describing the spread of the disease during the championship in the host country, while the second part models the spread of the disease by fans returning to their home countries.

We apply our model to the special case of measles epidemics in Ukraine during the Euro 2012. Four of the eight host cities of this championship are in Ukraine (Kiev, Kharkiv, Lviv and Donetsk); one of these, Lviv, is situated in the western region where the prevalence is the highest and vaccination coverage remained the lowest in the country. Games of the group phase took place in the four Ukrainian cities for groups B and D including Denmark, Germany, Netherlands, Portugal and Ukraine, England, France, Sweden [4]. Two of the quarterfinals, one of the semifinals as well as the final took place in Ukraine, so Spain and Italy also played some games in Ukraine. The suboptimal measles vaccination coverage in many European countries poses a risk of measles epidemics caused by fans returning from Euro 2012. Here we study the impact of different outcomes of Euro 2012 on the probability of post-tournament measles epidemics in

the participating countries and compare the effectiveness of different vaccination strategies by target host in reducing the risk of imported epidemics in other countries after Euro 2012. We discuss the applicability of our approach to other future events as well.

The rest of the paper is organized as follows. In Section 2 we describe the general mathematical model. In Section 3 we compute the probability of major epidemic in France after Euro 2012, while in Section 4 we compare the results with Euro 2008. Finally, we close with a discussion on the applicability of our model for other sports-related mass gathering events. In the appendix we calculate explicitly the expected number and the variance of infectious cases imported to the home country by supporters.

## 2. Methods

Since the supporter group spends a relatively short time in the infected area, it is possible that nobody gets infected, in which case there is no increased chance for epidemic in the home country. It is also clear that the risk of a huge epidemic is larger when five infected individuals arrive home (maybe to different parts of the country) than in the case when only one infectious supporter arrives. The fact that the number of infected supporters is zero, one, or five is just a matter of chance; thus, a deterministic model does not serve for our purposes in this case. It is well known (see [5]) that early stages of an epidemic in a large population can be approximated by branching processes, where having a descendant means infecting somebody. This exactly fits to our model, because in the host country the supporters spend short time (up to a month, say), and after returning to the home country we are only interested in the probability of a major epidemic, that is, in the early stage of a possible epidemic. To determine the final number of infectious individuals and/or the duration of the epidemic a mixture of a stochastic and a deterministic model is more appropriate. For general use of stochastic epidemic models we refer to a recent survey by Britton [5].

To describe the importation dynamics in the simplest manner, as a mathematical model, we propose a branching process with immigration. For simplicity, consider a single supporter population $S$ from a country $F$ which follows the matches of the team during the tournament, and ignore the interaction with other supporter groups. They can contract the disease from the local population or from each other. We define a discrete time Markov chain model, which is an adaptation of a multitype Galton-Watson process with immigration. We say that an individual is of type-$j$ if he/she contracted the infection exactly $j$ days ago. The model is divided into two phases: the first phase takes $T$ days and corresponds to the time spent in the host country, while the second phase describes the process upon returning to the home country. Let $m$ be the mean latent period and $k$ the mean infectious period of the disease (in days); that is, a newly infected individual becomes infectious only after $m$ days, and remains infectious for additional $k$ days. We assume these as constants. Denote by $\mathbf{X}_t$ the integer vector of infected individuals in population $S$ on day $t \le T$ where $X_t(j)$, the number of type-$j$ individuals, is the number of infected individuals in population $S$ who got infected $j$ days ago, $j = 1, 2, \ldots, m+k-1$. The evolution is the following. On day $t+1$ the newly infected individuals, that is, type-1 individuals, can originate from the local population (immigrants), or from an $S$-individual who is infectious on day $t+1$ (which means that he/she got infected at least $m+1$ and at most $m+k$ days ago) and thus is of type $\ge m$ on day $t$ (offsprings). We assume that the force of infection from the local population to $S$ is constant during the first phase, and that the daily incidence produced by an infectious member of $S$ is also constant. Finally, for $j > 1$, type-$j$ individuals arise only by getting one day older. After Phase 1, the infected vector $\mathbf{X}_T$ returns to the home country, and each infected individual independently starts a simple single type Galton-Watson process.

In the following we describe the exact mathematical model.

*2.1. Phase 1.* Let $\mathbf{X}_t = (X_t(1), \ldots, X_t(d))$ be a multitype Galton-Watson process with immigration, defined by

$$\mathbf{X}_t = \sum_{k=1}^{X_{t-1}(1)} \boldsymbol{\xi}_{t,k,1} + \cdots + \sum_{k=1}^{X_{t-1}(d)} \boldsymbol{\xi}_{t,k,d} + \boldsymbol{\varepsilon}_t, \quad t \in \{1, 2, \ldots, T\},$$
$$\mathbf{X}_0 = \mathbf{0}, \tag{1}$$

where $\{\boldsymbol{\xi}_{t,k,i}, \boldsymbol{\varepsilon}_t : t \in \{1, 2, \ldots, T\}, k \in \mathbb{N}, i \in \{1, 2, \ldots, d\}\}$ are independent random vectors with nonnegative integer coordinates, such that $\{\boldsymbol{\xi}_{t,k,i} : t \in \{1, 2, \ldots, T\}, k \in \mathbb{N}\}$ are identically distributed and $\{\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_t : t \in \{1, 2, \ldots, T\}\}$ are also identically distributed. Here the offsprings $\boldsymbol{\xi}$'s correspond to the new infections originated from an $S$-individual, while the immigrants $\boldsymbol{\varepsilon}$'s correspond to new infections originated from the local population. Introduce the generating functions:

$$F_t(\mathbf{z}) = \mathbf{E}\mathbf{z}^{\mathbf{X}_t}, \qquad G_i(\mathbf{z}) = \mathbf{E}\mathbf{z}^{\boldsymbol{\xi}_{1,1,i}},$$
$$\mathbf{G}(\mathbf{z}) = \left(G_1(\mathbf{z}), \ldots, G_d(\mathbf{z})\right), \qquad H(\mathbf{z}) = \mathbf{E}\mathbf{z}^{\boldsymbol{\varepsilon}}, \tag{2}$$

where $\mathbf{x}^{\mathbf{k}} = x_1^{k_1} \cdots x_d^{k_d}$. In the following boldface symbols $\mathbf{x}, \mathbf{y}, \mathbf{z}$ stand for $d$-dimensional vectors.

It is easy to show that the recursion $F_t(\mathbf{z}) = F_{t-1}(\mathbf{G}(\mathbf{z}))H(\mathbf{z})$ holds. Let $\mathbf{G}^k$ denote the $k$-fold iteration of $\mathbf{G}$, that is, $\mathbf{G}^0(\mathbf{z}) = \mathbf{z}$ and $\mathbf{G}^{t+1} = \mathbf{G}^t \circ \mathbf{G}$. Then an induction argument shows (see Quine [6]) that the generating function of the $t$th generation is

$$F_t(\mathbf{z}) = \prod_{k=0}^{t-1} H\left(\mathbf{G}^k(\mathbf{z})\right). \tag{3}$$

Up to now we did not use any particular property of the branching structure. However, note that in our case we have the following. The immigrants are always of type-1; thus, the generating function is in fact a one-variable function, that is, $H(\mathbf{z}) = H(z_1, \ldots, z_d) = H(z_1)$. We also have $d = m + k - 1$. For $j = 1, 2, \ldots, m-1$ every type-$j$ particle has

exactly one descendant of type-$(j+1)$ (the individual already infected is still not infectious, only getting one day older), thus $G_j(\mathbf{z}) = z_{j+1}$, while for $j = m, m+1, \ldots, m+k-1$ the type-$j$ individuals are already infecting and also getting one day older, so $G_j(\mathbf{z}) = z_{j+1}G_S(z_1)$, with $G_S(z)$ being the generating function of the infected individuals on one day by a single infectious individual in $S$. Without the vector notation we have

$$X_t(1) = \sum_{i=m}^{m+k-1} \sum_{j=1}^{X_{t-1}(i)} \xi_{t,i,j} + \varepsilon_t,$$
$$X_t(j) = X_{t-1}(j-1), \quad j = 2, \ldots, m+k-1, \tag{4}$$

where $\{\xi_{i,j,t} : i = m, \ldots, m+k-1, j \in \mathbb{N}, t = 1, 2, \ldots, T\}$ are iid random variables with generating function $G_S$.

*2.2. Phase 2.* Phase 2 starts with the infected vector $\mathbf{X}_T$ arriving home. In this stage there is no immigration, and since the infected individuals stay home there is no point on registering the different types; hence, instead of counting the days we count the generation: $Y_0$ is the number of individuals who are infected by $\mathbf{X}_T$, $Y_1$ is the number of individuals who are infected by $Y_0$, and so forth. That is, the process now can be described by a single type Galton-Watson process. However, the first step is different, because the different types have different meanings. Individuals of type-$j$, $j \le m$, spend all their infectious days in the home country, while individuals of type-$(m+j)$, $j = 1, 2, \ldots, k-1$, spend only $k-j$ infectious days in the home country. Let $G_F(z)$ denote the generating function of the infected individuals on one day by a single infectious individual in the home country. Let $Y_0$ be the number of individuals who were infected by $\mathbf{X}_T$, then we have

$$Y_0 = \sum_{i=1}^{m} \sum_{j=1}^{X_T(i)} \xi_{i,j} + \sum_{i=m+1}^{m+k-1} \sum_{j=1}^{X_T(i)} \xi_{i,j}, \tag{5}$$

where $\{\xi_{i,j} : i = 1, 2, \ldots, m+k-1, j \in \mathbb{N}\}$ are independent random variables, and $\{\xi_{i,j} : i = 1, 2, \ldots, m, j \in \mathbb{N}\}$ are iid with generating function $G_F^k$, and for $i \in \{m+1, \ldots, m+k-1\}$, $\{\xi_{i,j} : j \in \mathbb{N}\}$ are iid with generating function $G_F^{m+k-i}$. Using the representation above for the generating function of $Y_0$ we obtain

$$h(z) := \mathbf{E}z^{Y_0} = F_T\left(G_F^k(z), \ldots, G_F^k(z), G_F^{k-1}(z), \ldots, G_F(z)\right). \tag{6}$$

Now, all who were infected after this step spend their infectious days in the home country, so the process now is a simple single type Galton-Watson process with offspring generating function $g(z) = G_F^k(z)$, starting from random initial state $Y_0$.

If this simple Galton-Watson process is critical or subcritical, that is, $g'(1) \le 1$, then the process dies out almost surely, regardless of the distribution of $Y_0$; that is, there is no major epidemic in this case. In the supercritical case, when $g'(1) > 1$, the probability that starting from a single individual the process dies out is the unique root in $(0, 1)$ of the equation $g(x) = x$. Let $q$ denote this extinction probability. The process starting from $Y_0$ dies out if all the $Y_0$ branches die out, which has probability $q^{Y_0}$. That is, the probability of extinction of the whole process is

$$\mathbf{P}\{\text{extinction}\} = \mathbf{E}q^{Y_0} = h(q), \tag{7}$$

with $h$ as in (6).

## 3. Computations for the European Football Championship 2012

*3.1. Risk of Measles Outbreak Depends on the Results of the Football Games.* In this section we apply the results to the measles epidemic in Ukraine during the 2012 UEFA European Football Championship. For illustratory purposes, we have chosen France as a prototype for describing the results. In fact, as being amongst the four favourites for the European championship title [7], France was likely to be amongst the teams with the most supporters, while having low vaccination coverage against measles, posing an elevated risk of imported epidemic caused by supporter cases after Euro 2012. We compare the following three scenarios, one of which is the real situation in Euro 2012, while the two others are hypothetical cases representing the extremes for France by means of total time spent in Ukraine (see also Figure 1):

(a) France is eliminated in the group stage, thus playing only three games in Ukraine between June 11 and June 19 (hypothetical case);

(b) France finishes second in the group and is eliminated in the quarterfinals, playing four games in Ukraine between June 11 and June 23 (this is what actually happened);

(c) France finishes second in the group, and gets into the final, thus playing six games between June 11 and July 1, all in Ukraine (hypothetical case).

We assume that the supporter population is staying in Ukraine as long as the team continues to play games. The total length of stay would be the length of games plus one extra day due to international travel, and thus in the three cases we have $T = 10$ (a), $T = 14$ (b), and $T = 21$ (c).

For our computations we set $m = 9$, $k = 9$ [8]. Since measles is generally rare in Europe, the effective reproduction number in France $R_F = g'(1)$ is determined by the basic reproduction number $R_0$ of measles and the effective vaccination coverage $v_F$ in France (e.g., the fraction of population that is immunized and protected) due to $R_F = R_0(1 - v_F)$. The basic reproduction number $R_0$ is estimated between 12 and 18. For computations we use $R_0 = 15$. Thus, $R_F$ is realistically assumed to be in the range 1–3 [9–12]. The contact pattern within the supporter group might be different from the general population, but still the effective reproduction number in Ukraine $R_S = k\beta = kG_S'(1)$ is expected to be of the same magnitude as $R_F$ providing a reasonable range for $\beta$. Parameter $\lambda = H'(1)$, which represents the expected number of daily infected individuals infected by members of the local

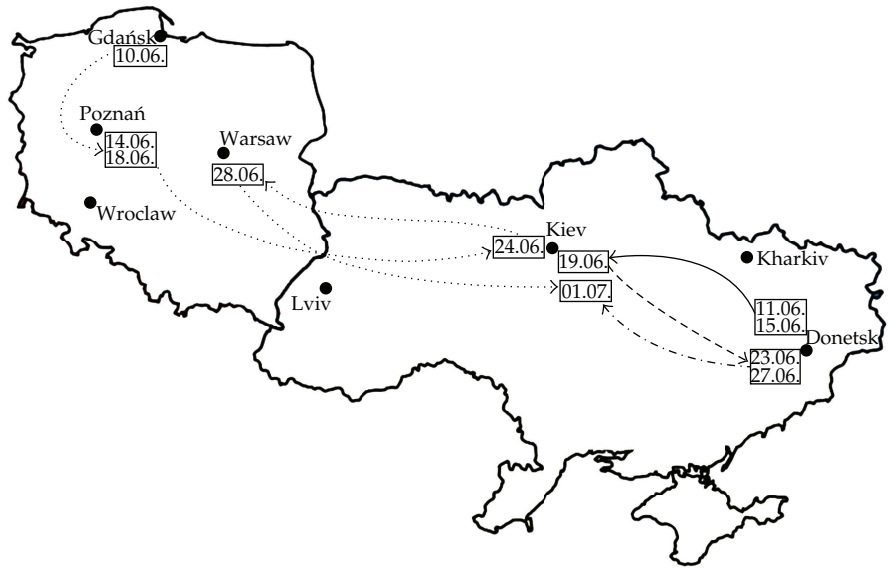


Figure 1: The movement of France during Euro 2012 and the dates of games. The solid arrow corresponds to the group stage, the dashed arrow corresponds to additional games in scenarios (b) and (c), and the dot-dashed arrow corresponds to the hypothetical case of getting into the final (scenario (c)). The dotted arrows represent the movement of Italy (chosen randomly for illustratory purposes) during the tournament.

population, is the most difficult to estimate, as this is given by a combination of several factors: the morbidity of measles in Ukraine during the tournament, the contact patterns between and within local and supporter populations, the size of the supporter group, and the level of susceptibility in this group. We scanned a large domain $[0, 0.125]$ for $\lambda$. A person having measles changes his mixing and contact patterns due to the infection, but it should be noted that generally such a change in social behaviour is accounted for the estimate of $R_0$. Here we assume that individuals in the host country, home country, and visitor populations modify their social behaviour similarly after contracting the disease; thus, our three key parameters $\lambda$, $\beta$, and $R_F$ are all proportional to $R_0$.

By the nature of the immigration and the offspring distributions it is natural to assume that these are Poisson, or compound Poisson distributed. We calculate the extinction probabilities in two cases: when the offspring and immigration distributions are Poisson distributions and when they are negative binomial distributions. In the appendix we explicitly calculate some relevant quantities. We assume that the expectations of the total number of daily new infections from the local population ($\lambda$), the expectation of daily new infections by one infectious individual from the supporter population ($\beta$), and the expectation of daily new infections by one infectious individual in the home country ($\gamma$) are known and choose the parameters of the generating functions accordingly. Note that $R_F = 9\gamma$.

Assuming that both the immigration and the offspring distributions are Poisson, we have

$$H(z) = e^{\lambda(z-1)}, \qquad G_S(z) = e^{\beta(z-1)},$$
$$G_F(z) = e^{\gamma(z-1)}, \qquad g(z) = e^{9\gamma(z-1)}. \tag{8}$$

A random variable $X$ has negative binomial distribution with parameters $r > 0$ and $p \in (0,1)$, if $\mathbf{P}\{X = k\} = \binom{k+r-1}{r-1}(1-p)^r p^k$, $k = 0, 1, 2, \ldots$, where the binomial coefficient is defined by $\binom{k+r-1}{r-1} = (k+r-1)(k+r-2)\cdots r/k!$. The generating function is

$$\mathbf{E}z^X = \left(\frac{1-p}{1-pz}\right)^r, \tag{9}$$

so the expectation is $\mathbf{E}X = rp/(1-p)$. In the special case, when $r = 1$ we obtain the geometric distribution. Assuming geometric distribution for the immigration and the one-day infections

$$H(z) = [1 + \lambda - \lambda z]^{-1}, \qquad G_S(z) = [1 + \beta - \beta z]^{-1},$$
$$G_F(z) = [1 + \gamma - \gamma z]^{-1}, \tag{10}$$

(the parameters are chosen to make the corresponding expectations to be $\lambda$, $\beta$, and $\gamma$ resp.), and using that $G_F^k = g$ we necessarily have

$$g(z) = [1 + \gamma - \gamma z]^{-9}. \tag{11}$$

Figure 2 shows that the risk in scenario $T = 21$ can be twice as large as in scenario $T = 10$.

Comparing Figures 2(a) and 2(b) we see that there is no much difference in the behaviour of the extinction probabilities. In the Poissonian case the extinction probability is slightly larger than in the negative binomial case, corresponding to the same parameter values. Therefore, in the following we assume the Poissonian setup.

The extinction probabilities cannot be computed explicitly. This is because $q$, the probability of extinction starting

from one individual, cannot be calculated explicitly neither in the Poissonian nor in the negative binomial case. We numerically solve the equation for different values of $\gamma$ running from 0.111 up to 0.334; that is, $R_F$ varies in the range 1–3 (recall that the expected value of offsprings in Phase 2 is $9\gamma$). Then we substitute these values into the explicitly known generating function $h$, given in (6).

*3.2. Vaccination and the Risk of Epidemics after Euro 2012.* We compare the effectiveness of three potential vaccination strategies in reducing the risk of imported major epidemic:

(i) vaccination of the general population in France;

(ii) vaccination of the general population in Ukraine;

(iii) vaccination of football-associated travellers between France and Euro 2012 venues.

To consider (i), note that increasing the vaccination rate $v_F$ decreases each of the parameters in our model. That is, $9\gamma = R_F = R_0(1 - v_F)$, $\lambda = \lambda_1(1 - v_F)$, and $\beta = \beta_1(1 - v_F)$. We plotted the risk of major epidemic $p$ as a function of $v_F$ in Figure 3, and the expected number of imported cases in Figure 4.

Figure 3(a) shows a milder measles epidemics in Ukraine, and Figure 3(b) corresponds to a more severe situation. We may notice that it is particularly worthwhile to increase the immunization rate in France if the epidemic is severe in Ukraine, because in this case we could observe a steep decline in the risk as $v_F$ increased beyond 0.84 (which is, roughly speaking, consistent with the reported present coverage in the country). Increasing $v_F$ has the benefit of decreasing the risk of outbreaks by imported cases unrelated to Euro 2012.

On the other hand, elevating the vaccination level $v_U$ of the local Ukrainian population decreases $\lambda = \lambda_2(1 - v_U)$. The current value of $v_U$ is reported to be about 0.5 [13]. Given the difference between the total populations of Ukraine and France, increasing $v_F$ by one unit requires the same amount of vaccines as increasing $v_U$ by 1.4 units. However, the computations show that $p$ is much less sensitive to $v_U$ than to $v_F$ (Figure 5), because small reduction in the risk of infection during travel may only slightly reduce the imported cases.

Targeted vaccination of football visitors reduces both $\lambda = \lambda_1(1 - v_T)$ and $\beta = \beta_1(1 - v_T)$ where $v_T$ is the level of immunization in $S$ (i.e., in the absence of targeted vaccination of travellers, it is assumed that $v_T = v_F$). Figure 6 shows the efficiency of this strategy in the case of a milder and a more severe Ukrainian measles epidemic. If the vaccination history was perfectly known, elevating from 0.84 to 0.94 would be achieved by vaccinating 10% of the travellers from France (targeting the unvaccinated ones). Such an intervention can halve the risk with relatively small efforts. It should be noted that elevating the coverage $v_T$ would require the smallest number of doses (as compared to conducting mass vaccinations in other scenarios) as vaccinating the supporters requires only a couple of thousands of doses.

## 4. Comparison with Euro 2008

In contrast to Euro 2012, here we descriptively review the measles outbreaks which are likely associated with Euro 2008 and other mass gathering events. The 2008 UEFA European Football Championship (Euro 2008) took place in Austria and Switzerland from 7 to 29 June 2008. Significant measles outbreaks were reported in both of the host countries before the championship [14]. However, that situation was different from this year's in several aspects. First of all, the vaccination coverage is much higher in Switzerland and Austria than in Ukraine, and consequently, as the morbidity data show, the measles outbreak in 2012 in Ukraine is of significantly larger scale than the one in the two host countries four years ago [15]. It is also likely that Euro 2008 did not elevate the relative number of travellers as much as Euro 2012 in Ukraine as the two host countries of Euro 2008 are close to several of the participating countries and most host cities are popular tourist destinations, hosting a large number of visitors even without the football championship. As it has been pointed out in [16], a large scale mass gathering can even discourage regular tourists to visit the given cities to avoid the crowdedness, as happened in 2008 during the Olympic Games in Beijing. The media reported a similar phenomenon in London during the 2012 Olympic Games. In other cases (e.g., Sydney 2000), there was a surge of travellers, and we can assume the same for Ukraine as well.

For Euro 2008 we chose Germany as the German national team reached the final of the championship, which means that their supporters spent 21 days in Austria and Switzerland, and WHO reports a suboptimal coverage of 83–89% for the second dose of measles-containing vaccine in Germany [17]. Taking into account the number of measles cases in Austria/Switzerland in 2008 and in Ukraine in 2012, and the population of these countries, we can expect the parameter $\lambda$ to be approximately ten times smaller for Euro 2008 than for Euro 2012. Assuming Poisson distribution, calculating with $\lambda = 0.004$ and $\beta = 0.27$, formula (A.4) says that the probability of no imported infection is 0.92; that is the probability of major epidemic is less than 0.08, which is significantly smaller than the probabilities for Euro 2012.

Data from 2008 show that in several participating countries (e.g., France, Germany, Spain, and Switzerland) there were increases in the number of measles cases after Euro 2008 compared to the same period of the year in 2007 [18–21]. However, based on available data, a direct link cannot be established between Euro 2008 and these outbreaks.

## 5. Other Sports-Related Mass Gatherings

As pointed out in [22], the last two European football championships are not unique in the sense that curiously, the football championships seem to coincide with measles outbreaks. Apart from the two cases mentioned above, during the FIFA World Cup 2006 a large measles outbreak was ongoing in Germany (host country), while there was an outbreak in South Africa during the FIFA World Cup 2010. Furthermore, the Winter Olympic Games in 2010, held in Vancouver, were followed by a measles outbreak in British Columbia of about

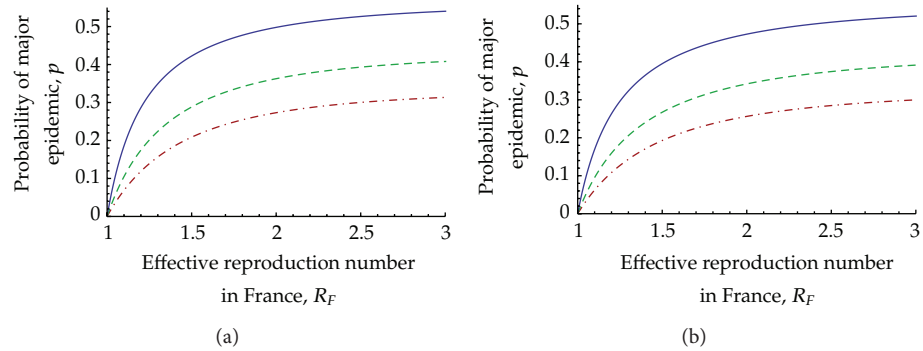


FIGURE 2: The probability of a major epidemic as the function of the effective reproduction number in France in the Poissonian case (a) and in the negative binomial case (b). The parameters are $\lambda = 0.04$, $\beta = 0.27$. The solid curve is for $T = 21$, the dashed is for $T = 14$, and the dot-dashed is for $T = 10$.

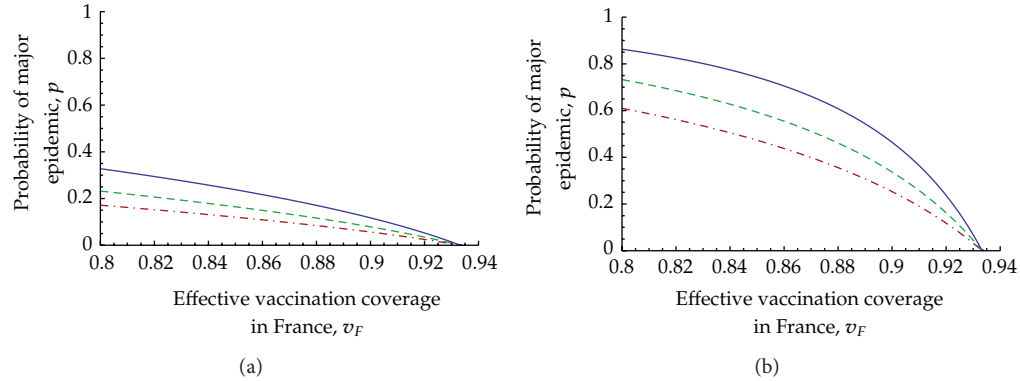


FIGURE 3: The probability of a major epidemic as the function of the immunization rate in France. The parameters are $\beta_1 = 1.8$, and $\lambda_1 = 0.1$ in (a) and $\lambda_1 = 0.5$ in (b). The solid curve is for $T = 21$, the dashed is for $T = 14$, and the dot-dashed is for $T = 10$.

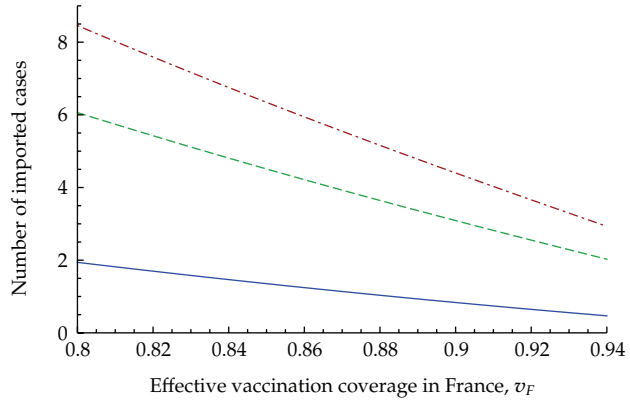


FIGURE 4: The solid curve is the expectation of the total number of imported cases in scenario (b) in the function of the immunization rate in France. At least with probability 0.75 the number of imported cases is smaller than the dashed curve and with probability 0.9 is smaller than the dot-dashed curve (calculated from Chebyshev's inequality). The parameter values are the same as in Figure 3(b).

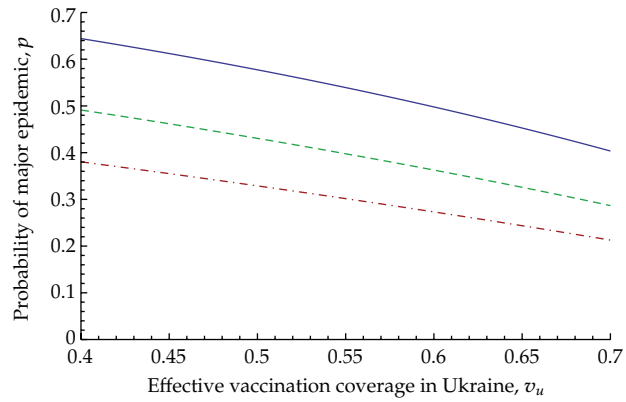


FIGURE 5: The probability of a major epidemic as the function of the immunization rate in Ukraine. The parameters are $R_F = 2$, $\beta_1 = 1.8$, and $\lambda_1 = 0.1$. The solid curve is for $T = 21$, the dashed is for $T = 14$, and the dot-dashed is for $T = 10$.

80 cases following three separate importations, two of which were linked to the Olympic Games [23].

After Euro 2012, another sports related mass gathering event followed, the Summer Olympic Games in London. There were several alerts about measles in connection with the Olympic Games [24, 25]. However, there are several differences between football championships and the Olympic

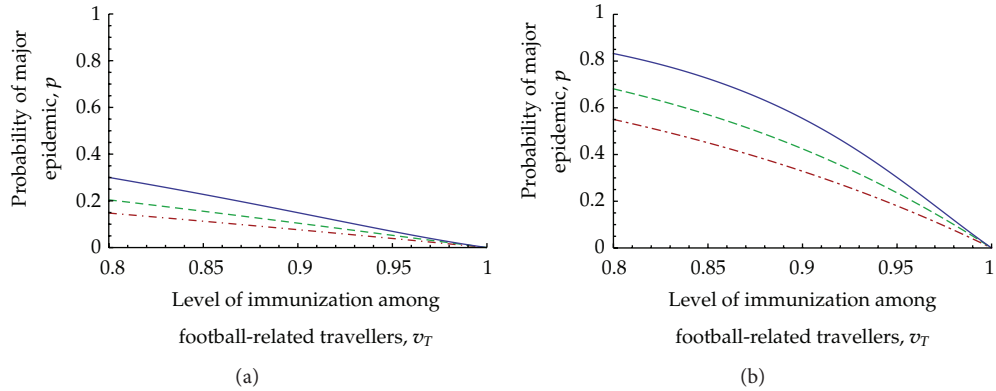


FIGURE 6: The probability of a major epidemic as the function of the immunization rate in the supporter group. The parameters are $R_F = 2$, $\beta_1 = 1.8$, and $\lambda_1 = 0.1$ in (a) and $\lambda_1 = 0.5$ in (b). The solid curve is for $T = 21$, the dashed is for $T = 14$, and the dot-dashed is for $T = 10$.

Games. Football championships have a special tournament structure and huge groups of fans moving together following their national teams, which is not typical for the Olympic Games. Football championships are hosted by several cities, while the Olympic Games are held (apart from some minor events) in one city. This means that our model fits rather for sport events which have the tournament structure like football World Cup and European Championship. With an efficient monitoring after Euro 2012, it may be possible to refine our parameters and prepare more realistic risk assessments using our approach for the forthcoming major championships such as FIFA World Cup 2014 in Brazil and Euro 2016 in France. The low vaccination rate and the recent and ongoing measles epidemics in France [26, 27] suggest that there will be a risk of measles during Euro 2016 as well.

## 6. Discussion

We constructed and applied a stochastic model to investigate the risk of imported epidemics caused by visitors returning from a sports related mass gathering event to their home countries after the tournament. For the sake of simplicity, we considered a single supporter population, while a realistic situation of course involves many additional complicating factors including movements within the host country and interactions between supporters and local population. We introduced a discrete time Markov chain model with two phases, which is an adaptation of a multitype Galton-Watson process with immigration as a mathematical model and derived several analytical relations for the expectations, variances and probabilities regarding key aspects of the process.

We applied our theoretical model to the measles epidemics in Ukraine during the 2012 UEFA European Football Championship, selecting the national team of France for illustratory purposes. Due to the uncertainties in social parameters, we considered a wide interval for the transmission rate between local and visitor populations. Our approach clearly demonstrated that the travel patterns depend on the schedule and the results of the football games, showing that the probability of a major measles epidemic in France could be greatly elevated by the successful outcomes of French games. Namely, the more successful the national team is in a football tournament, the higher the risk of a post-tournament imported measles epidemic would be in the home country. More importantly, we have compared different vaccination strategies and our study theoretically demonstrated that the risk of an imported measles epidemic by the visitors to Euro 2012 and other mass gatherings would be most efficiently reduced by vaccinating the visitors (travellers). Of course, vaccinating the entire French population would also be effective (which actually prevents the country from not only the risk from Euro 2012 but also any other epidemics to be imported), but in theory this option requires us to secure millions of doses. The optimal control by effectively targeting travellers is novel both in practical and theoretical sense, because the condensed interventions among travellers have been shown not to be very effective in preventing an epidemic (e.g., pandemic influenza) as long as there are arbitrarily large number of travellers. We have shown that it is worth focusing on travellers when the number is finite and in the manageable order. Unvaccinated travellers would likely be covered within a few thousand doses, and thus any country to respond to the associated risk is suggested to consider this option.

## Appendix

We compute explicitly the expectation and variance of the overall number of infectious individuals arriving home after day $T$, that is, $Z := \sum_{i=1}^{17} X_T(i)$, and also we compute the probability that there is no imported infection, that is, $\mathbf{P}\{Z = 0\}$. To do this we compute the generating function $F_T(\mathbf{z})$ given in (3) and then we use that

$$\mathbf{P}\{Z = 0\} = F_T(\mathbf{0}), \qquad \mathbf{E}Z = \sum_{i=1}^{17} \frac{\partial F_T}{\partial z_i}(\mathbf{1}),$$
$$\mathbf{Var}\, Z = \sum_{i,j} \left( \frac{\partial^2 F_T}{\partial z_i \partial z_j}(\mathbf{1}) - \mathbf{E}X_T(i)\,\mathbf{E}X_T(j) + \delta_{i,j}\mathbf{E}X_T(i) \right), \tag{A.1}$$

with $\delta_{i,j} = 1$ for $i = j$, and 0 otherwise. Note that one minus the probability of no imported infection is a trivial

upper bound for the probability of a major epidemic, and it is independent of $\gamma$.

Assuming Poissonian offspring and immigration distribution, we have the following:

(i) $T = 10$: in this case everything is relatively easy to compute. For example, there is no imported infection if and only if each day the number of immigrants is 0, which has probability $e^{-10\lambda}$. We have

$$\mathbf{P}\{Z = 0\} = e^{-10\lambda}, \qquad \mathbf{E}Z = \lambda(10 + \beta), \qquad \mathbf{Var}\, Z = \lambda\left(10 + 3\beta + \beta^2\right); \tag{A.2}$$

(ii) $T = 14$:

$$\mathbf{P}\{Z = 0\} = e^{-14\lambda}, \qquad \mathbf{E}Z = \lambda(14 + 15\beta), \qquad \mathbf{Var}\, Z = \lambda\left(14 + 45\beta + 55\beta^2\right); \tag{A.3}$$

(iii) $T = 21$:

$$\mathbf{P}\{Z = 0\} = e^{-17\lambda - 4(1-e^{-9\beta})\lambda}, \qquad \mathbf{E}Z = \lambda\left(17 + 72\beta + 10\beta^2\right), \qquad \mathbf{Var}\, Z = \lambda\left(17 + 144\beta + 558\beta^2 + 200\beta^3 + 46\beta^4\right). \tag{A.4}$$

In the negative binomial case for the different scenarios we have

(i) $T = 10$:

$$\mathbf{P}\{Z = 0\} = (1 + \lambda)^{-10}, \qquad \mathbf{E}Z = \lambda(10 + \beta), \qquad \mathbf{Var}\, Z = \lambda\left(10(1 + \lambda) + \beta^2(2 + \lambda) + \beta(3 + 2\lambda)\right); \tag{A.5}$$

(ii) $T = 14$:

$$\mathbf{P}\{Z = 0\} = (1 + \lambda)^{-14}, \qquad \mathbf{E}Z = \lambda(14 + 15\beta), \qquad \mathbf{Var}\, Z = \lambda\left(14(1 + \lambda) + 15\beta(3 + 2\lambda) + \beta^2(70 + 55\lambda)\right), \tag{A.6}$$

(iii) $T = 21$:

$$\mathbf{P}\{Z = 0\} = (1 + \lambda)^{-17}\left(1 + \lambda - \frac{\lambda}{(1 + \beta)^9}\right)^{-4}, \qquad \mathbf{E}Z = \lambda\left(17 + 72\beta + 10\beta^2\right), \qquad \mathbf{Var}\, Z = \lambda\left(17(1 + \lambda) + 72\beta(2 + \lambda) + 6\beta^2(105 + 88\lambda) + 10\beta^3(23 + 18\lambda) + \beta^4(66 + 46\lambda)\right). \tag{A.7}$$

Also note that in both cases the variance is large compared to the expectation, implying that the probability of no imported cases is large.

## Acknowledgments

A. Dénes and G. Röst were supported by the European Research Council Starting Investigator Grant no. 259559, the Hungarian Scientific Research Fund OTKA K75517, and Bolyai Scholarship of the Hungarian Academy of Sciences. P. Kevei was supported by the TÁMOP-4.2.1/B-09/1/KONV-2010-0005 Project and the Hungarian Scientific Research Fund OTKA PD106181. H. Nishiura received funding support from the JST PRESTO Program and The University of Hong Kong Seed Funding Program (Grant Code: 10208192).

## References

[1] "Outbreak of measles in Ukraine and potential for spread in the EU, Rapid Risk Assessment," ECDC, 13 March 2012, http://ecdc.europa.eu/en/publications/Publications/20120314_RA_Measles_Ukraine.pdf.

[2] "Communicable disease threats report," ECDC, Week 26, 24–30 June 2012, http://ecdc.europa.eu/en/publications/Publications/CDTR_web_2012_6_28.pdf.

[3] B. R. Humphreys and S. Prokopowicz, "Assessing the impact of sports mega-events in transition economies: EURO 2012 in Poland and Ukraine," *International Journal of Sport Management and Marketing*, vol. 2, no. 5-6, pp. 496–509, 2007.

[4] "UEFA webpage," http://www.uefa.com/uefaeuro/season=2012/tournament-calendar/index.html.

[5] T. Britton, "Stochastic epidemic models: a survey," *Mathematical Biosciences*, vol. 225, no. 1, pp. 24–35, 2010.

[6] M. P. Quine, "The multi-type Galton-Watson process with immigration," *Journal of Applied Probability*, vol. 7, no. 2, pp. 411–422, 1970.

[7] "Bwin," http://www.bwin.com/Euro2012.

[8] R. D. Feigin, J. Cherry, G. J. Demmler-Harrison, and S. L. Kaplan, *Feigin and Cherry's Textbook of Pediatric Infectious Diseases*, Saunders, 6th edition, 2009.

[9] L'Institut de Veille Sanitaire (InVS), http://www.invs.sante.fr/Dossiers-thematiques/Maladies-infectieuses/Maladies-a-prevention-vaccinale/Rougeole/Points-d-actualites.

[10] L. Fonteneau, J.-M. Urcun, C. Kerneur et al., "Couverture vaccinale des enfants gs de 11 ans scolariss en CM2, France, 2004-2005," *Bulletin Épidémiologique Hebdomadaire*, vol. 51-52, pp. 493–497, 2008 (French).

[11] I. P. du Châtelet, D. Antona, F. Freymuth et al., "Spotlight on measles 2010: update on the ongoing measles outbreak in france, 2008–2010," *Euro Surveillance*, vol. 15, no. 36, pp. 1–4, 2010 http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19656.

[12] I. Bonmarin and D. Lvy-Bruhl, "Measles in France: the epidemiological impact of suboptimal immunisation coverage," *Euro Surveillance*, vol. 7, no. 4, p. 322, 2002, http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=322.

[13] "WHO epidemiological brief," no. 21, 2012, http://www.euro.who.int/__data/assets/pdf_file/0004/159475/WHO_EPI_Brief_Feb_2012e.pdf.

[14] R. Strauss, P. Kreidl, M. Muscat et al., "The measles situation in Austria: a rapid risk assessment by an ECDC team and the outcome of an International Meeting in Vienna, Austria," *Euro Surveillance*, vol. 13, no. 17, 2008, http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=18852.

[15] P. Kreidl, P. Buxbaum, F. Santos-O'Connor et al., "2008 European Football Championship—ECDC epidemic intelligence support," *Euro Surveillance*, vol. 13, no. 32, 2008, http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=18946.

[16] K. Khan, S. J. N. McNabb, Z. A. Memish et al., "Infectious disease surveillance and modelling across geographic frontiers and scientific specialties," *The Lancet Infectious Diseases*, vol. 12, no. 3, pp. 222–230, 2012.

[17] WHO, "Reported estimates of MCV coverage," http://apps.who.int/immunization_monitoring/en/globalsummary/timeseries/tscoveragemcv.htm.

[18] I. Parent du Chtelet, D. Floret, D. Antona, and D. Lvy-Bruhl, "Measles resurgence in France in 2008, a preliminary report," *Euro Surveillance*, vol. 14, no. 6, 2009, http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19118.

[19] Robert Koch Institut SurvStatRKI, Abfrage der Meldedaten nach Infektionsschutzgesetz (IfSG) ber das Web, http://www3.rki.de/SurvStat/.

[20] "Resultados de la vigilancia epidemiolgica de las enfermedades transmisibles," Informe Anual, 2008, http://www.isciii.es/ISCIII/es/contenidos/fd-servicios-cientifico-tecnicos/fd-vigilancias-alertas/fd-enfermedades/Informeanual2008.pdf.

[21] E. Delaporte, E. Jeannot, P. Sudre, C. A. W. Lazarevic, J. L. Richard, and P. Chastonay, "Measles in Geneva between 2003 and 2010: persistence of measles outbreaks despite high immunisation coverage," *Euro Surveillance*, vol. 16, no. 39, 2011, http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19980.

[22] Editorial team, "Spotlight on measles 2010," *Euro Surveill*, vol. 15, no. 17, 2010, http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19559.

[23] "BC Centre for Disease Control," http://www.bccdc.ca/resourcematerials/newsandalerts/healthalerts/Before+travelling+ensure+your+measles+vaccination+is+up+to+date.htm.

[24] "Surveillance report," European Monthly Measles Monitoring (EMMO) Issue 10, 2012, http://ecdc.europa.eu/en/publications/Publications/1205-SUR-Measles-monthly-monitoring.pdf.

[25] "Centers for Disease Control and Prevention," http://wwwnc.cdc.gov/travel/notices/outbreak-notice/measles.htm.

[26] C. Huoi, J. S. Casalegno, T. Bnet et al., "A report on the large measles outbreak in Lyon, France, 2010 to 2011," *Euro Surveillance*, vol. 17, no. 36, 2012 http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20264.

[27] "Measles and rubella monitoring," Surveillance report, ECDC, 2012, http://www.ecdc.europa.eu/en/publications/Publications/2012Sept_SUR_measles-rubella-monitoring.pdf.

Hindawi Publishing Corporation
International Journal of Stochastic Analysis
Volume 2013, Article ID 963831, 10 pages
http://dx.doi.org/10.1155/2013/963831

*Research Article*

# Modeling Neutral Evolution Using an Infinite-Allele Markov Branching Process

**Xiaowei Wu[1] and Marek Kimmel[2,3]**

[1] *Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA*
[2] *Departments of Statistics and Bioengineering, Rice University, Houston, TX 77005, USA*
[3] *Systems Engineering Group, Silesian University of Technology, 44-100 Gliwice, Poland*

Correspondence should be addressed to Xiaowei Wu; xwwu@vt.edu

Received 6 November 2012; Revised 14 February 2013; Accepted 18 February 2013

Academic Editor: Peter Olofsson

Copyright © 2013 X. Wu and M. Kimmel. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We consider an infinite-allele Markov branching process (IAMBP). Our main focus is the frequency spectrum of this process, that is, the proportion of alleles having a given number of copies at a specified time point. We derive the variance of the frequency spectrum, which is useful for interval estimation and hypothesis testing for process parameters. In addition, for a class of special IAMBP with birth and death offspring distribution, we show that the mean of its limiting frequency spectrum has an explicit form in terms of the hypergeometric function. We also derive an asymptotic expression for convergence rate to the limit. Simulations are used to illustrate the results for the birth and death process.

## 1. Introduction

The infinite-allele branching process was first introduced by Griffiths and Pakes [1]. As a special type of branching process, this process allows individuals to mutate into infinitely many allelic variants, each of which is "new" in the sense of being different from all previously existing variants. This idealization is approximately correct for rare point mutations in long DNA sequences. Fundamental results for the discrete-time case (simple branching process) and for the continuous-time case (Markov branching process) have been obtained in [1, 2]. These include the number of alleles at a given generation or time, the generation number or time of the last mutation, and the limiting frequency spectrum. There exists an analogy between the results for the discrete-time and the continuous-time cases; however, the characteristics in the continuous case are relatively easier to derive [2]. Many evolutionary processes may be considered time continuous, and frequently we assume Markov property in modeling. A classical example is the discrete-time Wright-Fisher model, which is typically either approximated by a continuous-time diffusion or replaced by a continuous-time Markov chain, the so-called continuous-time Moran process [3]. Therefore, the time-continuous infinite-allele Markov branching process (TCIAMBP, or simply IAMBP) seems to be appropriate for modeling evolution in population genetics.

Consider a Markov branching process with neutral mutations. Suppose that the process starts from a group of individuals carrying the same allele, and individuals can mutate into new allelic variants. We assume that the mutation is independent of the previous history of the process, and the offspring distribution is independent of the allelic type, that is, the selection is neutral for all alleles. The process can be described as an "infinitely-many-alleles" model (IAM). Whenever a mutation happens, it yields a new allele, which differs from all the previously existing ones. In this paper, we are interested in the frequency spectrum of the IAMBP, which may be defined as the number or proportion of alleles present in a given number of individuals at a specified time point. Frequency spectrum in this paper refers to random allele frequencies, not their expected value as in Griffiths and Pakes [1] and in Pakes [2], since we also consider the variance of the allele frequency later on. Unless specified otherwise, we will use terms "mean frequency spectrum" and "variance frequency spectrum" in the remainder of this paper to denote the expected value and variance of the allele frequency. The frequency spectrum plays an important role in many genetic processes, such as DNA sequence evolution. As an example, Kimmel and

Mathaes [4] modeled the Alu sequence data using an infinite-allele simple branching process with linear-fractional offspring distribution, and the goodness of fit testing suggested that Alu sequences do not evolve neutrally and might be under selection. It has to be noted that the concept of the frequency spectrum is in some sense similar to the Ewens' sampling formula [5] in population genetics. We will return to this subject in the discussion, although analysis of the analogies and differences transcends the scope of the present paper.

The paper is organized as follows. In Section 2, we rigorously define the IAMBP and the mean frequency spectrum of the IAMBP. Then, we provide explicit expressions for the special case of the birth and death process. In Section 3, we derive the variance frequency spectrum and discuss its use in interval estimation for process parameters. We perform simulations to illustrate the results using the birth and death process example in Section 4. Section 5 is a summary.

## 2. IAMBP and Its Limiting Mean Frequency Spectrum

*2.1. Definition and Basic Properties in the Supercritical Case.* Let us consider a continuous-time Markov branching process consisting of individuals with exponential life spans with mean $a^{-1}$. Let us assume that upon death, each individual produces a random number of offspring. As usually assumed, the offspring counts are identically distributed according to probability generating function (pgf) $f(s)$, and they are independent conditional on the past process. The mean $f'(1^-)$ of the offspring distribution is $m$, regardless of the allelic type. We further assume that a newborn individual mutates into a new allelic type with probability $\mu$ independently of the previous history of the process. Let us denote by $h(s) = f(\mu+(1-\mu)s)$ the offspring pgf in a clone, started by the overall ancestor or any of mutants, containing only the like-type individuals. The entire process is a union over all individual types of such clones. The theory of the IAMBP has been developed by Griffiths and Pakes [1] in the discrete-time case and then by Pakes [2] in the continuous-time case. We will assume $m > 1$ and $M = h'(1^-) > 1$, although some results can be proved without this latter assumption.

Let $\alpha_t(j)$ be the number of alleles present in $j$ individuals at time $t$ and $\phi_{i,t}(j) = E_i[\alpha_t(j)]$, where subscript $i$ indicates that the process begins with $i$ individuals carrying the same allele. It has been shown that [2]

$$\phi_{i,t}(j) = q_{ij}(t) + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)\,dx, \quad j \geq 0, \tag{1}$$

where $\lambda = a(m-1)$ is the Malthusian parameter of the overall process and $q_{ij}(t)$ is the probability of observing $j$ individuals ($j \geq 1$) carrying the parental allele at time $t$ when starting from $i$ individuals with the parental allele at time $t = 0$. Consequently, for the number $K_t$ of alleles at time $t$, we have

$$E_i[K_t] = \sum_{j=1}^{\infty}\phi_{i,t}(j) = 1 - q_{i0}(t) + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x}\left[1 - q_{10}(x)\right]dx. \tag{2}$$

Let $G_j = \int_0^\infty e^{-\lambda t} q_{1j}(t)dt$, $j \geq 0$. If we define $\psi_{ij}(t) = \phi_{i,t}(j)/E_i[K_t]$ and

$$\psi_j = \lim_{t\to\infty}\psi_{ij}(t) = \lim_{t\to\infty}\frac{q_{ij}(t) + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)\,dx}{1 - q_{i0}(t) + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x}\left[1 - q_{10}(x)\right]dx} \tag{3}$$

as the limiting mean frequency spectrum, that is, the expected proportion of alleles present in $j$ individuals as $t \to \infty$, then we see that for the supercritical process such that $\lambda > 0$,

$$\psi_j = \frac{\lambda G_j}{1 - \lambda G_0}, \quad j \geq 1. \tag{4}$$

If $M > 1$, then the process of the like-type clones is supercritical, and as it is known [6], $q_{10}(t) \uparrow q_{10}(\infty) < 1$ and $q_{1j}(t) \to 0$, $j \geq 1$, as $t \to \infty$. Therefore, $e^{\lambda t}\left|\int_0^t e^{-\lambda x} q_{10}(x)dx - q_{10}(\infty)/\lambda\right| \to 0$ and $e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)dx \to 0$ as $t \to \infty$, for $j \geq 1$. This yields the following asymptotic equivalence:

$$\psi_{ij}(t) - \psi_j \underset{t\to\infty}{\sim} \frac{\lambda G_j\left[1 - q_{10}(\infty) - \left(\lambda\left(1 - q_{i0}(\infty)\right)/iam\mu\right)\right]}{\left(1 - \lambda G_0\right)^2} e^{-\lambda t}. \tag{5}$$

Details of the proof are omitted, since they appear elementary.

*2.2. IAMBP with Birth and Death Offspring Distribution.* For the IAMBP with birth and death offspring distribution $f(s) = \alpha + \beta s^2$, $\alpha + \beta = 1$, we are able to obtain an explicit form for $G_j$, $j \geq 0$; therefore, the limiting mean frequency spectrum $\psi_j$, $j \geq 1$ can be derived. The offspring pgf of the like-type individuals clone in the birth and death IAMBP is written as

$$h(s) = f\left(\mu + (1-\mu)s\right) = \alpha + \beta\left[\mu + (1-\mu)s\right]^2, \tag{6}$$

where $\alpha$, $\beta$ and $\mu$ stand for the death, birth, and mutation probabilities for every individual and $\alpha + \beta = 1$. Note that under another parameterization where the two newborn individuals die, live, and mutate independently, this pgf may be formulated differently as $h(s) = [\alpha + \beta\mu + \beta(1-\mu)s]^2$. Under either parameterization, $\lambda = a(2\beta - 1)$. If, as assumed, $M = m(1-\mu) > 1$, then parameters $\alpha$ and $\mu$ are subject to a constraint

$$(1-\alpha)(1-\mu) > \frac{1}{2}. \tag{7}$$

Let us write $A^2 = \alpha + \beta\mu^2$ and $B^2 = \beta(1-\mu)^2$ (note, for the other formulation, $A^2 = (\alpha+\beta\mu)^2$ and $B^2 = \beta^2(1-\mu)^2$). The explicit form of $G_j$ can be written as

$$\begin{aligned} G_0 &= \frac{1}{c}\frac{A^2}{B^2}\frac{\Gamma(\lambda/c)\,\Gamma(2)}{\Gamma(2+(\lambda/c))} \\ &\quad \times F\left(1, \frac{\lambda}{c}; 2+\frac{\lambda}{c}; \frac{A^2}{B^2}\right), \\ G_j &= \frac{1}{c}\left(1-\frac{A^2}{B^2}\right)^2 \frac{\Gamma(1+(\lambda/c))\,\Gamma(j)}{\Gamma(j+1+(\lambda/c))} \\ &\quad \times F\left(j+1, 1+\frac{\lambda}{c}; j+1+\frac{\lambda}{c}; \frac{A^2}{B^2}\right), \quad j \geq 1, \end{aligned} \tag{8}$$

where $c = a(B^2 - A^2) = a[2\beta(1-\mu) - 1]$ is the Malthusian parameter of the like-type clone and $F(\cdot,\cdot;\cdot;\cdot)$ is the Gauss hypergeometric function [7], defined as

$$\begin{aligned} &F(a,b;c;z) \\ &\quad = \frac{\Gamma(c)}{\Gamma(b)\,\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}dt, \\ &\qquad\qquad c > b > 0. \end{aligned} \tag{9}$$

For a detailed derivation, see Appendix A. Note that the supercritical condition also guarantees that the argument of the hypergeometric function remains within its region of definiteness.

It follows that

$$\begin{aligned} \psi_j &= \frac{\lambda G_j}{1-\lambda G_0} \\ &= \left(\frac{\lambda}{c}\left(1-\frac{A^2}{B^2}\right)^2 \right. \\ &\quad \times \left(\frac{\Gamma(1+\lambda/c)\,\Gamma(j)}{\Gamma(j+1+\lambda/c)}\right) \\ &\quad \left. \times F\left(j+1, 1+\frac{\lambda}{c}; j+1+\frac{\lambda}{c}; \frac{A^2}{B^2}\right)\right) \\ &\quad \times \left(1-\frac{\lambda}{c}\frac{A^2}{B^2}\frac{\Gamma(\lambda/c)\,\Gamma(2)}{\Gamma(2+\lambda/c)} \right. \\ &\qquad \left. \times F\left(1, \frac{\lambda}{c}; 2+\frac{\lambda}{c}; \frac{A^2}{B^2}\right)\right)^{-1}, \quad j \geq 1. \end{aligned} \tag{10}$$

Figure 1 shows an example of the limiting mean frequency spectrum for the birth and death process with parameters $a = 1$, $\alpha = 0.25$, and $\mu = 10^{-4}$, based on formula (10). To see how the spectrum varies with different parameter settings, we plot in Figure 2(a), the 3-D surface of a major component of the spectrum, $\psi_1$, for different $\alpha$'s and $\mu$'s. Figures 2(b) and 2(c) illustrate the effect of one parameter on $\psi_1$ given a fixed value of the other parameter.

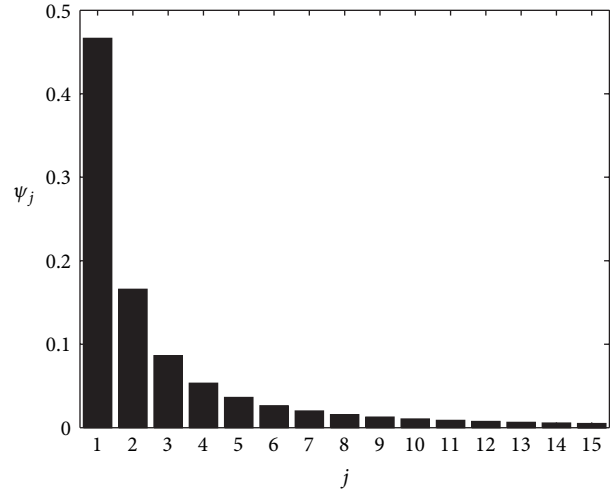


Figure 1: Limiting mean frequency spectrum of the infinite-allele birth and death process with $a = 1$, $\alpha = 0.25$, and $\mu = 10^{-4}$.

We see that for fixed $\alpha$, increasing $\mu$ causes an increase of $\psi_1$. This can be intuitively explained by the offspring pgf $h(s)$ of the like-type clone. From the pgf expression $h(s) = \alpha + \beta[\mu + (1-\mu)s]^2$, we see that the probability of obtaining one like-type individual in the offspring is $2(1-\alpha)\mu(1-\mu)$, which is an increasing function of $\mu$ for a given $\alpha$, under the constraint $(1-\alpha)(1-\mu) > 1/2$. Therefore, increasing $\mu$ will finally lead to an increase of $\psi_1$. The effect of $\alpha$ on $\psi_1$ when fixing $\mu$ is not so obvious, but we notice that when fixing $\mu$ very close to 0, as $\alpha$ approaches $1/2$, the process is approximately critical binary fission; therefore, $\psi_1$ drops down because of almost sure extinction of the process, as seen from the tail behavior of the solid thick line in Figure 2(c).

Arguably, the frequency spectrum can only be observed in finite time. The finite-time mean frequency spectrum can be obtained by computing $G_j(t) = \int_0^t e^{-\lambda x} q_{1j}(x)dx$, $j \geq 0$ numerically. For the birth and death process, this involves the computation of the incomplete hypergeometric function. The following is a valid question in this context. In order to safely use the limiting mean frequency spectrum, how long should the process history be? Figure 3(a) compares the limiting mean frequency spectrum with some long-term mean frequency spectra, for the birth and death process with parameters $a = 1$, $\alpha = 0.25$, and $\mu = 10^{-4}$. We see that under this setting, the long-term mean frequency spectrum is almost identical to the limiting mean frequency spectrum when $t \geq 28$. In general, this result depends strongly on parameters $a$, $\alpha$, and $\mu$, for example, small $\mu$ leads to longer $t$. This provides us with some intuitions concerning the sufficiently large $t$ for approximating the limiting mean frequency spectrum. Figure 3(b) illustrates the difference between the finite-time mean frequency spectrum and the limiting mean frequency spectrum as a function of $t$, for large $t$, $t \in [15, 35]$ and for $j = 1, 2$, where lines represent the true difference and markers represent the asymptotic approximation by formula (5). To emphasize the agreement for $t$ large, this figure is plotted in semilogarithmic scale. We

see that the true difference drops exponentially fast, and the asymptotic approximation is good for large $t$.

Given the observed long-term mean frequency spectrum, the parameters $\theta$ of the IAMBP, such as $\alpha$, $\mu$ in the birth and death process, can be estimated by equating the observed long-term mean frequency spectrum $\psi_{\text{obs}}$ from the sample to the expected limiting mean frequency spectrum $\psi_{\text{exp}}$ from formula (3) and solving for the process parameters. In the case of the birth and death process, we may estimate $\alpha$ and $\mu$ for example by solving

$$\begin{aligned}
&\left(\Gamma(j_1)\,\Gamma\left(k_1+1+\frac{\lambda}{c}\right)\right.\\
&\quad\times F\left(j_1+1,1+\frac{\lambda}{c};\right.\\
&\qquad\left.\left.j_1+1+\frac{\lambda}{c};\frac{A^2}{B^2}\right)\right)\\
&\quad\times\left(\Gamma(k_1)\,\Gamma\left(j_1+1+\frac{\lambda}{c}\right)\right.\\
&\qquad\times F\left(k_1+1,1+\frac{\lambda}{c};\right.\\
&\qquad\left.\left.\left.k_1+1+\frac{\lambda}{c};\frac{A^2}{B^2}\right)\right)^{-1}\right.\\
&\quad=\frac{\psi_{\text{obs}}(j_1)}{\psi_{\text{obs}}(k_1)}\\
&\left(\Gamma(j_2)\,\Gamma\left(k_2+1+\frac{\lambda}{c}\right)\right.\\
&\quad\times F\left(j_2+1,1+\frac{\lambda}{c};\right.\\
&\qquad\left.\left.j_2+1+\frac{\lambda}{c};\frac{A^2}{B^2}\right)\right)\\
&\quad\times\left(\Gamma(k_2)\,\Gamma\left(j_2+1+\frac{\lambda}{c}\right)\right.\\
&\qquad\times F\left(k_2+1,1+\frac{\lambda}{c};\right.\\
&\qquad\left.\left.\left.k_2+1+\frac{\lambda}{c};\frac{A^2}{B^2}\right)\right)^{-1}\right.\\
&\quad=\frac{\psi_{\text{obs}}(j_2)}{\psi_{\text{obs}}(k_2)}
\end{aligned}\tag{11}$$

for positive integers $j_1 \neq k_1$, $j_2 \neq k_2$, where $\lambda/c$ and $A^2/B^2$ are both functions of $\alpha$ and $\mu$.

There is no explicit solution for such estimator, but numerical search according to some criteria is feasible. Another possibility is to minimize the distance (such as the $l^2$ norm) between the observed long-term mean frequency spectrum and the expected limiting mean frequency spectrum, that is, $\widehat{\theta} = \arg\min_\theta \|\psi_{\text{obs}} - \psi_{\text{exp}}(\theta)\|_2$.

The estimated parameters can be used to check the goodness of fit of the IAMBP model. Another interesting problem is to test whether two sets of parameters are identical, given two observed mean frequency spectra. A simple approach is to use Pearson's $\chi^2$ test, such as in Kimmel and Mathaes [4]. However, there may be restrictions to applying the $\chi^2$ test, such as small cell counts and inappropriateness due to the finite length of the observed spectrum. This motivates us to develop an interval estimator for the IAMBP parameters.

## 3. Variance of the Frequency Spectrum

Moment estimators based on the mean frequency spectrum only give point estimates of the process parameters. In order to quantify the uncertainty of point estimates, an interval estimator is needed, which requires more information about the distribution of the statistic $\alpha_t(j)$. First, it can be seen that [2]

$$\alpha_t(j) = I_{0,j}(t) + \sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t - T_n), \tag{12}$$

where $T_1, T_2, \ldots$ are the successive split times of the process, $I_{0,j}(t)$, $I_{n,k,j}(t)$ are two indicators, and $I_{0,j}(t) = 1$ if there are $j$ individuals alive at time $t$ carrying the parental allele, and $I_{n,k,j}(t) = 1$, for $n, k \geq 1$ if the $k$th individual born at time $T_n$ ($T_n < t$) mutates to a novel allelic type and further produces $j$ individuals carrying this allele $t$ time units later. $N_t$ is the number of split times in $(0, t]$, and $U_n$ is the number of offspring produced at time $T_n$. Obtaining the distribution of $\alpha_t(j)$ is not elementary. However, it may still be possible to define a confidence interval (CI) based on the first and second moments of $\alpha_t(j)$.

Let $\eta_{i,t}(j) = \text{Var}_i(\alpha_t(j))$ be the variance frequency spectrum; by the law of total variance and independence between the indicators in the expression of $\alpha_t(j)$ (details in Appendix B), we have

$$\begin{aligned}
\eta_{i,t}(j) &= q_{ij}(t)\left[1 - q_{ij}(t)\right]\\
&\quad + im^2\mu^2\left[C(t) + (\lambda + a)\,e^{\lambda t}\int_0^t e^{-\lambda x}C(x)\,dx\right]\\
&\quad + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x}q_{1j}(x)\,dx + ia\left(\sigma^2 - m\right)\mu^2 e^{\lambda t}\\
&\quad\times\int_0^t e^{-\lambda x}q_{1j}^2(x)\,dx,
\end{aligned}\tag{13}$$

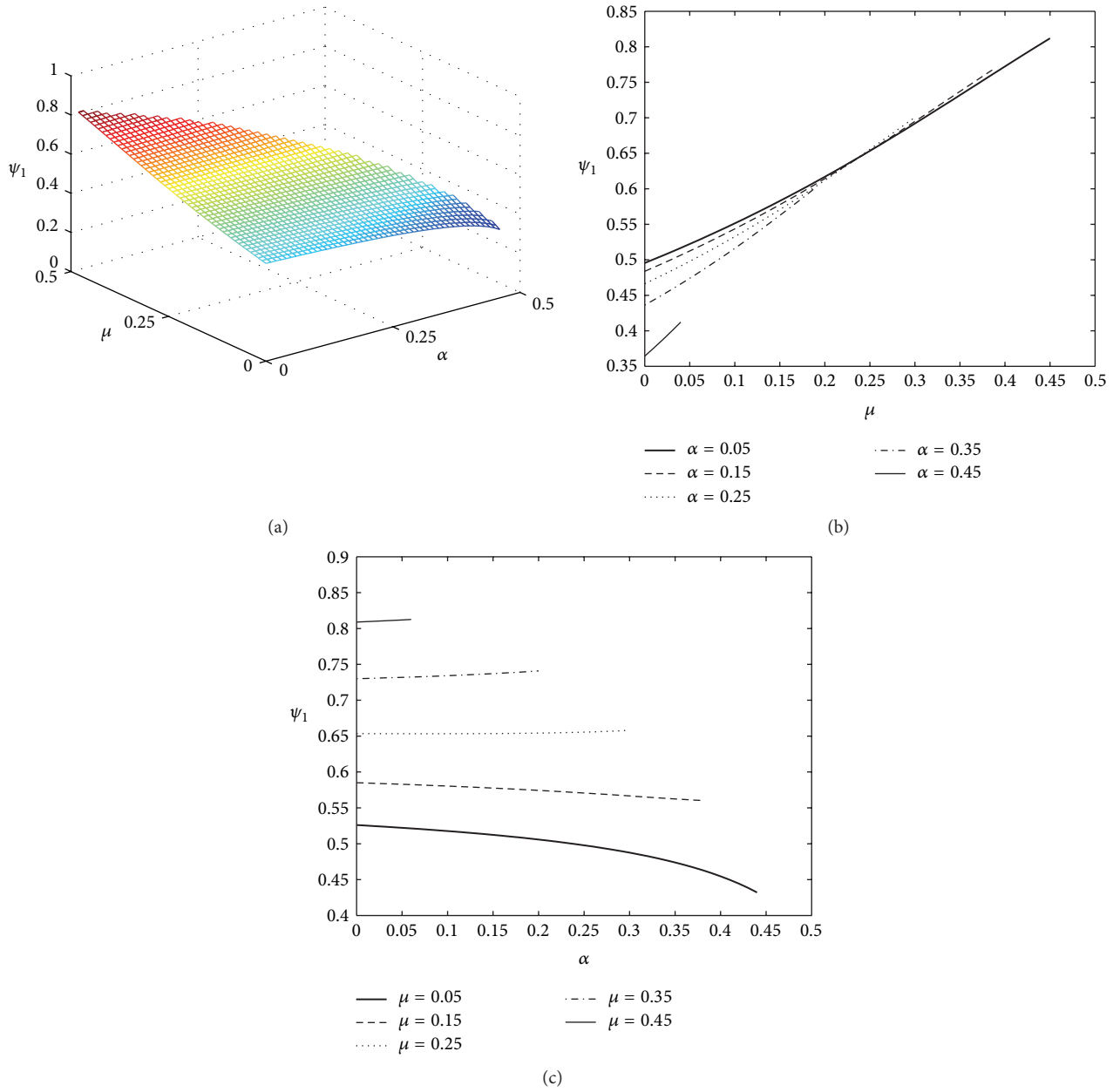


FIGURE 2: (a) Surface of $\psi_1$ at different $\alpha$ and $\mu$, for the infinite-allele birth and death process. (b) Relation between $\psi_1$ and $\mu$ when fixing $\alpha$. (c) Relation between $\psi_1$ and $\alpha$ when fixing $\mu$.

where

$$C(t) = a\int_0^t \left[q_{1j}^2(x) + \left(\sigma^2 + m^2\right)\beta_1^2(x)\right]e^{-a(t-x)}dx - \left[a\int_0^t q_{1j}(x)e^{-a(t-x)}dx\right]^2 - \left[am\int_0^t \beta_1(x)e^{-a(t-x)}dx\right]^2. \quad (14)$$

In Expression (14), $\beta_1(x) = ae^{\lambda x}\int_0^x e^{-\lambda u}q_{1j}(u)du$, and $\sigma^2$ is the variance of the offspring distribution, regardless of the allelic types.

Similarly as in Expression (3), we may define a limiting variance frequency spectrum $\xi_j = \lim_{t\to\infty}\eta_{i,t}(j)/(E_i[K_t])^2$. Expression (13) is complicated and usually does not assume an explicit form, even for the special case of the birth and death process. Therefore, we will only give numerical solutions for the finite-time variance frequency spectrum. Figure 4 shows an example of the "2$\sigma$"-bands of the finite-time frequency spectrum for the infinite-allele birth and death

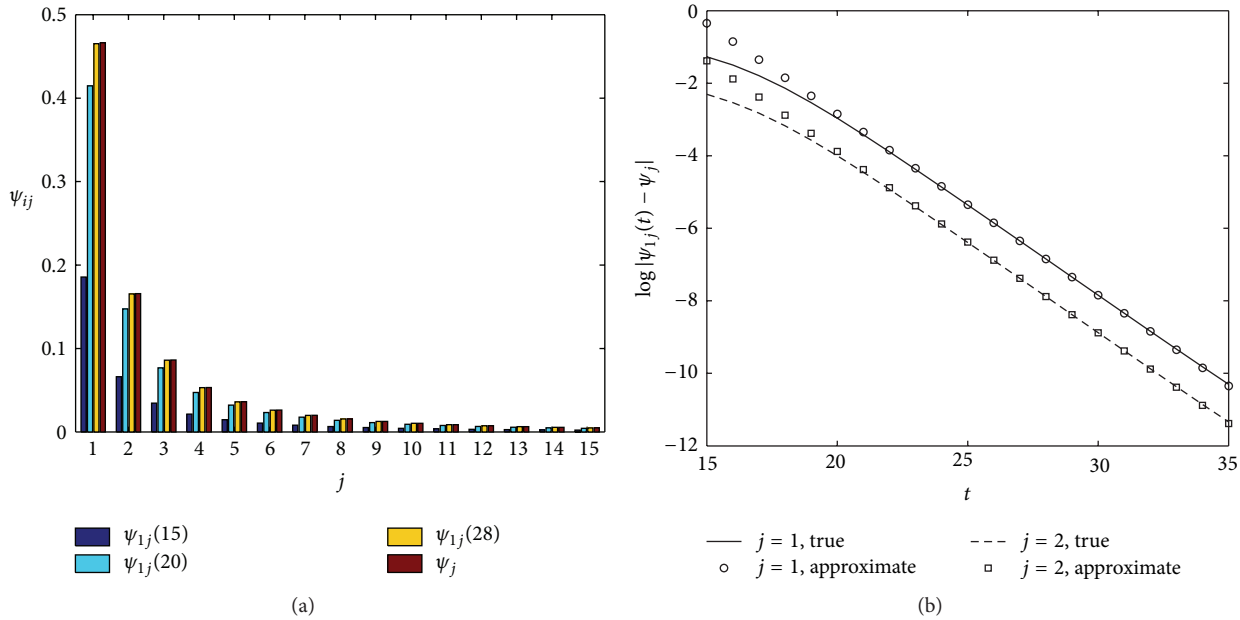


FIGURE 3: Comparison between the finite-time mean frequency spectrum and the limiting mean frequency spectrum of the infinite-allele birth and death process. (a) $\psi_{1j}(t)$ and $\psi_j$ for $1 \leq j \leq 15$, $t = 15, 20, 28$ and $\infty$. (b) Difference between $\psi_{1j}(t)$ and $\psi_j$ as a function of $t$, $t \in [15, 35]$ for $j = 1, 2$. Lines represent the true difference and markers represent asymptotic approximations.

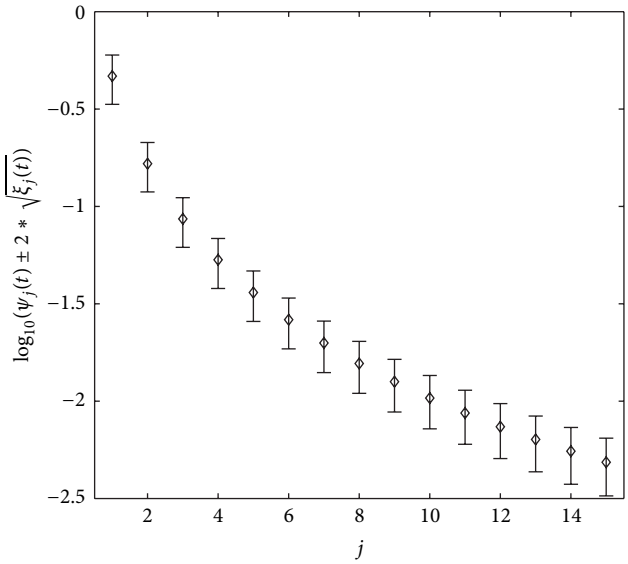


FIGURE 4: "$2\sigma$"-bands in semi-logarithmic scale of the finite-time ($t = 28$) frequency spectrum for the infinite-allele birth and death process with $i = 100$, $a = 1$, $\alpha = 0.25$, and $\mu = 10^{-4}$.

process with $i = 100$, $\alpha = 0.25$, $\mu = 10^{-4}$, $a = 1$, and $t = 28$. To emphasize the tail probabilities, we draw this plot in semilogarithmic scale.

From the finite-time variance frequency spectrum, it is possible to define a CI $[\theta_l, \theta_u]$ where the upper and lower bounds can be written as

$$\arg\min_{\theta} \left\| \psi_{\text{obs}}(t) - \left[ \psi_{\text{exp}}(\theta, t) \pm 2\sqrt{\xi_{\text{exp}}(\theta, t)} \right] \right\|_2. \quad (15)$$

This CI is useful for checking model validity and for testing whether two observed mean frequency spectra are from the same IAMBP model.

## 4. Simulation Study

We perform a simulation study of the birth and death process to illustrate the finite-time mean and variance frequency spectra. First we generate samples (genealogical trees) from an IAMBP with birth and death offspring distribution starting from 100 individuals carrying the same parental allele. Due to memory restrictions caused by forward simulation, we limit our simulations to 12 generations and a relatively large mutation probability $\mu = 0.01$. The other parameters of the process are set to be $a = 1$ and $\alpha = 0.25$. At time $t = 2$, we record the number of alleles $\alpha_t(j)$ represented by $j$ copies, for $j = 1, 2, \ldots$. Repeating the simulation 1000 times, we then obtain the simulated finite-time mean and variance frequency spectra from the replicates.

Figure 5 shows side-by-side bar plots of the simulated and expected finite-time mean and variance frequency spectra, for $j = 1, \ldots, 10$. We plot the mean frequency spectrum and the variance frequency spectrum in semi-logarithmic scale to emphasize the tail probabilities. In each bar plot, the first black bar represents the expected finite-time mean frequency spectrum $\psi_j$ or variance frequency spectrum $\xi_j$. The remaining ten white bars represent ten replicates of the simulated finite-time mean or variance frequency spectrum as described above. We see that for some classes, the expected mean or variance frequency spectrum is slightly different from the simulated spectrum. Beside sampling bias, this may be caused by the small scale of the simulations. For small mutation probability $\mu$, we have to set large initial population

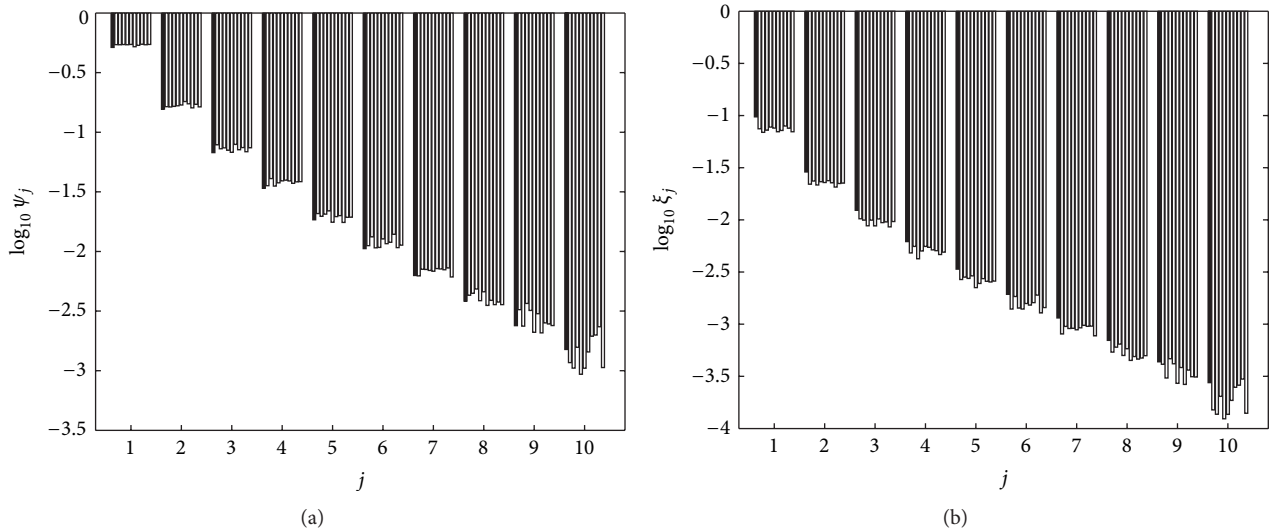


FIGURE 5: Comparison between the simulated and expected finite-time ($t = 2$) frequency spectra of the infinite-allele birth and death process with $i = 100$, $a = 1$, $\alpha = 0.25$, and $\mu = 0.01$. In each class $j$, $1 \le j \le 10$, the first black bar represents the expected frequency spectrum and the rest 10 white bars represent 10 replicates of the simulated frequency spectrum. (a) mean frequency spectrum $\psi_j$ in semi-logarithmic scale; (b) variance frequency spectrum $\xi_j$ in semi-logarithmic scale.

size $i$ and a long time $t$ to obtain acceptable values of $\psi_{i,t}(j)$ and $\xi_{i,t}(j)$ from the simulated genealogical trees. We note that if one tries to use a naive method to calculate the variance frequency spectrum, that is, assume the proportions $\psi_j$ of alleles having $j$ representatives to be the mean of some independent Bernoulli random variables (they are not independent) and employ $\xi_j = \psi_j(1 - \psi_j)$, such method performs much worse than $\xi_j$ based on the derivation of the variance frequency spectrum.

## 5. Summary

In this paper, we consider the frequency spectrum of the IAMBP of Pakes [2]. We develop an explicit expression for the limiting mean frequency spectrum for the special case of the birth and death process, which can be stated in terms of the hypergeometric function. We also derive an asymptotic expression for the rate of convergence of the finite-time mean frequency spectrum to the limiting mean frequency spectrum and illustrate the convergence using the birth and death process. We further state and prove a theorem concerning the variance frequency spectrum of the IAMBP, which helps to quantify uncertainty in parameter estimation and hypothesis testing. We illustrate the results using simulations of the birth and death process case.

As noted in the introduction, the frequency spectrum is similar to the Ewens' sampling formula [5] in population genetics, since they both concern the count or frequency spectrum based on an infinite-allele model under neutral selection. However, they differ in several aspects. (1) Our frequency spectrum describes population property under a branching process, whereas the Ewens' sampling formula describes allelic class count probabilities caused by a sampling procedure and further requires the sample size $n$ to be small compared to the size of the whole population which is assumed constant. (2) Our results concerning the frequency spectrum only provide the first and second moments and not the distribution function of the proportion of alleles having a given number of copies at a specified time point, whereas the Ewens' sampling formula gives the joint probability of all allelic classes. We also note that the Poisson-Dirichlet process [8] is usually used to describe the equilibrium behavior of the neutral infinite-allele model. Study of the relation between variations of the frequency spectrum under different models is of our future interest.

The question of validity of the Wright-Fisher and Moran models of population genetics [3], as compared to stochastic population processes such as the IAMBP or O'Connell process [9], has importance for estimation of parameters based on genetic data. As an example, Cyran and Kimmel [10] compared estimates of the age of the Mitochondrial Eve based on the various versions of the Wright-Fisher model with those based on various branching process models. The outcomes showed differences of about 10–15%.

## Appendices

## A. Derivation of $G_j$, $j \ge 0$ for the Birth and Death Process

For the birth and death process, the offspring pgf of the like-type individuals clone assumes either the form $h_1(s) = \alpha + \beta[\mu + (1 - \mu)s]^2$ or the form $h_2(s) = [\alpha + \beta\mu + \beta(1 - \mu)s]^2$. In both cases, the backward Kolmogorov equation gives a unified expression for the process pgf $F(s, t)$:

$$\begin{aligned} \frac{\partial F}{\partial t} &= a\left[B^2F^2 - \left(A^2 + B^2\right)F + A^2\right] \\ &= a\left(A^2 + B^2\right)\left[\frac{1}{A^2 + B^2}\left(A^2 + B^2F^2\right) - F\right], \end{aligned} \tag{A.1}$$

under different parameterizations, where for offspring pgf $h_1$, $A^2 = \alpha + \beta\mu^2$ and $B^2 = \beta(1-\mu)^2$, whereas for offspring pgf $h_2$, $A^2 = (\alpha + \beta\mu)^2$ and $B^2 = \beta^2(1-\mu)^2$. We see that this process is equivalent to a birth and death process with $\widetilde{a} = a(A^2 + B^2)$ and offspring pgf $\widetilde{h}(s) = (1/(A^2 + B^2))(A^2 + B^2 s^2)$. Using the known result of the birth and death process pgf [6], we obtain

$$F(s,t) = \frac{A^2(1-s) - \left(A^2 - B^2 s\right)e^{-ct}}{B^2(1-s) - \left(A^2 - B^2 s\right)e^{-ct}}, \tag{A.2}$$

where $c = a(B^2 - A^2) > 0$.

To obtain an explicit form for $G_j$, we may use two approaches. The first approach is to start from finding the pgf of $G_j$, which then leads to $G_j$. The second approach is to find $q_{1j}(t)$ directly and then obtain $G_j$. Both approaches lead to the same result. Here, we give derivation for the second approach only.

From the explicit form of $F(s,t)$, we can directly read $q_{10}(t)$ and $q_{1j}(t)$, $j \geq 1$. Consider

$$\begin{aligned} F(s,t) &= \frac{A^2\left(1-e^{-ct}\right)}{B^2 - A^2e^{-ct} + B^2\left(e^{-ct}-1\right)s} \\ &\quad + \frac{\left(B^2e^{-ct} - A^2\right)s}{B^2 - A^2e^{-ct} + B^2\left(e^{-ct}-1\right)s} \\ &= \frac{A^2\left(1-e^{-ct}\right)}{B^2 - A^2e^{-ct}} \cdot \frac{1}{1 - \left(B^2\left(1-e^{-ct}\right)/\left(B^2 - A^2e^{-ct}\right)\right)s} \\ &\quad + \frac{B^2e^{-ct} - A^2}{B^2 - A^2e^{-ct}} \cdot \frac{s}{1 - \left(B^2\left(1-e^{-ct}\right)/\left(B^2 - A^2e^{-ct}\right)\right)s}. \end{aligned} \tag{A.3}$$

Let $w_1 = A^2(1-e^{-ct})/(B^2 - A^2)e^{-ct}$, $w_2 = (B^2e^{-ct} - A^2)/(B^2 - A^2)e^{-ct}$, and $\widetilde{p} = (B^2 - A^2)e^{-ct}/(B^2 - A^2e^{-ct})$, the above expression becomes

$$F(s,t) = w_1\frac{\widetilde{p}}{1 - s\left(1-\widetilde{p}\right)} + w_2\frac{s\widetilde{p}}{1 - s\left(1-\widetilde{p}\right)}, \tag{A.4}$$

and $w_1 + w_2 = 1$.

This is a mixture of two geometric pgf's with the same parameter $\widetilde{p}$ but different supports, one is the set $\{0, 1, \ldots\}$, the other is the set $\{1, 2, \ldots\}$. Therefore,

$$q_{10}(t) = w_1\widetilde{p} = \frac{A^2\left(1-e^{-ct}\right)}{B^2 - A^2e^{-ct}}, \tag{A.5}$$

$$\begin{aligned} q_{1j}(t) &= w_1\widetilde{p}(1-\widetilde{p})^j + w_2\widetilde{p}(1-\widetilde{p})^{j-1} \\ &= \frac{\left(B^2 - A^2\right)^2\left[B^2\left(1-e^{-ct}\right)\right]^{j-1}e^{-ct}}{\left(B^2 - A^2e^{-ct}\right)^{j+1}}, \quad j \geq 1. \end{aligned} \tag{A.6}$$

Hence,

$$\begin{aligned} G_0 &= \int_0^\infty e^{-\lambda t}q_{10}(t)\,dt \\ &= \int_0^\infty \frac{A^2\left(1-e^{-ct}\right)}{\left(B^2 - A^2e^{-ct}\right)}e^{-\lambda t}dt \\ &= \frac{1}{c}\frac{A^2}{B^2}\int_0^1 \frac{v^{(\lambda/c)-1}(1-v)}{1 - \left(A^2/B^2\right)v}dv \\ &= \frac{1}{c}\frac{A^2}{B^2}\frac{\Gamma(\lambda/c)\,\Gamma(2)}{\Gamma(2 + (\lambda/c))}F\left(1, \frac{\lambda}{c};\ 2 + \frac{\lambda}{c}; \frac{A^2}{B^2}\right), \end{aligned} \tag{A.7}$$

$$\begin{aligned} G_j &= \int_0^\infty e^{-\lambda t}q_{1j}(t)\,dt \\ &= \int_0^\infty \frac{\left(B^2 - A^2\right)^2\left[B^2\left(1-e^{-ct}\right)\right]^{j-1}}{\left(B^2 - A^2e^{-ct}\right)^{j+1}}e^{-(\lambda+c)t}dt \\ &= \frac{\left(B^2 - A^2\right)^2}{cB^4}\int_0^1 \frac{v^{\lambda/c}(1-v)^{j-1}}{\left(1 - \left(A^2/B^2\right)v\right)^{j+1}}dv \\ &= \frac{1}{c}\left(1 - \frac{A^2}{B^2}\right)^2\frac{\Gamma(1 + (\lambda/c))\,\Gamma(j)}{\Gamma\left(j + 1 + (\lambda/c)\right)} \\ &\quad \times F\left(j+1, 1 + \frac{\lambda}{c}; j + 1 + \frac{\lambda}{c}; \frac{A^2}{B^2}\right), \quad j \geq 1. \end{aligned} \tag{A.8}$$

## B. Derivation of the Variance Frequency Spectrum for IAMBP

Let $T_1, T_2, \ldots$ be the successive split times, let $N_t$ be the number of split times till time $t$, and let $U_n$ be the number of offspring produced at split time $T_n$. Consider that at time $t$, the alleles which are represented by $j$ individuals are from two sources: the initial allele or the mutant alleles. Correspondingly, we define two indicator functions. $I_{0,j}(t) = 1$ if there are $j$ individuals carrying the initial allele alive at time $t$, and $I_{n,k,j}(t) = 1$, for $n, k \geq 1$ if the $k$th individual born at time $T_n$ ($T_n < t$) mutates to a new allelic type and further produces $j$ individuals carrying this allele $t$ units later. Then

$$\alpha_t(j) = I_{0,j}(t) + \sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}\left(t - T_n\right). \tag{B.1}$$

For each $n$, $I_{n,k,j}(t)$ is independent of $U_n$ and $T_n$, as well as $I_{0,j}(t)$, and it can be seen that $E_i[I_{0,j}(t)] = q_{ij}(t)$, $\mathrm{Var}_i(I_{0,j}(t)) = q_{ij}(t)[1 - q_{ij}(t)]$, $E[I_{n,k,j}(t) \mid U_n, T_n] = \mu q_{1j}(t)$, $\mathrm{Var}(I_{n,k,j}(t) \mid$

$U_n, T_n) = \mu q_{1j}(t)[1-\mu q_{1j}(t)]$. By the law of total variance, the variance frequency spectrum takes the form

$$\begin{aligned}
&\eta_{i,t}(j) \\
&\quad = \mathrm{Var}_i(\alpha_t(j)) \\
&\quad = \mathrm{Var}_i\left(I_{0,j}(t)\right) + \mathrm{Var}_i\left(\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\right) \\
&\quad = q_{ij}(t)\left[1-q_{ij}(t)\right] \\
&\qquad + \mathrm{Var}_i\left(E\left[\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\middle| N_t\right]\right) \\
&\qquad + E_i\left[\mathrm{Var}\left(\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\middle| N_t\right)\right]
\end{aligned} \tag{B.2}$$

For the second term, we know from independence among the indicator functions conditional on $N_t$ that

$$\begin{aligned}
&\mathrm{Var}_i\left(E\left[\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\middle| N_t\right]\right) \\
&\qquad = m^2\mu^2\mathrm{Var}_i\left(\sum_{n=1}^{N_t} q_{1j}(t-T_n)\right).
\end{aligned} \tag{B.3}$$

The variance on the right hand side can be obtained from the following theorem, which is an analogue to Lemma 3.1.1 in [2]: $\beta_i(t) = E_i[\sum_{n=1}^{N_t}\alpha(t-T_n)] = iae^{\lambda t}\int_0^t e^{-\lambda x}\alpha(x)dx$.

**Theorem B.1.** *Let $\alpha(t)$ be a bounded continuous function. Then*

$$\begin{aligned}
\gamma_i(t) &= \mathrm{Var}_i\left(\sum_{n=1}^{N_t}\alpha(t-T_n)\right) \\
&= iC(t) + i(\lambda+a)e^{\lambda t}\int_0^t e^{-\lambda x}C(x)\,dx,
\end{aligned} \tag{B.4}$$

*where*

$$\begin{aligned}
C(t) &= a\int_0^t\left[\alpha^2(x) + \left(\sigma^2+m^2\right)\beta_1^2(x)\right]e^{-a(t-x)}dx \\
&\quad - \left[a\int_0^t \alpha(x)e^{-a(t-x)}dx\right]^2 \\
&\quad - \left[am\int_0^t \beta_1(x)e^{-a(t-x)}dx\right]^2.
\end{aligned} \tag{B.5}$$

*Proof.* By independence of family lines, $\gamma_i(t) = i\gamma_1(t)$. By the law of total variance,

$$\begin{aligned}
\gamma_1(t) &= \mathrm{Var}_1\left(\sum_{n=1}^{N_t}\alpha(t-T_n)\right) \\
&= \mathrm{Var}\left(E\left[\sum_{n=1}^{N_t}\alpha(t-T_n)\mid T_1, U_1\right]\right) \\
&\quad + E\left[\mathrm{Var}\left(\sum_{n=1}^{N_t}\alpha(t-T_n)\mid T_1, U_1\right)\right] \\
&= \mathrm{Var}\left(\alpha(t-T_1) + U_1 E\left[\sum_{n=1}^{N_t'}\alpha\left(t-T_1-T_n'\right)\mid T_1\right]\right) \\
&\quad + E\left[U_1\,\mathrm{Var}\left(\sum_{n=1}^{N_t'}\alpha\left(t-T_1-T_n'\right)\mid T_1\right)\right],
\end{aligned} \tag{B.6}$$

where $T_n' = T_n - T_1$ and $N_t'$ is the number of split times in $(T_1, t]$. The right hand side can be further written as

$$\begin{aligned}
&E\left[\alpha^2(t-T_1)\right] - E^2\left[\alpha(t-T_1)\right] + \left(\sigma^2+m^2\right)E\left[\beta_1^2(t-T_1)\right] \\
&\quad - m^2E^2\left[\beta_1(t-T_1)\right] + mE\left[\gamma_1(t-T_1)\right] \\
&\quad = a\int_0^t\left[\alpha^2(t-x) + \left(\sigma^2+m^2\right)\beta_1^2(t-x)\right. \\
&\qquad \left. + m\gamma_1(t-x)\right]e^{-ax}dx \\
&\qquad - \left[a\int_0^t \alpha(t-x)e^{-ax}dx\right]^2 - \left[am\int_0^t \beta_1(t-x)e^{-ax}dx\right]^2 \\
&\quad = am\int_0^t \gamma_1(t-x)e^{-ax}dx + C(t).
\end{aligned} \tag{B.7}$$

Differentiating both sides and solving the resulting differential equation, we obtain

$$\begin{aligned}
\gamma_1(t) &= e^{\lambda t}\int_0^t e^{-\lambda x}\left[C'(x) + aC(x)\right]dx \\
&= C(t) + (\lambda+a)e^{\lambda t}\int_0^t e^{-\lambda x}C(x)\,dx.
\end{aligned} \tag{B.8}$$

□

Replacing $\alpha(x)$ by $q_{1j}(x)$ and replacing $\beta_1(x)$ by $ae^{\lambda x}\int_0^x e^{-\lambda u}q_{1j}(u)du$, we see that (B.3) becomes

$$\begin{aligned}
&\mathrm{Var}_i\left(E\left[\sum_{n=1}^{N_t}\sum_{k=1}^{U_n} I_{n,k,j}(t-T_n)\middle| N_t\right]\right) \\
&\qquad = im^2\mu^2\left[C(t) + (\lambda+a)e^{\lambda t}\int_0^t e^{-\lambda x}C(x)\,dx\right].
\end{aligned} \tag{B.9}$$

For the third term, we have the expression inside the expectation as

$$\begin{aligned}
&\sum_{n=1}^{N_t} \mathrm{Var}\left(\sum_{k=1}^{U_n} I_{n,k,j}\left(t-T_n\right)\right) \\
&\quad = \sum_{n=1}^{N_t}\left\{\mathrm{Var}\left(E\left[\sum_{k=1}^{U_n} I_{n,k,j}\left(t-T_n\right) \mid U_n, T_n\right]\right)\right. \\
&\qquad \left. + E\left[\mathrm{Var}\left(\sum_{k=1}^{U_n} I_{n,k,j}\left(t-T_n\right) \mid U_n, T_n\right)\right]\right\} \\
&\quad = \sum_{n=1}^{N_t}\left[m\mu q_{1j}\left(t-T_n\right) + \left(\sigma^2 - m\right)\mu^2 q_{1j}^2\left(t-T_n\right)\right],
\end{aligned} \tag{B.10}$$

where $\sigma^2$ is the variance of the offspring distribution, regardless of the allelic types. Therefore,

$$\begin{aligned}
&E_i\left[\sum_{n=1}^{N_t} \mathrm{Var}\left(\sum_{k=1}^{U_n} I_{n,k,j}\left(t-T_n\right)\right)\right] \\
&\quad = m\mu E_i\left[\sum_{n=1}^{N_t} q_{1j}\left(t-T_n\right)\right] \\
&\qquad + \left(\sigma^2 - m\right)\mu^2 E_i\left[\sum_{n=1}^{N_t} q_{1j}^2\left(t-T_n\right)\right] \\
&\quad = iam\mu e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)\,dx + ia\left(\sigma^2 - m\right)\mu^2 e^{\lambda t} \\
&\qquad \times \int_0^t e^{-\lambda x} q_{1j}^2(x)\,dx
\end{aligned} \tag{B.11}$$

By (B.2), (B.9) and (B.11), the final expression of $\eta_{i,t}(j)$ is then

$$\begin{aligned}
\eta_{i,t}(j) &= q_{ij}(t)\left[1 - q_{ij}(t)\right] \\
&\quad + im^2\mu^2\left[C(t) + (\lambda + a)e^{\lambda t}\int_0^t e^{-\lambda x} C(x)\,dx\right] \\
&\quad + iam\mu e^{\lambda t}\int_0^t e^{-\lambda x} q_{1j}(x)\,dx + ia\left(\sigma^2 - m\right)\mu^2 e^{\lambda t} \\
&\quad \times \int_0^t e^{-\lambda x} q_{1j}^2(x)\,dx.
\end{aligned} \tag{B.12}$$

## Acknowledgments

The authors thank an anonymous referee for suggesting an asymptotic expression, which following some revisions, became (5). Part of M. Kimmel's work has been carried out in Fall 2011 when he was visiting the Institute of Advance Study at the Warwick University, supported by grant from IAS and EPSCR. His work was also supported by the Polish NCN Grant NN519579938.

## References

[1] R. C. Griffiths and A. G. Pakes, "An infinite-alleles version of the simple branching process," *Advances in Applied Probability*, vol. 20, no. 3, pp. 489–524, 1988.

[2] A. G. Pakes, "An infinite alleles version of the Markov branching process," *Journal of Australian Mathematical Society A*, vol. 46, no. 1, pp. 146–169, 1989.

[3] W. J. Ewens, *Mathematical Population Genetics*, vol. 9, Springer, New York, NY, USA, 1979.

[4] M. Kimmel and M. Mathaes, "Modeling neutral evolution of Alu elements using a branching process," *BMC Genomics*, vol. 11, supplement 1, article S11, 2010.

[5] W. J. Ewens, "The sampling theory of selectively neutral alleles," *Theoretical Population Biology*, vol. 3, pp. 87–112, 1972.

[6] K. B. Athreya and P. E. Ney, *Branching Processes*, Springer, Berlin, Germany, 1972.

[7] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, NY, USA, 1972.

[8] J. F. C. Kingman, S. J. Taylor, A. G. Hawkes et al., "Random discrete distribution," *Journal of the Royal Statistical Society B*, vol. 37, pp. 1–22, 1975.

[9] N. O'Connell, "The genealogy of branching processes and the age of our most recent common ancestor," *Advances in Applied Probability*, vol. 27, no. 2, pp. 418–442, 1995.

[10] K. A. Cyran and M. Kimmel, "Alternatives to the Wright-Fisher model: the robustness of mitochondrial Eve dating," *Theoretical Population Biology*, vol. 78, no. 3, pp. 165–172, 2010.

Hindawi Publishing Corporation
International Journal of Stochastic Analysis
Volume 2013, Article ID 826321, 23 pages
http://dx.doi.org/10.1155/2013/826321

*Review Article*

# Simulating the Emergence of Mutations and Their Subsequent Evolution in an Age-Structured Stochastic Self-Regulating Process with Two Sexes

**Charles J. Mode,[1] Candace K. Sleeman,[2] and Towfique Raj[3]**

[1] *Department of Mathematics, Drexel University, Philadelphia, PA 19104, USA*
[2] *Nokia Corporation, 5 Wayside Road, Burlington, MA 01803, USA*
[3] *Division of Genetics, Harvard Medical School, Brigham and Women's Hospital, Boston, MA 02115, USA*

Correspondence should be addressed to Charles J. Mode; cjmode@comcast.net

Received 28 August 2012; Revised 30 November 2012; Accepted 3 December 2012

Academic Editor: Peter Olofsson

Copyright © 2013 Charles J. Mode et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The stochastic process under consideration is intended to be not only part of the working paradigm of evolutionary and population genetics but also that of applied probability and stochastic processes with an emphasis on computer intensive methods. In particular, the process is an age-structured self-regulating multitype branching process with a genetic component consisting of an autosomal locus with two alleles for females and males. It is within this simple context that mutation will be quantified in terms of probabilities that a given allele mutates to the other per meiosis. But, unlike many models that are currently being used in mathematical population genetics, in which natural selection is often characterized in terms of parameters called fitness by genotype or phenotype, in this paper the parameterization of submodules of the model provides a framework for characterizing natural selection in terms of some of its components. One of these modules consists of reproductive success that is quantified in terms of the total expected number of offspring a female contributes to the population throughout her fertile years. Another component consists of survival probabilities that characterize an individual's ability to compete for limited environmental resources. A third module consists of a parametric function that expresses the probabilities of survival in a birth cohort of individuals by age for both females and males. A forth module of the model as an acceptance matrix of conditional probabilities such female may show a preference for the genotype or phenotype as her male sexual partner. It is assumed that any force of natural selection acts at the level of the three genotypes under consideration for each sex. By assigning values of the parameters in each of the modules under consideration, it is possible to conduct Monte Carlo simulation experiments designed to study the effects of each component of selection separately or in any combination on a population evolving from a given initial population over some specified period of time.

## 1. Introduction

Comparatively little attention has been given to models on the evolution of age-structured populations in the extensive literature on evolutionary and population genetics. Much of the work on this class of models has, however, been reviewed and extended in the book by Charlesworth [1] on evolution in age structured populations. This book also contains a long list of references citing papers that are quite well known by, among others, Norton, Haldane, and Fisher, but it is beyond the scope of this paper to provide an overview of the content of these early works. It is of interest to note, however, that Norton began work on models of age structured populations as early as 1910, soon after Mendel's laws of inheritance had become widely known. For the most part, models on the evolution of age structured populations that have been presented in this literature belong to the deterministic paradigm, but in this paper a special class stochastic models on the evolution of age structured populations will be the focus of attention.

This special class of stochastic processes is rooted in ideas from branching processes that also have an extensive literature dating back to about 100 years. Rather than attempting to cite papers from this extensive literature, books on branching processes will be cited which do contain extensive lists of literature. During the second half of the 20th century, a book

by Harris [2] provided a stimulus for other workers enter the field of branching processes. In the period that followed the publication of this book, other books on this subject were published. Included in these books are those of Athreya and Ney [3], Jagers [4], and Mode [5]. In particular, the class of branching process under consideration in this paper is an extension of the general class of branching processes presented in Jagers, that contains a one type formulation of this process and Mode, which contains a multitype version of the general branching process. The multitype case of the general processes is the most useful when formulating processes applicable to evolutionary and population genetics, because this case accommodates not only mutations among the types but also differences in reproductive success among the types as well as the age structure of an evolving population. Other more recent books on branching processes are Asmussen and Hering [6], Kimmel and Axelrod [7], and Haccou et al. [8]. Each of these books has an extensive list of references that an interested reader may wish to explore.

The working paradigm underlying the class of branching processes to be formulated and studied in this paper differs markedly from that used in the books cited above. In these books, as well as the literature on branching processes published elsewhere, emphasis was placed on finding threshold conditions under which the population either became extinct or its total number, total population size, would increase without bound. Perhaps the most significant departure from the working paradigm of classical branching processes considered in this paper is that the process is self-regulating in the sense that its total size is dependent on the carrying capacity of the environment expressed in terms of parameters of survival functions, in which the parameter values may differ among genotypes. Because the survival of individuals during each time period under consideration is a non-linear function of total population size, the formulation being considered has nonlinear properties which arise in connection with the interaction of individuals in the population, as they compete for food and other resources. There are also other non-linear properties of the formulation under consideration that arise in connection with females selecting male sexual partners by genotypes that leads to the production of offspring that are necessary for the long-term survival of a population. Given a formulation that accommodates selection of mates by genotype, it is possible to study the effects of sexual selection on the long-term evolution of an age-structured population. For the most part, interactions among individuals were not accommodated in the formulations of classical branching processes in the books cited above, but the more recent book by Haccou et al. [8], where a one type a density dependent process is formulated in terms of, among other things, the well-known logistic function that played a role in the development of chaos theory that may be of interest to readers, but it is beyond the scope of this paper to present further details here. It should also be mentioned that, unlike the classical analysis of branching processes that often focuses on limit theorems, it is possible to study the long transient period of evolution a process before the process converges to some limit by using Monte Carlo simulation methods.

There is another aspect of the formulation to be considered in this paper which is also a significant departure from the working paradigm of classical branching processes, namely, the embedding of systems of nonlinear difference equations in the stochastic processes, which will be discussed in detail in a subsequent section and will be referred to as the embedded deterministic model. After software had been written to provide a computer implementation of the embedded deterministic model, given a set of numerical assignments to the parameters, it was possible to conduct computer simulation experiments, such that 10,000 to 20,000 years of evolution of an age structured population may be simulated with a computer running time from one to four minutes. Given this rapid computer response time, it was easy to determine whether a particular set of assigned numerical values of the parameters belong to a set of points in the parameter space, such that the population becomes extinct or grows to a total size where environmental resources limit further growth. Consequently, in this paper the task of finding criteria that partition a multidimensional parameter space into regions such that, according to the embedded deterministic model, the population either becomes extinct or its total size approaches some limit determined by the environment will not be under taken, but it is hoped that some analysts may be stimulated to undertake this task. After interesting regions of the parameter space have been found while using the embedded deterministic model, a Monte Carlo simulation experiment is conducted in which sample realizations of the sample functions of the stochastic process are computed using the same parameter values as those used in an experiment with embedded deterministic model to test the extent to which the predictions of deterministic are consistent with those of the process. As will be shown by examples, the predictions of the embedded deterministic process are not always consistent with those of the process, which is another facet of how the working paradigm of this paper differs from that of classical branching processes. It should also be mentioned that the formulation of an age structured process is essentially a complete revision of the age structured process considered in chapter 12 of Mode and Sleeman [9], where all reported computer experiments were based on applications of the embedded deterministic model.

## 2. Genotypes, Gametes, Mutation, and Types of Matings in an Age Structured Population

Only three genotypes with respect to two alleles, $A$ and $a$, at some autosomal locus will be considered, and let

$$\mathbb{G} = \{AA, Aa, aa\} \tag{1}$$

denote the set of three genotypes. To simplify the notation, elements of $\mathbb{G}$ will be denoted by the symbol $\tau$, and to distinguish the sexes, female and male genotypes will be denoted by $\tau_f$ and $\tau_m$, respectively. In what follows, the genotypes in the set $\mathbb{G}$ will sometimes be denoted by 1, 2, and 3 and will be made clear what sex is under consideration. The number of age classes under consideration is $r$ for both sexes, and the ages of females will be denoted by $x = 0, 1, 2, \ldots, r$, where

$x = 0$ denotes infants or newborns. Similarly, ages of males will be denoted by $y = 0, 1, 2, \ldots, r$. Let $x_m$ denote the minimum age at which fertile females may bear children, and let $x_{\max}$ denote the maximum age at which they fertile. Let $y = y_m, \ldots, r$ denote the ages of males, such that they may sire offspring. From now on such males will be referred to as fertile. It will be assumed for the sake of simplicity that $x_m = y_m$.

Mutations among the alleles $A$ and $a$, which will be assumed to occur during meiosis in both sexes, will be formulated in terms of conditional probabilities in the $2 \times 2$ matrix

$$\mathbb{M} = \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{pmatrix}, \tag{2}$$

where $\mu_{12}$ is the conditional probability that allele $A$ mutates to allele $a$ per meiosis. The conditional probability that allele does not mutate is $\mu_{11} = 1 - \mu_{12}$. Similarly, $\mu_{21}$ is the conditional probability that allele $a$ mutates to $A$ per meiosis and $\mu_{22} = 1 - \mu_{21}$. It will be assumed that these conditional probabilities hold for both sexes as well as throughout the fertile ages for both females and males.

The next step in the formulation of the model is to derive a gametic distribution for each of the three genotypes, which will be symbolized as follows. Let $i = 1$ denote allele $A$, and let $j = 2$ denote allele $a$, then the set of two alleles under consideration may be denoted by $\mathbb{A} = \{1, 2\}$. Similarly, the set of three genotypes will be denoted by $\mathbb{G} = \{(1, 1), (1, 2), (2, 2)\} = \{1, 2, 3\}$. When there are mutations among the two alleles, each of these genotypes may produce a gamete containing an allele $j \in \mathbb{A}$. Let $p_g(i; j)$ denote the conditional probability that genotype $i \in \mathbb{G}$ produces the gamete $j \in \mathbb{A}$. By way if an illustration, a formula for the conditional probability $p_g(1; 1)$ will be derived. It will be assumed that the allele on the left in the symbol $AA$ was contributed by the maternal parent of an individual and the one on the right was contributed by the paternal parent. The conditional probability that the allele on the left in the genotype $AA$ is contributed to the gamete pool of the population is $\mu_{11}/2$. Similarly, the conditional probability that the allele on the right is contributed to the gamete pool is $\mu_{11}/2$. By adding these probabilities it follows that $p_g(1; 1) = \mu_{11}$. By a similar argument, it can be shown that $p_g(1; 2) = \mu_{12}$. Furthermore, by continuing this line of reasoning, it can be shown that $p_g(2; 1) = (\mu_{11} + \mu_{21})/2$, $p_g(2; 2) = (\mu_{12} + \mu_{22})/2$, $p_g(3; 1) = \mu_{21}$, and $p_g(3; 2) = \mu_{22}$.

A mating between a female and male of genotypes $i$ and $j$, respectively, will be denoted by the mating type $\kappa = (i, j)$. Because there are three genotypes of each sex, it follows that there are 9 types of matings. Throughout this paper the set of mating types will be denoted by

$$\begin{aligned} \mathbb{T} = \{&(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), \\ &(2, 3), (3, 1), (3, 2), (3, 3)\} \end{aligned} \tag{3}$$

in the order indicated. For every mating type $\kappa \in \mathbb{T}$, let $p(\kappa; \nu)$ denote the conditional probability that a mating of type $\kappa = (i_1, i_2)$ produces an offspring of genotype $\nu = (i, j) \in \mathbb{G}$. By assumption females and males have the same gametic distributions, and it also assumed that female and male gamete combine independently. From these assumptions, it follows that

$$p(\kappa; \nu) = p_g(i_1, i)\, p_g(i_2, j), \tag{4}$$

for all couple types $\kappa = (i_1, i_2) \in \mathbb{T}$ and genotypes $\nu = (i, j) \in \mathbb{G}$. In computer implementations of the model under consideration, these conditional probabilities are computed in the order indicated in the set $\mathbb{T}$.

As discussed in Mode and Sleeman [9] in chapter 12, a module for couple formation, depending on the ages of the female and male, is an integral part of the formulation of an age dependent stochastic population process, then the number of couple types may become very large, which leads to the necessity of processing large arrays in computer implementations of the process. The processing of large arrays in a computer, in turn, becomes problematic, unless some scheme is adopted to reduce the number of couple types. It is of interest, therefore, to formulate a two sex age dependent population process in which sexual contacts may occur between females and males but without partnerships or couples which may last for long time periods. Accordingly, the purpose of this section is to formulate a process that includes sexual contacts between females and males but not partnerships of females and males of long duration but are transient and may vary from year to year among age groups of fertile females and males. This approach mating system modelled by this approach may be justified by assuming this type of polygamy was part of the evolution of the human species before the idea of long-term marital partnerships evolved.

## 3. Births in a Two Sex Age Dependent Population Process without Couple Formation

When an age structured population is under consideration, individuals in a population will also be classified by type. If for example, a female of genotype $\tau_f = i$ is of age $x$, then her type will be denoted by $\mathbf{t}_f = (i, x)$. Similarly, a male type will be denoted by $\mathbf{t}_m = (j, y)$, where $y$ is his age and $i$ and $j$ belong to the set $\mathbb{G}$ of genotypes. At this point in defining the components of the model, it should be mentioned that age and time will expressed in terms of a lattice: $T = (0, h, 2h, 3h, \ldots)$, where $h$ is some time unit. For example, if the evolution of a human population is under consideration, then usually $h$ will be a year. Hence, from now on $h = 1$. For every $t \in T$, let $\mathbf{X}(t)$ denote 3 by $r + 1$ matrix valued function with 3 rows and $r + 1$ columns, where $X(t; i, x)$ denotes the number of females to type $\mathbf{t}_f = (i, x)$ at time $t \in T$ in a population. The 3 by $r + 1$ matrix $\mathbf{Y}(t)$ is defined similarly for males. To reduce the size of the arrays to be processed in a computer, it will be assumed that sexual contacts between females and males do not depend on the age of the male but only on the frequency of the male's genotype. For $y = y_m, \ldots, r$, let the random function $Y(t; i, y)$ denote the number of males of type $\mathbf{t}_m = (i, y)$ in the population at time $t$. Observe that the random function $Y(t; i, y)$ is the number of individuals of

genotype $i$ and age $y$ in the males population at time $t \in T$. Then it follows that, the random function

$$Y(t;i) = \sum_{y=y_m}^{r} Y(t;i,y) \tag{5}$$

is the total number of fertile males of genotype $i$ in the population at time $t$. Therefore, the frequency of fertile males of genotype $i$ in the population at time $t$ is given by the random function

$$U_m(t;i) = \frac{Y(t;i)}{Y(t;\circ)}, \tag{6}$$

for $Y(t;\circ) > 0$, where

$$Y(t;\circ) = \sum_i Y(t;i) \tag{7}$$

and $U_m(t;i) = 0$ if $Y(t;\circ) = 0$, where $i = 1, 2, 3$.

With a view towards minimizing the size of arrays that need to be processed in a computer, it will be assumed that for each fertile female of age $x$ expresses preferences for males sexual partners by genotype. To simplify the notation, denote the three genotypes under consideration by $1 = AA$, $2 = Aa$, and $3 = aa$, and let

$$\mathbf{A}_f = \begin{pmatrix} \alpha_f(1,1) & \alpha_f(1,2) & \alpha_f(1,3) \\ \alpha_f(2,1) & \alpha_f(2,2) & \alpha_f(2,3) \\ \alpha_f(3,1) & \alpha_f(3,2) & \alpha_f(3,3) \end{pmatrix} \tag{8}$$

denote the $3\times3$ matrix of acceptance probabilities for females. For example, $\alpha_f(1,1)$ is the conditional probability that a female of genotype 1 finds a male of genotype 1 acceptable as a sexual partner, and in general, given a female of genotype $i$, $\alpha_f(i,j)$ is the conditional probability that she finds a male of genotype $j$ acceptable as a sexual partner. For the sake of simplicity, it will be assumed that this matrix does not depend on the age $x$ of a fertile female. If $\alpha_f(i,j) = 1$ for all pairs $(i,j)$, then, by definition, females select male genotypes at random for sexual contacts. But, if, for example, $\alpha_f(i,3)$ is greater than the other two probabilities for all $i = 1, 2, 3$, then all females prefer males of genotype 3 as sexual partners. This is an example of what is called sexual selection.

Given these definitions and an application of Bayes's rule, it follows that

$$\gamma_f(t;i,j) = \frac{U_m(t;j)\,\alpha_f(i,j)}{\sum_j U_m(t;j)\,\alpha_f(i,j)} \tag{9}$$

is the conditional probability that a female of genotype $i$ has a sexual contact with a male of genotype $j$ during the any time interval $[t;t+1)$, and let

$$\boldsymbol{\gamma}_f(t;i) = \left(\gamma_f(t;i,,j) \mid j = 1,2,3\right) \tag{10}$$

denote a $1\times3$ vector of these conditional probabilities. At time $t$, for $x = x_m, \ldots, x_{\max}$, let the random function $X(t;\tau_f,x)$ denote the number of fertile females of type $\mathbf{t}_f = (i,x)$ in the population at time $t$, and let the random function $Z(t;\mathbf{t}_f,\tau_m)$ denote the number of females of type $\mathbf{t}_f = (i,x)$ who have a sexual contacts with a males of genotype $\nu$ during the time interval $[t,t+h)$. Then, let

$$\mathbf{Z}(t;i,x) = (Z(t;i,x,\nu) \mid \nu = 1,2,3) \tag{11}$$

denote a $1\times3$ random vector whose elements are the indicated random functions. It will be assumed that this random vector has a conditional multinomial distribution with index $X(t;i,x)$ and probability vector $\boldsymbol{\gamma}_f(t;i)$. In symbols,

$$\mathbf{Z}(t;i,x) \sim \text{CMultinom}\left(X(t;i,x), \boldsymbol{\gamma}_f(t;i)\right), \tag{12}$$

for $x = x_m, \ldots, x_{\max}$ and $i = 1, 2, 3$. More precisely, the components of the vector are

$$\mathbf{Z}(t;i,x) = (Z(t;i,x,1), Z(t;i,x,2), Z(t;i,x,3)). \tag{13}$$

Observe for each $i = 1, 2, 3$ the elements in the vector $Z(t;i,x,j)$ for $j = 1, 2, 3$ are the random number of females of genotype $i$ and age $x$ that have sexual contacts with a males of genotype $j$. Altogether there are 9 types of sexual contacts when there are 3 genotypes of each sex as was shown in the previous section. The random numbers of these 9 types of contacts during any time interval $[t,t+1)$ is given by the random vector

$$\mathbf{Z}(t;x) = (\mathbf{Z}(t;i,x) \mid i = 1,2,3). \tag{14}$$

Sexual selection, whereby females of all genotypes prefer males of one genotype over the others, is thought to be one of the driving forces of natural selection. The process just described accommodates this component of natural selection. It should also be mentioned that this simple mating process just described not only simplifies the computer implementation of the two-sex model under consideration, but it may also be a plausible model for the evolution of a population prior to the time long-term monogamous sexual partnerships evolved.

Reproductive success is also thought to be a component of natural selection. For each fertile female, who has sexual contacts with a fertile male, this component will be characterized by the parameters $\lambda(i,x)$ denoting the expected number of offspring each fertile female type $\mathbf{t}_f = (i,x)$ contributes to the population during a time interval $[t,t+1)$ for $i = 1, 2, 3$. Human females are most fertile during their twenties, and the value of $\lambda(i,x)$ starting with $\lambda(i,x_m)$ will increase with $x$ up to about age 25 and then decreases to some value $\lambda(i,x_{\max})$ at age $x_{\max}$, the age at which a female is no longer fertile. Let $\lambda_i$ denote the expected number of offspring produced by a female of genotype $i$ throughout her fertile ages. Then, for every $i = 1, 2, 3$,

$$\sum_{x=x_m}^{x_{\max}} \lambda(i,x) = \lambda_i. \tag{15}$$

When assigning values to the parameters of a model in the process of desiging a computer experiment, suppose one assigns a value to each $\lambda_i$ for $i = 1, 2, 3$. For example, for the case of a population with high fertility and no selection among the three genotypes, then one could assign the value

$\lambda_i = 10$ for $i = 1, 2, 3$. Whereas, by way of an illustrative example, if it is assumed that selection is such that $\lambda_i$ increases with $i$, then the assigned values of the parameters could be $\lambda_1 = 10$, $\lambda_2 = 11$, and $\lambda_3 = 12$, so that, from the point of view of natural selection, genotype 3 would have a selective advantage over the other two genotypes.

Given (15), a question that naturally arises is how should one distribute the assigned value $\lambda_i$ over the fertile ages, such that (15) holds? One approach to answer to this question is to find a probability density $f(x)$, such that

$$\sum_{x=x_m}^{x_{\max}} f(x) = 1. \tag{16}$$

Then, if $\lambda(i, x) = f(x)\lambda_i$, it follows that

$$\sum_{x=x_m}^{x_{\max}} \lambda(i, x) = \lambda_i, \tag{17}$$

for all $i = 1, 2, 3$, so that (15) is satisfied. The density function $f(x)$ may be called the fertility distribution of females during the fertile ages.

Among the many choices of the density function $f(x)$ is a truncated Poisson distribution with parameter $\lambda_0 > 0$. It is well known that the mode of a Poisson density function is at the value $x = \lambda_0$ when $\lambda_0$ is a positive integer. If, for example, a female becomes fertile at age $x_m = 15$, then, under the assumption that her peak in fertility is at age 25, it follows that one should choose $\lambda_0 = 10$, so that the mode of the distribution is age at age $x = 15 + 10 = 25$. Next suppose that the maximum age a female is fertile is $x_{\max} = 40$. Then the number of numbers between $x_m$ and $x_{\max}$ starting at $x_m = 15$ and ending in 40 is $40 - 14 = 26$. To create a truncated Poisson distribution, compute the numerical values of the density function

$$g(y) = \exp[-\lambda_0] \frac{(\lambda_0)^y}{y!} \tag{18}$$

for $y = 0, 1, 2, \ldots, 25$. The next is to compute the normalizing constant

$$C = \sum_{y=0}^{25} g(y). \tag{19}$$

Let $f(x; \lambda_0, 40)$ denote the probability density function of Poisson distribution truncated at age 40. Then, for $x = 15 + y$

$$f(x; \lambda_0, 40) = \frac{g(y)}{C} \tag{20}$$

for $y = 0, 1, \ldots, 25$. Other choices of the density of the distribution of the age of child bearing could be made, but, for the sake of simplicity, in all the computer experiments reported in this paper, this distribution will be that of the truncated Poisson described above.

In what follows, a well-known property of the family of Poisson distributions will be used repeatedly. Let $X_i$ for $i = 1, 2, \ldots, n$ be a collection of independent Poisson random variables with expectations $\alpha_i > 0$ for all $i$. Furthermore, suppose each of these random variables takes values in the set of nonnegative integers $\{x \mid x = 0, 1, 2, 3 \ldots\}$. Then, the random variable $Y = X_1 + X_2 + \cdots + X_n$ has a Poisson distribution with expectation (parameter) $\beta = \alpha_1 + \alpha_2 + \cdots + \alpha_n$. It should be remarked that this property is easy to prove using probability generating functions. In what follows, it will be assumed that the number of females $Z(t, i, x, j)$ of age $x$ with sexual contacts of type $(i, j)$ produce offspring independently, and, moreover, females of different ages also produce offspring independently. It should also be mentioned that all conditional distributions and expectations are conditioned on the state of $(\mathbf{X}(t), \mathbf{Y}(t))$ at time $t$, so that the idea of conditional independence is in force.

In principle, under the assumption that the number $N_i$ of offspring produced by an individual female of genotype $i$ follows a Poisson distribution with expectation $\lambda_i$, then given a realization of the random variable $Z(t; i, x, j)$ and the expected value $\lambda(i, x) = \lambda_i f(x; \lambda_0, 40)$, it follows that the random number of offspring produced by females of genotype $i$ and age $x$ that had contacts with a males of genotype $j$ would follow a Poisson distribution with expectation $Z(t; i, x, j)\lambda(i, x)$. Then, because it is assumed that females of different ages also produce offspring independently, it follows that the random $W(t; i, j)$ denoting the total number of offspring produced by females of genotype $i$ having male sexual partners of genotype $j$ follows a Poisson distribution with parameter

$$\begin{aligned} &\gamma(t; i, j) \sum_{x=x_m}^{x_{\max}} Z(t; i, x, j) \lambda(i, x) \\ &\quad = \left( \sum_{x=x_m}^{x_{\max}} Z(t; i, x, j) f(x; \lambda_0, 40) \right) \lambda_i \end{aligned} \tag{21}$$

during the time interval $[t, t + 1)$ for all nine combinations of sexual partnerships $(i, j) \in \mathbb{T}$.

To simplify the notation, denote the sum in parentheses by $V(t; i, j)$. Then $\gamma(t; i, j)$ has the simpler form

$$\gamma(t; i, j) = V(t; i, j) \lambda_i. \tag{22}$$

From this formula, it can be seen that if $V(t; i, j) > 1$, then the expected number of total number of offspring produced by partnerships of type $(i, j)$ during the time interval $[t, t + 1)$ would exceed $\lambda_i$, indicating that if this relationship held forward in time, then the expected number of offspring produced by partnerships of type $(i, j)$ would increase over time. On the other hand, if $0 \leq V(t; i, j) < 1$ and if this relationship held forward in time, then eventually mating types of the form $(i, j)$ would disappear from the population. It should also be noted that if there are no individuals of genotype $i$ in the population at time $t$, then $V(t; i, j) = 0$ for all $j = 1, 2, 3$. It is of interest to observe that if the initial population consists only of individuals of genotype $i = 1 = AA$ for both sexes, then initially $V(t; i, j) = 0$ for $i = 2, 3$, until a sufficient numbers of mutant genotypes $Aa$ and $aa$ have arisen as a result of mutations before $V(t; i, j) > 0$ for $i = 2, 3$ and $j = 1, 2, 3$. Moreover, if the initial number of females of genotype $AA$ is

small, then population may become extinct before the mutant genotypes $Aa$ and $aa$ appear in the population. It is also of interest to observe that if for some $t \geq 0$ all individuals, the female and male population, are of genotype $1 = AA$, then $\gamma(t; i, j) = 0$ except for the case $i = j = 1$.

Therefore, when writing software to simulate realizations of Poisson random variables, zero valued parameters need to be accommodated. Symbolically, let CPois($\alpha$) denote a procedure for simulating a realization of a Poisson random variable with parameter $\alpha \geq 0$, given the state of the population at time $t$. Let the random variable $W(t; i, j)$ denote the number of matings of a female of genotype $i$ with a male of genotype $j$ during the time interval $[t, t+1)$. Then, if $\gamma(t; i, j) = 0$, then $W(t; i, j) = 0$, but if $\gamma(t; i, j) > 0$, then

$$W\left(t; i, j\right) \sim \text{CPois}\left(\gamma\left(t; i, j\right)\right), \tag{23}$$

indicating that the random variable $W(t; i, j)$ has a conditional Poisson distribution with parameter $\gamma(t; i, j)$. When the parameter $\gamma(t; i, j)$ is large, then it is well known that the distribution of $W(t; i, j)$ may be approximated by a normal distribution with expectation $\gamma(t; i, j)$ and variance $\sigma^2 = \gamma(t; i, j)$.

In general, consider a random variable $W$ with a Poisson distribution with parameter $\lambda > 0$, and suppose $\lambda$ is large. Then,

$$W \sim S = \lambda + \sqrt{\lambda} Z, \tag{24}$$

where $Z$ is a standard normal random variable with expectation 0 and variance 1 and ~means approximately in distribution. When $Z > 0$, the sum on the right is always positive, but if $Z < 0$, then this sum may be negative, and, therefore, $W$ may be negative. But, every realization of a Poisson random variable must take a value in the set nonnegative integers. A question that naturally arises is if we set $W = [\lambda + \sqrt{\lambda} Z]$, where $[\circ]$ is the greatest integer function, then for what values of $\lambda$ will the sum $S = \lambda + \sqrt{\lambda} Z > 0$ with sufficiently high probability? Observe that $S = \sqrt{\lambda}(\sqrt{\lambda} + Z)$. Thus, $S < 0$, if and only if $Z < -\sqrt{\lambda}$.

For any value of $z > 0$, it follows that

$$P\left[Z < -z\right] = 1 - P\left[Z \leq z\right] = P\left[Z > z\right], \tag{25}$$

because the standard normal distribution is symmetric about 0. Let $z = 2.8$ be a trial value. Then, it can be shown by either computing $P[Z \leq 2.8]$ or by consulting a reliable table for the standard normal distribution that $P[Z \leq 2.8] = 0.9974$. Thus,

$$P\left[Z < -2.8\right] = 1 - 0.9974 = 0.0026 = P\left[Z > 2.8\right]. \tag{26}$$

Therefore, for all $\lambda > 0$ such that $\sqrt{\lambda} > 2.8$, it follows that

$$P\left[Z > \sqrt{\lambda}\right] < P\left[Z > 2.8\right] = 0.0026. \tag{27}$$

It seems reasonable, therefore, for all $\lambda > 7.84 = (2.8)^2$ to use the central limit theorem approximation when simulating a realization of a Poisson random variable. When simulating realizations of many Poisson random variables, it is most efficient to use the normal approximation whenever it is reasonable, because each realization requires only one call to a procedure for simulating a realization of a standard normal random variable. Consequently, the procedure just described was implemented in the software, but, of course any investigator would be free to choose any trial value of $z$ that suits his goals.

For those readers who may be uncomfortable with this choice of the Poisson distribution for biological reasons, it would be straight forward if, for example, an investigator chooses to use negative binomial distribution for the offspring distribution in a branching process formulation. A procedure for simulating realizations of random variables would be more complicated using this two-parameter distribution than that described for the one parameter Poisson distribution. It could, however, be accomplished with a little more effort from some investigator by writing the software to accomplish the task in a programming language of his choosing. But, just as in the case of using a Poisson distribution, at some point in the development of the software, a central limit theorem would need to be invoked to simulate realizations of sums of conditionally independent random variables. If a reader is interested in a review of parametric offspring distributions that have been used extensively in demographic studies of human populations, it is suggested that the book Mode [10] be consulted.

Given a realization of the random variable $W(t; i, j)$, let the $1 \times 3$ random vector

$$\mathbf{O}\left(t; i, j\right) = \left(O\left(t; i, j; k\right) \mid k = 1, 2, 3\right) \tag{28}$$

denote the random number of offspring produced partnerships of type $(i, j)$ of each of the genotypes $k = 1, 2, 3$. Then, the vector $\mathbf{O}(t; i, j)$ has a conditional multinomial distribution with index, sample size, $W(t; i, j)$, and probability vector $\mathbf{p}(i, j) = (p(i, j; k) \mid k = 1, 2, 3)$ derived in the previous section. In symbols, for every mating type $\kappa = (i, j)$

$$\mathbf{O}\left(t; i, j\right) \sim \text{CMultinom}\left(W\left(t; i, j\right), \mathbf{p}\left(i, j\right)\right). \tag{29}$$

Let the random variable

$$O\left(t; \circ, k\right) = \sum_{(i,j) \in \mathbb{T}} O\left(t; i, j, k\right) \tag{30}$$

denote the total number of offspring of genotype $k$ produced by females during the time interval $[t, t+1)$ for $k = 1, 2, 3$. Given that the probability that an offspring is a girl is $p_f$, then

$$B_f\left(t; k\right) \sim \text{CBinom}\left(O\left(t; \circ, k\right), p_f\right), \tag{31}$$

and the number of females with genotype $k$ born during this time interval $[t.t+1)$. Therefore, the number of boys of genotype $k$ born during this time interval is

$$B_m\left(t; k\right) = O\left(t; \circ, k\right) - B_f\left(t; k, 0\right) \tag{32}$$

for $k = 1, 2, 3$.

In general, for $x = 1, 2, \ldots, r$ and $y = 1, 2, \ldots, r$ at time $t$, let the random functions $X(t; j, x)$ and $Y(t; j, y)$ denote,

respectively, the number of females of type $t_f = (j, x)$ and the number of males of type $t_m = (j, y)$ in the population at time $t$ for $j = 1, 2, 3$. Let $X(t; 0)$ denote a $3 \times 1$ vector with the components $B_f(t; j)$ for $j = 1, 2, 3$ for females, and define the $3 \times 1$ vector $Y(t; 0)$ similarly for males. These $3 \times 1$ column vectors contain the number of females and males by genotype, respectively, born during the time interval $[t, t + 1)$. Then let $X(t; x \geq 1)$ denote a $3 \times r$ matrix with the column vectors $X(t; x)$ for $x = 1, 2, \ldots, r$, and define the vector $Y(t, y \geq 1)$ for males similarly. Then, the $3 \times (r + 1)$ matrix for females at time $t + 1$ is given by the matrix

$$\mathbf{X}(t) = \mathbf{X}(t; 0), \mathbf{X}(t; x \geq 1), \tag{33}$$

where the, indicates that the $3 \times r$ matrix is catenated onto the $3 \times 1$ vector $\mathbf{X}(t; 0)$ to construct a $3 \times (r + 1)$ matrix. The matrix $Y(t)$ is constructed for males in the same way. Observe that at time $t$ females of age $r$ are in column $r + 1$ of the $3 \times (r + 1)$ matrix $\mathbf{X}(t)$. Thus, in the construction of the $3 \times (r + 1)$ matrix $\mathbf{X}(t)$, females of age $r$ at time $t$ are dropped from the population. But, because very few females will survive to age $r$, this loss will be negligible. This remark also applies to males. In the next section, a parametric model of a birth cohort survival function, governing the survival of individuals in a birth cohort as a function of their age will be described.

The next component of the age structured process under consideration is that of taking mortality into account in terms of condition survival probabilities by age. Before moving on to an overview of the survival component of the age structured process under consideration, it is appropriate to mention that the book Mode [10] also contains a description of stochastic models of human reproduction expressed in terms monthly estrous cycle of human females along with such factors as abortions, spontaneous and induced age of marriage that other factors involving the control of the number of births experienced by a female during her fertile period. Even though the inclusion of these details was useful in the quantifying of the short-term effects of family planning programs to include them in age structured evolutionary process under consideration with a long-term evolutionary perspective seems inappropriate for model that is in an early stage of development.

## 4. Survival of Individuals in a Birth Cohort as a Function of Age

Another component of natural selection that can be accommodated in the evolution of the age structured population under consideration is that of mortality of individuals. Let $S(x)$ denote the probability that an individual in a birth cohort, who by definition was age $x = 0$ at birth, survives to an age of at least $x > 0$. In symbols, if $X$ is a random variable taking values in the set $\mathbb{R}^+ = \{x \in \mathbb{R} \mid x \geq 0\}$ of nonnegative real numbers, where $\mathbb{R}$ is the set of real numbers, denoting the life span of an individual, then

$$S(x) = P[X > x]. \tag{34}$$

This function has the property that $S(0) = 1$ and $S(x)$ is a monotone decreasing function, as $x$ increases $x \uparrow \infty$ and in the limit

$$\lim_{x \uparrow \infty} S(x) = 0. \tag{35}$$

The purpose of this section is to derive a formulas for a survival function in terms of parametric risk functions.

For many species of animals, offspring are at high risk of death following hatching or birth, but as age increases the risk of death decreases. Therefore, the risk function for deaths of individuals for whom this risk function applies will be assumed to have a latent risk function of the exponential form as follows:

$$\theta_0(x) = \alpha_0 \beta_0 \exp[-\beta_0 x], \tag{36}$$

where $x \geq 0$, and $\alpha_0$ and $\beta_0$ are positive parameters. The integral of this latent risk function is

$$H_0(x) = \int_0^x \theta_0(s)\, ds = \alpha_0 (1 - \exp[-\beta_0 x]) \tag{37}$$

for $x \geq 0$, and, by using well known methods for expressing a survival function in terns of an integral of the risk function, it follows that the survival function corresponding to this risk function is

$$S_0(x) = \exp[-H_0(x)]. \tag{38}$$

In what follows, it will be assumed that this survival function is in force for ages $x = 0, 1, 2, \ldots, 30$, because after age 30 the risk of death function will be assumed to be increasing. From now on all parameters of survival functions will depend on genotypes and sex of individuals, but simply the notation the genotype and sex of individuals will be suppressed except in cases when clarity is an issue.

Observe that $H_0(0) = 0$, so that $S(0) = 0$ as it should. The integral $H_0(x)$ will need to be modified to accommodate the conditions that $x = 0, 1, 2, \ldots, 30$. Let $F_0(x) = 1 - \exp(-\beta_0 x)$ and let the modification of $H_0(x)$ be defined as

$$H_0^*(x) = \frac{\alpha_0 F_0(x)}{F_0(30)}. \tag{39}$$

Let $S^*(x) = \exp(-H_0^*(x))$ denote the modified survival function. Then it follows that $H_0^*(30) = \alpha_0$, so that $S^*(30) = \exp(-\alpha_0)$, which is the probability that an individual born at age $x = 0$ survives to at least age 30. Thus, if one has some prior knowledge of the probability $S^*(30) = p(30)$, then $\alpha_0 = -\ln(p(30))$. Given this estimate of $\alpha_0$, a task that remains is that of finding a value for the parameter $\beta_0$. Next observe that $S^*(1) = \exp(-H_0^*(1))$ is the conditional probability that an infant born at $x = 0$ survives to age $x = 1$ year. Therefore, if one has some prior knowledge of the probability $S^*(1) = \exp(-H_0^*(1)) = p(1)$, by knowing $\alpha_0$ the parameter $\beta_0$ could, in principle, be determined by finding a value of $\beta_0 > 0$ that would satisfy the equation $S^*(1) = \exp(-H_0^*(1)) = p(1)$. This is a nonlinear equation in the unknown parameter $\beta_0$, and it may be possible by plotting $S^*(1; \beta_0)$ as a function of $\beta_0 > 0$ to find a value $\beta_0^*$, such that $S^*(1; \beta_0^*) = p(1)$.

But, there is at least one other approach to find a value of $\beta_0$. Observe that

$$F_0(x) = 1 - \exp[-\beta_0 x] \tag{40}$$

is the distribution function of a exponentially distributed random variable $X_0$ with expectation

$$E[X_0] = \frac{1}{\beta_0}. \tag{41}$$

Thus, the parameter $\beta_0$ will determine the speed at which deaths occur. Small values of $\beta_0$ will correspond to longer infant survival times, while large values of $\beta_0$ will correspond to shorter survival times. Such ideas may be quantified by assigning trial values to $E[X_0]$ to find an estimate of the form $\beta_0 = 1/E[X_0]$. For example, a plausible value of this expectation may be chosen as $E[X_0] = x_m/2$, which yields $\beta_0 = 2/x_m$ as a preliminary estimate of $\beta_0$. The parameter $\beta_0$ may also depend on the sex and genotype of an individual and will be denoted by the symbols $\beta_0(\tau_f)$ and $\beta_0(\tau_m)$ for females and males of genotypes $\tau_f$ and $\tau_m$, respectively. Given a value of $\beta_0$, one would have to check whether $S^*(1;\beta_0^*) = S^*(1;\beta_0^*) = p(1)$ is a plausible estimate of $p(1)$, the conditional probability that an infant born at age $x = 0$ survives to age $x = 1$ year. If the estimate of $p(1)$ is not plausible, too low for example, then it would be necessary to try another estimate of $\beta_0$.

To accommodate accidents that may occur throughout the life span of an individual, the latent risk function for this component of a survival function will have the simple form $\theta_1(x) = \alpha_1$ for all $x \geq 0$, where $\alpha_1$ is a positive constant. In this case, the integral of the risk function is

$$H_1(x) = \int_0^x \theta_1(s)\, ds = \alpha_1 x \tag{42}$$

for $x \geq 0$. Therefore, for $x \geq 0$, the latent survival function has the simple form

$$S_1(x) = \exp[-\alpha_1 x]. \tag{43}$$

A useful trial value of $\alpha_1$, based on period studies of human mortality, is about 0.001. This component of the model is often referred to as the Makeham component and is named after a 19-century investigator that introduced this component.

A third two-parameter latent risk function, due to Gompertz (19th century), deals with risks of deaths at the older ages. Let $\alpha_2$ and $\beta_2$ be positive parameters. Then, it will be assumed that for $x \geq 0$ the latent risk function $\theta_2(x)$ has the form

$$\theta_2(x) = \alpha_2\beta_2 \exp[\beta_2 x]. \tag{44}$$

It will be assumed that this risk function is in force for ages $x = 31, 32, \ldots, r$. Observe that, as it should, this risk function increases as age $x$ of an individual increases, and, by assumption, the risk of death increases exponentially with increasing age. The integral of this risk function has the form

$$H_2(x) = \int_0^x \theta_2(s)\, ds = \alpha_2 (\exp[\beta_2 x] - 1) \tag{45}$$

for $x \geq 0$. Therefore, the latent survival function for this component is

$$S_2(x) = \exp[-\alpha_2 (\exp[\beta_2 x] - 1)] \tag{46}$$

for $x \geq 0$.

By applying a general but standard formula that a density is the risk function times the survival function, it can be seen that the probability density function of the Gompertz distribution has the form

$$\begin{aligned} f_2(x) &= \theta_2(x) S_2(x) \\ &= \alpha_2\beta_2 \exp[\beta_2 x] \exp[-\alpha_2 (\exp[\beta_2 x] - 1)] \end{aligned} \tag{47}$$

for $x \geq 0$. Although this distribution may be derived from intuitively appealing assumptions, it is more difficult to handle from a mathematical point of view than some other distributions that arise in probability and statistics. Nevertheless, because many advanced mathematical functions are now available to research workers in such software packages as MAPLE, MATHEMATICA., and MATLAB, an outline of the mathematics used in analyzing the Gompertz distribution seems appropriate.

Because the parameters $\alpha_2$ and $\beta_2$ do not have obvious statistical interpretations, such as an expectation or variance, it is difficult to assign tentative values to them. Quite often, however, there is some feeling about the modal age of death for those who survive to old age. Let $m_2$ denote the mode of the Gompertz distribution. Then by using elementary calculus to find the maximum of the density $f_2(x)$, it can be shown that the equation

$$\alpha_2 = \exp[-\beta_2 m_2] \tag{48}$$

formalizes a connection among the parameters $\alpha_2$, $\beta_2$, and $m_2$. In particular, if $m_2$ is assigned a value and $\beta_2$ is known, then $\alpha_2$ is determined. But, to find a plausible value of $\beta_2$, more input is needed.

Let $X_2$ denote a random variable with a Gompertz distribution. Then, after considerable analysis, it can be shown that the exact formula for the expectation is

$$E[X_2] = e^{\alpha_2} \left[ m_2 - \frac{C}{\beta_2} + \frac{1}{\beta_2} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu \alpha_2^{\nu+1}}{\nu!(\nu+1)^2} \right], \tag{49}$$

where $C \simeq 0.57721\cdots$ is Euler's constant. Furthermore, the exact formula for the second moment is

$$E[X_2^2] = e^{\alpha_2} \left[ \left( m_2 - \frac{C}{\beta_2} \right)^2 + \frac{\pi^2}{\beta_2^2 6} - \frac{2}{\beta_2^2} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu \alpha_2^{\nu+1}}{\nu!(\nu+1)^3} \right]. \tag{50}$$

When $\alpha_2 > 0$ is small, then $\exp[\alpha_2] \simeq 1$ and the above infinite series may be neglected. Thus, the approximations

$$\begin{aligned} E[X_2] &\simeq m_2 - \frac{C}{\beta_2}, \\ E[X_2^2] &\simeq \left( m_2 - \frac{C}{\beta_2} \right)^2 + \frac{\pi^2}{\beta_2^2 6} \end{aligned} \tag{51}$$

hold for small $\alpha_2$. Therefore, a formula for the approximate variance of the Gompertz distribution is

$$\sigma_2^2 \simeq \frac{\pi^2}{\beta_2^2 6}. \tag{52}$$

Equivalently,

$$\beta_2 \simeq \frac{\pi}{\sigma_2 \sqrt{6}}. \tag{53}$$

It will be instructive to present a simple numerical example, illustrating the use of these approximations. Suppose, for example, one wished to construct a Gompertz distribution with a mode $m_2 = 60$ years, and suppose the standard deviation of this distribution is $\sigma_2 = 10$. Then, this formula yields the approximation $\beta_2 \simeq 0.12825498301686$. Furthermore, by using the formula connecting $\alpha_2$, $\beta_2$, and $m_2$, it can be seen that this formula yields the approximation $\alpha_2 = 4.54960943585548 \times 10^{-4}$. Moreover,

$$\exp\left(4.54960943585548 \times 10^{-4}\right) = 1.00045506445401 \simeq 1. \tag{54}$$

Because of the approximation of the variance is $\sigma_2^2 \simeq \pi^2/\beta_2^2 6$, it appears that the parameter $\beta_2$ will belong to the interval $(0, 1)$ for plausible values of $\sigma_2$, so that it can be seen that when the mode $m_2$ of the distribution is sufficiently large, the parameter $\alpha_2 > 0$ will be small. As can be seen form the above discussion, when $\alpha_2$ is small, the approximation for $\sigma_2^2$ will yield good results. A more extensive account of parametric forms of risk functions for death as well as historical references has be given in Mode and Sleeman [11, Section 13.2], but the details will be omitted here.

In the class of self regulating branch process under consideration, the effects on the mortality of individuals are formulated as a function of total population size $Z_{\text{tot}}(t)$ at time $t$. The formula for computing a realization of the random function $Z_{\text{tot}}(t)$ will depend on the class of stochastic process under consideration. For the class of branching processes under consideration, as stated in a previous section, let $X(t; \tau_f, x)$ denote the number of females of type $\mathbf{t}_f = (\tau_f, x)$ in the population at time $t$, and let the random function $Y(t; \tau_m, y)$ be defined similarly for males. Then, let the random function

$$X(t; \circ, \circ) = \sum_{\tau_f \in \Im} \sum_{x=0}^{r} X\left(t; \tau_f, x\right) \tag{55}$$

denote the total number of females in the population at time $t$, and let $Y(t; \circ, \circ)$ denote the corresponding random function for males. Then, $Z_{\text{tot}}(t) = X(t; \circ, \circ) + Y(t; \circ, \circ)$ is total population size at time $t$.

It will be assumed that the probability for all members of the population at time $t$ survive to time $t + 1$ is governed by a Weibull type survival function depending on total population size of the form

$$S_3\left(Z_{\text{tot}}(t)\right) = \exp\left(-\left(\beta_{\text{tot}} Z_{\text{tot}}(t)^{\alpha_{\text{tot}}}\right)\right). \tag{56}$$

At this point in the discussion, it is appropriate to discuss the rational for choosing values of the parameter $\beta_{\text{tot}}$. For example, suppose the carrying capacity of the environment is $10^7$ individuals. Then, $\beta_{\text{tot}}$ may be chosen as $\beta_{\text{tot}} = 10^{-7}$. As indicated with regard to all parameters under consideration, the positive parameters $\alpha_{\text{tot}}$ and $\beta_{\text{tot}}$ will depend on sex and genotype. It follows, therefore, if one genotype has a smaller value of the parameter $\beta_{\text{tot}}$, it will be able to survive in larger populations and will thus have a selective advantage over the others. From now on the function $(\beta_{\text{tot}} Z_{\text{tot}}(t))^{\alpha_{\text{tot}}}$ will play a role similar to that of an integral of a risk function as illustrated above.

But, just as the total fertility of a female $\lambda$ was distributed over her reproductive ages, a density function will needed to distribute $(\beta_{\text{tot}} Z_{\text{tot}}(t))^{\alpha_{\text{tot}}}$ over the ages of individuals $x = 1, 2, \ldots, r$ in the population at time $t > 0$, which gives rise to the question as to how such a distribution may be chosen. A plausible answer to this question is that it may be constructed from the risks functions as functions of age as illustrated by the parametric risk functions introduced above. Observe that the larger the values of a total risk functions of an individual at these ages, the greater will be the proportion assigned to the total contribution of the Weibull component.

Thus, if the Makeham component is included in the risk function for the ages $x = 1, 2, \ldots, 30$, then the total risk function for these ages is

$$\theta_{\text{tot}}(x) = \alpha_0 \beta_0 \exp\left[-\beta_0 x\right] + \alpha_1. \tag{57}$$

For ages $x > 30$ the total risk function is

$$\theta_{\text{tot}}(x) = \alpha_2 \beta_2 \exp\left[\beta_2 x\right] + \alpha_1. \tag{58}$$

Let $\theta_{\text{tot}}$ be defined by the sum

$$\theta_{\text{tot}} = \sum_{x=1}^{r} \theta_{\text{tot}}(x). \tag{59}$$

Then, the density for distributing $(\beta_{\text{tot}} Z_{\text{tot}}(t))^{\alpha_{\text{tot}}}$ over the ages may be chosen as

$$f_{\text{tot}}(x) = \frac{\theta_{\text{tot}}(x)}{\theta_{\text{tot}}} \tag{60}$$

for $x = 1, 2, \ldots, r$. Given these definitions, the proportion, $H(Z_{\text{tot}}(t), x)$, for age $x$ of the total $(\beta_{\text{tot}} Z_{\text{tot}}(t))^{\alpha_{\text{tot}}}$ will be chosen as

$$H\left(Z_{\text{tot}}(t), x\right) = f_{\text{tot}}(x) \times \left(\beta_{\text{tot}} Z_{\text{tot}}(t)\right)^{\alpha_{\text{tot}}} \tag{61}$$

for $x = 1, 2, \ldots, r$.

The next step in the formulation of the mortality component of the process is to express the final survival function in terms of integrals of risk functions. Thus, for ages $x = 1, 2, \ldots, 30$ the sum of these integrals is

$$H_{\text{tot}}\left(Z_{\text{tot}}(t), x\right) = H_0^*(x) + H_1(x) + H\left(Z_{\text{tot}}(t), x\right). \tag{62}$$

Recall that $H_1(x)$ is the integral of the constant Makeham risk function, namely, $\alpha_1 x$. For ages $x > 30$,

$$H_{\text{tot}}\left(Z_{\text{tot}}(t), x\right) = H_1(x) + H_2(x) + H\left(Z_{\text{tot}}(t), x\right). \tag{63}$$

Therefore, the survival function for all ages $x = 1, 2, \ldots, r$ is

$$S\left(Z_{\text{tot}}(t), x\right) = \exp\left(-H_{\text{tot}}\left(Z_{\text{tot}}(t), x\right)\right). \tag{64}$$

The last step in formulating the mortality component of the process is to derive a formula for the conditional probability that an individual that is alive at age $x$ survives to age $x + 1$. Let $p(Z_{\text{tot}}(t), x)$ denote this probability. Then,

$$\begin{aligned} &p\left(Z_{\text{tot}}(t), x\right) \\ &\quad = \frac{S\left(Z_{\text{tot}}(t), x+1\right)}{S\left(Z_{\text{tot}}(t), x\right)} \\ &\quad = \exp\left(-\left(\left(H_{\text{tot}}\left(Z_{\text{tot}}(t), x+1\right) - H_{\text{tot}}\left(Z_{\text{tot}}(t), x\right)\right)\right)\right). \end{aligned} \tag{65}$$

From this formula, it follows that in order that $0 \leq p(Z_{\text{tot}}(t), x) < 1$, the condition

$$H_{\text{tot}}\left(Z_{\text{tot}}(t), x+1\right) - H_{\text{tot}}\left(Z_{\text{tot}}(t), x\right) > 0 \tag{66}$$

must be satisfied for all $x = 1, 2, \ldots, r$. Equivalently, the function $H_{\text{tot}}(Z_{\text{tot}}(t), x)$ must be strictly monotone increasing as a function of $x$. It seems likely that because all parameters $\beta$ will be small, the component $H(Z_{\text{tot}}(t), x)$ for population density will be small in relation to the others in $H_{\text{tot}}(Z_{\text{tot}}(t), x)$. It seems reasonable, therefore, that if an investigator in a preliminary experiment ignores the component for population density, then it is likely that strict monotonicity of this function will be preserved when the population density component is included. To complete of the definition of survival function $S(Z_{\text{tot}}(t), x)$ for $x = 0$, by definition $H_{\text{tot}}(Z_{\text{tot}}(t), 0) = 0$, so that $S(Z_{\text{tot}}(t), 0) = 1$.

When checking the plausibility of a conditional survival probability

$$p\left(Z_{\text{tot}}(t), x\right) \tag{67}$$

in a preliminary experiment, the probabilities $p(Z_{\text{tot}}(t), 0)$ and $p(Z_{\text{tot}}(t), 30)$ require special attention. For example,

$$p\left(Z_{\text{tot}}(t), 0\right) = \frac{S\left(Z_{\text{tot}}(t), 1\right)}{S\left(Z_{\text{tot}}(t), 0\right)} = \exp\left(-\left(\left(H_{\text{tot}}\left(Z_{\text{tot}}(t), 1\right)\right)\right)\right) \tag{68}$$

so that the probability on the right should be consistent with a plausible probability that an infant dies in his first year of life. Similarly, because the risk function is increasing for all ages $x \geq 31$, conditional probability

$$\begin{aligned} &p\left(Z_{\text{tot}}(t), 30\right) \\ &\quad = \exp\left(-\left(\left(H_{\text{tot}}\left(Z_{\text{tot}}(t), 31\right) - H_{\text{tot}}\left(Z_{\text{tot}}(t), 30\right)\right)\right)\right) \end{aligned} \tag{69}$$

should satisfy the condition $0 < p(Z_{\text{tot}}(t), 30) < 1$, so that the condition $(H_{\text{tot}}(Z_{\text{tot}}(t), 31) - H_{\text{tot}}(Z_{\text{tot}}(t), 30)) > 0$ should also be satisfied. If this condition holds when the component for population density is ignored, then it seems likely, for reasons stated above, that this condition will also hold when the population density factor is taken into account.

The component of natural selection that needs to be considered in formulating the mortality component of the stochastic process under consideration is that of describing a procedure for simulating the number of individuals classified by age, sex, and genotype that are alive at time $t$ and survive to time $t + 1$. For example, let $p_f(Z_{\text{tot}}(t); i, x)$ denote the conditional probability that a female of genotype $i$ and age $x$ at time $t$ survives to age $x + 1$ at time $t + 1$. At time $t$ as before let $X(t; i, x)$ denote the number of females of $\mathbf{t} = (i, x)$, and let $X(t + 1; i, x + 1)$ denote the number of these females who survive to age $X + 1$ at time $t + 1$. Then, by assumption, $X(t + 1; i, x + 1)$has a conditional binomial distribution with index, sample size $X(t; i, x)$, and probability $p_f(Z_{\text{tot}}(t); i, x)$. In symbols,

$$X(t+1; i, x+1) \sim \text{CBinom}\left(X(t; i, x); p_f\left(Z_{\text{tot}}(t); i, x\right)\right). \tag{70}$$

Similarly, if $Y(t; i, x)$ is the number of males of type $\mathbf{t} = (i, x)$ alive at time $t$, then

$$Y(t+1; i, x+1) \sim \text{CBinom}\left(Y(t; i, x); p_m\left(Z_{\text{tot}}(t); i, x\right)\right) \tag{71}$$

for $x = 0, 1, 2, \ldots, r$.

## 5. An Embedded Deterministic Model in an Age-Structured Stochastic Process

As demonstrated in recent publications, Mode and Sleeman [9], Mode et al. (2011), [12, 13], by deriving formulas for the conditional expectation of any random variable at time $t$, given the evolution of the process prior to $t$, it is possible to derive a set of recursive nonlinear difference equations, such that, given the initial values of the random functions at time $t = 0$, estimates of the sample functions of the process can be derived for all $t > 1$. In previous publications dating back at least a decade, this derived set of equations has been called the deterministic model embedded in the stochastic process under consideration. The purpose of this section is to derive formulas for a deterministic model embedded in the age structured stochastic process under consideration. As will be seen in what follows, the derivation of formulas for the embedded deterministic model is closely related to procedures for computing Monte Carlo realizations of the process.

To fix ideas, suppose at time $t = 0$ the elements of the random $3 \times (r + 1)$ matrices $\mathbf{X}(0)$ and $\mathbf{Y}(0)$, denoting, respectively, the assigned numbers of females and males classified by age and genotype. All these initial numbers are nonnegative integers. Next consider the matrix valued conditional expectation

$$E\left[\mathbf{X}(1) \mid \mathbf{X}(0)\right]. \tag{72}$$

Because $\mathbf{X}(0)$ is known, $E[\mathbf{X}(1) \mid \mathbf{X}(0)]$ is known. Thus,

$$\widehat{\mathbf{X}}(1) = E\left[\mathbf{X}(1) \mid \mathbf{X}(0)\right] \tag{73}$$

is, as is well known that $\widehat{\mathbf{X}}(1)$, the best estimate of $\mathbf{X}(1)$ in the sense of minimum mean square error. This procedure may be

continued recursively by considering the conditional matrix valued conditional expectation

$$E\left[\mathbf{X}\left(t\right) \mid \mathbf{X}\left(t-1\right)\right]. \tag{74}$$

But, this conditional matrix valued expectation is a random valuable, because $\mathbf{X}(t-1)$ is a random variable. However, if one has the estimate

$$\widehat{\mathbf{X}}\left(t-1\right) \tag{75}$$

that has been calculated recursively, then

$$\widehat{\mathbf{X}}\left(t\right) = E\left[\mathbf{X}\left(t\right) \mid \widehat{\mathbf{X}}\left(t-1\right)\right] \tag{76}$$

is a reasonable estimate of the matrix valued random function $\mathbf{X}(t)$ for $t = 1, 2, \ldots$. This idea is the essential concept underling the procedure for embedding a recursive deterministic system in a stochastic process whose sample functions take values in matrices of non-negative integers. Deterministic estimates of the matrix valued random functions for males $\mathbf{Y}(t)$ may be derived by an identical procedure for $t \geq 1$. Observe that the stochastic process and the embedded deterministic equations have the same parameter space, which is usually multidimensional.

Given these definitions, suppose that all the random matrix valued functions for males in (5), (6), and (7) are all replaced by estimates of the form $\widehat{\mathbf{Y}}(t; i, y)$ starting in (5). Then, the estimate of the vector of the conditional probabilities in (9) have the form

$$\widehat{\gamma}_f\left(t; i, j\right) = \frac{\widehat{U}_m\left(t; j\right)\alpha_f\left(i, j\right)}{\sum_\nu \widehat{U}_m\left(t; j\right)\alpha_f\left(i, \nu\right)} \tag{77}$$

for $j = 1, 2, 3$. Observe that, because $\alpha_f(i, j)$ is constant for all nine pairs $(i, j)$ of matings types, the $\alpha_f(\cdot, \cdot)$ parameters do not need to be estimated. Thus, the estimate of the probability vector in (10) is

$$\widehat{\boldsymbol{\gamma}}_f\left(t; i\right) = \left(\widehat{\gamma}_f\left(t; i, j\right) \mid j = 1, 2, 3\right). \tag{78}$$

The next step in deriving the embedded deterministic model is to describe a procedure for estimating the random functions $Z(t; i, x, j)$, which by definition is the number of matings of females of age $x$ and genotype $i$ with males of genotype $j$, during the time interval $[t, t+1)$. According to (11), the random vector

$$\mathbf{Z}\left(t; i, x\right) = \left(Z\left(t; i, x, j\right) \mid j = 1, 2, 3\right) \tag{79}$$

has a conditional multinomial distribution index, sample size, $X(t; i, x)$, and probability vector $\boldsymbol{\gamma}_f(t; i)$ in (10). As is well known, given the state of the population $\mathbb{S}(t) = (\mathbf{X}(t), \mathbf{Y}(t))$ at time $t$, the conditional expectation of the vector $\mathbf{Z}(t; i, x)$ is the vector

$$E\left[\mathbf{Z}\left(t; i, x\right) \mid \mathbb{S}\left(t\right)\right] = X\left(t; i, x\right)\boldsymbol{\gamma}_f\left(t; i\right). \tag{80}$$

Therefore, let $\widehat{Z}(t; i, x, j)$ denote the estimate of the random function $Z(t; i, x, j)$. Then,

$$\widehat{Z}\left(t; i, x, j\right) = \widehat{X}\left(t; i, x\right)\widehat{\gamma}\left(t; i, j\right) \tag{81}$$

is an estimate of the random function $Z(t; i, x, j)$ for all pairs of mating types $(i, j)$ of females of age $x$, where $\widehat{X}(t; i, x)$ and $\widehat{\gamma}(t; i, j)$ are estimates, respectively, of the random functions $X(t; i, x)$ and $\boldsymbol{\gamma}_f(t; i, j)$ at time $t$. From these definitions it follows that

$$\widehat{\gamma}\left(t; i, j\right) = \sum_{x_m}^{x_{\max}} \widehat{Z}\left(t; i, x, j\right)\lambda\left(i, x\right) \tag{82}$$

is an estimate of the random function $\gamma(t, i, j)$ at time $t$, see (21). Let

$$\widehat{V}\left(t; i, j\right) = \left(\sum_{x_m}^{x_{\max}} \widehat{Z}\left(t; i, x, j\right) f\left(x; \lambda_0, 40\right)\right), \tag{83}$$

where $f(x; \lambda_0, 40)$ is the density of the Poisson distribution defined in (20). Then, at time $t$

$$\widehat{\gamma}\left(t; i, j\right) = \widehat{V}\left(t; i, j\right)\lambda_i \tag{84}$$

is an estimate of $\gamma(t, i, j)$ for all mating pairs of type $(i, j)$, where $\lambda_i$ is the expected potential number of total offspring produced by a female of genotype $i$ throughout her life span.

Let the random function $W(t; i, j)$ denote the number of offspring produced by matings of type $(i, j)$ during the time interval $[t, t+1)$. When it is not zero, then according to (23) this random function has a conditional Poisson distribution with parameter $\boldsymbol{\gamma}_f(t; i, j)$ and sample size 1. Therefore, the estimate of this random function is

$$\widehat{W}\left(t; i, j\right) = \widehat{\gamma}\left(t; i, j\right). \tag{85}$$

The next step in deriving formulas for the embedded deterministic model is to estimate the random functions $O(t; i, j; k)$ defined in the vector (28). Because, by assumption (29), the elements of this vector have a conditional multinomial distribution with index $W(t; i, j)$ and probability vector $\mathbf{p}(i, j) = (p(i, j; k) \mid k = 1, 2, 3)$ derived in a previous section, it follows that an estimate of the random function $O(t; i, j; k)$ is

$$\widehat{O}\left(t; i, j; k\right) = \widehat{\gamma}\left(t; i, j\right)p\left(i, j; k\right) \tag{86}$$

for all mating types $(i, j) \in \mathbb{T}$ and genotypes $k \in \mathbb{G}$. Observe that the probability $p(i, j; k)$ depends only on the probabilities of mutation, which, by assumption, are constant and, therefore, there is no need to estimate them during any time interval. From these results, it follows that

$$\widehat{O}\left(t; \circ, k\right) = \sum_{(i,j)\in\mathbb{T}} \widehat{O}\left(t; i, j; k\right) \tag{87}$$

is an estimate of the total number of offspring of genotype $k$ produced by females during the time interval $[t, t+1)$.

Let $p_f$ denote the conditional probability that an offspring is a female. Then, according to (31) an estimate of $B_f(t; k)$, the number of females of genotype $k$ born during the time interval $[t, t+1)$, is

$$\widehat{B}_f\left(t; k\right) = \widehat{O}\left(t; \circ, k\right)p_f. \tag{88}$$

From this result it follows that

$$\widehat{B}_m(t;k) = \widehat{O}(t;\circ,k) - \widehat{B}_f(t;k) \tag{89}$$

is an estimate of the number of males if genotype $k$ born during the time interval $[t, t+1)$. The estimate for the $3 \times (r+1)$ matrix $\mathbf{X}(t)$ at time $t$ is constructed by using the formula

$$\widehat{\mathbf{X}}(t) = \widehat{\mathbf{X}}(t;0),\ \widehat{\mathbf{X}}(t;x \geq 1), \tag{90}$$

where $\widehat{\mathbf{X}}(t;0)$ is a $3 \times 1$ vector with elements $\widehat{B}_f(t;k)$ for $k = 1, 2, 3$. If a reader is interested in more details, the discussion pertaining to (33) should be consulted, because the procedure for constructing estimates of the random matrix $\mathbf{X}(t)$ involves replacing a call for random numbers by a conditional expectation. The construction of an estimate of the $3 \times (r+1)$ random matrix for males $\mathbf{Y}(t)$ is identical of that for females.

The last step in constructing estimates of the state of the population at time $t+1$, given estimates at time $t$, is to replace the calls for random variables in (70) and (71) with conditional expectations of random variables with binomial distributions for those individuals who survive from $t$ to $t+1$. Thus,

$$\begin{aligned} \widehat{X}(t+1;i,x+1) &= \widehat{X}(t;i,x)\, p_f\left(\widehat{Z}_{\text{tot}}(t);i,x\right), \\ \widehat{Y}(t+1;i,x+1) &= \widehat{Y}(t;i,x)\, p_m\left(\widehat{Z}_{\text{tot}}(t);i,x\right) \end{aligned} \tag{91}$$

for all genotypes $i \in \mathbb{G}$ and ages $x = 0, 1, 2, \ldots, r$.

## 6. Overview of Procedures and Issues Involved in Designing Computer Experiments

Among the issues that arise as one begins the process of designing a computer simulation experiment is that of choosing a numerical value for each of the parameters in the multidimensional parameter space of the age structured stochastic population process under consideration. As can be seen from Section 2, the minimal age $x_m$ at which a female becomes fertile and may bear children as well as the maximum age $x_{\max}$ that she is fertile must be assigned numbers. For most human populations, the age at which females become fertile varies around 15 years. In all computer experiments reported in this paper, however, variation among the ages when a female becomes fertile will be ignored, and for the sake of simplicity that parameter $x_m$ will be assigned the value $x_m = 15$. For similar reasons, the maximum fertile age for females will be assigned the value $x_{\max} = 40$ years. Another parameter that must be assigned a numerical value is $r$, the greatest age an individual, female or male, may attain. For most of the computer experiments reported in this paper, this parameter was assigned the value $r = 50$ years. Part of the rationale for choosing this value was the thought that in ancient populations the life span of individuals was short when compared with those of some modern populations, so that age 50 years would be a plausible maximum life span for an ancient human population. The other issue underlying this choice was the time required to complete a computer experiment covering an evolutionary time span of thousands of years. Simply put, the larger the value chosen for the parameter $r$, the greater the time span required to complete a Monte Carlo simulation experiment consisting of some number $M \leq 100$ of replications. It was found in preliminary experiments with the assigned value $r = 50$ could be completed in reasonable lengths of time, given personal computer resources that were available to conduct the experiments.

The evolutionary driving force of mutation will play a basic role in all computer experiments reported in this paper, and, as can be seen from (2), the probability $\mu_{12}$ that allele $A$ mutates to allele $a$ per meiosis must be assigned a numerical value. Similarly, the probability $\mu_{21}$ that allele $a$ mutates to allele $A$ must be assigned a numerical value. The assigned values for these mutations will differ among the experiments reported in this paper; thus, in what follows, the values used in the experiments reported in subsequent sections will be sated in the initial computer input for each experiment.

In Section 3 on simulating the number of births that occur in a population in any year, it was shown that the birth process depends on two sets of parameters. The first set of these parameters is set forth equation (8) where a $3 \times 3$ matrix of acceptance probabilities $\alpha_f(i, j)$ displayed that characterizes preferences of females of genotype $i$ for males of genotype $j$ as sexual partners for all pairs $(i, j)$ of sexual partnerships. The second set of parameters is $(\lambda_1, \lambda_2, \text{and } \lambda_3)$, where $\lambda_i$ the expected total number of offspring produced by a female of genotype $i = 1, 2, 3$ during her fertile period from ages $x_m = 15$ up to and including age $x_{\max} = 40$. The values chosen for these two sets of parameters will also vary among experiments, so that no specific values will be given here. In all experiments reported in this paper, it was assumed that the probability that a baby was a girl was assigned the value $p_f = 100/205$, and by assumption, the probability a baby was a boy was assigned the value $p_m = 1 - p_f$.

In Section 4, which is devoted to the survival of individuals in a population by age and from year to year, the two-parameter Weibull distribution was chosen to provide a framework to model the regulationof total population size. As may be seen from (61), a function of the form $(\beta Z_{\text{tot}}(t))^{\alpha}$, where $Z_{\text{tot}}(t)$ is total population size in year $t$, and $\alpha$ and $\beta$ are positive parameters that may differ by sex and genotype of an individual, must be assigned numerical values. Because two sexes and three genotypes are under consideration, the total number of alpha and beta parameters is 12. For the sake of simplicity, in all experiments reported in this paper, the values of all alpha parameters were chosen as $\alpha = 1$, which was equivalent to assuming that all survival functions had a simple exponential form. This simplification of the formulation reduced the dimension of the parameter space to 6, but in order to provide a framework for simulating the evolution of a population that was, in part, dependent on the carrying capacity of the environment, it would still be necessary to assign to six beta parameters to accommodate the notion that one genotype may have a selective advantage over others by being a superior competitor for resources and was able to survive and reproduce in higher population densities. To take into account that individuals of genotypes may be more competitive users of environmental resources, and in all computer experiments reported in this paper, the values of

the beta parameters will be given. Even though the survival capabilities may differ by sex, it was assumed, for the sake of simplicity, in all experiments reported in this paper that the assigned value of each parameter in the parameter space was equal for males and females.

In the component of survival function described in the first paragraphs of Section 4 for ages $x = 0, 1, 2, \ldots, 30$, the parameter $\alpha_0$, which may differ among the three genotypes and the two sexes, was among the parameters that must be assigned numerical values when designing a computer simulation experiment. As indicated in Section 4, this parameter has the interpretation that in a birth cohort for any year $\alpha_0$ may be interpreted as potential probability that an individual in this cohort would survive to age 30. In all computer experiments reported in the experiments reported in this paper, this parameter was assigned the value $\alpha_0 = 0.99$ for each sex and genotype. For the sake of simplicity, it was also assumed that the parameter $\beta_0$ had the same value, 2/30, for both sexes and genotypes within sexes. Thus, by assumption, in the computer experiments reported in this paper, no genotype or sex had a selective advantage over others with respect to the component of natural selection expressed in terms of this survival component of the formulation for ages $x = 0, 1, 2, \ldots, 30$.

Another aspect of the formulation governing the survival of individuals was that of the one-parameter Makeham component, taking into account deaths due to accidents at all ages. The positive parameter of this component was denoted by $\alpha_1$, and for each sex and genotypes within sexes, this parameter was assigned numerical value $\alpha_1 = 0.001$, which has also been used by others in connection with parametric models of human mortality. The technical details concerning this component are expressed in (42) and (43). The survival component in the formulation for ages $x = 31, \ldots, r$ was described in terms of the two parameter risk function in (44). As was shown in the discussion following this equation, the positive parameters $\alpha_2$ and $\beta_2$ of the Gompertz distribution may be expressed in terms of the mode and standard deviation of this distribution. For each genotype within sexes, the mode and standard deviation of this distribution were assigned the values $m_2 = 50$ and $\sigma_2 = 10$ years, respectively. In summary, the procedure just described, entailing the assignment of values of the parameters that were the same for both sexes and genotypes within each sex, was the implementation of an approach to quantifying the idea that forces governing the evolution of the age structured population were neutral with respect to survival by age.

The last component of the formulation that requires assignments of values, expressed in terms of non-negative integers, is the elements of the 3 by $r + 1$ matrices $\mathbf{X}(0)$ and $\mathbf{Y}(0)$, denoting the number of individuals by genotype $i$ and age $x$ in the initial population at time $t = 0$. In all experiments reported in this paper, it was assumed that the initial population for both sexes consisted only of individuals of genotype $1 = AA$, so that the genotypes $2 = Aa$ and $3 = aa$ could arise in an evolving population only as a result of the allele $A$ mutating to allele $a$, see (2). According to the "Out of Africa Hypothesis," the people who now make up the current populations of Europe, Asia, Australia, and the Americas are descendents of a group of our species who emigrated out of Africa about 60,000 years ago. If a reader is interested in a detailed account of this hypothesis by a paleoanthropologist, it is suggested that the video course and book are Hawks [14] of consulted. It is also thought by many that this group of emigrants consisted of a rather small number of individuals with a total population size ranging from the hundreds to at most a few thousands. To simulate this initial founder population, whose size was uncertain, a random number generator was used to estimate the number of individuals of each age and sex for genotype 1, and the number 0 was assigned to all ages for individuals of genotypes 2 and 3. A broad objective of the computer simulation experiments reported in this paper is to study the evolution of the mutant genotypes $Aa$ and $aa$ in an age structured population evolving from the ancestral genotype $AA$.

Various procedures will be used to statistically summarize data generated in Monte Carlo simulation experiments. These methods will not be described here, but if a reader is interested in a detailed overview of these methods, it is suggested that the paper of Mode and Gallop [15] is consulted. With a view towards complete openness as to the generation of random numbers used in the Monte Carlo simulation experiments reported in this paper, it is appropriate to mention the random number generator used in these experiments is that for 32-bit word computers designed by Deng and described in Mode and Gallop, see Section 2 of that paper. Further discussion of issues concerning random number generator may be found in the papers by Deng and Lin [16], L'Ecuyer and Touzin [17], and Deng [18]. But, unlike the 32-bit word computers that the Deng random number generator was designed for, the word length for computer used in the Monte Carlo simulation experiments reported in this paper was a 64-bits. Because the developers of the version 11 of the APL 2000 software used to write programs to implement the age structured process under consideration did not include a default 64-bit word random number generator, it was necessary for users to implement such a generator. However, to implement one of the 64-bit random number generators described in Deng would have required a rather long period of time to develop the necessary APL code in way, such that the times taken to complete Monte Carlo simulation experiments would be satisfactory. Consequently, this developmental task will be postponed to some future date. It is felt, however, that even if the software for a 64 bit random number generator would have been developed, from a qualitative point of view, the results of the Monte Carlo simulation experiments reported in this paper would not have been significantly different, due to, among other things, the long period of the random number generator used in these experiments, and it verified randomness properties.

Collections of human mortality data are also sometimes useful in choosing numerical input to age structured stochastic processes. In this connection, the collection of data in the book by Alderson [19] may be useful. A search for historical human mortality data on the internet may also yield informative information.

## 7. On the Emergence of Rare Mutations-Differing Deterministic and Stochastic Predictions

The objective of the computer simulation experiment reported in this section was to compare the predictions of the embedded deterministic model with those of the stochastic process with respect to the emergence of rare mutations. To simulate the occurrence of rare mutations, it was assumed that the conditional probability that the ancestral allele $A$ mutated to allele $a$ per meiosis was $\mu_{12} = 10^{-10}$, and the conditional probability of the back mutation $a \rightarrow A$ was assigned the value $\mu_{21} = 10^{-11}$. These assignments were based on anecdotal evidence gleaned in conversations with microbiologists, who studied the evolution of a bacterial population for thousands of generations. The mutation studied in the experiment were rare in the sense that in previous experiments it was assumed that probabilities of mutation were in the interval $(10^{-8}, 10^{-5})$ per cell division or, in the case of diploids, per meiosis. It was also assumed that genotypes of individuals carrying the mutant allele had a reproductive advantage over individuals with the ancestral genotype. To quantify this idea, the potential number of offspring of females of each of the three genotypes ($1 = AA$, $2 = Aa$, and $3 = aa$) throughout their fertile years were assigned the values $(\lambda_1, \lambda_2, \lambda_3) = (6, 8, 10)$. Given these assignments, it was expected that the population would grow rapidly so as to increase the likelihood of a rare mutation being realized in a Monte Carlo simulation experiment.

It was also assumed that mating was random in the sense that females of any genotype when choosing a male sexual partner had no preferences with regard to the genotype of her male partner. To quantify this assumption it was assumed that all the elements on the $3 \times 3$ matrix $\mathbf{A}_f$ of acceptance probabilities in (8) were assigned value 1, but it may be noted that any constant probability would suffice for the case that female mating preferences were neutral with respect to the genotype of the male. With regard to the carrying capacity of the environment, it was assumed that the beta parameters for each genotype within each sex had the constant value $\beta^{-7}$. It is of interest to note that this value is greater than the probability $\mu_{12} = 10^{-10}$ of a rare mutation, which raised the question as to whether the carrying capacity of the environment was sufficiently large to assure that at least one rare mutation would actually be observed in a Monte Carlo simulation experiment.

Given these assigned values of the parameters of the model, a preliminary experiment was conducted using the embedded deterministic model to simulate the evolution of an age structured population for 5,000 years. The numerical operations entailed in an experiment with the embedded deterministic model consist essentially of multiplications and additions, so that even for mutations with small probabilities of occurrence, the number of mutant genotypes will eventually exceed one for the case of high fertility considered in the experiment. As presented in Figure 1 there are the trajectories of total population size for each of the three genotypes in the female population for 5,000 years of evolution.

As can be seen form this figure, the total size of the initial ancestral population of females, consisting only of individuals of ancestral genotype $1 = AA$, reached the carrying capacity of the environment within 500 years of evolution. But, after about 3,000 years of evolution, the total size of population of individuals of the ancestral genotype began a steep decline as the total numbers of the females of the mutant genotypes $2 = Aa$ and $3 = aa$ grew to significant numbers. Sometime before 4,500 years of evolution, the total number of individuals of genotype 2 starts a decline, as the total number of individuals of genotypes 3 rose to predominance in the population. This rise was due to the assumption that females of this genotype had a selective advantage over the other genotypes due to a higher level of reproductive success as quantified in terms of the numerical values of the parameters $(\lambda_1, \lambda_2, \lambda_3) = (6, 8, 10)$. That the trajectory for the total number of individuals of genotype 3 that was still rising near the carrying capacity of the environment suggests that this limit would be reached shortly after 5,000 years of evolution.

To check whether the predictions of a sample of the process would be such that the trajectories computed using the embedded deterministic model could be interpreted as measures of central tendency for the sample functions of the stochastic processes, a Monte Carlo simulation experiment was run with an evolutionary time span of 2,000 years with 50 replications. In preliminary experiments with 5 replications, it had been observed that within 2,000 years of evolution genotypes carrying rare mutant alleles had been realized in the simulated data. A decision was, therefore, made to run a confirmatory Monte Carlo simulation experiment with 50 replications of 2,000 years of evolution, which 28 took about 30 hours of computer time to complete the experiment. The reason for computing only 50 realizations of the process was purely practical, for it had been decided to compute 100 realizations of the process, and it would have taken about 60 hours of computing time to complete the experiment. Moreover, the risk of power outage, resulting in the loss of simulated data, during such a time period was significant for the home computer used to carry out the experiments reported in this paper. If a reader is interested in a more in depth discussion of the issues that arise in choosing the number of replications for a Monte Carlo simulation experiment, the paper by Mode et al. [13] may be consulted.

Whenever a Monte Carlo simulation experiment is carried out based on an age structured framework with several genotypes, a plethora of simulated data is produced, which gives rise to problems of presenting the results of such experiments in an informative manner. In this section, attention will be focused on the evolution of the population of individuals of genotype $3 = aa$ in which each individual carries two copies of the rare mutant allele $a$. As presented in Figure 2 there are the Min, $Q_{50}$, and Max trajectories of the total number of individuals of genotype 3 for 2,000 years of evolution from a founder population consisting only of individuals of the ancestral genotype $1 = AA$.

As can be seen from this figure, significant numbers of individuals of mutant genotype $3 = aa$ appeared between 1,000 and 1,200 years of evolution, but, as can be seen from Figure 1, significant numbers of this genotype were not

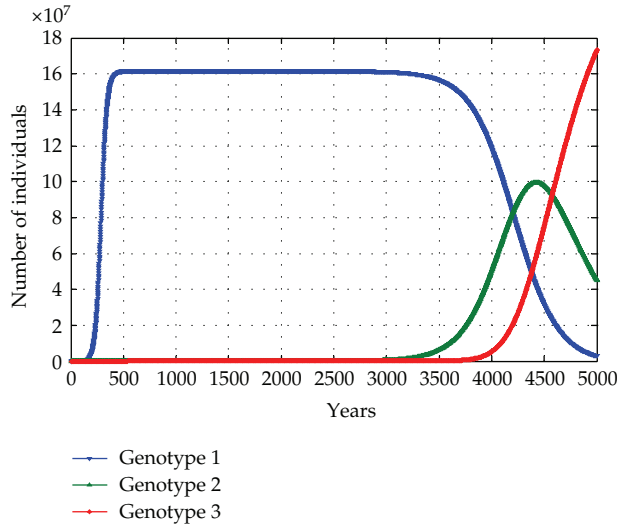


Figure 1: Trajectories of the total population size of the three female genotypes as computed using the embedded deterministic model.

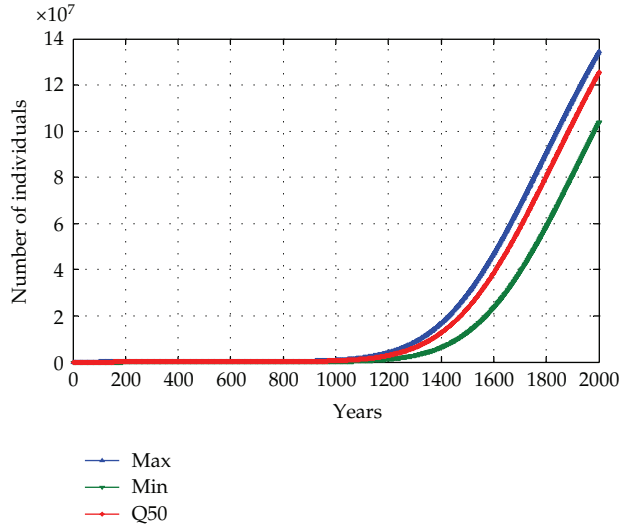


Figure 2: Estimated quantile trajectories of the total number of females of mutant genotype 3.

realized in the population until somewhere between 3,000 and 4,000 years of evolution according to the predictions of the embedded deterministic model. Consequently, this is a clear cut case in which the predictions of the timing of the emergence of a mutant genotype differs significantly as can be seen by comparing the trajectories of the stochastic process and the embedded deterministic model. If a reader is interested in other examples in which the predictions of the stochastic process differ significantly from those of the embedded deterministic model, the papers of Mode et al. (2011) [12, 13] may be consulted.

When an age structured model is under consideration, most of the simulated data arising in a Monte Carlo simulation experiment is in the form of arrays in which the number of individuals of each age is recorded. Furthermore, when many replications of realizations of the stochastic process are computed, it becomes necessary to statistically summarize the data in an informative manner. In the experiment under consideration the mean number of individuals for each age group and its standard deviation were computed. For example, let $\widehat{\mu}_f(t;x)$ denotethe estimated mean number of individuals of age $x$ in the female population in year $t$ of a Monte Carlo simulation experiment, and let $\widehat{\sigma}_f(t;x)$ denote the estimated standard deviation. Then for year $t = 300$, the graph of the estimates $\widehat{\mu}_f(t;x)$ is plotted in the upper panel of Figure 3 for ages $x = 0, 1, 2, \ldots, 50$ and females of genotype 3. Similarly, in the lower panel of this figure the estimates of the standard deviations by age for this genotype are plotted.

From these two graphs, it can be seen that the graph of number of individuals by age declines rapidly as age $x$ increases, which is a signature of rapidly growing population with quite high levels of mortality. It is also of interest to note that the number of individuals of the mutant genotype 3 = *aa* was present in the population, with numbers ranging from

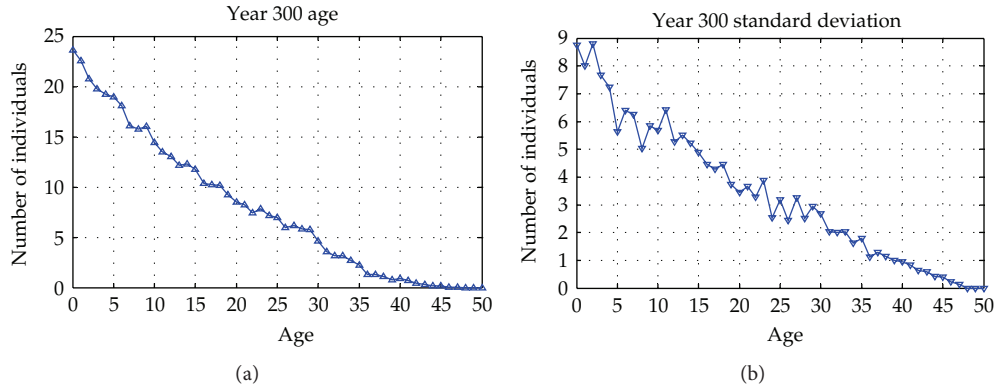


Figure 3: Graphs of the mean and standard deviation by age for females of genotype 3 in year 300.

the low single digits to 25, within at least 300 years of simulated evolution, which illustrates that if one replied solely on the observing the graph of the total number of females with this genotype, this earlier appearance of the mutant genotype would have been missed. From the lower panel of this figure, it can be seem from the graph of the standard deviation by age is quite irregular for those ages representing the young, but these fluctuations level off age increases. The magnitude of the values of the standard deviation by age are indicators of the variations among the 50 replications of the process, particularly for ages less than 10 years. As presented in Figure 4 there are the graphs for the mean number of individuals by age along with the corresponding standard deviation for year $t = 2,000$ in the Monte Carlo simulation experiment for females of mutant genotype 3.

From these graphs, it can be seen in the upper panel of the figure that, in year 2,000 of the Monte Carlo simulation experiment, the population of infants, $x = 0$, and ages near that of infants had reached a population size of nearly $9 \times 10^6$, that is, a size approaching the carrying capacity of the environment. In the lower panel, it can be seen that the standard deviations for these early years of life were near the large value $6 \times 10^5$, which is indicative of a rather high level of variation among the 50 replications of Monte Carlo realizations of the process for these ages. As age increases, the numerical values depicted in the graphs of the mean and standard deviations for the number of individuals decrease rather rapidly, which again was a signature of high levels of mortality in the population across the ages. It should be noted, however, that there is a steep drop in the graphs for both the means and standard deviations at age 30, where it was assumed that mortality beyond this age mortality was governed by a Gompertz type risk function. There are methodological reasons for this drop. A reader may recall that the potential probability that an infant survives to age 30 was assigned number 0.99. It seems plausible, therefore, that this potential probability should be lowered. On the other hand, this drop also suggests that different values for the two parameters of the Gompertz distribution may also help level of f the steep decline in these graphs at age 30. While developing the formulation of the age structured stochastic process described in this paper, the most difficult problem encountered in this initial period of model development was the choices made in constructing and choosing the parametric components of the model governing the survival or mortality of individuals. It should also be noted that for ancient populations of man considered in the experiments reported in this paper, there is little or no data on mortality of individuals apart from the fossil record that suggest through measurements of radioactive that maximum ages of those individuals whose skeletons survived was in the range from 40 to 50 years. In passing, it should be mentioned that it is doubtful whether the changes in mortality parameters just suggested would have changed the differences in estimated timings for the emergence of rare mutations as observed in the realized predictions of the embedded deterministic and the stochastic process used in the simulation experiments reported in this section.

## 8. Sexual Selection in Evolving Age-Structured Populations

Among the many questions addressed by Darwin in his seminal work on evolution was that of the impact of sexual selection on the evolution of two-sex-species such as man, other mammals, and birds. A historical sketch of these ideas is beyond the scope of this paper, but if a reader is interested in the subject, a search of the internet may yield useful results, or one could also consult text books on population and evolutionary genetics, which are available in university libraries as well as for sale on the internet. For example, the book on population genetics by Hartl and Clarke [20] devotes some space to sexual selection. Essentially all the literature on population genetics in which models of sexual selection are presented are confined to deterministic paradigm, but in the book by Mode and Sleeman [9] a formulation of a two sex self-regulating branching process is the subject of chapter 11 along with reported results of some Monte Carlo simulation experiments. A limitation of the model considered in that chapter was that the age structure of an evolving population was not included in the formulation. Also this chapter contains the results of a Monte Carlo simulation experiment designed

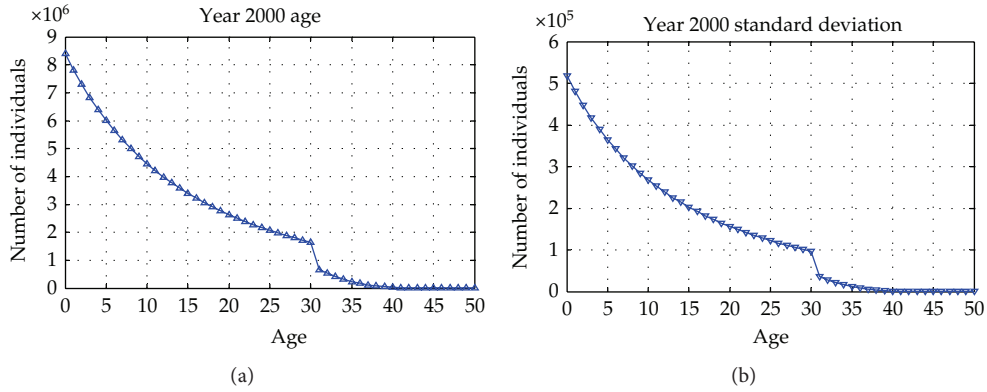


FIGURE 4: Graphs of mean and standard deviation by age for females of genotype 3 in year 2,000.

to study the rise of a recessive mutant allele in an evolving population when the character or trait it expressed was subject to sexual selection. In that chapter, sexual selection was characterized in terms of matrices of acceptance probabilities regrading both females and males as in (8). In Monte Carlo simulation experiments, it was shown that when it was assumed that the mutant genotype 3 = *aa* was preferred as a sexual partner over the other two genotypes, it would eventually arise to predominance in a population, but its steep rise to predominance would not occur until the total population size of individuals of genotype *aa* had reached a threshold. In this experiment the embedded deterministic model was a fair predictor of the central tendencies of Monte Carlo realization of the sample functions of the process, but in a related experiment, Mode et al. [12], the embedded deterministic model predicted a rare mutation that had sexual selective advantage would eventually become predominant in an evolving population, but in the corresponding Monte Carlo simulation experiment, the mutant recessive genotype did not appear, so that the predictions of the deterministic and stochastic models were not consistent.

In the Monte Carlo experiment reported in this section, the number of years of evolution under consideration was 10,000. Among the reasons underlying this choice was that about 10,000 years ago, in geographical regions of the world, such as Egypt and Mesopotamia, man began to change from a hunter-gather means of living to a more sedentary form of life, the development of agriculture that resulted in a more dependable source of food. At the outset it was known that if 50 Monte Carlo replication of an experiment were computed, the span if time taken to complete the experiment would be too long, given the private computer facilities available to carry out the experiment. So a decision was made to compute only 10 replications of 10,000 years of simulated evolution, which took about 30 to 40 hours of computer time. Because it was observed in some experiments not reported in this section that a significant number of individuals may be alive at age 50, a decision was made to set the maximum age at 80 years, which required the computer to process lager arrays that in turn would increase the expected time needed to complete an experiment. To take into account that sexual selection was operative in the experiment, the matrix of acceptance probabilities for females was chosen as

$$\mathbf{A}_f = \begin{pmatrix} 0.1 & 0.9 & 0.9 \\ 0.1 & 0.9 & 0.9 \\ 0.1 & 0.9 & 0.9 \end{pmatrix}. \tag{92}$$

According to this selection of numerical values, the lower case allele *a*, which arose from the mutation $A \rightarrow a$, was dominant with respect to sexual selection. Interestingly, in preliminary experiment with the embedded deterministic model, it was observed that if the values of the matrix $\mathbf{A}_f$ were chosen, such that the values in the first and second columns were assigned the constant 0.1, and the values in the third column were assigned the constant value 0.9, then individuals of the mutant genotype *aa* would eventually become predominate in the population, but if these values were used in a Monte Carlo simulation experiment, then the computer time required to complete an experiment would be too long. Thus, for this reason the values in the matrix $\mathbf{A}_f$ were chosen, because the study of the emergence of a dominant mutation would also be of interest in its own right.

By assumption, selection was neutral with respect to reproductive success so that the potential expected number of offspring contributed to the population by each female through out her lifespan was assigned the number 4 for all three genotypes. Observe that under this assumption the pace of evolution was expected to be slower than in the experiment reported in Section 6. To assure that an individual with a mutant gene would appear in a simulated population within an acceptable number of years, the probability of the mutation $A \rightarrow a$ per meiosis was assigned the value $\mu_{12} = 10^{-5}$, and the probability of the back mutation $a \rightarrow A$ per meiosis was assigned the value $\mu_{21} = 10^{-6}$. The carrying capacity of the environment for each genotype for both sexes was assigned the value, expressed in terms of the beta parameters, and had the constant value $10^{-9}$. All parameters not mentioned in the above discussion were assigned the same values as those in the experiment reported in Section 6. Moreover, it was assumed that the initial population of females and males was all of genotype *AA* and assigned the same initial numbers as those used in the experiment reported in Section 7.

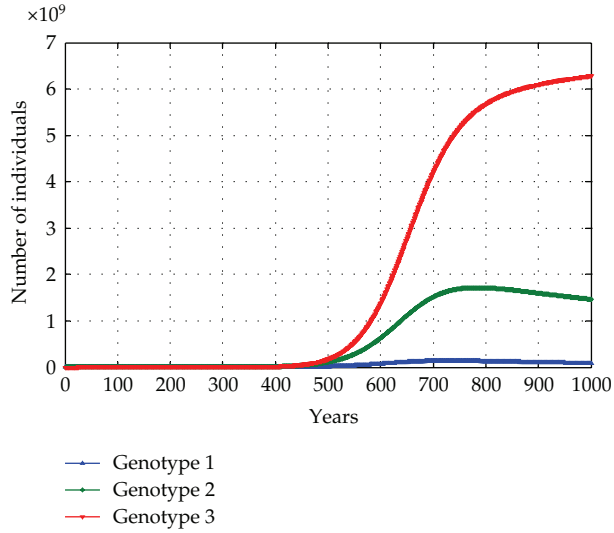


FIGURE 5: Deterministic trajectories of total numbers for each of the three male genotype under sexual selection.

As displayed in Figure 5 there are the trajectories of the three genotypes, expressed in terns of the numbers of individuals per genotype as functions of time, for the first 1,000 years as computed using the embedded deterministic model.

As can be seen from these trajectories, from an initial population of 2,693 males of genotype $AA$, within 500 years of simulated evolution the mutant genotype $aa$ starts to become predominant in the population, and by 1,000 years the number of individuals of this genotype has risen to over $6 \times 10^9$ but has not yet reached the maximum carrying capacity of the environment. As one would expect, the number of individuals of genotype $Aa$ occurs in significant numbers in the time interval from 500 to 1,000 years but is far below the number of individuals of genotype 3 throughout this time period. An inspection of the simulated data in the years preceding 10,000 years indicates that a balanced polymorphism had been reached in which the genotype $aa$ was predominant with genotypes $AA$ and $Aa$ present in smaller numbers. Unlike the cases in which the recessive genotype $aa$ is favored by sexual selection presented in Mode and Sleeman [9] and Mode et al. [12], there is gradual rise but no steep rise to predominance in the population of individuals of genotype $aa$.

From now on, due to limitations on space, attention will be focused on the predominant genotype 3 with respect to total population size for males of this genotype. As presented in Figure 6 there are graphs of the Min, $Q_{50}$, and Max trajectories as estimated from the Monte Carlo realizations of the process as well as the deterministic trajectory (DET) for the total number of individuals of mutant genotype $aa$ for the first 200 years of 10,000 years of simulated evolution.

As can be seen from this graph, after about 140 years of evolution, the number of individuals of genotype 3 = $aa$ had risen to significant numbers, but after 150 years of evolution, the variation among the 10 realizations of the process increased as can be seen from the distances among the Min, $Q_{50}$, and Max trajectories at year 200. Interestingly, however, at year 200, there is a large distance between the deterministic trajectory (DET) and the Max of the stochastic process. This observation demonstrates that even with the high probabilities of mutation used in the experiment under consideration, the embedded deterministic model tends to underestimate the number of mutations that actually arise among the simulated realizations of the process in the first 200 years of simulated evolution. But, as shown in the next figure, after a sufficient amount of evolutionary time, the four trajectories in Figure 6 tend to coalesce.

As can be seen from this figure on the scale of the vertical axis, the distance between the Min and DET trajectories is close up to about 2,500 years of evolution. But, in the years following 2,500, the four trajectories coalesce and remain near one another at 4,000 years at evolution at least on the scale of the vertical axis. In the years following 4,000, it was seen, by inspecting the statistical summarizations of the Monte Carlo realizations of the process as well as that for embedded deterministic model, that the pattern observed at 4,000 years of evolution continued to 10,000 years. For the class of stochastic processes under consideration, the trajectory of the embedded deterministic model was quite typical in the sense that in the early years of an experiment the predictions of the deterministic model may differ significantly from those of the process, but in the long term, with the exception of a few cases, the trajectories of the stochastic process and that of the deterministic model tend to coalesce. But, the waiting time to coalescence may indeed be very long when expressed in terms of human life spans (Figure 7).

Because we are dealing with an age structured process, it is appropriate to view graphs of a simulated age distribution at some selected year of the experiment. As presented in the upper panel of Figure 8 ther is a graph for males of the mutant genotype 3 = $aa$, at year 200 of the experiment, and the lower panel contains the graph of the standard deviations for each age as estimated from the simulated data.

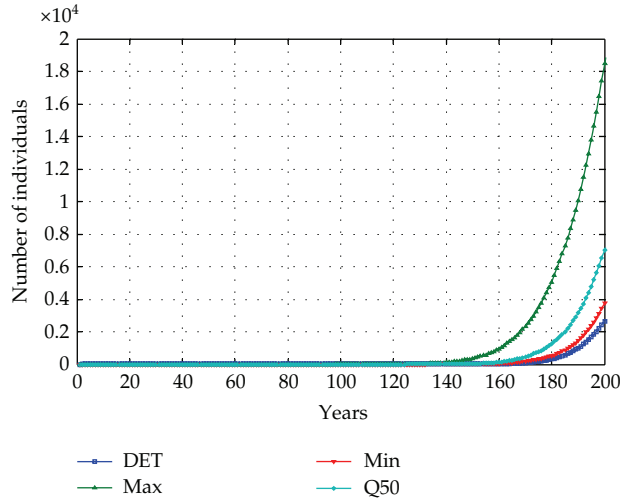


Figure 6: Estimated Min, $Q^{50}$ and Max trajectories of the process and the deterministic trajectory (DET) for the total number males of genotype 3.

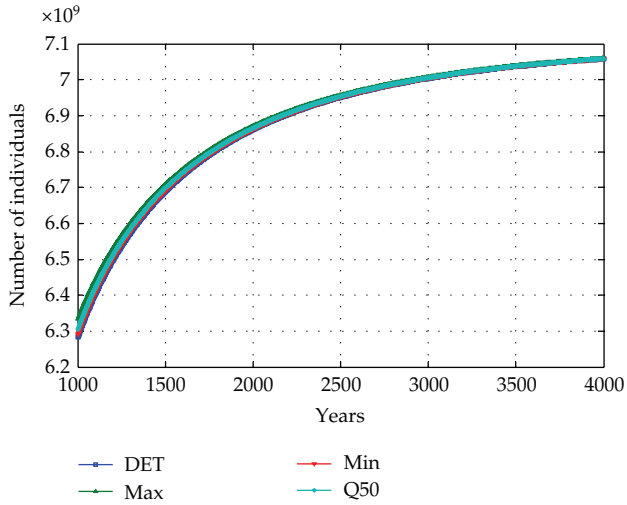


Figure 7: Stochastic and deterministic trajectories for the total numbers of males of genotype 3 for years from 1,000 to 4,000 of evolution.

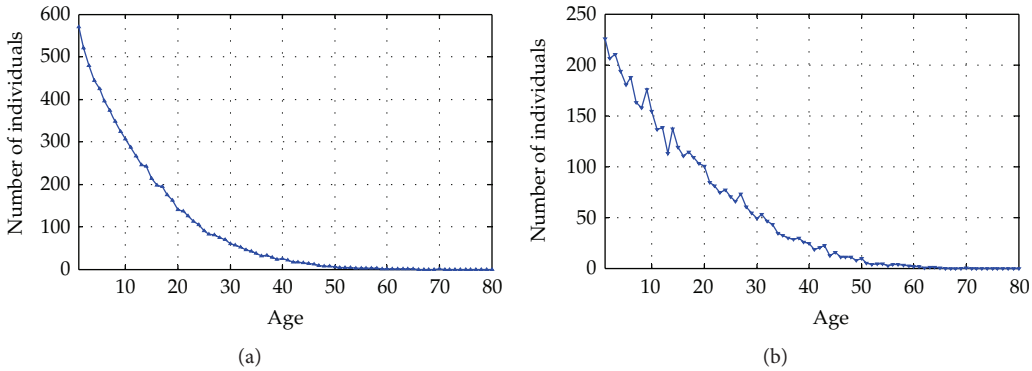


Figure 8: Estimated mean age distribution and standard deviations by age for the number of males of genotype 3 in year 200.

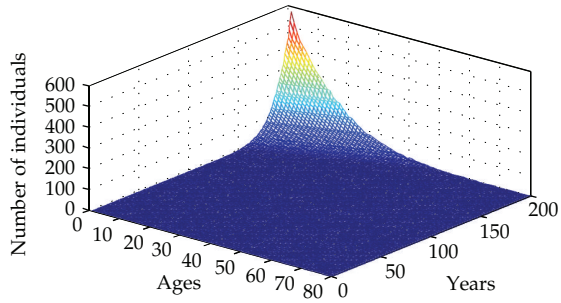


Figure 9: Three dimensional view of the evolution of the mean age distribution for males of genotype 3.

From the upper panel of the figure, it can be seen that by year 200 of the experiment the number of males of the mutant genotype 3 had risen to sufficiently high numbers to form a nearly monotone decreasing mean age distribution that one would expect in a population evolving with high levels of mortality and fertility as realized in a large number of births and deaths each year. But as can be seen from the lower panel for the graph of standard deviations by age, there was a considerable amount of variation by age around the mean age distribution.

It is also of interest to view a three-dimensional version of the evolving mean age distribution for males of mutant genotype 3 = *aa*. As presented in Figure 9 there is a three dimensional view of the evolving age distribution for males of genotype 3 for the first 200 years of evolution.

As can be seen from this figure, the $x$-axis denotes age, the $y$-axis denoted time in years, and the $z$-axis denotes the number of individuals. Unlike the conventional displays of an evolving age distribution, in which the numbers are normalized such that they all are in the interval $[0, 1]$, in this figure the number of individuals for each age is depicted. From an inspection of this figure, it can be seen that for the first of about 140 years of evolution the age distribution is a flat plane, indicating that the number of individuals of mutant genotype 3 was either zero at each group or a number too small to depict in a graph. However, sometime after 140 years of evolution the number of individuals of mutant genotype 3 increases in the population, and by year 200 the age distribution is typical of a fast growing population with high levels of mortality as is also shown in the upper panel of Figure 4.

A question that naturally arises is what is the three-dimension form of the three dimension standard deviation surface corresponding to the graph presented in Figure 9? Presented in Figure 10 is a graph of this evolving surface.

The $x$- and $y$-axes are the same as those in Figure 9, but on the $z$-axis the standard deviations are expressed in terms of units of individuals. Observe that shape of the standard deviation surface is very similar to that for the age distribution in Figure 9, but the range of values for the $z$ in Figure 10 is from 0 to 200; whereas in Figure 9 the range of the $z$ axis is from 0 to 600. Thus, at its maximum, the $z$ axis in Figure 10 is about only one-third as that for that axis in Figure 9.

A question that naturally arises is to what extent would the above figures change, if 50 to 100 realizations of the process had computed rather than the small sample of ten? It seems likely, based on other Monte Carlo simulation experiments, that from a qualitative perspective the estimated trajectories, as estimated from a Monte Carlo simulation experiment, would not change significantly, but if a larger number of realizations of the process were computed, it is likely more outliers would appear in the simulated data. Thus, for example, the Max at 200 years may have risen to a larger number than that displayed in Figure 6. Moreover, the standard deviations by age in the lower panel of Figure 8 would very likely be lager, indicating a greater range of variation among the simulated realizations of the process.

Another question that arises is what is the mathematical basis underlying the apparent stochastic-balanced polymorphism that was observed in the latter years of 10,000 years of simulated evolution? It can be shown that the two-sex age structured process under consideration is a Markov chain in discrete time with a state space consisting of pairs $(x, y)$ if $3 \times (r + 1)$ matrices of non-negative integers, where the rows represent genotypes and columns ages of individuals and $x$ and $y$ females and males, respectively. In the absence of immigration, the state $(0, 0)$ is an absorbing state, indicating extinction of the population. Given that extinction that did not occur, it appears that the process had converged to a quasi-stationary distribution of the Markov chain. No attempt will be made here to prove this assertion, but if the reader is interested in a derivation of a quasi-stationary distribution for a Wright-Fisher process with two absorbing sates, chapters 5 and 6 of Mode and Sleeman [9] may be consulted.

## 9. Discussion, Further Reading, and Future Experiments

In the two experiments reported in this paper dealing with selection, only one component of selection was dealt with in each experiment. For example, in one experiment dealing with the component of reproductive success, if females of one genotypes contributed to average a greater expected number of offspring than females of other genotypes, this selective advantage was sufficient for individuals of this genotype to become predominant in the population in the long run. On the other hand, if individuals of one genotype or a dominant phenotype in males was preferred by females as sexual partners for the case of sexual selection, this preference was

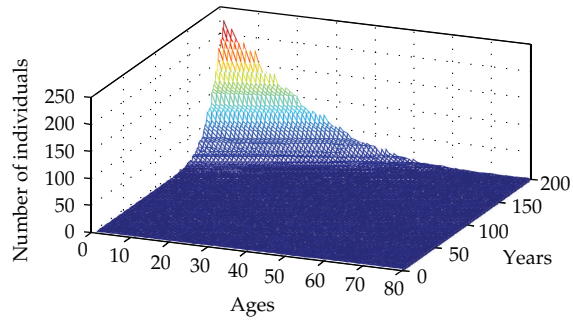


Figure 10: Three dimensional view of the evolving standard deviation surface for males of genotype 3.

sufficient for that genotype or phenotype to become predominant in a population in the long run. In both these experiments, all other components of selection were neutral in the sense that for each component not under consideration each genotype was assigned the same parameter values. In an experiment not reported in this paper, it was also shown that if individuals of genotype $3 = aa$ for either sex had a competitive advantage over the other two with respect to competition for resources, then individuals of genotype 3 would eventually become predominant in an evolving population.

A component of selection that has so far not been studied using the formulation under consideration is that of different survival patterns by age as formulated in terms of the parametric birth cohort survival function described in a previous section. For example, if females of one of the genotypes were very good mothers in the sense that most of their children survived to reproduce, then individuals of that genotype would have a selective advantage over individuals with other genotypes. It would be of interest to study this component of selection, but such computer experiments will be deferred to future research. Experiments dealing with combinations of components of selection would also be of interest. There is at least one other class of experiments that would be of considerable interest, namely, a case in which a human population evolves during a phase of evolution in which life is maintained by hunting animals, fishing, and gathering plant foods. Under such conditions the carrying capacity of the environment would be less than that considered in the experiments reported in this paper, in which it was tacitly assumed that agriculture had advanced to the point at which the environment could support population sizes of the order $10^9$. In a hunter-gather society, however, a plausible estimate of the carrying capacity of the environment may be one or two people per square mile. As an illustrative example, suppose that the environment could support two people, a female and male, per square mile, then a population of 1,000 females and males would require an area of 500 hundred square miles to support such a population. Under such conditions, one would expect that the pace of evolution would be very slow, because a population would not attain a total population size that would make it probable that rare beneficial mutations would occur and become predominant in a population.

Even though it was mentioned that genetic evidence suggests that the present world human population are all descendants of a small band of our species that migrated out of Africa about 60,000 years ago, little attention in this paper was given to the evolution of the world population over time. In this connection, the book by McEvedy and Jones [21] would be of interest in which estimates of population size are given from about 400 years before the common era (BCE) to 1975. Another book of interest is that of McKeown [22], which traces the evolution of the European population, and in particular the population of England and Wales, where the rise modern population did not start to occur until about 1750. As pointed out in this book, the evolution of modern medicine played a significant role recent rise of population. The book by Hostetler [23] is also of interest, because it provides a documentation of the Hutterite society over a little more than a century, since they migrated to South Dakota, USA, in the 19th century and have expanded their population to various communities in Montana and in neighboring prairie provinces of Canada. The Hutterites are of particular interest, because, as a population with records spanning over a century, they experienced one of the highest levels of fertility in recorded human history.

As this paper is being written, the estimated size of the world human population is at about 7 billion, and as this population grows, a question that naturally arises is how many people can the earth support? In a definitive book with this title, Cohen [24] addresses this question in depth. Among the many issues addressed in this book are questions regarding the carrying capacity of the earth. It is very difficult to provide usable answers to this question, but it may turn out that the approach used to formulate the notion of a carrying capacity suggested in this paper for an age structured population may be useful in obtaining experimental answers to this question. But it should also be mentioned that other approaches yet to be invented may also be useful in an experimental approach to estimate the carrying capacity of the earth.

It should also be mentioned that climate change has also played an important role in human evolution, and moreover, catastrophic events during the past have led to an abrupt climate change. An example of such a change was the eruption of the mega Toba volcano on the island of Sumatra about 70,000 to 75,000 years ago. According to Rampino and

Ambrose [25] during this mega eruption vast quantities of materials emanating from the earth's crust were thrust into the earth's atmosphere resulting in a "nuclear winter" that cooled the earth to the extent that there were frosts in the tropics for five to ten years that resulted in population crash of animals and plants, including populations of humans. According to this view, all human on the earth today are descendants of a relatively small population that survived the Toba eruption. A common term used by evolutionists to describe small population size during evolution of a species is "bottleneck."

The short list of references cited after that are informative but not current. It is, therefore, of interest to search for more recent publications on the evolution of world population. When this phrase was typed into an internet search engine, a total of about 29,100,000 sites were listed. Within this large listing of sites, a large number of technical papers were also listed that were beyond the scope of this paper. But, there were two significant papers that were within the scope of this paper. One on these papers was that of Hawks et al. [26] in which evidence for a population bottleneck was investigated during the Pleistocene period of human evolution that included the Toba eruption. Another paper was that of Stearns et al. [27] in which the measurement of selection was investigated, using data from contemporary populations. In comparison with an average human lifespan, evolution of a population occurs over very lager periods of time expressed in terms of generations. Therefore, when a contemporary set of data is used in attempts to measure selection, it is necessary to investigate the heritability of a trait measured from generation to generation, and as this paper is read it is recommended that a reader be mindful of the concept of heritability. For if an expression genotype or phenotype is not governed by some region (or regions) of a genome, the trait cannot be passed on from generation to generation and thus cannot be subject to natural selection.

In the evolving literature on evolutionary and population genetics, two themes have been received considerable attention, namely, genome wide searches for signatures of selection and methods for simulating models of genomes. In genome wide searches, a problem that often arises is that of developing statistical criteria to differentiate among regions of a genome, such that selection has acted in the recent evolutionary past from regions that presumedly have evolved under neutral evolution. Two papers describing methods for detecting signatures of positive selection are those of Sabeti et al. [28] and Grossman et al. [29]. Additional references on this subject may be found in the cited literature in Mode and Gallop [15]. As yet methods have not been developed to model the evolution region of a genome with respect to mutation and selection within the framework of the age structured process under consideration, but if a reader is interested in a brief review of published papers dealing with simulating the evolution of a genomic region, the papers cited in chapter 14 of Mode and Sleeman [9] may be consulted. Among these papers there are Chadeau-Hyam et al. [30], Hoggart et al. [31], and Hoggart et al. [32]. Preliminary algorithms for simulation various types of mutations, such as nucleotide substitutions, deletions, insertions, and repeats are also included in chapter 14 of Mode and Sleeman.

## References


[1] B. Charlesworth, *Evolution in Age-Structured Populations*, vol. 1 of *Cambridge Studies in Mathematical Biology*, Cambridge University Press, Cambridge, Mass, USA, 1980.

[2] T. E. Harris, *The Theory of Branching Processes*, Die Grundlehren der Mathematischen Wissenschaften, Bd. 119, Springer, Berlin, Germany, 1963.

[3] K. B. Athreya and P. E. Ney, *Branching Processes*, Springer, New York, NY, USA, 1972.

[4] P. Jagers, *Branching Processes with Biological Applications*, John Wiley & Sons, London, UK, 1975.

[5] C. J. Mode, *Multitype Branching Processes: Theory and Applications*, American Elsevier, New York, NY, USA, 1971.

[6] S. Asmussen and H. Hering, *Branching Processes*, vol. 3 of *Progress in Probability and Statistics*, Birkhäuser, Boston, Mass, USA, 1983.

[7] M. Kimmel and D. E. Axelrod, *Branching Processes in Biology*, vol. 19 of *Interdisciplinary Applied Mathematics*, Springer, New York, NY, USA, 2002.

[8] P. Haccou, P. Jagers, and V. A. Vatutin, *Branching Processes: Variation, Growth, and Extinction of Populations*, Cambridge Studies in Adaptive Dynamics, Cambridge University Press, Cambridge, UK, 2007.

[9] C. J. Mode and C. K. Sleeman, *Stochastic Processes in Genetics and Evolution: Computer Experiments in the Quantification of Mutation and Selection*, World Scientific Publishing, Hackensack, NJ, USA, 2012.

[10] C. J. Mode, *Stochastic Processes in Demography and Their Computer Implementation*, vol. 14 of *Biomathematics*, Springer, Berlin, Germany, 1985.

[11] C. J. Mode and C. K. Sleeman, *Stochastic Processes in Epidemiology, HIV/AIDS, Other Infectious Diseases and Computers*, World Scientific, Singapore, 2000.

[12] C. J. Mode, T. Raj, and C. K. Sleeman, "Monte Carlo implementations of two sex density dependent branching processes and their applications in evolutionary genetics," in *Applications of Monte Carlo Methods in Biology, Medicine and Other Fields of Science*, C. J. Mode, Ed., pp. 273–296, INTECH, 2011.

[13] C. J. Mode, T. Raj, and C. K. Sleeman, "Simulating the emergence and survival of mutations using a self regulating multitype branching processes," *Journal of Probability and Statistics*, vol. 2011, Article ID 867493, 39 pages, 2011.

[14] J. Hawks, *The Rise of Humans: Great Scientific Debates*, The Teaching Company, Chantilly, Va, USA, 2011, http://www.thegreatcourses.com/.

[15] C. J. Mode and R. J. Gallop, "A review on Monte Carlo simulation methods as they apply to mutation and selection as formulated in Wright-Fisher models of evolutionary genetics," *Mathematical Biosciences*, vol. 211, no. 2, pp. 205–225, 2008.

[16] L. Deng and J. Lin, "Random number generation of a new century," *The American Statistician*, vol. 54, no. 2, pp. 145–150, 2000.

[17] P. L'Ecuyer and R. Touzin, "On the Deng-Lin random number generators and related methods," *Statistics and Computing*, vol. 14, no. 1, pp. 5–9, 2004.

[18] L. Deng, "Efficient and portable multiple recursive generators of large order," *ACM Transactions on Modeling and Computer Simulation*, vol. 15, no. 1, pp. 1–13, 2005.

[19] M. Alderson, *International Mortality Statistics*, Facts on File, Inc., New York, NY, USA, 1981.

[20] D. L. Hartl and A. G. Clark, *Principles of Population Genetics*, Sinauer Associates, INC., Sunderland, Mass, USA, 1989.

[21] C. McEvedy and R. Jones, *Atlas of World Population History*, Facts on File, New York, NY, USA, 1979.

[22] T. McKeown, *The Modern Rise of Population*, Academic Press, New York, NY, USA, 1976.

[23] J. Hostetler, *Hutterite Society*, The Johns Hopkins University Press, Baltimore, Md, USA, 1974.

[24] J. E. Cohen, *How Many People Can the Earth Support?* W.W. Norton & Company, New York, NY, USA, 1995.

[25] M. Rampino and S. Ambrose, "Volcanic winter in the Garden of Eden: the Toba supereruption and the late Pleistocene human population crash," in *Symposium: Catastrophism, Natural Disasters and Cultural Change*, World Archaeological Congress 4, University of Cape Town, 1999.

[26] J. Hawks, K. Hunley, S. Lee, and M. Wolpoff, "Population Bottlenecks and Pleistocene evolution," *Molecular Biology and Evolution*, vol. 17, no. 1, pp. 2–22, 2000.

[27] S. C. Stearns, S. G. Byars, D. R. Govindaraju, and D. Ewbank, "Measuring selection in contemporary human populations," *Nature Reviews Genetics*, vol. 11, pp. 611–622, 2010.

[28] P. Sabeti et al., "Positive selection in the human linage," *Science*, vol. 312, pp. 1614–1620, 2006.

[29] S. Grossman et al., "A composite of multiple signals distinguishes causal variants in regions of positive selection," *Science*, vol. 327, pp. 883–886, 2010.

[30] M. Chadeau-Hyam et al., "Fregene: simulation of realistic sequencelevel-data in populations and ascertained samples," *BMC Bioinformatics*, vol. 9, article 364, 2008.

[31] C. Hoggart et al., "Sequence level population simulations over large genomic regions," *Genetics*, vol. 177, pp. 1725–1731, 2007.

[32] C. Hoggart et al., "Genome-wide significance for dense *SNP* and resequencing data," *Genetic Epidemiology*, vol. 32, pp. 179–185, 2008.

Hindawi Publishing Corporation
International Journal of Stochastic Analysis
Volume 2012, Article ID 694285, 24 pages
doi:10.1155/2012/694285

*Research Article*

# A Decomposable Branching Process in a Markovian Environment

**Vladimir Vatutin,[1] Elena Dyakonova,[1] Peter Jagers,[2] and Serik Sagitov[2]**

[1] *Steklov Mathematical Institute, Russian Academy of Sciences, Moscow 119991, Russia*
[2] *Chalmers University of Technology and University of Gothenberg, 412 96 Gothenburg, Sweden*

Correspondence should be addressed to Serik Sagitov, serik@chalmers.se

Received 28 September 2012; Accepted 18 December 2012

Academic Editor: Peter Olofsson

Copyright © 2012 Vladimir Vatutin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A population has two types of individuals, with each occupying an island. One of those, where individuals of type 1 live, offers a variable environment. Type 2 individuals dwell on the other island, in a constant environment. Only one-way migration ($1 \to 2$) is possible. We study then asymptotics of the survival probability in critical and subcritical cases.

## 1. Introduction

Multitype branching process in random environment is a challenging topic with many motivations from population dynamics (see, e.g., [1–3]). Very little is known in the general case and in this paper we consider a particular two-type branching process with two key restrictions: the process is decomposable and the final type individuals live in a constant environment.

The subject can be viewed as a stochastic model for the sizes of a geographically structured population occupying two islands. Time is assumed discrete, so that one unit of time represents a generation of individuals, some living on island 1 and others on island 2. Those on island 1 give birth under influence of a randomly changing environment. They may migrate to island 2 immediately after birth, with a probability again depending upon the current environmental state. Individuals on island 2 do not migrate and their reproduction law is not influenced by any changing environment. Our main concern is the survival probability of the whole population.

An alternative interpretation of the model under study might be a population (type 1) subject to a changing environment, say in the form of a predator population of stationary

but variable size. Its individuals may mutate into a second type, no longer exposed to the environmental variation (the predators do not regard the mutants as prey). Our framework may be also suitable for modeling early carcinogenesis, a process in which mutant clones repeatedly arise and disappear before one of them becomes established [4, 5]. See [6] for yet another possible application.

The model framework is that furnished by Bienaymé-Galton-Watson (BGW) processes with individuals living one unit of time and replaced by random numbers of offspring which are conditionally independent given the current state of the environment. We refer to such individuals as particles in order to emphasize the simplicity of their lives. Particles of type 1 and 2 are distinguished according to the island number they are occupying at the moment of observation. Our main assumptions are

(i) particles of type 1 form a critical or subcritical branching process in a random environment,

(ii) particles of type 2 form a critical branching process which is independent of the environment.

Let $X_n$ and $Z_n$ be the numbers of particles of type 1 and of type 2, respectively, present at time $n$. Throughout this paper it is assumed (unless otherwise specified) that $X_0 = 1$ and $Z_0 = 0$. We investigate asymptotics of the survival probability $\mathbb{P}[X_n + Z_n > 0]$ as $n \to \infty$. In all cases addressed here we have

$$\mathbb{P}[X_n > 0] = o(\mathbb{P}[X_n + Z_n > 0]). \tag{1.1}$$

Therefore, in view of

$$\mathbb{P}[Z_n > 0] \le \mathbb{P}[X_n + Z_n > 0] \le \mathbb{P}[X_n > 0] + \mathbb{P}[Z_n > 0], \tag{1.2}$$

we focus on the asymptotic behavior of $\mathbb{P}[Z_n > 0]$.

In Section 2 we recall known facts for constant environments. They will then be compared to the results of this paper on random environments. In Section 3.1 we describe IID environments (Independent and Identically Distributed environmental states) and then in Section 3.2 Markovian environments. The main results of the paper are

(i) Theorem 4.1 in Section 4 on the critical case with an IID environment,

(ii) Theorem 5.1 in Section 5 on the subcritical case with an IID environment,

(iii) Theorem 6.1 in Section 6 on the critical case with a Markovian environment,

(iv) Theorem 7.1 in Section 7 on subcritical case with a Markovian environment.

Theorems 6.1 and 7.1 treating the case of Markovian environment are extensions of Theorems 4.1 and 5.1 obtained under rather restrictive conditions and yielding qualitatively the same asymptotic behavior as in the case of IID environment.

*Notation 1.* In asymptotic formulae constants denoted by the same letter $c$ are always assumed to be fixed and independent of the parameter that tends to infinity (or zero).

## 2. Two-Type Decomposable Branching Processes

Consider a two-type BGW process initiated at time zero by a single individual of type 1. We focus on the decomposable case where type 1 particles may produce particles of types 1 and 2, while type 2 particles can give birth only to type 2 particles. Put

(i) $Y_n :=$ the number of type 2 daughters produced by the particles of type 1 present at time $n$, in particular, $Y_0 = Z_1$,

(ii) $T :=$ the first time $n$ when $X_n = 0$, so that $\{T > n\} = \{X_n > 0\}$,

(iii) $S_n := \sum_{k=0}^{n-1} X_k$, so that $S_{\mathrm{T}}$ gives the total number ever of type 1 particles,

(iv) $W_n := \sum_{k=0}^{n-1} Y_k$, so that $W_{\mathrm{T}}$ gives the total number of type 2 daughters produced by all $S_{\mathrm{T}}$ particles of type 1.

The aim of this section is to summarize what is already known about such branching processes in the case of a constant environment. This will pave our way in terms of notation and basic manipulation with generating functions towards branching processes in IID random and then Markovian environments.

If the environment is constant from generation to generation, two-type decomposable BGW processes are fully described by a pair of probability generating functions

$$\begin{aligned} f(s_1, s_2) &:= \mathbb{E}\left[s_1^{\xi_1} s_2^{\xi_2}\right], \\ h(s) &:= \mathbb{E}[s^{\eta}], \end{aligned} \tag{2.1}$$

where $\xi_1$ and $\xi_2$ represent the numbers of daughters of type 1 and 2 of a mother of type 1, while $\eta$ stands for the number of daughters (necessarily of type 2) of a mother of type 2. Let

$$\begin{aligned} \mu_1 &:= \mathbb{E}[\xi_1] = \left.\frac{\partial f(s_1, s_2)}{\partial s_1}\right|_{s_1=s_2=1}, \\ \mu_2 &:= \mathbb{E}[\xi_1(\xi_1 - 1)] = \left.\frac{\partial^2 f(s_1, s_2)}{\partial s_1^2}\right|_{s_1=s_2=1}, \\ \theta_1 &:= \mathbb{E}[\xi_2] = \left.\frac{\partial f(s_1, s_2)}{\partial s_2}\right|_{s_1=s_2=1}, \\ \theta_2 &:= \mathbb{E}[\xi_2(\xi_2 - 1)] = \left.\frac{\partial^2 f(s_1, s_2)}{\partial s_2^2}\right|_{s_1=s_2=1}, \\ m_1 &:= \mathbb{E}[\eta] = h'(1), \\ m_2 &:= \mathbb{E}[\eta(\eta - 1)] = h''(1) \end{aligned} \tag{2.2}$$

be the first two moments of the reproduction laws. Concerning the second type of particles we assume that

$$m_1 = 1, \qquad m_2 \in (0, \infty), \tag{2.3}$$

implying that the probability of extinction

$$Q_n := \mathbb{P}[Z_n = 0 \mid X_0 = 0, Z_0 = 1] \tag{2.4}$$

(of a single-type BGW process evolving in constant environment with the probability generating function $h(s)$) satisfies [7, Chapter I.9]

$$1 - Q_n \sim \frac{2}{m_2 n}, \quad n \longrightarrow \infty. \tag{2.5}$$

It follows that

$$a_n := -\log f(1, Q_n) \sim 1 - f(1, Q_n) \sim \frac{2\theta_1}{m_2 n}, \quad n \longrightarrow \infty. \tag{2.6}$$

We will be interested in two kinds of reproduction regimes for particles of type 1, critical and subcritical. In the constant environment setting with $\mu_2 \in (0, \infty)$, the critical case corresponds to $\mu_1 = 1$, and the subcritical case is given by $\mu_1 \in (0, 1)$. In the critical case with a constant environment we have

$$\mathbb{P}[X_n > 0] = \mathbb{P}[T > n] \sim \frac{2}{\mu_2 n}, \quad n \longrightarrow \infty, \tag{2.7}$$

and according to [8, Theorem 1]

$$\mathbb{P}[X_n + Z_n > 0] \sim \mathbb{P}[Z_n > 0] \sim \frac{2\sqrt{\theta_1}}{\sqrt{m_2 \mu_2 n}}, \quad n \longrightarrow \infty. \tag{2.8}$$

Next we outline a proof of (2.8) based on the representation

$$\begin{aligned} \mathbb{P}[Z_n > 0] &= \mathbb{E}\left[1 - \prod_{k=0}^{n-1} Q_{n-k}^{Y_k}\right] = \mathbb{E}\left[1 - \prod_{k=0}^{n-1} f^{X_k}(1, Q_{n-k})\right] \\ &= \mathbb{E}\left[1 - e^{-\sum_{k=0}^{n-1} X_k a_{n-k}}\right], \end{aligned} \tag{2.9}$$

preparing for the proof in the random environment case, to be given in Section 4. Thanks to (2.7) and (1.2), it is enough to verify that

$$\mathbb{P}[Z_n > 0] \sim \frac{2\sqrt{\theta_1}}{\sqrt{m_2 \mu_2 n}}, \quad n \longrightarrow \infty, \tag{2.10}$$

in order to prove (2.8). However, by the branching property the total progeny of a single-type branching process $S_T$ is 1 plus $\xi_1$ independent daughter copies of $S_T$. In terms of the Laplace transform

$$\phi(\lambda) = e^{-\lambda} f(\phi(\lambda), 1), \tag{2.11}$$

where $\phi(\lambda) := \mathbb{E}[e^{-\lambda S_T}]$. As $\lambda \to 0$, a Taylor expansion of $f(\phi(\lambda), 1)$ as a function of $1 - \phi(\lambda)$ yields

$$1 - \phi(\lambda) = 1 - e^{-\lambda} + e^{-\lambda}\mu_1\big(1 - \phi(\lambda)\big) - e^{-\lambda}\frac{\mu_2}{2}\big(1 - \phi(\lambda)\big)^2(1 + o(1)). \tag{2.12}$$

For $\mu_1 = 1$, after removing the negligible terms, we get a quadratic equation whose solution shows that

$$1 - \phi(\lambda) \sim \sqrt{\frac{2\lambda}{\mu_2}}, \quad \lambda \longrightarrow 0. \tag{2.13}$$

Replacing $\lambda$ by $a_n$ and using (2.6), we obtain

$$\mathbb{E}\Big[1 - e^{-S_T a_n}\Big] \sim \frac{2\sqrt{\theta_1}}{\sqrt{m_2\mu_2 n}}, \quad n \longrightarrow \infty. \tag{2.14}$$

It remains to verify, see (2.9), that

$$\mathbb{E}\Big[1 - e^{-\sum_{k=0}^{n-1} X_k a_{n-k}}\Big] \sim \mathbb{E}\Big[1 - e^{-S_T a_n}\Big], \quad n \longrightarrow \infty. \tag{2.15}$$

This holds, indeed, since by (2.7) and for any fixed $\epsilon > 0$ the probability $\mathbb{P}[T > n\epsilon]$ is much smaller than the target value of order $c/\sqrt{n}$. (In [9, 10] infinite second moments in decomposable two-type critical processes were allowed.)

On the other hand, in the subcritical case (2.12) implies that

$$1 - \phi(\lambda) \sim \frac{\lambda}{(1 - \mu_1)}, \quad \lambda \longrightarrow 0, \tag{2.16}$$

so that by (2.6)

$$\mathbb{E}\Big[1 - e^{-S_T a_n}\Big] \sim \frac{2\theta_1}{m_2(1 - \mu_1)n}, \quad n \longrightarrow \infty. \tag{2.17}$$

In view of $\mathbb{P}[X_n > 0] \sim c\mu_1^n$ we conclude that in the subcritical case

$$\mathbb{P}[X_n + Z_n > 0] \sim \mathbb{P}[Z_n > 0] \sim \frac{2\theta_1}{m_2(1 - \mu_1)n}, \quad n \longrightarrow \infty. \tag{2.18}$$

See [11] for a comprehensive study of subcritical decomposable branching processes in a constant environment.

## 3. Branching Processes in a Random Environment

A randomly changing environment for BGW processes is modeled by a random sequence of probability generating functions for the offspring distributions of consecutive generations. Throughout this paper we assume that the offspring distribution for type 2 particles is the same across the different states of the environment and characterized by the same generating function $h(s)$. This restriction that greatly simplifies analysis still allows new interesting asymptotic regimes.

We consider two types of stationarily changing environments: IID and Markovian.

### 3.1. IID Environment

Our description of the IID environment case starts with a simple illustration based on just two alternative bivariate generating functions $f^{(1)}(s_1, s_2)$ and $f^{(2)}(s_1, s_2)$ with mean offspring numbers $(\mu_1^{(1)}, \theta_1^{(1)})$ and $(\mu_1^{(2)}, \theta_1^{(2)})$, respectively. We assume that at each time $n$ the environment is say "good" with probability $\pi_1$, so that type 1 particles reproduce independently according to the $f^{(1)}(s_1, s_2)$, and with probability $\pi_2 = 1 - \pi_1$ the environment is "bad" and particles of type 1 reproduce according to the $f^{(2)}(s_1, s_2)$ law. In other words, the generating function $f(s_1, s_2)$ should be treated as a random function having distribution

$$\mathbb{P}\left[f(s_1, s_2) = f^{(1)}(s_1, s_2)\right] = \pi_1, \qquad \mathbb{P}\left[f(s_1, s_2) = f^{(2)}(s_1, s_2)\right] = \pi_2. \tag{3.1}$$

In particular, the vector of the mean offspring numbers $(\mu_1, \theta_1)$ takes values $(\mu_1^{(1)}, \theta_1^{(1)})$ and $(\mu_1^{(2)}, \theta_1^{(2)})$ with probabilities $\pi_1$ and $\pi_2$.

More generally, our two-type branching process in an IID random environment is characterized (besides the fixed reproduction law $h(s)$ for the type 2 particles) by a sequence of generating functions $\{f_n(s_1, s_2)\}_{n=0}^{\infty}$ independently drawn from a certain distribution over probability generating functions so that

$$f_n(s_1, s_2) \stackrel{d}{=} f(s_1, s_2). \tag{3.2}$$

In this setting the respective conditional moments $\mu_1$, $\mu_2$, $\theta_1$, and $\theta_2$ should be treated as random variables. An important role is played by the random variable $\zeta := \log \mu_1$ representing the step size of the so-called associated random walk [12] formed by the partial sums $\zeta_0 + \cdots + \zeta_{n-1}$ with $\zeta_i \stackrel{d}{=} \zeta$.

*Notation 2.* Characteristics of the reproduction law in generation $n$ are denoted by adding an extra lower index $n$ to the generic notation, like in (3.2).

### 3.2. Markovian Environment

One way to relax the IID assumption on the environment is to allow for Markovian dependence among its consecutive states. We implement this by modelling changes in terms of an irreducible aperiodic positive recurrent Markov chain $\{e_n\}_{n=0}^{\infty}$ with countably many states $\{1, 2, \ldots\}$. Assuming a stationary initial distribution $(\pi_1, \pi_2, \ldots)$, we associate

with each state $i$ of this chain a probability generating function $f^{(i)}(s_1, s_2)$, so that the changing environment for the branching process is governed by the sequence of identically reproduction laws

$$f_n(s_1, s_2) := f^{(e_n)}(s_1, s_2), \quad n = 0, 1, \ldots \tag{3.3}$$

with Markovian dependence. Due to the stationarity we can again write (3.2) and use the same notation for the marginal moments of the reproduction laws as in the IID case.

To build a bridge to the IID environment case we use an embedding through a sequence of regeneration moments $\{\tau_k\}_{k=0}^{\infty}$ defined as

$$\tau_0 := 0, \qquad \tau_{k+1} := \min\{n > \tau_k : e_n = e_0\}. \tag{3.4}$$

The times $\tau_{k+1} - \tau_k$ between consecutive regenerations are independent and all distributed as $\tau := \tau_1$. The embedded process $(\widehat{X}_n, \widehat{Z}_n)$ defined as

$$\left(\widehat{X}_n, \widehat{Z}_n\right) := (X_{\tau_n}, Z_{\tau_n}), \quad n = 0, 1, \ldots \tag{3.5}$$

is a decomposable branching process in an IID environment with two types of particles $\widehat{1}$ and $\widehat{2}$ and conditional reproduction generating functions

$$\widehat{f}(s_1, s_2) := f^{(e_0)}\left(f^{(e_1)}\left(\cdots\left(f^{(e_{\tau-1})}(s_1, s_2), h(s_2)\right)\cdots\right), h_{\tau-1}(s_2)\right), \tag{3.6}$$

$$\widehat{h}(s) := h(h(\cdots h(s)\cdots)) = h_\tau(s), \tag{3.7}$$

where $h_k(s)$ stands for the $k$-fold iteration of $h(s)$.

*Notation 3.* For all characteristics of the embedded process $(\widehat{X}_n, \widehat{Z}_n)$ and related constants appearing in the asymptotic formulae we use the same notation as for the process $(X_n, Z_n)$ in the IID case just adding the hat sign.

The key difference from the IID case is that the reproduction law for the $\widehat{2}$-type particles is dependent on the random environment. However, this dependence is of specific nature which we are able to manage using the law of large numbers for renewal processes. Notice that on its own the $\widehat{2}$-type particles form a so-called degenerate critical branching process in an IID random environment [12]: its conditional offspring mean is deterministic $\widehat{m}_1 = 1$. Meanwhile, the conditional variance is random $\widehat{m}_2 = \tau m_2$.

Taking the first- and second-order derivatives of (3.6), we can express the moments of the reproduction law of the embedded process in terms of the moments of the consecutive reproduction laws with Markovian dependence. In what follows we use Notation 2 again,

while keeping in mind that the sequence $(\mu_{1,k}, \mu_{2,k}, \theta_{1,k}, \theta_{2,k})_{k=0}^{\tau-1}$ now consists of dependent random vectors. It can be shown that

$$\begin{aligned} \widehat{\mu}_1 &= \prod_{k=0}^{\tau-1} \mu_{1,k}, \qquad \widehat{\mu}_2 = \widehat{\mu}_1 \sum_{k=0}^{\tau-1} \frac{\mu_{2,k}}{\mu_{1,k}} \prod_{i=k+1}^{\tau-1} \mu_{1,i}, \\ \widehat{\theta}_1 &= \sum_{k=0}^{\tau-1} \theta_{1,k} \prod_{i=0}^{k-1} \mu_{1,i}, \end{aligned} \tag{3.8}$$

where, as usual, the product of the elements of an empty set is one. Furthermore, setting

$$A_{k,n} = \sum_{j=k}^{n} \theta_{1,j} \prod_{i=k}^{j-1} \mu_{1,i}, \tag{3.9}$$

we can write

$$\widehat{\theta}_2 = \sum_{k=0}^{\tau-1} \theta_{2,k} \prod_{i=0}^{k-1} \mu_{1,i} + \sum_{k=0}^{\tau-2} \left\{ \mu_{2,k} A_{k+1,\tau-1}^2 + 2\mu_{1,k}\theta_{1,k} A_{k+1,\tau-1} + \sigma^2(\tau - 1 - k)\theta_{1,k} \right\} \prod_{i=0}^{k-1} \mu_{1,i}. \tag{3.10}$$

**Lemma 3.1.** *Let*

$$\sum_{k=0}^{\infty} \mathbb{E}\left[ |\zeta_k| 1_{\{\tau \geq k+1\}} \right] < \infty. \tag{3.11}$$

*For the following sum of a random number of random variables*

$$\widehat{\zeta} := \sum_{k=0}^{\tau-1} \zeta_k, \quad \zeta_k := \log \mu_{1,k} = \zeta(e_k), \tag{3.12}$$

*a version of the Wald identity holds:* $\mathbb{E}[\widehat{\zeta}] = \mathbb{E}[\tau]\mathbb{E}[\zeta]$.

*Proof.* For any state $j$ consider the function

$$\mu_j(i) := \sum_{n=0}^{\infty} \mathbb{P}\left[ e_n = i, \tau > n \mid e_0 = j \right]. \tag{3.13}$$

According to [13, Theorem 6.5.2] this defines a stationary measure which is necessarily of the form $\mu_j(i) = c_j \pi_i$. The constants $c_j$ are such that

$$\begin{aligned} \sum_{j=1}^{\infty} \pi_j c_j &= \sum_{j=1}^{\infty} \pi_j \sum_{i=1}^{\infty} \mu_j(i) = \sum_{j=1}^{\infty} \pi_j \sum_{n=0}^{\infty} \sum_{i=1}^{\infty} \mathbb{P}\left[ e_n = i, \tau > n \mid e_0 = j \right] \\ &= \sum_{j=1}^{\infty} \pi_j \mathbb{E}\left[ \tau \mid e_0 = j \right] = \mathbb{E}[\tau]. \end{aligned} \tag{3.14}$$

It follows that

$$\begin{aligned}\mathbb{E}\left[\widehat{\zeta}\right] &= \sum_{k=1}^{\infty}\sum_{n=0}^{k-1}\mathbb{E}\left[\zeta_n 1_{\{\tau=k\}}\right] = \sum_{n=0}^{\infty}\mathbb{E}\left[\zeta(e_n)1_{\{\tau>n\}}\right] \\ &= \sum_{n=0}^{\infty}\sum_{i=1}^{\infty}\zeta(i)\mathbb{P}[e_n = i, \tau > n] = \sum_{i=1}^{\infty}\zeta(i)\sum_{j=1}^{\infty}\pi_j\mu_j(i) \\ &= \sum_{i=1}^{\infty}\zeta(i)\pi_i\mathbb{E}[\tau] = \mathbb{E}[\tau]\mathbb{E}[\zeta].\end{aligned} \tag{3.15}$$

□

Developing the example of two environmental states from Section 3.1, let us consider a Markov chain $\{e_n\}_{n=0}^{\infty}$ with transition probabilities

$$\begin{pmatrix} 1-d\pi_2 & d\pi_2 \\ d\pi_1 & 1-d\pi_1 \end{pmatrix}, \quad 0 < d < \min\left(\frac{1}{\pi_1}, \frac{1}{\pi_2}\right) \tag{3.16}$$

and a stationary distribution $(\pi_1, \pi_2)$. (Notice that $d = 1$ corresponds to the IID case.) Under stationarity the regeneration time satisfies

$$\begin{aligned}\mathbb{P}[\tau = 1] &= \pi_1(1-d\pi_2) + \pi_2(1-d\pi_1) = 1 - 2\pi_1\pi_2 d, \\ \mathbb{P}[\tau = k] &= \pi_1 d\pi_2(1-d\pi_1)^{k-2}d\pi_1 + \pi_2 d\pi_1(1-d\pi_2)^{k-2}d\pi_2 \\ &= d\pi_1\pi_2\left(d\pi_1(1-d\pi_1)^{k-2} + d\pi_2(1-d\pi_2)^{k-2}\right), \quad k \geq 2,\end{aligned} \tag{3.17}$$

implying that

$$\mathbb{E}[\tau - 1] = 1, \qquad \mathbb{E}[\tau(\tau-1)] = \frac{2}{d\pi_1\pi_2} - \frac{4}{d}. \tag{3.18}$$

If $(b_1, b_2)$ are the two possible values for $\zeta$, we can write $\mathbb{E}[\zeta] = \pi_1 b_1 + \pi_2 b_2$ and

$$\begin{aligned}\mathbb{E}\left[\widehat{\zeta}\right] &= \mathbb{E}\left[\mathbb{E}\left[\widehat{\zeta} \mid \tau\right]; \tau = 1\right] + \mathbb{E}\left[\mathbb{E}\left[\widehat{\zeta} \mid \tau\right]; \tau \geq 2\right] \\ &= \pi_1(1-d\pi_2)b_1 + \pi_2(1-d\pi_1)b_2 \\ &\quad + \pi_1 d\pi_2\left(b_1 + \frac{b_2}{d\pi_1}\right) + \pi_2 d\pi_1\left(b_2 + \frac{b_1}{d\pi_2}\right) = 2\mathbb{E}[\zeta],\end{aligned} \tag{3.19}$$

in full agreement with Lemma 3.1.

## 4. Critical Processes in IID Environment

The single-type critical branching process with an IID environment displays an asymptotic behavior that is in stark contrast with the constant environment formula (2.7). According to [14, Theorem 1], if

$$\mathbb{E}[\zeta] = 0, \qquad \mathrm{Var}[\zeta] \in (0, \infty), \tag{4.1}$$

$$\mathbb{E}\left[\mu_2 \mu_1^{-2} \left(1 + \max\left(0, \log \mu_1\right)\right)\right] < \infty, \tag{4.2}$$

then for some positive constant $c$

$$\mathbb{P}[X_n > 0] = \mathbb{P}[T > n] \sim \frac{c}{\sqrt{n}}, \quad n \longrightarrow \infty. \tag{4.3}$$

(A much more general limit theorem is obtained in [12].) The following theorem shows that in the decomposable case the difference between constant and random environments is even more striking. For constant environments the survival probability decays as $c/\sqrt{n}$, see (2.8), but in random environments the decay is like $c/\log n$.

**Theorem 4.1.** *Consider a critical decomposable branching process in an IID environment satisfying* (2.3) (4.1), (4.2), *and*

$$\mathbb{E}\left[\mu_1^{-1}\right] < \infty. \tag{4.4}$$

*If for some positive* $\alpha$

$$\begin{aligned} \mathbb{P}\left[\theta_1 < \frac{1}{x}\right] &= o\left(\left(\log x\right)^{-3-\alpha}\right), \quad x \longrightarrow \infty, \\ \mathbb{P}[\theta_1 > x] &= o\left(\left(\log x\right)^{-3-\alpha}\right), \quad x \longrightarrow \infty, \\ \mathbb{P}[\theta_2 > x\theta_1] &= o\left(\left(\log x\right)^{-3-\alpha}\right), \quad x \longrightarrow \infty, \end{aligned} \tag{4.5}$$

*then there exists a constant* $K_0$ *such that*

$$\mathbb{P}[Z_n > 0] \sim \frac{K_0}{\log n}, \quad n \longrightarrow \infty. \tag{4.6}$$

Before turning to the proof, we make some comments on the conditions and statement of this theorem.

*Notation 4.* We will often use the abbreviations $x_a := (\log x)^{2+a}$ and $n_a := (\log n)^{2+a}$.
Condition (4.5) is needed for the following properties to hold for any fixed $\varepsilon > 0$; recall notation agreements Notations 2 and 4,

$$\mathbb{P}\left[\min_{0\le k\le x_\alpha} \theta_{1,k} < x^{-\varepsilon}\right] = o\left(\frac{1}{\log x}\right), \quad x \longrightarrow \infty, \tag{4.7}$$

$$\mathbb{P}\left[\max_{0\le k\le x_\alpha} \theta_{1,k} > x^{\varepsilon}\right] = o\left(\frac{1}{\log x}\right), \quad x \longrightarrow \infty, \tag{4.8}$$

$$\mathbb{P}\left[\max_{0\le k\le x_\alpha} \left(\frac{\theta_{2,k}}{\theta_{1,k}}\right) > x^{\varepsilon}\right] = o\left(\frac{1}{\log x}\right), \quad x \longrightarrow \infty. \tag{4.9}$$

Each of them is proven via an intermediate step like

$$\mathbb{P}\left[\min_{0\le k\le x_\alpha} \theta_{1,k} < x^{-\varepsilon}\right] \le x_\alpha \mathbb{P}\left[\theta_1 < x^{-\varepsilon}\right], \tag{4.10}$$

relying on the IID assumption for consecutive environmental states. The constant $K_0$ in the statement of Theorem 4.1 is the same as in the asymptotic formula from [15] concerning the total number $S_{\mathrm{T}}$ of particles of type 1 ever appeared in the process:

$$\mathbb{P}[S_{\mathrm{T}} > x] \sim \frac{K_0}{\log x}, \quad x \longrightarrow \infty. \tag{4.11}$$

This constant has a complicated nature and is not further explained here. It is necessary to mention that the representation (4.11) has been proved in [15] under conditions (4.1), (4.2), and (4.4) only for the case when the probability generating functions $f_n(s, 1)$ are linear fractional with probability 1. However, the latter restriction is easily removed using the results established later on for the general case in [12, 14].

Our proof of Theorem 4.1 uses the next lemma.

**Lemma 4.2.** *Consider conditional moments of the entity $W_n$ defined at the beginning of Section 2:*

$$S_n^{(i)} := \sum_{k=0}^{n-1} X_k \theta_{i,k}, \quad i = 1, 2. \tag{4.12}$$

*Under conditions* (4.1), (4.5),

$$\mathbb{P}\left[S_{\mathrm{T}}^{(1)} > x\right] \sim \frac{K_0}{\log x}, \quad x \longrightarrow \infty. \tag{4.13}$$

*For any fixed $\epsilon > 0$, in Notation 4,*

$$\mathbb{P}\left[S_{\mathrm{T}}^{(2)} > n^{\varepsilon} S_{\mathrm{T}}^{(1)}; T \le n_\alpha\right] = o\left(\frac{1}{\log n}\right), \quad n \longrightarrow \infty. \tag{4.14}$$

*Proof.* For any fixed $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\left[S_{\mathrm{T}}^{(1)} > x\right] &\geq \mathbb{P}\left[S_{\mathrm{T}}^{(1)} > x; T \leq x_\alpha; \min_{0\leq k\leq T} \theta_{1,k} > x^{-\varepsilon}\right] \\ &\geq \mathbb{P}\left[S_{\mathrm{T}} > x^{1+\varepsilon}\right] - \mathbb{P}[T > x_\alpha] - \mathbb{P}\left[\min_{0\leq k\leq x_\alpha} \theta_{1,k} \leq x^{-\varepsilon}\right].\end{aligned} \tag{4.15}$$

Notice that according to (4.3)

$$\mathbb{P}\left[T > (\log x)^{2+\varepsilon}\right] = o\left(\frac{1}{\log x}\right), \quad x \longrightarrow \infty, \text{ for any fixed } \varepsilon > 0. \tag{4.16}$$

Thus, using (4.7) and (4.11) we get

$$\liminf_{x\to\infty}\left\{\log x \cdot \mathbb{P}\left[S_{\mathrm{T}}^{(1)} > x\right]\right\} \geq \liminf_{x\to\infty}\left\{\log x \cdot \mathbb{P}\left[S_{\mathrm{T}} > x^{1+\varepsilon}\right]\right\} \geq \frac{K_0}{(1+\epsilon)}. \tag{4.17}$$

To obtain a similar estimate from the previously mentioned we write, recalling Notation 4,

$$\begin{aligned}\mathbb{P}\left[S_{\mathrm{T}}^{(1)} > x\right] &\leq \mathbb{P}\left[S_{\mathrm{T}}^{(1)} > x; T \leq x_\alpha; \max_{0\leq k\leq T} \theta_{1,k} \leq x^{\varepsilon}\right] \\ &\quad + \mathbb{P}[T > x_\alpha] + \mathbb{P}\left[T \leq x_\alpha; \max_{0\leq k\leq T} \theta_{1,k} > x^{\varepsilon}\right] \\ &\leq \mathbb{P}\left[S_{\mathrm{T}} > x^{1-\varepsilon}\right] + \mathbb{P}[T > x_\alpha] + \mathbb{P}\left[\max_{0\leq k\leq x_\alpha} \theta_{1,k} > x^{\varepsilon}\right],\end{aligned} \tag{4.18}$$

which together with (4.8), (4.11), and (4.16) yields

$$\begin{aligned}\limsup_{x\to\infty}\left\{\log x \cdot \mathbb{P}\left[S_{\mathrm{T}}^{(1)} > x\right]\right\} &\leq \limsup_{x\to\infty}\left\{\log x \cdot \mathbb{P}\left[S_{\mathrm{T}} > x^{1-\varepsilon}\right]\right\} \\ &\leq \frac{K_0}{(1-\epsilon)}.\end{aligned} \tag{4.19}$$

Finally, according to (4.9)

$$\begin{aligned}\mathbb{P}\left[S_{\mathrm{T}}^{(2)} > n^{\varepsilon} S_{\mathrm{T}}^{(1)}; T \leq n_\alpha\right] &\leq \mathbb{P}\left[\max_{1\leq k\leq T}\left(\frac{\theta_{2,k}}{\theta_{1,k}}\right) > n^{\varepsilon}; T \leq n_\alpha\right] \\ &= o\left(\frac{1}{\log n}\right), \quad n \longrightarrow \infty.\end{aligned} \tag{4.20}$$

□

*Proof of Theorem 4.1.* We will show that

$$\limsup_{n\to\infty}\{\log n\cdot\mathbb{P}[Z_n>0]\}\le K_0\le\liminf_{n\to\infty}\{\log n\cdot\mathbb{P}[Z_n>0]\} \tag{4.21}$$

using a counterpart of (2.9)

$$\begin{aligned}\mathbb{P}[Z_n>0]&=\mathbb{E}\left[1-\prod_{k=0}^{n-1}Q_{n-k}^{Y_k}\right]=\mathbb{E}\left[1-\prod_{k=0}^{n-1}f_k^{X_k}(1,Q_{n-k})\right]\\&=\mathbb{E}\left[1-\exp\left\{\sum_{k=0}^{n-1}X_k\log f_k(1,Q_{n-k})\right\}\right]\end{aligned} \tag{4.22}$$

and Lemma 4.2.

First we prove the second inequality in (4.21). It follows from (4.22) and the monotonicity of $Q_n$ that for any fixed $\epsilon\in(0,1)$

$$\mathbb{P}[Z_n>0]\ge\mathbb{E}\left[1-\exp\left\{\sum_{k=0}^{T-1}X_k\log f_k(1,Q_n)\right\};S_{\mathrm{T}}^{(2)}\le n^{\epsilon}S_{\mathrm{T}}^{(1)},T\le n_{\alpha}\right]. \tag{4.23}$$

Recall that $\log(1-x)\le -x$ and

$$f(1,s)\le 1+\theta_1(s-1)+\frac{\theta_2}{2}(1-s)^2, \tag{4.24}$$

with the latter inequality being valid thanks to the monotonicity of the second derivative of the generating function. Therefore,

$$\begin{aligned}\log f(1,s)&\le -\theta_1(1-s)+\left(\frac{\theta_2}{2}\right)(1-s)^2,\\\sum_{k=0}^{T-1}X_k\log f_k(1,Q_n)&\le -(1-Q_n)\sum_{k=0}^{T-1}X_k\theta_{1,k}+\frac{(1-Q_n)^2}{2}\sum_{k=0}^{T-1}X_k\theta_{2,k}\\&\le -c_1n^{-1}\sum_{k=0}^{T-1}X_k\theta_{1,k}+c_2n^{-2}\sum_{k=0}^{T-1}X_k\theta_{2,k},\end{aligned} \tag{4.25}$$

where the last inequality is due to (2.5). It follows that given $S_{\mathrm{T}}^{(2)}\le n^{\epsilon}S_{\mathrm{T}}^{(1)}$,

$$\sum_{k=0}^{T-1}X_k\log f_k(1,Q_n)\le -cn^{-1}S_{\mathrm{T}}^{(1)} \tag{4.26}$$

for sufficiently large $n$. As a result, we see that for large $n$

$$\begin{aligned}\mathbb{P}[Z_n > 0] &\geq \mathbb{E}\left[1 - e^{-cn^{-1}S_{\mathrm{T}}^{(1)}}; S_{\mathrm{T}}^{(2)} \leq n^{\varepsilon} S_{\mathrm{T}}^{(1)}, T \leq n_{\alpha}\right] \\ &\geq \mathbb{E}\left[1 - e^{-cn^{-1}S_{\mathrm{T}}^{(1)}}\right] - \mathbb{P}[T > n_{\alpha}] - \mathbb{P}\left[S_{\mathrm{T}}^{(2)} > n^{\varepsilon} S_{\mathrm{T}}^{(1)}; T \leq n_{\alpha}\right].\end{aligned} \tag{4.27}$$

Now, to finish the proof of the second inequality in (4.21) it remains to use (4.16), Lemma 4.2, and

$$\mathbb{E}\left[1 - e^{-\lambda S_{\mathrm{T}}^{(1)}}\right] \sim \frac{K_0}{\log(1/\lambda)}, \quad \lambda \longrightarrow 0, \tag{4.28}$$

which again due to Lemma 4.2 follows from the Tauberian theorem [16, Chapter XIII.5, Theorem 4] applied to the right hand side of

$$\lambda^{-1}\mathbb{E}\left[1 - e^{-\lambda S_{\mathrm{T}}^{(1)}}\right] = \int_0^{\infty} \mathbb{P}\left[S_{\mathrm{T}}^{(1)} > x\right] e^{-\lambda x} dx. \tag{4.29}$$

Next, we verify the first inequality in (4.21). From the estimates $\log(1-x) \geq -2x$, valid for $x \in (0, 1/2)$, and $f(1,s) \geq 1 + \theta_1(s-1)$, we conclude that for all sufficiently large $n$

$$\begin{aligned}&\mathbb{E}\left[1 - \exp\left\{\sum_{k=0}^{T-1} X_k \log f_k(1, Q_{n-T})\right\}; T \leq n_{\alpha}; \max_{0\leq k\leq T} \theta_{1,k} \leq n^{\varepsilon}\right] \\ &\quad \leq \mathbb{E}\left[1 - \exp\left\{\sum_{k=0}^{T-1} X_k \log\left(1 - \frac{c\theta_{1,k}}{n}\right)\right\}; T \leq n_{\alpha}; \max_{0\leq k\leq T} \theta_{1,k} \leq n^{\varepsilon}\right] \\ &\quad \leq \mathbb{E}\left[1 - \exp\left\{-2cn^{-1}\sum_{k=0}^{T-1} X_k \theta_{1,k}\right\}; T \leq n_{\alpha}; \max_{0\leq k\leq T} \theta_{1,k} \leq n^{\varepsilon}\right] \\ &\quad \leq \mathbb{E}\left[1 - e^{-2cn^{\varepsilon-1}S_{\mathrm{T}}^{(1)}}\right].\end{aligned} \tag{4.30}$$

Thus,

$$\begin{aligned}\mathbb{P}[Z_n > 0] &\leq \mathbb{E}\left[1 - \exp\left\{\sum_{k=0}^{T-1} X_k \log f_k(1, Q_{n-T})\right\}\right] \\ &\leq \mathbb{E}\left[1 - e^{-2cn^{\varepsilon-1}S_{\mathrm{T}}^{(1)}}\right] + \mathbb{P}[T > n_{\alpha}] + \mathbb{P}\left[\max_{0\leq k\leq n_{\alpha}} \theta_{1,k} > n^{\varepsilon}\right],\end{aligned} \tag{4.31}$$

and (4.21) follows due to (4.8) and (4.16). □

## 5. The Subcritical Case with an IID Environment

We continue studying BGW processes in IID environment, but now assume that

$$\mathbb{E}[\zeta] < 0, \qquad \mathrm{Var}[\zeta] \in (0, \infty) \tag{5.1}$$

instead of (4.1). Results rely upon a theorem from [17] giving the asymptotics for $\mathbb{P}[W_{\mathrm{T}} > x]$ as $x \to \infty$. It requires an important technical assumption, namely, the existence of a constant $\kappa$ such that

$$\mathbb{E}\left[e^{\kappa\zeta}\right] = \mathbb{E}\left[\mu_1^{\kappa}\right] = 1, \quad 0 < \kappa < \infty. \tag{5.2}$$

If, in addition, for some $\delta > 0$

$$0 < \mathbb{E}\left[\xi_2^{\kappa+\delta}\right] < \infty, \qquad \mathbb{E}\left[\theta_1^{\kappa}\right] < \infty, \tag{5.3}$$

and either

$$\kappa > 1, \quad \mathbb{E}\left[|\xi_1|^{\kappa}\right] < \infty, \tag{5.4}$$

or

$$0 < \kappa \leq 1, \quad \mathbb{E}\left[\left|\mu_2 - \mu_1^2\right|^{\kappa} + \left|\theta_2 - \theta_1^2\right|^{\kappa}\right] < \infty, \tag{5.5}$$

then, according to [17], there exists a constant $C_\kappa \in (0, \infty)$ such that

$$\mathbb{P}[W_{\mathrm{T}} > x] \sim C_\kappa x^{-\kappa}, \quad x \longrightarrow \infty. \tag{5.6}$$

It is also known [18–20] that under (5.1) and (5.2)

$$\mathbb{P}[X_n > 0] = \mathbb{P}[T > n] = o(A^n) \ \text{ for some constant } A \in (0, 1)\ . \tag{5.7}$$

**Theorem 5.1.** *If conditions* (5.1), (5.2), (5.3), *and either* (5.4) *or* (5.5) *hold, then*

$$\mathbb{P}[Z_n > 0] \sim K_\kappa \cdot q_\kappa(n), \quad n \longrightarrow \infty, \tag{5.8}$$

*for some positive constant* $K_\kappa$, *given by* (5.15) *latter, where*

$$q_\kappa(n) = \begin{cases} n^{-\kappa} & \text{if } \kappa < 1, \\ n^{-1} \log n & \text{if } \kappa = 1, \\ n^{-1} & \text{if } \kappa > 1. \end{cases} \tag{5.9}$$

*Proof.* Referring to (5.7), put $B_n := 2\log n/\log(A^{-1})$ and notice that

$$\mathbb{P}[T > B_n] = o\left(n^{-2}\right), \quad n \longrightarrow \infty. \tag{5.10}$$

From the first equality in (4.22) and the evident inequality $Q_{n-k} \le Q_n$ we obtain for $n \ge B_n$

$$\begin{aligned}\mathbb{P}[Z_n > 0] &\ge \mathbb{E}\left[1 - \prod_{k=0}^{n-1} Q_n^{Y_k}; T \le B_n\right] \\ &= \mathbb{E}\left[1 - e^{W_{\mathrm{T}} \log Q_n}; T \le B_n\right] \ge \mathbb{E}\left[1 - e^{W_{\mathrm{T}} \log Q_n}\right] - \mathbb{P}[T > B_n].\end{aligned} \tag{5.11}$$

On the other hand, we have a similar upper bound

$$\begin{aligned}\mathbb{P}[Z_n > 0] &\le \mathbb{E}\left[1 - Q_{n-T}^{W_{\mathrm{T}}}; T \le B_n\right] + \mathbb{P}[T > B_n] \\ &\le \mathbb{E}\left[1 - e^{W_{\mathrm{T}} \log Q_{n - N\log n}}\right] + \mathbb{P}[T > B_n].\end{aligned} \tag{5.12}$$

It remains to observe that due to (2.5) and (5.6), the same Tauberian theorem [16, Chapter XIII.5, Theorem 4] applied to the right-hand side of

$$\lambda^{-1}\mathbb{E}\left[1 - e^{-\lambda W_{\mathrm{T}}}\right] = \int_0^\infty \mathbb{P}[W_{\mathrm{T}} > x] e^{-\lambda x} dx \tag{5.13}$$

yields

$$\mathbb{E}\left[1 - e^{W_{\mathrm{T}} \log Q_n}\right] \sim K_\kappa \cdot q_\kappa(n), \quad n \longrightarrow \infty, \tag{5.14}$$

with

$$K_\kappa = \begin{cases} \Gamma(1-\kappa) C_\kappa \left(\dfrac{2}{m_2}\right)^\kappa & \text{if } \kappa < 1, \\ \dfrac{2}{m_2} C_1 & \text{if } \kappa = 1, \\ \dfrac{2}{m_2} \displaystyle\int_0^\infty \mathbb{P}(W_{\mathrm{T}} > x) dx & \text{if } \kappa > 1. \end{cases} \tag{5.15}$$

□

## 6. The Critical Case with a Markovian Environment

As compared to the IID case, Markovian environments require extra conditions on the underlying Markov chain. First we assume that the two-type critical process $(X_n, Z_n)$ evolves

in a stationary Markovian random environment as defined in Section 3.2. Besides, we suppose the validity of (3.11) and that for some $\rho > 0$

$$\mathbb{P}[\tau > x] = o\Big(x^{-1}(\log x)^{-1-\rho}\Big), \quad x \longrightarrow \infty. \tag{6.1}$$

This implies that $a := \mathbb{E}[\tau] < \infty$ and due to Lemma 3.1 conditions $\mathbb{E}[\widehat{\zeta}] = 0$ and $\mathbb{E}[\zeta] = 0$ become equivalent. Moreover, under condition (6.1) the sequence of regeneration times (3.4) satisfies

$$\mathbb{P}\Big[\Big|k^{-1}\tau_k - a\Big| > \varepsilon\Big] = o\Big((\log k)^{-1-\rho}\Big), \quad k \longrightarrow \infty, \tag{6.2}$$

for an arbitrarily small $\varepsilon > 0$, cf. [21].

**Theorem 6.1.** *Assume* (2.3), (3.11), (6.1), *and*

$$\begin{gathered} \mathbb{E}[\zeta] = 0, \qquad \operatorname{Var}\Big[\widehat{\zeta}\Big] \in (0, \infty), \\ \mathbb{E}\Big[\widehat{\mu}_2\widehat{\mu}_1^{-2}(1 + \max(0, \log\widehat{\mu}_1))\Big] < \infty, \qquad \mathbb{E}\Big[\widehat{\mu}_1^{-1}\Big] < \infty. \end{gathered} \tag{6.3}$$

*Further, for some positive* $\boldsymbol{\alpha}$ *let*

$$\begin{aligned} \mathbb{P}\Big[\widehat{\theta}_1 < \frac{1}{x}\Big] &= o\Big((\log x)^{-3-\alpha}\Big), \quad x \longrightarrow \infty, \\ \mathbb{P}\Big[\widehat{\theta}_1 > x\Big] &= o\Big((\log x)^{-3-\alpha}\Big), \quad x \longrightarrow \infty, \\ \mathbb{P}\Big[\widehat{\theta}_2 > x\widehat{\theta}_1\Big] &= o\Big((\log x)^{-3-\alpha}\Big), \quad x \longrightarrow \infty. \end{aligned} \tag{6.4}$$

*Then* $\mathbb{P}[X_n > 0] = O(n^{-1/2})$, *and there exists a constant* $\widehat{K}_0 > 0$ *such that*

$$\mathbb{P}[Z_n > 0] \sim \frac{\widehat{K}_0}{\log n}, \quad n \longrightarrow \infty. \tag{6.5}$$

*Proof.* The statement is divided into two steps: first

$$\mathbb{P}\Big[\widehat{Z}_r > 0\Big] \sim \frac{\widehat{K}_0}{\log r}, \quad r \longrightarrow \infty \tag{6.6}$$

and then

$$\mathbb{P}[X_n + Z_n > 0] \sim \frac{\widehat{K}_0}{\log n}, \quad n \longrightarrow \infty, \tag{6.7}$$

together with

$$\mathbb{P}[X_n > 0] = O\left(n^{-1/2}\right), \quad n \longrightarrow \infty. \tag{6.8}$$

Fix $\delta \in (0, 1/4)$, and write

$$\begin{aligned}\mathbb{P}\left[\widehat{Z}_r > 0\right] &= \mathbb{P}\left[\widehat{Z}_r > 0; \tau_{r^\delta} \le r^{2\delta}, |\tau_r - ra| \le r^{1-\delta}\right] \\ &\quad + O\left(\mathbb{P}\left[\tau_{r^\delta} > r^{2\delta}\right]\right) + O\left(\mathbb{P}\left[\left|r^{-1}\tau_r - a\right| > r^{-\delta}\right]\right).\end{aligned} \tag{6.9}$$

Here the last two terms are treated with the help of $\mathbb{P}(\tau_{r^\delta} > r^{2\delta}) \le ar^{-\delta}$ and (6.2), while the main term is analyzed by means of ideas from the proof of Theorem 4.1. Letting $\widehat{Y}_k$ be the number of type $\widehat{2}$ daughters produced by $\widehat{X}_k$ particles of type $\widehat{1}$ and putting $\widehat{T} := \min\{r : \widehat{Z}_r = 0\}$, we deduce from

$$\mathbb{P}\left[\widehat{Z}_r > 0; \tau_{r^\delta} \le r^{2\delta}, |\tau_r - ra| \le r^{1-\delta}\right] = \mathbb{E}\left[1 - \prod_{k=0}^{r-1} Q^{\widehat{Y}_k}_{\tau_r - \tau_k}; \tau_{r^\delta} \le r^{2\delta}, |\tau_r - ra| \le r^{1-\delta}\right] \tag{6.10}$$

a lower bound

$$\begin{aligned}&\mathbb{P}\left[\widehat{Z}_r > 0; \tau_{r^\delta} \le r^{2\delta}, |\tau_r - ra| \le r^{1-\delta}\right] \\ &\quad \ge \mathbb{E}\left[1 - \prod_{k=0}^{\widehat{T}-1} Q^{\widehat{Y}_k}_{\tau_r - \tau_k}; \widehat{T} \le r^\delta, \tau_{r^\delta} \le r^{2\delta}, |\tau_r - ra| \le r^{1-\delta}\right] \\ &\quad \ge \mathbb{E}\left[1 - \prod_{k=0}^{\widehat{T}-1} f_k^{\widehat{X}_k}(1, Q_{2ra}); \widehat{T} \le r^\delta, \tau_{r^\delta} \le r^{2\delta}, \left|r^{-1}\tau_r - a\right| \le r^{-\delta}\right] \\ &\quad \ge \mathbb{E}\left[1 - \prod_{k=0}^{\widehat{T}-1} f_k^{\widehat{X}_k}(1, Q_{2ra})\right] - \mathbb{P}\left(\widehat{T} > r^\delta\right) - \mathbf{P}\left(\tau_{r^\delta} > r^{2\delta}\right) - \mathbb{P}\left(\left|r^{-1}\tau_r - a\right| > r^{-\delta}\right) \\ &\quad = \mathbb{E}\left[1 - \prod_{k=0}^{\widehat{T}-1} f_k^{\widehat{X}_k}(1, Q_{2ra})\right] + o\left(\frac{1}{\log^{1+\rho} r}\right).\end{aligned} \tag{6.11}$$

Hence, applying arguments used to derive (4.21) in Theorem 4.1 one can show that for some $\widehat{K}_0 > 0$

$$\limsup_{r\to\infty}\left\{\log r \cdot \mathbb{P}\left[\widehat{Z}_r > 0\right]\right\} \le \widehat{K}_0 \le \liminf_{r\to\infty}\left\{\log r \cdot \mathbb{P}\left[\widehat{Z}_r > 0\right]\right\} \tag{6.12}$$

proving (6.6).

To demonstrate that (6.7) follows from (6.6) observe first that due to

$$\mathbb{P}\left[\widehat{X}_r > 0\right] = O\left(r^{-1/2}\right), \quad r \longrightarrow \infty, \tag{6.13}$$

we have

$$\mathbb{P}\left[\widehat{X}_r + \widehat{Z}_r > 0\right] \sim \frac{\widehat{K}_0}{\log r}, \quad r \longrightarrow \infty. \tag{6.14}$$

Setting $N_n := \max\{k : \tau_k \le n\}$ we obtain

$$\mathbb{P}\left[\widehat{X}_{N_n+1} + \widehat{Z}_{N_n+1} > 0\right] \le \mathbb{P}[X_n + Z_n > 0] \le \mathbb{P}\left[\widehat{X}_{N_n} + \widehat{Z}_{N_n} > 0\right], \tag{6.15}$$

and for any $\varepsilon \in (0,1)$ we get

$$\begin{aligned}\mathbb{P}\left[\widehat{X}_{N_n} + \widehat{Z}_{N_n} > 0\right] &= \mathbb{P}\left[\widehat{X}_{N_n} + \widehat{Z}_{N_n} > 0; N_n \ge a^{-1}n(1-\varepsilon)\right] \\ &\quad + O\left(\mathbb{P}\left[N_n < a^{-1}n(1-\varepsilon)\right]\right).\end{aligned} \tag{6.16}$$

It follows that

$$\mathbb{P}\left[\widehat{X}_{N_n} + \widehat{Z}_{N_n} > 0; N_n \ge a^{-1}n(1-\varepsilon)\right] \le \mathbb{P}\left[\widehat{X}_{a^{-1}n(1-\varepsilon)} + \widehat{Z}_{a^{-1}n(1-\varepsilon)} > 0\right]. \tag{6.17}$$

On the other hand, again by (6.2) as $n \to \infty$

$$\mathbb{P}\left[N_n < a^{-1}n(1-\varepsilon)\right] = \mathbb{P}\left[\tau_{a^{-1}n(1-\varepsilon)} > n\right] = o\left((\log n)^{-1-\rho}\right). \tag{6.18}$$

Thus,

$$\limsup_{n\to\infty}\{\log n \cdot \mathbb{P}[X_n + Z_n > 0]\} \le \limsup_{n\to\infty}\left\{\log n \cdot \mathbb{P}\left[\widehat{X}_{a^{-1}n(1-\varepsilon)} + \widehat{Z}_{a^{-1}n(1-\varepsilon)} > 0\right]\right\} \le \widehat{K}_0. \tag{6.19}$$

A similar estimate shown latter follows from

$$\begin{aligned}\mathbb{P}\left[\widehat{X}_{N_n+1} + \widehat{Z}_{N_n+1} > 0\right] &\ge \mathbb{P}\left[\widehat{X}_{a^{-1}n(1+\varepsilon)} + \widehat{Z}_{a^{-1}n(1+\varepsilon)} > 0; N_n + 1 \le a^{-1}n(1+\varepsilon)\right] \\ &= \mathbb{P}\left[\widehat{X}_{a^{-1}n(1+\varepsilon)} + \widehat{Z}_{a^{-1}n(1+\varepsilon)} > 0\right] + o\left((\log n)^{-1-\rho}\right).\end{aligned} \tag{6.20}$$

Finally, relation (6.8) is derived from (6.13) by the law of large numbers argument. □

## 7. Subcritical Processes with a Markovian Environment

Assume now that the two-type subcritical process $(X_n, Z_n)$ evolves in a stationary Markovian random environment as defined in Section 3.2. Here, similarly to Section 6 the auxiliary branching process $(\widehat{X}_r, \widehat{Z}_r)$ in IID environment with probability generating functions (3.6) and (3.7) plays an important role.

Single-type subcritical processes with a Markovian environment were recently studied in [22]. According to [22] under the conditions of our next theorem one has, similarly to (5.7), that

$$\mathbb{P}[X_n > 0] = o(A^n) \quad \text{for some constant } A \in (0,1)\ . \tag{7.1}$$

**Theorem 7.1.** *Assume that assumption* (2.3) *holds,*

$$\mathbb{E}[\zeta] < 0, \qquad \operatorname{Var}\left[\widehat{\zeta}\right] \in (0,\infty), \tag{7.2}$$

*and conditions* (5.2), (5.3) *and either* (5.4) *or* (5.5) *are valid for the corresponding random variables related to the embedded process* $(\widehat{X}_r, \widehat{Z}_r)$ *with the key constant* $\kappa$ *replaced by* $\widehat{\kappa} > 0$*. Suppose, in addition, that*

$$\mathbb{P}[\tau > x] = o\left(x^{-1-\min(\widehat{\kappa},1)}\right), \quad x \longrightarrow \infty. \tag{7.3}$$

*Then, there exists a constant* $\widehat{K} \equiv \widehat{K}_{\widehat{\kappa}} > 0$*, given by* (7.16) *and* (7.20) *latter, such that, see* (5.9),

$$\mathbb{P}[Z_n > 0] \sim a^{\min(1,\widehat{\kappa})}\widehat{K}q_{\widehat{\kappa}}(n), \quad n \longrightarrow \infty. \tag{7.4}$$

*Proof.* Our main arguments here are similar to that used in the proof of Theorem 6.1. Fix $\varepsilon \in (0,1)$ and a sufficiently large $N$ and with $a = \mathbb{E}(\tau)$ write

$$\mathbb{P}\left[\widehat{Z}_r > 0\right] = \mathbb{P}\left[\widehat{Z}_r > 0; \mathcal{B}(r,\varepsilon)\right] + O\left(\mathbb{P}\left[\tau_{N\log r} > r^{\widehat{\kappa}}\log^3 r\right] + \mathbb{P}[|\tau_r - ra| > r\varepsilon]\right), \tag{7.5}$$

where

$$\mathcal{B}(r,\varepsilon) := \left\{\tau_{N\log r} \le r^{\widehat{\kappa}}\log^3 r, |\tau_r - ra| \le \varepsilon r\right\}. \tag{7.6}$$

Clearly,

$$\mathbb{P}\left[\tau_{N\log r} > r^{\widehat{\kappa}}\log^3 r\right] \le \frac{Na}{r^{\widehat{\kappa}}\log^2 r} = o\left(r^{-\widehat{\kappa}}\right). \tag{7.7}$$

Further, if $\widehat{\kappa} < 1$ then according to [21] under condition (7.3) we have

$$\mathbb{P}[|\tau_r - ra| > \varepsilon r] = o\left(r^{-\widehat{\kappa}}\right). \tag{7.8}$$

Thus,

$$\mathbb{P}\left[\widehat{Z}_r > 0\right] \geq \mathbb{P}\left[\widehat{Z}_r > 0; \mathcal{B}(r,\varepsilon)\right] = \mathbb{E}\left[1 - \prod_{k=0}^{r-1} Q^{\widehat{Y}_k}_{\tau_r - \tau_k}; \mathcal{B}(r,\varepsilon)\right], \tag{7.9}$$

and therefore, denoting by $\overline{\mathcal{B}}(r,\varepsilon)$ the event complementary to $\mathcal{B}(r,\varepsilon)$, we get

$$\begin{aligned}
\mathbb{P}\left[\widehat{Z}_r > 0\right] &\geq \mathbb{E}\left[1 - \prod_{k=0}^{\widehat{T}-1} Q^{\widehat{Y}_k}_{\tau_r - \tau_k}; \widehat{T} \leq N \log r; \mathcal{B}(r,\varepsilon)\right] \\
&\geq \mathbb{E}\left[1 - \prod_{k=0}^{\widehat{T}-1} Q^{\widehat{Y}_k}_{ra+2\varepsilon r}; \widehat{T} \leq N \log r; \mathcal{B}(r,\varepsilon)\right] \\
&= \mathbb{E}\left[1 - e^{\widehat{W}_{\mathrm{T}} \log Q_{ra+2\varepsilon r}}\right] - \mathbb{P}\left[\widehat{T} > N \log r\right] - \mathbb{P}\left[\overline{\mathcal{B}}(r,\varepsilon)\right].
\end{aligned} \tag{7.10}$$

Due to (1.2) and (7.2)

$$\mathbb{P}\left[\widehat{X}_r > 0\right] = \mathbb{P}\left[\widehat{T} > r\right] = o(A^r) \quad \text{for some } A < 1. \tag{7.11}$$

It follows that in view of (1.2), (7.7), and (7.8) one can find $N$ such that

$$\mathbb{P}\left[\widehat{T} > N \log r\right] + \mathbb{P}\left[\overline{\mathcal{B}}(r,\varepsilon)\right] = o\left(r^{-\widehat{\kappa}}\right). \tag{7.12}$$

On the other hand, using (2.5) and

$$\mathbb{P}\left[\widehat{W}_{\mathrm{T}} > y\right] \sim \widehat{C} y^{-\widehat{\kappa}}, \quad \widehat{C} \in (0,\infty), \tag{7.13}$$

one can show, arguing as in Theorem 5.1, that for $\widehat{\kappa} < 1$

$$\begin{aligned}
&\liminf_{\varepsilon \downarrow 0} \liminf_{r \to \infty} r^{\widehat{\kappa}} \mathbb{E}\left[1 - e^{\widehat{W}_{\mathrm{T}} \log Q_{ra+2\varepsilon r}}\right] \\
&\quad = \liminf_{\varepsilon \downarrow 0} \liminf_{r \to \infty} \frac{r^{\widehat{\kappa}}}{-\log Q_{ra+2\varepsilon r}} \int_0^\infty \mathbb{P}\left[\widehat{W}_{\mathrm{T}} > x\right] e^{x \log Q_{ra+2\varepsilon r}} dx \\
&\quad = \widehat{C}\Gamma(1-\widehat{\kappa}) \liminf_{\varepsilon \downarrow 0} \left(\frac{2}{m_2 a(1+2\varepsilon)}\right)^{\widehat{\kappa}}
\end{aligned} \tag{7.14}$$

giving

$$\liminf_{r \to \infty} r^{\widehat{\kappa}} \mathbb{P}\left[\widehat{Z}_r > 0\right] \geq \widehat{K} \tag{7.15}$$

with

$$\widehat{K} = \widehat{C}\Gamma(1-\widehat{\kappa})\left(\frac{2}{m_2 a}\right)^{\widehat{\kappa}} \quad \text{for } \widehat{\kappa} < 1. \tag{7.16}$$

A similar upper bound in view of (1.2) and (7.11) yields

$$\lim_{r\to\infty} r^{\widehat{\kappa}}\mathbb{P}\left[\widehat{X}_r + \widehat{Z}_r > 0\right] = \widehat{K}. \tag{7.17}$$

If $\widehat{\kappa} \geq 1$ then condition (7.3) entails

$$\mathbb{P}[|\tau_r - ra| > \varepsilon r] = o\left(r^{-1}\right), \tag{7.18}$$

and, as before, this implies

$$\lim_{r\to\infty} \left(q_{\widehat{\kappa}}(r)\right)^{-1}\mathbf{P}\left[\widehat{X}_r + \widehat{Z}_r > 0\right] = \widehat{K}, \tag{7.19}$$

where

$$\widehat{K} = \frac{2}{m_2 a} \cdot \begin{cases} \widehat{C}, & \text{if } \widehat{\kappa} = 1, \\ \displaystyle\int_0^\infty \mathbb{P}\left[\widehat{W}_{\mathrm{T}} > x\right]dx, & \text{if } \widehat{\kappa} > 1. \end{cases} \tag{7.20}$$

We proceed by recalling (6.15). For any $\varepsilon \in (0,1)$

$$\begin{aligned} \mathbb{P}\left[\widehat{X}_{N_n} + \widehat{Z}_{N_n} > 0\right] &= \mathbb{P}\left[\widehat{X}_{N_n} + \widehat{Z}_{N_n} > 0; N_n \geq a^{-1}n(1-\varepsilon)\right] \\ &\quad + O\left(\mathbb{P}\left[N_n < a^{-1}n(1-\varepsilon)\right]\right), \end{aligned} \tag{7.21}$$

and as $n \to \infty$

$$\begin{aligned} &\mathbb{P}\left[\widehat{X}_{N_n} + \widehat{Z}_{N_n} > 0; N_n \geq a^{-1}n(1-\varepsilon)\right] \\ &\quad \leq \mathbb{P}\left[\widehat{X}_{a^{-1}n(1-\varepsilon)} + \widehat{Z}_{a^{-1}n(1-\varepsilon)} > 0\right] \sim q_{\widehat{\kappa}}\left(a^{-1}n(1-\varepsilon)\right)\widehat{K}. \end{aligned} \tag{7.22}$$

It follows from [21] and our conditions that

$$\mathbb{P}\left[N_n < a^{-1}n(1-\varepsilon)\right] = \mathbb{P}\left[S_{a^{-1}n(1-\varepsilon)} > n\right] = o\left(n^{-\min(\widehat{\kappa},1)}\right), \quad n \longrightarrow \infty. \tag{7.23}$$

Thus,

$$\limsup_{n\to\infty}\left(q_{\widehat{\kappa}}(n)\right)^{-1}\mathbb{P}[X_n + Z_n > 0] \leq \lim_{\varepsilon\downarrow 0}\limsup_{n\to\infty}\left(q_{\widehat{\kappa}}(n)\right)^{-1}\mathbb{P}\left[\widehat{X}_{N_n} + \widehat{Z}_{N_n} > 0\right], \tag{7.24}$$

so that

$$\limsup_{n\to\infty}\left(q_{\widehat{\kappa}}(n)\right)^{-1}\mathbb{P}[X_n + Z_n > 0] \le a^{\min(1,\widehat{\kappa})}\widehat{K}. \tag{7.25}$$

The corresponding lower bound is obtained similarly. □

## Acknowledgments

E. Dyakonova was supported in part by the Program of RAS "Theoretical Problems of Contemporary Mathematics" and by INTAS Grant no. 03-51-5018. S. Sagitov was supported by Swedish Research Council Grant no. 621-2010-5623. V. Vatutin was supported in part by the Program of RAS "Theoretical Problems of Contemporary Mathematics" and by a Wenner-Gren Foundation Visiting Researcher Scholarship.

## References

[1] M. Benaim and S. J. Schreiber, "Persistence of structured population in random environments," *Theoretical Population Biology*, vol. 76, no. 1, pp. 19–34, 2009.
[2] P. Haccou and Y. Iwasa, "Optimal mixed strategies in stochastic environments," *Theoretical Population Biology*, vol. 47, no. 2, pp. 212–243, 1995.
[3] S. J. Schreiber, M. Benaïm, and K. A. S. Atchadé, "Persistence in fluctuating environments," *Journal of Mathematical Biology*, vol. 62, no. 5, pp. 655–683, 2011.
[4] S. Sagitov and M. C. Serra, "Multitype Bienaymé-Galton-Watson processes escaping extinction," *Advances in Applied Probability*, vol. 41, no. 1, pp. 225–246, 2009.
[5] C. Tomasetti, "On the probability of random genetic mutations for various types of tumor growth," *Bulletin of Mathematical Biology*, vol. 74, no. 6, pp. 1379–1395, 2012.
[6] G. Alsmeyer and S. Grottrup, "A host-parasite model for a two-type cell population," *Advances in Applied Probability*, vol. 45, 2013.
[7] K. B. Athreya and P. E. Ney, *Branching Processes*, vol. 196 of *Die Grundlehren der mathematischen Wissenschaften*, Springer, New York, NY, USA, 1972.
[8] J. Foster and P. Ney, "Decomposable critical multi-type branching processes," *Sankhyā. The Indian Journal of Statistics A*, vol. 38, no. 1, pp. 28–37, 1976.
[9] V. A. Vatutin and S. M. Sagitov, "A decomposable critical branching process with two types of particles. Probabilistic problems of discrete mathematics," *Proceedings of the Steklov Institute of Mathematics*, vol. 4, pp. 1–19, 1988.
[10] A. M. Zubkov, "The limit behavior of decomposable critical branching processes with two types of particles," *Theory of Probability & Its Applications*, vol. 27, no. 2, pp. 235–237, 1983.
[11] Y. Ogura, "Asymptotic behavior of multitype Galton-Watson processes," *Journal of Mathematics of Kyoto University*, vol. 15, no. 2, pp. 251–302, 1975.
[12] V. I. Afanasyev, J. Geiger, G. Kersting, and V. A. Vatutin, "Criticality for branching processes in random environment," *The Annals of Probability*, vol. 33, no. 2, pp. 645–673, 2005.
[13] R. Durrett, *Probability: Theory and Examples*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK, 4th edition, 2010.
[14] J. Geiger and G. Kersting, "The survival probability of a critical branching process in random environment," *Theory of Probability & Its Applications*, vol. 45, no. 3, pp. 607–615, 2000.
[15] V. I. Afanasyev, "On the maximum of a critical branching process in a random environment," *Discrete Mathematics and Applications*, vol. 9, pp. 267–284, 1999.
[16] W. Feller, *An Introduction to Probability Theory and Its Applications. II*, John Wiley & Sons, New York, NY, USA, 1966.
[17] V. A. Vatutin, "Polling systems and multitype branching processes in a random environment with a final product," *Theory of Probability & Its Applications*, vol. 55, no. 4, pp. 631–660, 2011.

[18] V. I. Afanasyev, J. Geiger, G. Kersting, and V. A. Vatutin, "Functional limit theorems for strongly subcritical branching processes in random environment," *Stochastic Processes and Their Applications*, vol. 115, no. 10, pp. 1658–1676, 2005.

[19] V. I. Afanasyev, C. Böinghoff, G. Kersting, and V. A. Vatutin, "Limit theorems for weakly subcritical branching processes in random environment," *Journal of Theoretical Probability*, vol. 25, no. 3, pp. 703–732, 2012.

[20] V. I. Afanasyev, C. Boeinghoff, G. Kersting, and V. A. Vatutin, "Conditional limit theorems for intermediately subcritical branching processes in random environment," *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, In press, http://arxiv.org/abs/1108.2127.

[21] C. C. Heyde and V. K. Rohatgi, "A pair of complementary theorems on convergence rates in the law of large numbers," *Proceedings of the Cambridge Philosophical Society*, vol. 63, pp. 73–82, 1967.

[22] E. E. Dyakonova, "Multitype Galton-Watson branching processes in Markovian random environment," *Theory of Probability and Its Applications*, vol. 56, pp. 508–517, 2011.

Hindawi Publishing Corporation
International Journal of Stochastic Analysis
Volume 2012, Article ID 569081, 20 pages
doi:10.1155/2012/569081

*Research Article*

# Birth and Death Processes with Neutral Mutations

**Nicolas Champagnat,[1,2] Amaury Lambert,[3] and Mathieu Richard[3,4]**

[1] *IECN, Université de Lorraine, Campus Scientifique, B.P. 70239, 54506 Vandœuvre-lès-Nancy Cedex, France*
[2] *Inria, 54600 Villers-lès-Nancy, France*
[3] *Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599 CNRS and UPMC Université Paris 06, Case courrier 188, 4 Place Jussieu, 75252 Paris Cedex 05, France*
[4] *CMAP, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France*

Correspondence should be addressed to
Mathieu Richard, mathieu.richard@cmap.polytechnique.fr

Received 27 September 2012; Accepted 28 November 2012

Academic Editor: Fima Klebaner

Copyright © 2012 Nicolas Champagnat et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We review recent results of ours concerning branching processes with general lifetimes and neutral mutations, under the infinitely many alleles model, where mutations can occur either at the birth of particles or at a constant rate during their lives. In both models, we study the allelic partition of the population at time $t$. We give closed-form formulae for the expected frequency spectrum at $t$ and prove a pathwise convergence to an explicit limit, as $t \to +\infty$, of the relative numbers of types younger than some given age and carried by a given number of particles (small families). We also provide the convergences in distribution of the sizes or ages of the largest families and of the oldest families. In the case of exponential lifetimes, population dynamics are given by linear birth and death processes, and we can most of the time provide general formulations of our results unifying both models.

## 1. Introduction

We consider a general branching model, where particles have i.i.d. (not necessarily exponential) life lengths and give birth at a constant rate $b$ during their lives to independent copies of themselves. The genealogical tree thus produced is called *splitting tree* [1–3]. The process that counts the number of the alive particles through time is a *Crump-Mode-Jagers process* (or general branching process) [4] which is binary (births occur singly) and homogeneous (constant birth rate).

We enrich this genealogical model with mutations. In Model I, each child is a clone of her mother with probability $1-p$ and a mutant with probability $p$. In Model II, independently of other particles, each particle undergoes mutations during her life at constant rate $\theta$ (and births are always clonal). For both models, we are working under the infinitely many alleles model; that is, a mutation yields a type, also called *allele*, which was never encountered before. Moreover, mutations are supposed to be neutral; that is, they do not modify the way particles die and reproduce. For any type and any time $t$, we call *family* the set of all particles that share this type at time $t$.

Branching processes (and especially birth and death processes) with mutations have many applications in biology. In carcinogenesis [5–10], they can model the evolution of cancerous cells. In [11], Kendall modeled carcinogenesis by a birth and death process where mutations occur during life according to an inhomogeneous Poisson process. In [8, 10], cancerous cells are modeled by a multitype branching process where a cell is of type $k$ if it has undergone $k$ mutations and where the more a cell has undergone mutations, the faster it grows. The object of this study is the time $\tau_k$ of appearance of the first cell of type $k$. In [7], the authors study the arrival time of the first resistant cell and the number of resistant cells, in a model of cancerous cells undergoing a medical treatment and becoming resistant after having experienced a certain number of mutations.

Branching processes with mutations are also used in epidemiology. Epidemics, and especially their onset, can be modeled by birth and death processes, where particles are infected hosts, births are disease transmissions, and deaths are recoveries or actual deaths. In [12], Stadler provides a statistical method for the inference of transmission rates and of the reproductive value of epidemics in a birth and death model with mutations. In [13], Lambert and Trapman enriched the transmission tree with Poissonian marks modeling detection events of hospital patients infected by an antibiotic-resistant pathogen. They provided an inference method based on the knowledge of times spent by patients at the hospital at the detection of the outbreak.

Let us also mention the existence of models, for example, [14], of phage reproduction within a bacterium by a (possibly time inhomogeneous) birth and death process with Poissonian mutations, where particles model phage in the vegetative phase (DNA strands in the bacterium without protein coating) and death is interpreted as phage maturing (reception of protein coating).

In ecology, the neutral theory of biodiversity [15] gives a prediction of the diversity patterns, in terms of species abundance distributions, that are generated by individual-based models where speciation is caused by mutation or by immigration from mainland. Usually, the underlying genealogical models are assumed to keep the population size constant through time, as in the Moran or Wright-Fisher models, and so have the same well-known properties as models in mathematical population genetics (e.g., Ewens sampling formula), with a different interpretation. See [16, 17] for cases where this assumption is relaxed in favor of the branching property.

In this paper, we are first interested in the allelic partition of the population and more precisely in properties about the *frequency spectrum* $(M_t^{i,a},\ i \geq 1)$, where $M_t^{i,a}$ is the number of distinct types younger than $a$ (i.e., whose original mutation appeared after $t-a$) carried by exactly $i$ particles at time $t$. This kind of question was first studied by Ewens [18] who discovered the well-known "sampling formula" named after him and which describes the law of the allelic partition for a Wright-Fisher model with neutral mutations.

In our models, it is not possible to obtain a counterpart of Ewens sampling formula but we obtain different kinds of results concerning the frequency spectrum $(M_t^{i,a},\ i \geq 1)$. First, we

get a closed-form formula for the expected frequency spectrum, even in the non-Markovian cases. Second, we get pathwise convergence results as $t \to +\infty$ on the survival event, of the relative abundances of types. Third, we investigate the order of magnitude of the sizes of the largest families at time $t$ and of the ages of oldest types at time $t$, as $t \to +\infty$, and show the convergence in distribution of these quantities properly rescaled. Several regimes appear, depending on whether the *clonal process*, which is the process counting particles of a same type, is subcritical, critical, or supercritical.

We do not know any previous mathematical studies, other than ours, on branching processes with Poissonian mutations, but there are several existing mathematical results on branching models with mutations at birth that we now briefly review.

In discrete time, Griffiths and Pakes [19] studied the case of a Bienaymé-Galton-Watson (BGW) process where at each generation, all particles mutate independently with some probability $u$. The authors obtained properties about the number of alleles/types in the population, about the time of last mutation in the (sub)critical case and about the expected frequency spectrum. In [20, 21], Bertoin considers an infinite alleles model with neutral mutations in a subcritical or critical BGW process where particles independently give birth to a random number of clonal and mutant children according to the same joint distribution. In [20], the tree of alleles is studied, where all particles of a common type are gathered in clusters and the law of the allelic partition of the total population is given by describing the joint law of the sizes of the clusters and of the numbers of their mutant children. In [21], Bertoin obtains the joint convergence of the sizes of allelic families in the limit of a large initial population size and a small mutation rate.

In continuous time, Pakes [22] studied Markovian branching processes and gave the counterpart in the time-continuous setting, of properties found in the previously cited paper [19]. In particular, his results about the frequency spectrum and the "limiting frequency spectrum" are similar to ours, stated in Section 3. Recently, Maruvka et al. [23, 24] have considered the linear birth and death process with Poissonian mutations. Actually, they rather studied a PDE satisfied by a concentration $n(x,t)$ which can be seen as (but is not proved to be) a deterministic approximation to the number of families of size $x$ at time $t$. It is remarkable that this PDE has a steady concentration $\overline{n}(x)$, whose behavior as $x \to \infty$ is comparable to the asymptotic behavior of the relative numbers of families of size $m$ as $m \to +\infty$ in the discrete model studied here and in [19]. In the monography [25], Taïb is interested in general branching processes known as Crump-Mode-Jagers processes (see [4, 26] and references therein) where mutations still occur at birth but with a probability that may depend, for example, on the age of the mother. He obtained limited theorems about the frequency spectrum by using random characteristics techniques but in most cases, limits cannot be explicitly computed. Some of our results in Model I are applications of Taïb's, but use techniques specific to splitting trees to yield explicit formulae. We have refrained to apply results of Taïb on the convergence in distribution of properly rescaled sizes of the largest families, on the validity of which we have doubts in the case of supercritical clonal processes (see last section).

The paper is organized as follows. In Section 2, we define the models and give some of their properties that will be useful to state the main results. Section 3 is devoted to the study of the frequency spectrum (small families). Finally, in Section 4, we give the results about ages of the oldest families and about sizes of the largest ones.

Notice that in this paper, most of the results are stated for linear birth and death processes in order to simplify the notation. Most of them are also true with general life length distributions and are proved in Chapter 3 of the Ph.D. thesis [27] for Model I and in [28, 29]

for Model II. Specific effort has been put on finding a unifying formulation for our results as soon as it seemed possible.

## 2. The Models

### 2.1. Model without Mutations

We first define the model without mutations and give some of its properties. Afterwards, we will explain the two mutation mechanisms that we consider in this paper.

As a population model, we consider *splitting trees* [1–3]; that is,

(i) at time $t = 0$, the population starts with one progenitor;

(ii) all particles have i.i.d. reproduction behaviors;

(iii) conditional on her birth date $\alpha$ and her life length $\zeta$, each particle gives birth at a constant rate $b \in (0, \infty)$ during $(\alpha, \alpha + \zeta)$, to a single particle at each birth event.

It is important to notice that the common law of life lengths can be as general as possible. Let $Z = (Z(t), t \geq 0)$ be the process counting the number of extant particles through time. We denote the lifespan distribution by $\Lambda(\cdot)/b$, where $\Lambda$ is a finite positive measure on $(0, +\infty]$ with total mass $b$ and is called a *lifespan measure* [3].

The total population process $Z$ belongs to a large class of branching processes called *Crump-Mode-Jagers* or *CMJ processes*. In these processes, also called general branching processes [4, 26], one associates with each particle $x$ in the population a non-negative r.v. $\lambda_x$ (her life length) and a point process $\xi_x$ called birth point process. One assumes that the sequence $(\lambda_x, \xi_x)_x$ is i.i.d. but $\lambda_x$ and $\xi_x$ are not necessarily independent. Then, the CMJ process is defined as

$$Z(t) = \sum_x \mathbf{1}_{\{\sigma_x \leq t < \sigma_x + \lambda_x\}}, \quad t \geq 0, \tag{2.1}$$

where for any particle $x$ in the population, $\sigma_x$ is her birth time.

In our particular case, the common distribution of lifespans is $\Lambda(\cdot)/b$ and conditional on her lifespan, the birth point process of a particle is distributed as a Poisson point process during her life. We can say that the CMJ process $Z$ is *homogeneous* (constant birth rate) and *binary* (births occur singly). We will say that $Z$ is *subcritical, critical, or supercritical* according to whether the mean number of children per particle

$$m := \int_0^\infty u\Lambda(\mathrm{d}u), \tag{2.2}$$

is less than, equal to, or greater than 1.

The advantage of homogeneous, binary CMJ processes is that they allow for explicit computations, for example, about one-dimensional marginals of $Z$ (see forthcoming Proposition 2.1). More precisely, for $\lambda \geq 0$, define

$$\psi(\lambda) := \lambda - \int_{(0,\infty]} \left(1 - e^{-\lambda u}\right)\Lambda(\mathrm{d}u), \tag{2.3}$$

and let $r$ be the greatest root of $\psi$. Notice that $\psi$ is convex, $\psi(0) = 0$, and $\psi'(0) = 1 - m$. As a consequence,

$$\begin{aligned} r &= 0 \quad \text{if } Z \text{ is subcritical or critical,} \\ r &> 0 \quad \text{if } Z \text{ is supercritical.} \end{aligned} \tag{2.4}$$

Let $W$ be the so-called *scale function* [30, page 194] associated with $\psi$, that is, the unique increasing continuous function $(0, \infty) \to (0, \infty)$ satisfying

$$\int_0^\infty W(x) e^{-\lambda x} \mathrm{d}x = \frac{1}{\psi(\lambda)}, \quad \lambda > r. \tag{2.5}$$

**Proposition 2.1** (Lambert [3, 17]). *The one-dimensional marginals of $Z$ are given by*

$$\mathbb{P}(Z(t) = 0) = 1 - \frac{W'(t)}{bW(t)} \tag{2.6}$$

*and for $n \geq 1$,*

$$\mathbb{P}(Z(t) = n) = \left(1 - \frac{1}{W(t)}\right)^{n-1} \frac{W'(t)}{bW(t)^2}. \tag{2.7}$$

*In other words, conditional on being nonzero, $Z(t)$ is distributed as a geometric r.v. with success probability $1/W(t)$.*

If $\mathrm{Ext} := \{Z(t) \xrightarrow[t \to \infty]{} 0\}$ denotes the extinction event of $Z$, according to [3], as a consequence of the last proposition,

$$\mathbb{P}(\mathrm{Ext}) = 1 - \frac{r}{b}. \tag{2.8}$$

Thus, thanks to (2.4), extinction occurs a.s. when $Z$ is (sub)critical and $\mathbb{P}(\mathrm{Ext}^c) > 0$ when it is supercritical.

The following proposition justifies the fact that $r$ is called the *Malthusian parameter* of the population in the supercritical case.

**Proposition 2.2** (Lambert [3]). *If $m > 1$, conditional on the survival event $\mathrm{Ext}^c$,*

$$e^{-rt} Z(t) \xrightarrow[t \to \infty]{} \mathcal{E} \quad a.s., \tag{2.9}$$

*where $\mathcal{E}$ is exponential with parameter $\psi'(r)$.*

In fact, convergence in distribution is proved in [3] and a.s. convergence holds according to [31] (see [32, page 285]).

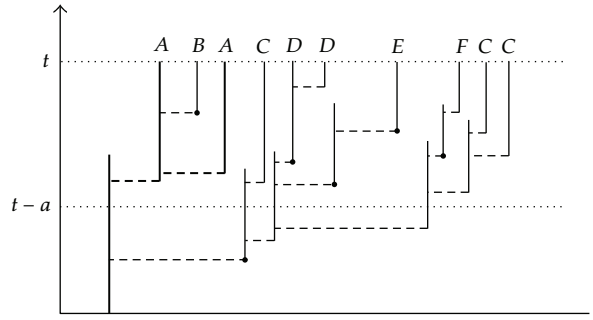


**Figure 1:** An example of a splitting tree in Model I and of the allelic partition of the whole extant population at time $t$. Vertical axis is time and horizontal axis shows filiation (horizontal lines have zero length). Full circles represent mutations at birth and thick lines, the clonal splitting tree of the ancestor up to time $t$. The different letters are the alleles of alive particles at time $t$.

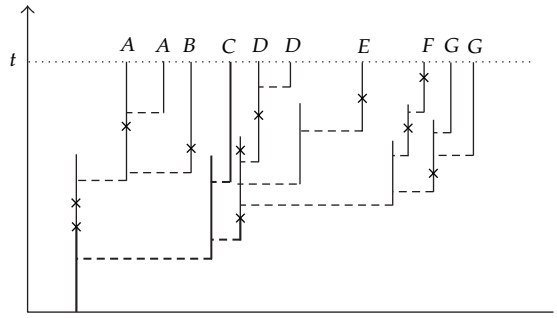


**Figure 2:** An example of a splitting tree with mutations in Model II and of the allelic partition of the whole extant population at time $t$. Crosses represent mutations and thick lines represent the clonal splitting tree of the ancestor up to time $t$. The different letters are the alleles of alive particles at time $t$.

## 2.2. *Two Mutation Models I and II*

We now assume that particles in the population carry types, also called *alleles*. We consider two population models where mutations appear in different ways. In each case, we will make the assumption of *infinitely many alleles*; that is, to every mutation event is associated a different type, so that every type appears only once. We will also assume that mutations are *neutral*; that is, they do not change the way particles die and reproduce.

In Model I, mutations occur at birth. More precisely, there is some $p \in (0,1)$ such that at each birth event, independently of all other particles, the newborn is a clone of her mother with probability $p$ and a mutant with probability $1-p$. An illustration is given in Figure 1.

In Model II, particles independently experience mutations during their lives at constant rate $\theta > 0$. In particular, in contrast with Model I, particles can change type several times during their lifetime, but always bear at birth the same type as their mother at this very time. An illustration is given in Figure 2.

In what follows, an important role will be played by the *clonal process*, generically denoted $Z_\star$, counting, as time passes, the number of particles bearing the same type as the progenitor of the population at time 0. It can easily be seen that the genealogy of a clonal population is again a splitting tree, so that $Z_\star$ is also a homogeneous, binary CMJ process. We denote by $b_\star$ its birth rate, by $\psi_\star$ the associated convex function as in (2.3), and by $W_\star$ the

nonnegative function with Laplace transform $1/\psi_\star$. Furthermore, when the clonal population is supercritical, that is, when $\psi'_\star(0+) < 0$, we denote by $r_\star$ its Malthusian parameter, which is the only nonzero root of $\psi_\star$. We will sometimes need to have this generic notation depending on the model considered: $Z_p$, $\psi_p$, $W_p$, and $r_p$ for Model I and $Z_\theta$, $\psi_\theta$, $W_\theta$, and $r_\theta$ for Model II.

Concerning Model I, it can be seen [27] that the clonal splitting tree has the same life lengths as the original splitting tree and birth rate $b_p = b(1-p)$, so that its lifespan measure is $(1-p)\Lambda$ and

$$\psi_p(\lambda) = p\lambda + (1-p)\psi(\lambda), \quad \lambda > 0. \tag{2.10}$$

In particular, as in (2.2), the clonal population is subcritical, critical, or supercritical according to whether $m(1-p)$ is less than, equal to, or greater than 1. It should be noted that there is no closed-form formula for $W_p$.

Concerning Model II, it can be seen [28] that the clonal splitting tree has birth rate $b_\theta = b$ and life lengths distributed as $\min(X, Y)$ where $X$ has the probability distribution $\Lambda(\cdot)/b$ and $Y$ is an independent exponential r.v. with parameter $\theta$. Then we get

$$\psi_\theta(\lambda) = \frac{\lambda\psi(\lambda+\theta)}{\lambda+\theta}, \quad \lambda > 0. \tag{2.11}$$

In particular, $r_\theta = r - \theta$ and the clonal population is subcritical, critical, or supercritical according to whether $r$ is less than, equal to, or greater than $\theta$. It can also be proved that $W$ and $W_\theta$ are differentiable and that their derivatives are related via

$$W'_\theta(x) = e^{-\theta x}W'(x), \quad x \geq 0, \tag{2.12}$$

with the requirement that $W_\theta(0) = 1$.

### 2.3. Exponential Case

An interesting case that we will focus on is the *exponential (or Markovian) case*, when the common distribution of life lengths is exponential with parameter $d$ (with the convention that lifespans are a.s. infinite if $d = 0$), that is, $\Lambda(\mathrm{d}u) = bde^{-\mathrm{d}u}\mathrm{d}u$ or $\Lambda(\mathrm{d}u) = b\delta_\infty(\mathrm{d}u)$. In that case, $Z$ is, respectively, a linear birth and death process with birth rate $b$ and death rate $d$ or a pure birth process (or Yule process) with parameter $b$.

In this case, $Z$ and $Z_\star$ are Markov processes and the quantities defined in Section 2.1 are computable. Indeed, we have

$$\begin{aligned} \psi(\lambda) &= \frac{\lambda(\lambda - b + d)}{\lambda + d}, \quad r = b - d, \\ m = 1 - \psi'(0) &= \frac{b}{d}, \quad \psi'(r) = 1 - \frac{d}{b}. \end{aligned} \tag{2.13}$$

It is also possible to compute the function $W$, defined by (2.5), while it is generally unknown. From [3, page 393], we have

$$W(x) = \begin{cases} \dfrac{be^{rx} - d}{r}, & \text{if } b \neq d, \\ 1 + bx, & \text{if } b = d, \end{cases} \quad x \geq 0, \tag{2.14}$$

and in all cases

$$W'(x) = be^{rx}, \quad x \geq 0. \tag{2.15}$$

The same results hold for $W_\star$, by, respectively, replacing $b$, $d$, and $r$ by

$$b_\star := b(1-p), \qquad d_\star := d, \qquad r_\star := r - bp, \tag{2.16a}$$

in Model I and by

$$b_\star := b, \qquad d_\star := d + \theta, \qquad r_\star := r - \theta, \tag{2.16b}$$

in Model II.

We will sometimes state results in the total generality of splitting trees, in which case an equation numbered (-a) (resp. (-b)) refers to Model I (resp. Model II), as done previously. However, we will most of the time focus on the exponential case, in which we will as soon as possible use the unified notation using $\star$'s. We will notify when the results can be generalized and will give precise references.

*Remark 2.3.* In the exponential case, notice that Models I and II are two (incompatible) cases of a more general class of linear birth and death processes with mutations, where particles mutate spontaneously at rate $\theta$, die at rate $d$, and give birth at rate $b$ and at each birth event: with probability $p_2$, the mother and the daughter both mutate (and bear either the same new type or two different new types); with probability $p_1$, the daughter (only) mutates; with probability $p_0 = 1 - p_1 - p_2$, none of them mutates. Then Model I corresponds to the case when $\theta = p_2 = 0$ and Model II to the case when $p_1 = p_2 = 0$. The case studied by Pakes in [22] corresponds to $\theta = 0$, $p_0 = u^2$, $p_1 = 2u(1-u)$, and $p_2 = (1-u)^2$. It is still an open question to check whether, when our results hold for both Models I and II with the unified notation, they hold for all linear birth and death processes with mutations.

## 3. Small Families

Recall that a *family* is a maximal set of particles bearing the same type at the same given time. In this section, we are interested in results about *small families*; that is, families whose sizes and ages are fixed, in opposition to those of Section 4 which concern asymptotic properties of the largest and oldest ones.

More precisely, we give properties of the allelic partition of the entire population by studying the *frequency spectrum* $(M_t^{i,a},\ i \geq 1)$, where $M_t^{i,a}$ denotes the number of distinct types,

whose ages are less than $a$ at time $t$, carried by exactly $i$ particles at time $t$. Notice that $M_t^{i,t}$ is simply the number of alleles carried by $i$ particles at time $t$ (regardless of their ages).

For instance, in Figure 1, the frequency spectrum $(M_t^{i,t},\ i \geq 1)$ is $(3,2,1,0,\ldots)$ because three alleles ($B$, $E$, and $F$) are carried by one particle, $A$ and $D$ are carried by two particles, and $C$ is the only allele carried by three particles. Moreover, if we only consider families with ages less than $a$, $(M_t^{i,a},\ i \geq 1)$ equals $(3,1,0,\ldots)$ because alleles $A$ and $C$ appear in the population before time $t-a$. Similarly, in Figure 2, the frequency spectrum in Model II is $(4,3,0,\ldots)$.

In the case of branching processes, there is no closed-form formula available for the law of the frequency spectrum as it is the case for the Wright-Fisher model thanks to Ewens sampling formula [18]. Nevertheless, we obtained for both mutation models an exact computation of the expected frequency spectrum and an almost sure asymptotic behavior of this frequency spectrum as $t \to +\infty$.

### 3.1. Expected Frequency Spectrum

We first give an exact expression of the expected frequency spectrum at any time $t$.

For $0 < a < t$ and $i \geq 1$, we denote by $M_t^{i,\mathrm{d}a}$ the number of types carried by $i$ particles at time $t$ and with ages in $[a-\mathrm{d}a, a]$. The following proposition yields its expected value.

**Proposition 3.1.** *For $0 < a < t$ and $i \geq 1$, one has*

$$\mathbb{E}\left[M_t^{i,\mathrm{d}a}\right] = \frac{p}{b(1-p)} W'(t-a)\left(1-\frac{1}{W_p(a)}\right)^{i-1}\frac{W_p'(a)}{W_p^2(a)}\mathrm{d}a, \tag{3.1a}$$

$$\mathbb{E}\left[M_t^{i,\mathrm{d}a}\right] = \frac{\theta W'(t)}{b}\left(1-\frac{1}{W_\theta(a)}\right)^{i-1}\frac{e^{-\theta a}}{W_\theta^2(a)}\mathrm{d}a. \tag{3.1b}$$

*In the exponential case, both expressions are read as*

$$\mathbb{E}\left[M_t^{i,\mathrm{d}a}\right] = (r-r_\star)e^{rt}\left(1-\frac{1}{W_\star(a)}\right)^{i-1}\frac{e^{-(r-r_\star)a}}{W_\star^2(a)}\mathrm{d}a. \tag{3.2}$$

In [27], (3.1a) is proved in the general case. Its proof uses the branching property and basic properties about Poisson processes. The main argument is that conditional on $Z(t-a)$, $M_t^{i,\mathrm{d}a}$ is the sum of $Z(t-a)$ independent r.v. distributed as the number of mutants that appear in the population in a time interval $\mathrm{d}a$ and with $i$ clonal alive descendants at time $a$. The proof of the general case of (3.1b) in [28] is based on coalescent point processes.

The expected frequency spectrums $\mathbb{E}[M_t^{i,a}]$ can be obtained by integrating (3.1a) and (3.1b) over ages. Taking into account the contribution of the type of the progenitor, we can prove the following result.

**Corollary 3.2.** *For $a \leq t$ and $i \geq 1$,*

$$\begin{aligned}\mathbb{E}\left[M_t^{i,a}\right] &= \frac{p}{b(1-p)} \int_0^a W'(t-x)\left(1-\frac{1}{W_p(x)}\right)^{i-1} \frac{W_p'(x)}{W_p^2(x)}\mathrm{d}x \\ &\quad + \frac{1}{b(1-p)}\left(1-\frac{1}{W_p(t)}\right)^{i-1} \frac{W_p'(t)}{W_p^2(t)}\mathbf{1}_{\{a=t\}},\end{aligned} \tag{3.3a}$$

$$\begin{aligned}\mathbb{E}\left[M_t^{i,a}\right] &= \frac{\theta W'(t)}{b} \int_0^a \left(1-\frac{1}{W_\theta(x)}\right)^{i-1} \frac{e^{-\theta x}}{W_\theta^2(x)}\mathrm{d}x \\ &\quad + W(t)\left(1-\frac{1}{W_\theta(t)}\right)^{i-1} \frac{e^{-\theta t}}{W_\theta^2(t)}\mathbf{1}_{\{a=t\}}.\end{aligned} \tag{3.3b}$$

*In the exponential case,*

$$\mathbb{E}\left[M_t^{i,a}\right] = (r-r_\star)e^{rt} \int_0^a \left(1-\frac{1}{W_\star(x)}\right)^{i-1} \frac{e^{-(r-r_\star)x}}{W_\star^2(x)}\mathrm{d}x + \mathbb{P}(Z_\star(t)=i)\mathbf{1}_{\{a=t\}}. \tag{3.4}$$

The second terms that appear in the r.h.s. correspond to the probabilities that the progenitor has $i$ alive clonal descendants at time $t$. In the exponential case, we left this probability as such, since its expression depends on the model. It is also possible to get similar equations for the number of families with ages less than $a$ (resp. with size $i$) by summing over $i$ (resp. by taking $a = t$) in the last expressions.

*Remark 3.3.* In the exponential case, when the process $Z$ is critical, that is, when $r = b - d = 0$, for $a < t$,

$$\mathbb{E}\left[M_t^{i,a}\right] = \frac{|r_\star|}{b_\star}\frac{1}{i}\left(1-\frac{1}{W_\star(a)}\right)^i, \tag{3.5}$$

which is reminiscent of Fisher log-series of species abundances [17]. Surprisingly, this expression is independent of $t \in (a, \infty)$.

From Corollary 3.2, we deduce the asymptotic behavior of $\mathbb{E}[M_t^{i,a}]$ in the supercritical case.

**Proposition 3.4.** *We suppose that $m > 1$. In the general case,*

$$\lim_{t\to+\infty} e^{-rt}\mathbb{E}\left[M_t^{i,a}\right] = \frac{r}{b}\frac{1}{\psi'(r)}J^{i,a}, \tag{3.6}$$

*where, for Model I,*

$$J^{i,a} := \frac{p}{(1-p)} \int_0^a \left(1-\frac{1}{W_p(u)}\right)^{i-1} \frac{e^{-ru}W_p'(u)}{W_p^2(u)}\mathrm{d}u \tag{3.7a}$$

*and, for Model II,*

$$J^{i,a} := \theta \int_0^a \left(1 - \frac{1}{W_\theta(u)}\right)^{i-1} \frac{e^{-\theta u}}{W_\theta^2(u)} \mathrm{d}u. \tag{3.7b}$$

*In the exponential case, one gets the simpler formula:*

$$J^{i,a} = (r - r_\star) \int_0^a \left(1 - \frac{1}{W_\star(u)}\right)^{i-1} \frac{e^{-(r-r_\star)u}}{W_\star^2(u)} \mathrm{d}u. \tag{3.8}$$

Notice that $\mathbb{E}[M_t^{i,a}]$ grows exponentially with parameter $r$, as does $Z$ on its survival event.

### 3.2. Convergence Results

In this section and in all following ones, we are interested in long-time behaviors in the two models we consider. Then, from now on, *we assume that the process $Z$ is supercritical.*

This paragraph deals with the improvements of the convergence results (3.6) regarding the expected frequency spectrum. The following results yield the asymptotic behavior as $t \to +\infty$ of the frequency spectrum $(M_t^{i,a},\ i \geq 1)$, conditional on the survival event.

The main technique we use to prove them is CMJ processes counted with random characteristics (see [4] and Appendix A in [25]). It enables us to obtain several pathwise convergence results regarding some processes embedded in the supercritical splitting tree.

A characteristic is a random nonnegative function on $[0, +\infty)$. To each particle $x$ in the population is associated a characteristic $\chi_x$, which can be viewed as a score or a weight. It must satisfy that $(\lambda_x, \zeta_x, \chi_x)_x$ is an i.i.d. sequence, where we recall that $\lambda_x$ is the life length of $x$ and $\zeta_x$ its birth process. Then, the process counted with the characteristic $\chi$ is defined as

$$Z^\chi(t) := \sum_x \chi_x(t - \sigma_x)\mathbf{1}_{\{\sigma_x \leq t\}}. \tag{3.9}$$

For instance, if $\chi(t) = \mathbf{1}_{\{t \leq \lambda_x\}}$, $Z^\chi$ equals $Z$ and if $\chi(t) = \mathbf{1}_{[t \leq \lambda_x \wedge a]}$, $Z^\chi(t)$ is the number of extant particles at time $t$ with ages less than $a$, then, provided technical conditions about $\chi$ are satisfied, the convergences of $e^{-rt}Z^\chi(t)$ and of $Z^\chi(t)/Z(t)$ as $t \to +\infty$ hold a.s. on the survival event. In our case, when $\chi$ is appropriately chosen, we can use this result to obtain the following statements.

**Proposition 3.5.** *Let $M_t$ be the number of extant types at time $t$. Almost surely, on the survival event of $Z$,*

$$\begin{aligned} \lim_{t \to +\infty} e^{-rt} M_t &= J\mathcal{E}, \\ \lim_{t \to +\infty} e^{-rt} M_t^{i,a} &= J^{i,a}\mathcal{E}, \end{aligned} \tag{3.10}$$

*where in Model I,*

$$J := \frac{rp}{1-p}\int_0^\infty e^{-ru}\ln\big(W_p(u)\big)\mathrm{d}u, \tag{3.11a}$$

*while in Model II,*

$$J := \theta\int_0^\infty \frac{e^{-\theta x}}{W_\theta(x)}\mathrm{d}x \tag{3.11b}$$

*and where $\mathcal{E}$ is the r.v. defined by* (2.9).

*In the exponential case, one has*

$$J = (r-r_\star)\int_0^\infty \frac{e^{-(r-r_\star)u}}{W_\star(u)}\mathrm{d}u. \tag{3.12}$$

Notice that (3.10) is consistent with (3.6) since $\mathbb{P}(\mathrm{Ext}^c) = r/b$ and $\mathbb{E}[\mathcal{E}] = 1/\varphi'(r)$. Moreover, (3.10) still holds after $M_t^{i,a}$ is replaced by $M_t^{i,t}$ and $J^{i,a}$ by $J^{i,\infty}$.

### 3.3. Asymptotic Behavior of the Limiting Frequency Spectrum

Thanks to Proposition 3.5, the proportion $M_t^{i,a}/M_t$ of types carried by $i$ particles and with ages less than $a$ converges a.s. to $J^{i,a}/J$ as $t \to +\infty$. This limit is called "the limiting frequency spectrum" by Pakes in [22]. This paragraph is devoted to the asymptotic behavior, as $i \to +\infty$, of $J^i := J^{i,\infty}$, obtained by taking $a = \infty$ in (3.7a) and (3.7b). In the exponential case,

$$J^i = (r-r_\star)\int_0^\infty \left(1-\frac{1}{W_\star(u)}\right)^{i-1}\frac{e^{-(r-r_\star)u}}{W_\star^2(u)}\mathrm{d}u. \tag{3.13}$$

#### 3.3.1. Supercritical Case

In this paragraph, we only treat the exponential case. Let us assume that the clonal process is supercritical, that is, $r_\star > 0$. Define

$$\nu := \frac{r}{r_\star}, \qquad \mu := \frac{b_\star}{r_\star}, \qquad \gamma := \frac{r-r_\star}{b_\star}. \tag{3.14}$$

We have $\gamma = p/(1-p)$ in Model I and $\gamma = \theta/b$ in Model II. Recall that $J^i$ is the proportion of types carried by $i$ particles in the large time asymptotic.

**Proposition 3.6.** *In the exponential case, one has for both models*

$$J^i \underset{i\to+\infty}{\sim} i^{-1-\nu}\gamma\Gamma(\nu+1)\mu^\nu. \tag{3.15}$$

Notice that this result is consistent with [24] where Maruvka et al. use an approximation of the frequency spectrum by a concentration driven by a PDE and with [22] where Pakes considers Markov branching processes with multiple simultaneous births, binomial mutations at birth, and no Poissonian mutations.

*Remark 3.7.* The following proof of Proposition 3.6 easily extends to any life length distributions since it is based on Proposition 2.2 which holds in the general case.

*Proof of Proposition 3.6.* Since $W_\star(t) \geq 1$ for $t \geq 0$, the sequence $(J^i)_{i\geq 1}$ is positive and nonincreasing. Then, according to a Tauberian theorem about series, to prove Proposition 3.6, it is sufficient to prove that $\sum_{i\geq j} J^i$ is equivalent to $i^{-\nu}\gamma\Gamma(\nu)\mu^\nu$ as $j \to +\infty$.

Recalling (3.13), we have

$$\sum_{i\geq j} J^i = (r - r_\star)\int_0^\infty e^{-rt}\mathbb{P}\big(Z_\star(t) \geq j\big)\mathrm{d}t \tag{3.16}$$

and from now on, we follow the proof of [22, Theorem 3.2.1]. Let $s > 0$ be such that $j = e^{r_\star s}$. Then $j^\nu = e^{rs}$ and

$$\begin{aligned} j^\nu \int_0^\infty e^{-rt}\mathbb{P}\big(Z_\star(t) \geq j\big)\mathrm{d}t &= \int_{-s}^\infty e^{-rt}\mathbb{P}(Z_\star(t+s) \geq e^{r_\star s})\mathrm{d}t \\ &= \int_{-s}^\infty e^{-rt}\mathbb{P}\Big(e^{-r_\star(t+s)}Z_\star(t+s) \geq e^{-r_\star t}\Big)\mathrm{d}t. \end{aligned} \tag{3.17}$$

Using Proposition 2.2 and (2.13), $\mathbb{P}(e^{-r_\star(t+s)}Z_\star(t+s) \geq e^{-r_\star t}) \xrightarrow[s\to+\infty]{} \mathbb{P}(E_\star \geq e^{-r_\star t})$, where $E_\star$ is a nonnegative r.v. such that

$$\mathbb{P}(E_\star = 0) = 1 - \frac{r_\star}{b_\star}, \tag{3.18}$$

and conditional on $\{E_\star > 0\}$, $E_\star$ is an exponential r.v. with parameter $\psi_\star'(r_\star) = 1 - d_\star/b_\star = 1/\mu$. Moreover, using Markov inequality, for $\varepsilon > 0$,

$$\mathbb{P}\Big(e^{-r_\star(t+s)}Z_\star(t+s) \geq e^{-r_\star t}\Big) \leq e^{r_\star(\nu+\varepsilon)t}\mathbb{E}\Big[\Big(Z_\star(s+t)e^{-r_\star(s+t)}\Big)^{\nu+\varepsilon}\Big] \leq Ce^{r_\star(\nu+\varepsilon)t}, \tag{3.19}$$

using again the a.s. convergence in Proposition 2.2. Then, for $s > 0$ and $t \in \mathbb{R}$, we have

$$e^{-rt}\mathbb{P}\Big(e^{-r_\star(t+s)}Z_\star(t+s) \geq e^{-r_\star t}\Big) \leq e^{-rt}\mathbf{1}_{\{t>0\}} + Ce^{r_\star\varepsilon t}\mathbf{1}_{\{t<0\}}, \tag{3.20}$$

and thanks to the dominated convergence theorem,

$$j^\nu \int_0^\infty e^{-rt}\mathbb{P}\big(Z_\star(t) \geq j\big)\mathrm{d}t \xrightarrow[j\to+\infty]{} \int_{\mathbb{R}} e^{-rt}\mathbb{P}\big(E_\star \geq e^{-r_\star t}\big)\mathrm{d}t = \frac{1}{\mu}\int_{\mathbb{R}} e^{-rt}e^{-(1/\mu)e^{-r_\star t}}\mathrm{d}t. \tag{3.21}$$

The change of variables $x = e^{-r_\star t}$ in the last integral leads to

$$j^\nu \int_0^\infty e^{-rt}\mathbb{P}(Z_\star(t) \geq j)\mathrm{d}t \xrightarrow[j\to+\infty]{} \frac{1}{b_\star}\int_0^\infty x^{\nu-1}e^{-x/\mu}\mathrm{d}x = \frac{\Gamma(\nu)}{b_\star}\mu^\nu, \tag{3.22}$$

which terminates the proof. □

### 3.3.2. Critical Case

We want to obtain a similar result to Proposition 3.6 when the clonal population is critical. It seems that this is not possible in a general setting due to the non explicit expression of the functions $W_p$ and $W_\theta$. However, in the exponential and critical case, we have the simpler expression $W_\star(t) = 1 + b_\star t$ and $r_\star = 0$. Then, we have

$$J^i = r\int_0^\infty e^{-rt}\left(1 - \frac{1}{1+b_\star t}\right)^{i-1}\frac{1}{(1+b_\star t)^2}\mathrm{d}t. \tag{3.23}$$

**Proposition 3.8.** *In the exponential case, one has*

$$J^i_\star \underset{i\to+\infty}{\sim} C(\gamma)i^{-3/4}e^{-2\sqrt{\gamma i}}, \tag{3.24}$$

*where one recalls that here* $\gamma = r/b_\star = r/d_\star$ *and one has set* $C(x) = \sqrt{\pi}e^{x/2}x^{5/4}$ *for* $x > 0$.

*Proof.* By a change of variables, we have set

$$J^i = \frac{r}{b_\star}\int_0^\infty e^{rt/b_\star}\frac{t^{i-1}}{(1+t)^{i+1}}\mathrm{d}t = \gamma\Gamma(i)U(i,0,\gamma), \tag{3.25}$$

where $U$ is known as a confluent hypergeometric function (see [33, Chapter 13]). Then, using [22, Theorem 3.3.2] with $B = -2$, we have the result. □

## 4. Asymptotic Results about Large and Old Families

We now state results about ages of the oldest families and about sizes of the largest ones. We mainly focus on the case when clonal populations are subcritical. Then, in Section 4.3, we explain which results hold in the critical and supercritical cases.

We need some notation. For $t \geq 0$,

(i) for $a \geq 0$, let $O_t(a)$ be the number of extant families at time $t$, with ages greater than $a$ ($O$ for "old"); for convenience, we set $O_t(a) = 0$ if $a < 0$,

(ii) for $x \in \mathbb{R}$, let $L_t(x)$ be the number of families with sizes greater than $x$ at time $t$ ($L$ for "large").

In this section, we are interested in finding the orders of magnitudes of the ages and of the sizes of the families; that is, in finding numbers $c_t$ and $x_t$ such that $\mathbb{E}[O_t(c_t)]$ and $\mathbb{E}[L_t(x_t)]$ converge to positive and finite real numbers as $t \to +\infty$.

### 4.1. Ages of Old Families in the Subcritical Case

In this section, we suppose that the clonal processes are subcritical and we are interested in ages of old families. Although we only state the results in the exponential case, they also hold in the general case and are proved in [27, Chapter 3] and [29]. However, to obtain the general results in Model I, additional assumptions about the lifespan measure $\Lambda$ are required, which are easily satisfied in the exponential case (for instance, we need the existence of a negative root of $\psi_p$, which, with easy computations, is $b(1-p)-d$ in the exponential case).

In the first result, which is a result in expectation, we show that in both models, the ages are of order of magnitude

$$c_t := \frac{r}{r - r_\star} t. \tag{4.1}$$

**Proposition 4.1** (see [27, 29]). *One supposes that $Z_\star$ is subcritical. For $a \in \mathbb{R}$, one has*

$$\mathbb{E}[O_t(a + c_t)] \xrightarrow[t\to\infty]{} \frac{|r_\star|}{d_\star} e^{-(r-r_\star)a}. \tag{4.2}$$

This result is a consequence of the expected spectrum formula (3.2), summed over $i \geq 1$ and integrated on $(a + c_t, t)$. We also obtain a more precise result about the convergence in distribution of $O_t(a + c_t)$ as $t \to +\infty$.

**Proposition 4.2** (see [27, 29]). *With the same assumptions as in Proposition 4.1, for $a \in \mathbb{R}$, conditional on the survival event, as $t \to \infty$, $O_t(a + c_t)$ converges in distribution to an r.v. $\mathcal{O}$, distributed as a mixed Poisson r.v. whose parameter of mixture is*

$$\frac{b}{r}\frac{|r_\star|}{d_\star}\, e^{-(r-r_\star)a} E, \tag{4.3}$$

*where $E$ is an exponential r.v. with mean 1. Equivalently, $\mathcal{O}$ is geometric on $\{0, 1, \ldots\}$ with success probability*

$$\frac{1}{1 + (b/r)(|r_\star|/d_\star)e^{-(r-r_\star)a}}. \tag{4.4}$$

The proof of this proposition in the general case and for Model I, given in [27], follows arguments of Taïb in [25] and uses the notion of CMJ processes counted with *time-dependent* random characteristics developed by Jagers and Nerman in [26, 34]. The difference with (3.9) is that here the characteristics are allowed to depend on time. This theory provides convergences in distribution, as $t \to +\infty$, of quantities of the form

$$Z_t^{\chi^t} := \sum_x \chi_x^t(t - \sigma_x)\mathbf{1}_{\{\sigma_x \leq t\}}, \tag{4.5}$$

under technical conditions about the family of characteristics $(\chi^t(\cdot), t \geq 0)$. The proof of Proposition 4.2 for Model II is given in [29] and does not make use of random characteristics.

The last result deals with the convergence in distribution of the sequence of the ranked ages of extant families. Let $\mathcal{M}(\mathbb{R})$ be the set of nonnegative $\sigma$-finite measures on $\mathbb{R}$ and finite on $\mathbb{R}^+$, equipped with the *left-vague topology* induced by the maps $\nu \mapsto \int_{\mathbb{R}} f(x)\nu(\mathrm{d}x)$ for all bounded continuous functions $f$ such that there exists $x_0 \in \mathbb{R}$ satisfying for all $x \leq x_0$, $f(x) = 0$.

**Theorem 4.3** (see [27, 29]). *With the same assumptions as previously, let $X_t$ be the point process defined by*

$$X_t(\mathrm{d}x) := \sum_{k\geq 1} \delta_{A_t^k - c_t}(\mathrm{d}x), \tag{4.6}$$

*where $A_t^1 \geq A_t^2 \geq \cdots$ is the decreasing sequence of ages of alive families at $t$. Then, conditional on the survival event, $X_t$ converges as $t \to \infty$ in $\mathcal{M}(\mathbb{R})$ equipped with the left-vague topology to a mixed Poisson point process with an intensity measure*

$$\frac{b}{r}\frac{|r_\star|}{d_\star} E\ (r - r_\star)e^{-(r-r_\star)x}\mathrm{d}x, \tag{4.7}$$

*where $E$ is an exponential r.v. with mean 1.*

## 4.2. Sizes of the Largest Families in the Subcritical Case of Model II

In this paragraph, we still suppose that the clonal process is subcritical and we are interested in similar results as those of Section 4.1 about the sizes of the largest families. The aim is to find a number $x_t$ such that $L_t(x_t)$ converges to a finite and positive limit as $t \to +\infty$.

Concerning Model I, this problem is still open. On the contrary, it is possible to obtain in Model II the sizes of the largest families. In [29], they are given for any life length distribution but to simplify the results, we only state them in the exponential case. The following result is a consequence of (3.3b) applied with $a = t$ and summed over $i \geq x_t + c$. Recall that the clonal process is assumed to be subcritical, so that $\theta > r$.

**Proposition 4.4** (see [29]). *One sets*

$$x_t := \frac{rt - (\theta/(\theta - r))\log t}{-\log(b/(\theta + d))}. \tag{4.8}$$

*Then, for $c \in \mathbb{R}$,*

$$\mathbb{E}[L_t(x_t + c)] \underset{t\to+\infty}{\sim} A(b,d,\theta)\left(\frac{b}{\theta + d}\right)^{c-1+\{-x_t-c\}}, \tag{4.9}$$

*where $\{x\}$ denotes the fractional part of a real number $x$ and where $A(b,d,\theta)$ is an explicit constant that only depends on $b$, $d$, and $\theta$.*

For $t \geq 0$ and $k \geq 1$, we denote by $S_t^k$ the size of the $k$th largest family in the whole population at time $t$. Let

$$X_t := \sum_{k\geq 1} \delta_{S_t^k - x_t}, \tag{4.10}$$

be the point measure of the renormalized sizes of the population. To get rid of fractional parts, the following theorem gives the convergence in distribution of $L_t(x_t + c)$ and $X_t$ along a subsequence. More precisely, for $n \geq 1$, let $t_n$ be such that $x_{t_n} = n$; this equation has a unique solution for any $n$ greater than some integer $n_0$. It satisfies

$$t_n \underset{n\to+\infty}{\sim} \frac{\theta - r}{\theta} \log\left(\frac{\theta + d}{b}\right) n. \tag{4.11}$$

We now state the convergence of the sequence $(X_{t_n},\ n \geq n_0)$.

**Theorem 4.5** (see [29]). *Conditional on the survival event, the sequence $(X_{t_n},\ n \geq n_0)$ of point processes on $\mathbb{Z}$ converges as $n \to +\infty$ on the set $\mathcal{M}(\mathbb{R})$ equipped with the left-vague topology to a mixed Poisson point measure on $Z$ with an intensity measure*

$$A(b,d,\theta)\frac{b-d}{b}\left(1 - \frac{b}{\theta + d}\right) E \sum_{k\in\mathbb{Z}} \left(\frac{b}{\theta + d}\right)^{k-1} \delta_k, \tag{4.12}$$

*where the mixture coefficient $E$ is an exponential r.v. with mean 1.*

## 4.3. Other Results

### 4.3.1. Critical Case in Model I

The case of a critical clonal process $Z_p$ for a general supercritical splitting tree is treated in Section 3.5.1 of [27] where the counterparts of Propositions 4.1 and 4.2 and Theorem 4.3 are proved.

If $(1 - p)m = 1$, provided that the second moment $\sigma^2 := \int_0^\infty \Lambda(\mathrm{d}u)u^2$ is finite and that a condition about the tail distribution of $\Lambda$ holds, ages of oldest families are of order

$$c_t = t - \frac{\log t}{r}. \tag{4.13}$$

Notice that these conditions about $\Lambda$ are trivially satisfied in the exponential case. These results were also proved in [25, Chapter 4] for any CMJ process $Z$, that is, with a birth point process as general as possible, but in that case, limits were not explicit.

Similarly to the subcritical case, the problem of sizes of the largest families is still open. Nevertheless, we can state the following conjecture about their order of magnitude.

**Conjecture 4.6.** *If*

$$x_t := \frac{2}{\sigma^2}\Big(rt^2 - t\log t\Big), \tag{4.14}$$

*as $t \to \infty$, on the survival event, $L_t(x_t)$ converges in distribution to a nondegenerate geometric r.v.*

### *4.3.2. Critical Case in Model II*

The general case when $Z_\theta$ is critical ($\theta = r$) can be found in Sections 3.4 and 5 in [29]. For both ages and sizes, the counterparts of the results of Sections 4.1 and 4.2 hold.

As in Model I, ages of the oldest families are of order $c_t = t - \log t/r$. Moreover, sizes of the largest ones are of order

$$x_t = \frac{r^2}{4\varphi'(r)}\left(t - \frac{\log t}{2r}\right)^2, \tag{4.15}$$

and the point measure,

$$\sum_{k\geq 1}\delta_{\sqrt{S_t^k}-\sqrt{x_t}}, \tag{4.16}$$

converges to a mixed Poisson measure as $t \to +\infty$ but contrary to Theorem 4.5, it does not only hold along a subsequence.

### *4.3.3. Sizes of the Largest Families in Supercritical Cases*

In [27, Chapter 3], general splitting trees in Model I are considered. When the clonal process $Z_p$ is supercritical, that is, when $(1-p)m > 1$, a result about the sizes of the largest families is proved. First notice that, as in (2.9),

$$\mathbb{P}\big(Z_p(t) \longrightarrow 0\big) = 1 - \frac{r_p}{(1-p)b}, \tag{4.17}$$

and on $\{Z_p(t) \to 0\}^c$, $e^{-r_p t}Z_p(t)$ a.s. converges as $t \to \infty$ to an exponential random variable. Hence, the sizes of alive families at time $t$ must be of order $e^{r_p t}$ as $t \to +\infty$. We proved this in [27] by showing that $\mathbb{E}[L_t(e^{r_p t})]$ converges as $t \to +\infty$ to an explicit limit.

Notice that we cannot obtain similar results to Proposition 4.2 and Theorem 4.3 concerning the convergence in distribution of $L_t(e^{(b(1-p)-d)t})$ and the convergence of the associated point measure of the decreasing sequence of family sizes.

In [25], Taïb considers a more general model than our Model I; mutation mechanism is the same but $Z_p$ can be any supercritical CMJ process. In his Theorem 4.6, by using a time-dependent characteristic argument, he proved the convergence in distribution of $L_t(e^{r_p t})$ (to a nonexplicit random variable). However, we have doubts about the application of Theorem A.7, since the technical requirements of this theorem do not seem to hold in his case. These technical requirements are neither proved to hold in [25] nor in [34].

In Model II, for a general supercritical splitting tree, if $Z_\theta$ is supercritical, that is, $r > \theta$, $Z_\theta(t)$ asymptotically grows like $e^{(r-\theta)t}$. In [29, Proposition 3.2], it is proved that $\mathbb{E}[L_t(e^{(r-\theta)t})]$ converges as $t \to +\infty$, but we were unable to obtain any convergence in distribution in that case.

## Acknowledgments

This work was supported by project MANEGE ANR-09-BLAN-0215 (French National Research Agency). The authors want to thank an anonymous referee for his/her careful check of this paper.

## References

[1] J. Geiger, "Size-biased and conditioned random splitting trees," *Stochastic Processes and their Applications*, vol. 65, no. 2, pp. 187–207, 1996.

[2] J. Geiger and G. Kersting, "Depth-first search of random trees, and Poisson point processes," in *Classical and Modern Branching Processes (Minneapolis, MN, 1994)*, vol. 84 of *IMA Volumes in Mathematics and its Applications*, pp. 111–126, Springer, New York, NY, USA, 1997.

[3] A. Lambert, "The contour of splitting trees is a Lévy process," *The Annals of Probability*, vol. 38, no. 1, pp. 348–395, 2010.

[4] P. Jagers, *Branching Processes with Biological Applications*, Wiley-Interscience, London, UK, 1975, Wiley Series in Probability and Mathematical Statistics—Applied Probability and Statistic.

[5] M. A. Nowak, F. Michor, and Y. Iwasa, "The linear process of somatic evolution," *Proceedings of the National Academy of Sciences*, vol. 100, no. 25, pp. 14966–14969, 2003.

[6] Y. Iwasa, M. A. Nowak, and F. Michor, "Evolution of resistance during clonal expansion," *Genetics*, vol. 172, no. 4, pp. 2557–2566, 2006.

[7] S. Sagitov and M. C. Serra, "Multitype Bienaymé-Galton-Watson processes escaping extinction," *Advances in Applied Probability*, vol. 41, no. 1, pp. 225–246, 2009.

[8] R. Durrett and S. Moseley, "Evolution of resistance and progression to disease during clonal expansion of cancer," *Theoretical Population Biology*, vol. 77, no. 1, pp. 42–48, 2010.

[9] R. Durrett and J. Mayberry, "Traveling waves of selective sweeps," *The Annals of Applied Probability*, vol. 21, no. 2, pp. 699–744, 2011.

[10] K. Danesh, R. Durrett, L. J. Havrilesky, and E. Myers, "A branching process model of ovarian cancer," *Journal of Theoretical Biology*, vol. 314, pp. 10–15, 2012.

[11] D. G. Kendall, "Birth-and-death processes, and the theory of carcinogenesis," *Biometrika*, vol. 47, pp. 13–21, 1960.

[12] T. Stadler, "Inferring epidemiological parameters based on allele frequencies," *Genetics*, vol. 188, no. 3, pp. 663–672, 2011.

[13] A. Lambert and P. Trapman, "Splitting trees stopped when the first clock rings and Vervaat's transformation," *Journal of Applied Probability*. In press.

[14] J. Gani and G. F. Yeo, "Some birth-death and mutation models for phage reproduction," *Journal of Applied Probability*, vol. 2, pp. 150–161, 1965.

[15] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography*, Princeton University Press, Princeton, NJ, USA, 2001.

[16] B. Haegeman and R. S. Etienne, "Relaxing the zero-sum assumption in neutral biodiversity theory," *Journal of Theoretical Biology*, vol. 252, no. 2, pp. 288–294, 2008.

[17] A. Lambert, "Species abundance distributions in neutral models with immigration or mutation and general lifetimes," *Journal of Mathematical Biology*, vol. 63, no. 1, pp. 57–72, 2011.

[18] W. J. Ewens, "The sampling theory of selectively neutral alleles," *Theoretical Population Biology*, vol. 3, pp. 87–112, 1972, erratum, ibid. vol. 3, p. 240, 1972; erratum, ibid. vol. 3, p. 376, 1972.

[19] R. C. Griffiths and A. G. Pakes, "An infinite-alleles version of the simple branching process," *Advances in Applied Probability*, vol. 20, no. 3, pp. 489–524, 1988.

[20] J. Bertoin, "The structure of the allelic partition of the total population for Galton-Watson processes with neutral mutations," *The Annals of Probability*, vol. 37, no. 4, pp. 1502–1523, 2009.
[21] J. Bertoin, "A limit theorem for trees of alleles in branching processes with rare neutral mutations," *Stochastic Processes and their Applications*, vol. 120, no. 5, pp. 678–697, 2010.
[22] A. G. Pakes, "An infinite alleles version of the Markov branching process," *Australian Mathematical Society A*, vol. 46, no. 1, pp. 146–169, 1989.
[23] Y. E. Maruvka, N. M. Shnerb, and D. A. Kessler, "Universal features of surname distribution in a subsample of a growing population," *Journal of Theoretical Biology*, vol. 262, no. 2, pp. 245–256, 2010.
[24] Y. E. Maruvka, D. A. Kessler, and N. M. Shnerb, "The birth-death-mutation process: a new paradigm for fat tailed distributions," *PLoS ONE*, vol. 6, no. 11, article e26480, 2011.
[25] Z. Taïb, *Branching Processes and Neutral Evolution*, vol. 93 of *Lecture Notes in Biomathematics*, Springer, Berlin, Germany, 1992.
[26] P. Jagers and O. Nerman, "The growth and composition of branching populations," *Advances in Applied Probability*, vol. 16, no. 2, pp. 221–259, 1984.
[27] M. Richard, *Arbres, Processus de branchement non markoviens et Processus de Lvy [Ph.D. thesis]*, UPMC, Paris, France, 2011.
[28] N. Champagnat and A. Lambert, "Splitting trees with neutral Poissonian mutations I: small families," *Stochastic Processes and their Applications*, vol. 122, no. 3, pp. 1003–1033, 2012.
[29] N. Champagnat and A. Lambert, "Splitting trees with neutral Poissonian mutations II:," *Largest and Oldest families*. In press, http://arxiv.org/abs/1108.4812.
[30] J. Bertoin, *Lévy Processes*, vol. 121 of *Cambridge Tracts in Mathematics*, Cambridge University Press, Cambridge, UK, 1996.
[31] O. Nerman, "On the convergence of supercritical general (C-M-J) branching processes," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, no. 3, pp. 365–395, 1981.
[32] M. Richard, "Limit theorems for supercritical age-dependent branching processes with neutral immigration," *Advances in Applied Probability*, vol. 43, no. 1, pp. 276–300, 2011.
[33] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, vol. 55 of *National Bureau of Standards Applied Mathematics Series*, Superintendent of Documents, U.S. Government Printing Office, Washington, DC, USA, 1964.
[34] P. Jagers and O. Nerman, "Limit theorems for sums determined by branching and other exponentially growing processes," *Stochastic Processes and their Applications*, vol. 17, no. 1, pp. 47–71, 1984.

Hindawi Publishing Corporation
International Journal of Stochastic Analysis
Volume 2012, Article ID 598701, 32 pages
doi:10.1155/2012/598701

*Research Article*

# Stochastic Methodology for the Study of an Epidemic Decay Phase, Based on a Branching Model

**Sophie Pénisson[1] and Christine Jacob[2]**

[1] *Université Paris-Est, LAMA (UMR 8050), UPEMLV, UPEC, CNRS, 94010 Créteil, France*
[2] *INRA, MIA (UR 341), 78352 Jouy-en-Josas, France*

Correspondence should be addressed to Sophie Pénisson, sophie.penisson@u-pec.fr

Received 16 July 2012; Revised 10 October 2012; Accepted 10 October 2012

Academic Editor: Charles J. Mode

Copyright © 2012 S. Pénisson and C. Jacob. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We present a stochastic methodology to study the decay phase of an epidemic. It is based on a general stochastic epidemic process with memory, suitable to model the spread in a large open population with births of any rare transmissible disease with a random incubation period and a Reed-Frost type infection. This model, which belongs to the class of multitype branching processes in discrete time, enables us to predict the incidences of cases and to derive the probability distributions of the extinction time and of the future epidemic size. We also study the epidemic evolution in the worst-case scenario of a very late extinction time, making use of the $Q$-process. We provide in addition an estimator of the key parameter of the epidemic model quantifying the infection and finally illustrate this methodology with the study of the Bovine Spongiform Encephalopathy epidemic in Great Britain after the 1988 feed ban law.

## 1. Introduction

Outbreaks of infectious diseases of animals or humans are subject, when possible, to control measures aiming at curbing their spread. Effective measures should force the epidemic to enter its decay phase and to reach extinction. The decay phase can then be simply detected by a decrease of the number of cases, when this decrease is obvious. However this is not always the case, and this rough qualitative information might not be sufficient to evaluate accurately the effectiveness of the proposed measures to reduce the final size and duration of the outbreak. The goal of this paper is to present a stochastic methodology in discrete time to study more accurately the decay phase of an epidemic. Our framework is the spread, in a large open population, of a rare transmissible disease such that the infection process may be

assumed to follow a Reed-Frost type model, with a probability for a susceptible to become infected by a given dose of pathogens inversely proportional to the total population size. Moreover the latent period (during which an individual is infected but not yet infectious) may be random and long compared to the generation time. Questions about the decay phase include the following: which quantitative criteria can ensure that the disease has entered an extinction phase? What is the probability distribution of the epidemic extinction time, of the epidemic final size, and of the incidence of infected individuals? Finally, what would be the evolution of the epidemic in the event of a very late extinction of the disease?

From a practical point of view, it is generally impossible to observe all infections. Susceptible and infected but not yet infectious individuals are most often not distinguishable, being both apparently healthy. This leads to the fact that the only available observations correspond to the incidence of individuals with clinical symptoms. One way to deal with this lack of information was proposed in [1] by Panaretos, who used a model taking into account two types of infected individuals, the observed and the unobserved. In order to answer the previous questions, we choose here a different approach, considering a stochastic model depending on the sole incidences $\{X_n\}_n$ of infectives at each time. We assume that an infective can transmit the disease during one given time unit at most. Therefore the incidence of infectives corresponds to the incidence of cases. The process then describes in a recursive way how one single infective can indirectly generate new infectives (so-called "secondary cases") $k$ time units later, where $1 \leqslant k \leqslant d$. We assume that this number of secondary cases follows a Poisson probability distribution with parameter $\Psi_k > 0$. The recursive formula defining $\{X_n\}_n$ is then the following:

$$X_n = \sum_{k=1}^{d} \sum_{i=1}^{X_{n-k}} Y_{n-k,n,i}, \tag{1.1}$$

where the variable $Y_{n-k,n,i}$ is the incidence of secondary cases produced at time $n$ with a delay $k$ (latent period) by individual $i$ infectious at time $n-k$. The $\{Y_{n-k,n,i}\}_{i,k}$ are assumed independent given $\mathcal{F}_{n-1} := \sigma(\{X_{n-k}\}_{k\geqslant 1})$, and the $\{Y_{n-k,n,i}\}_i$ are assumed i.i.d. (identically and independently distributed) given $\mathcal{F}_{n-1}$, with a common Poisson distribution with parameter $\Psi_k$. This model is therefore time homogeneous and is in this sense less general than the one introduced in [2], which describes the spread of infectious animal diseases in a varying environment. However since we focus on the extinction phase only, the assumption of a constant environment with no new control measure is well founded and enables us to describe more accurately the decay phase. This process is the generalization of the well-known single-type BGW (Bienaymé-Galton-Watson) branching process, which is the limit, as the total population size tends to infinity, of the process describing the spread in a closed population of an infectious disease with a negligible latent period and a probability to become infected following a Reed-Frost model (see, e.g., [3, 4] and citations therein).

The core of the paper lies in Section 2, where the whole methodology is presented. We first formulate the epidemic model $\{X_n\}_n$ as a multitype branching process with Poissonian transitions, the types representing the memory of the process. This formulation provides useful analytical results such as the extinction criteria, and the distributions of the extinction time and of the epidemic size (Sections 2.1–2.3). Then, in order to investigate the worst-case scenario of an extreme late extinction of the epidemic, we introduce in Section 2.4 the $Q$-process $\{X_n^*\}_n$, obtained by conditioning $\{X_n\}_n$ on a very late extinction. Using this process, we focus the study on the early behavior of the decay phase in the worst-case scenario, rather

than on its long range behavior, which would have little meaning in our setting. Motivated by practical applications to real epidemics, for which we want to predict the processes $\{X_n\}_n$ and $\{X_n^*\}_n$, as well as the derived distributions above, we need to know the values of the parameters $\{\Psi_k\}_k$. We may write $\Psi_k = \sum_{a=1}^{a_m-k} \theta_a P_{\text{inc},a}(k) P_{\text{age}}(a+k)$, where $\theta_a$ is the mean number of individuals infected at age $a$ by an infective by direct or indirect transmission, $a_m$ is the largest survival age, $P_{\text{inc},a}(k)$ is the probability for the individual aged $a$ at infection to have a latent period equal to $k$ (given his survival), and $P_{\text{age}}(a+k)$ is the probability to be aged $a+k$ at the end of the latent period. Parameters $\{\theta_a\}_a$ are the key quantities for the spread of the disease and can be subject to changes due to control measures during the epidemic. We assume here that $\theta_a = \theta_0 + p_a$, where $p_a$ is constant over time, while $\theta_0$ may change with control measures. A typical example is when $\theta_0$ is the mean number of individuals infected by an infective by horizontal route at age $a$, assumed independent of $a \geqslant 2$, and $p_1$ represents the maternal transmission probability $p_{\text{mat}}$. In this case

$$\Psi_k = \theta_0 \sum_{a=k+1}^{a_m} P_{\text{inc},a-k}(k) P_{\text{age}}(a) + p_{\text{mat}} P_{\text{inc},1}(k) P_{\text{age}}(k+1). \tag{1.2}$$

So we assume here that, except for $\theta_0$, the other parameters of the $\{\Psi_k\}_k$ are constant over time and are known (generally estimated) from previous experiences or from the study of the whole epidemic evolution, in particular its growth phase. We moreover assume that each $\Psi_k$ depends affinely on $\theta_0$ (see Section 3.1). In Sections 2.5 and 2.6, we provide optimal WCLSE (Weighted Conditional Least Squares Estimators) of $\theta_0$ in the decay phase, in the frame of $\{X_n\}_n$ as well as in the frame of the associated conditioned process $\{X_n^*\}_n$, and prove the strong consistency and the asymptotic normality of these estimators.

The final Section 3 is devoted to the application of this method to real epidemics. We first present in Section 3.1 some general conditions under which the spread of a SEIR disease (susceptible, exposed (latent), infectious, removed) can be approximated by our epidemic process defined by (1.1) and give an explicit derivation of the parameters $\{\Psi_k\}_k$. We then illustrate the methodology in Section 3.2 with the decay phase of the BSE (Bovine Spongiform Encephalopathy) epidemic in Great Britain. According to the available data [5], the epidemic is obviously fading out. We assume that the $\{\Psi_k\}_k$ satisfy (1.2). Then thanks to the stochastic tools developed here, we provide in addition to this rough information, short- and long-term predictions about the future spread of the disease as well as an estimation of a potentially remaining horizontal infection route after the 1988 feed ban law.

## 2. Methodology for the Study of an Epidemic Decay Phase

In this section we present a general methodology to study the decay phase of a SEIR disease in a large population, modeled by the process (1.1) defined in Section 1. Our main goal is to provide analytical tools to evaluate the efficiency of the last control measures taken prior to the considered time period. Most of our results are derived from the fact that this epidemic model can be seen as a multitype branching process. Indeed, $\{X_n\}_n$ defined by (1.1) is a Markovian process of order $d$. Consequently, the $d$-dimensional process $\{\mathbf{X}_n\}_n$ defined by

$$\mathbf{X}_n := (X_n, X_{n-1}, \ldots, X_{n-d+1}) \tag{2.1}$$

is Markovian of order 1, and it stems directly from (1.1) that $\{\mathbf{X}_n\}_n$ is a multitype Bienaymé-Galton-Watson (BGW) process with $d$ types (see, e.g., [6]). Note that the $d$ types in this branching process do not correspond to any attribute of the individuals in the population, which is usually the case in mathematical biology (see, e.g., [7]), but simply correspond to the memory of the process $\{X_n\}_n$. The information provided by the $d$-dimensional Markovian process $\{\mathbf{X}_n\}_n$ is therefore the same as the one given by the 1-dimensional $d$-Markovian process $\{X_n\}_n$, but the multitype branching process setting gives us powerful mathematical tools and results stemming from the branching processes theory [6]. The first basic tool is the generating function of the offspring distribution of $\{\mathbf{X}_n\}_n$, $\mathbf{f} := (f_1, \ldots, f_d)$, defined on $[0,1]^d$ by $f_i(\mathbf{r}) := \mathbb{E}[\mathbf{r}^{\mathbf{X}_1} \mid \mathbf{X}_0 = \mathbf{e}_i]$, where $\mathbf{e}_i := (0, \ldots, 1, \ldots, 0)$ denotes the $i$th basis vector of $\mathbb{N}^d$ and $\mathbf{u}^{\mathbf{v}} := \prod_{i=1}^{d} u_i^{v_i}$ for $\mathbf{u}, \mathbf{v} \in \mathbb{N}^d$. For all $\mathbf{r} \in [0,1]^d$, we have here

$$\begin{aligned} f_i(\mathbf{r}) &:= e^{-(1-r_1)\Psi_i} r_{i+1}, \quad i = 1 \cdots d-1, \\ f_d(\mathbf{r}) &:= e^{-(1-r_1)\Psi_d}. \end{aligned} \tag{2.2}$$

The second basic tool is the mean matrix $\mathbf{M}$ defined by $\mathbb{E}(\mathbf{X}_n \mid \mathcal{F}_{n-1}) = \mathbf{X}_{n-1}\mathbf{M}$, which is here

$$\mathbf{M} = \begin{pmatrix} \Psi_1 & 1 & 0 & \cdots & 0 \\ \Psi_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ \Psi_{d-1} & 0 & \cdots & \cdots & 1 \\ \Psi_d & 0 & \cdots & \cdots & 0 \end{pmatrix}. \tag{2.3}$$

Let us notice that, since $\Psi_k > 0$ for each $k = 1 \cdots d$, then $\{\mathbf{X}_n\}_n$ is nonsingular, positive regular (see [6]) and satisfies the $X \log X$ condition,

$$\mathbb{E}[\|\mathbf{X}_1\| \ln \|\mathbf{X}_1\| \mid \mathbf{X}_0 = \mathbf{e}_i] < \infty, \quad i = 1, \ldots, d, \tag{2.4}$$

where $\|\cdot\|$ denotes the sup norm in $\mathbb{R}^d$.

## 2.1. Extinction of the Epidemic

### 2.1.1. Almost Sure Extinction

Since the single-type process $\{X_n\}_n$ has a memory of size $d$, it becomes extinct when it is null at $d$ successive times, or equivalently as soon as the $d$-dimensional process $\{\mathbf{X}_n\}_n$ reaches the $d$-dimensional null vector $\mathbf{0}$. According to the theory of multitype positive regular and nonsingular BGW processes [6], the extinction of the process $\{\mathbf{X}_n\}_n$ occurs almost surely (a.s.), if and only if $\rho \leqslant 1$, where $\rho$ is the dominant eigenvalue (also called the Perron's root) of the mean matrix $\mathbf{M}$. Thus $\rho$ is solution of $\sum_{k=1}^{d} \Psi_k \rho^{-k} = 1$. In general for $d > 1$, $\rho$ has no explicit expression. However, $\sum_{k=1}^{d} \Psi_k \rho^{-k} = 1$ leads directly to the following explicit threshold criteria.

**Proposition 2.1.** *The epidemic becomes extinct almost surely if and only if $R_0 \leqslant 1$, where $R_0 := \sum_{k=1}^{d} \Psi_k$ is the total mean number of secondary cases generated by one infective in a SEIR disease. We call $R_0$ the basic reproduction number.*

*Moreover, when $\rho \leqslant 1$, then $R_0 \leqslant \rho$ with equality if and only if either $\rho = 1$ or $d = 1$, and when $\rho > 1$, then $R_0 \geqslant \rho$ with equality if and only if $d = 1$.*

Note that when $d > 1$, $R_0$ only provides information about the threshold level, whereas $\rho$ provides an additional information about the speed of extinction of the process, as shown in the next two paragraphs.

### 2.1.2. Speed of Extinction

Thanks to well-known results in the literature about multitype branching processes and more particularly to the Perron-Frobenius theorem (see, e.g., [6]), we can deduce the expected incidence of infectives in the population at time $n$, for $n$ large. Denoting by $\mathbf{u}$ and $\mathbf{v}$ the right and left eigenvectors of $\mathbf{M}$ associated to the Perron's root $\rho$; that is, $\mathbf{M}\mathbf{u}^T = \rho\mathbf{u}^T$ and $\mathbf{v}\mathbf{M} = \rho\mathbf{v}$, with the normalization convention $\mathbf{u}\cdot\mathbf{1} = \mathbf{u}\cdot\mathbf{v} = 1$, where $\mathbf{u}\cdot\mathbf{v}$ stands for the usual scalar product in $\mathbb{R}^d$ and where the superscript $T$ denotes the transposition, then $\mathbb{E}(\mathbf{X}_n \mid \mathbf{X}_0) = \mathbf{X}_0\mathbf{M}^n \underset{n\to\infty}{\sim} \rho^n\mathbf{X}_0\mathbf{u}^T\mathbf{v}$. The first coordinate in the latter formula becomes for the epidemic process $\rho^n \sum_{i=1}^{d} X_{-i+1}u_iv_1$. Computing explicitly $\mathbf{u}$ and $\mathbf{v}$, we obtain that for all $i = 1\cdots d$,

$$u_i = \frac{\sum_{k=i}^{d}\Psi_k\rho^{-k+i}}{\sum_{j=1}^{d}\sum_{k=j}^{d}\Psi_k\rho^{-k+j}}, \qquad v_i = \rho^{-i}\frac{\sum_{j=1}^{d}\sum_{k=j}^{d}\Psi_k\rho^{-k+j}}{\sum_{j=1}^{d}\sum_{k=j}^{d}\Psi_k\rho^{-k}}, \tag{2.5}$$

which leads to the following asymptotic result:

$$\mathbb{E}(X_n \mid \mathbf{X}_0) \underset{n\to\infty}{\sim} \rho^n\sum_{i=1}^{d}X_{-i+1}\frac{\sum_{k=i}^{d}\Psi_k\rho^{-k+i-1}}{\sum_{j=1}^{d}\sum_{k=j}^{d}\Psi_k\rho^{-k}}. \tag{2.6}$$

Hence if $\rho < 1$, the mean number of infectives decreases exponentially at the rate $\rho$. In the following section, we provide a much finer result on the estimation of the disease extinction time in the population.

### 2.1.3. Extinction Time of the Epidemic

The extinction time distribution can be derived as a function of the offspring generating function. As usual in stochastic processes, this quantity is calculated conditionally on the initial value $\mathbf{X}_0 = (X_0, X_{-1}, \ldots, X_{-d+1})$, but for the sake of simplicity we do not let it appear in the notations. Note that since we are building tools for the prediction of the spread of the disease, the time origin 0 corresponds here to the time of the last available data (generally the current date). Let $T := \inf\{n \geqslant 1, \mathbf{X}_n = \mathbf{0}\}$ denote the extinction time of the process $\{\mathbf{X}_n\}_n$, and let $\mathbf{f}_n := \mathbf{f}\circ\mathbf{f}_{n-1}$ be the $n$th iterate of the generating function $\mathbf{f}$ given by (2.2). We denote $\mathbf{f}_n := (f_{n,1}, \ldots, f_{n,d})$. Then, by the branching property of the process ([6]), the probability of extinction of the epidemic before time $n$ is immediately given by $\mathbb{P}(T \leqslant n) = \mathbb{P}(\mathbf{X}_n = \mathbf{0}) = \mathbf{f}_n(\mathbf{0})^{\mathbf{X}_0}$, that is to say,

$$\mathbb{P}(T \leqslant n) = \big(f_{n,1}(\mathbf{0})\big)^{X_0}\big(f_{n,2}(\mathbf{0})\big)^{X_{-1}}\cdots\big(f_{n,d}(\mathbf{0})\big)^{X_{-d+1}}. \tag{2.7}$$

It can be immediately deduced from convergence results for $\mathbf{f}_n(\mathbf{0})$ as $n \to \infty$ [8], that if $\rho = 1$, $\mathbb{P}(T \leqslant n) \sim 1 - (n\eta)^{-1}\mathbf{X}_0 \cdot \mathbf{u}$, while if $\rho < 1$, $\mathbb{P}(T \leqslant n) \sim 1 - \rho^n \gamma \mathbf{X}_0 \cdot \mathbf{u}$, for some constants $\eta, \gamma > 0$. As a consequence, the closer $\rho$ is to unity, the longer the time to extinction will be in most realizations. More specifically, (2.7) enables the exact computation (resp., estimation) of $\mathbb{P}(T \leqslant n)$ for any $n$ by the iterative computation of $\mathbf{f}_n$, $\mathbf{X}_0$ being given, when the parameters $\Psi_k$ of (1.1) are known (resp., estimated). Moreover, since for $\rho \leqslant 1$ the epidemic becomes extinct in an a.s. finite time and $\mathbb{P}(T \leqslant n) \nearrow_{n\to\infty} 1$, then for any given probability $p \in [0,1[$ there exists $n \in \mathbb{N}$ such that $\mathbb{P}(T \leqslant n) \geqslant p$. So in practice, for any $p \in [0,1[$, (2.7) enables us to compute the $p$-quantile $n_p^T$ of the extinction time,

$$n_p^T := \min\{n \geqslant 1 : \ \mathbb{P}(T \leqslant n) \geqslant p\}. \tag{2.8}$$

### 2.2. Total Size of the Epidemic

Under the assumption $\rho \leqslant 1$ and the independence of the $\{Y_{n-l,n,i}\}_{i,l,n}$ (we previously assumed the independence of the $\{Y_{n-l,n,i}\}_{i,l}$, for each $n$), we derive the distribution of the future total size $N := \sum_{n=1}^{T} X_n$ of the epidemic until its extinction, that is, the future total number of infectives until the extinction of the disease. It can be shown [9] that, given the initial value $\mathbf{X}_0$, the time origin being the same as in Section 2.1, the probability distribution of $N$ is

$$N \overset{\mathfrak{D}}{=} \sum_{k=1}^{d} \sum_{i=1}^{X_{-k+1}} \left( \sum_{j=1}^{Y_{k,i}} N_{k,i,j} \right), \qquad Y_{k,i} := \sum_{l=k}^{d} Y_{-k+1,-k+1+l,i}, \tag{2.9}$$

where $\overset{\mathfrak{D}}{=}$ denotes the equality in distribution, an empty sum is by convention 0, the $\{Y_{k,i}\}_{k,i}$ are independent, and the $\{Y_{k,i}\}_i$ and the $\{N_{k,i,j}\}_{k,i,j}$ are i.i.d. with

$$Y_{k,i} \overset{\mathfrak{D}}{=} \text{Poiss}\left( \sum_{l=k}^{d} \Psi_l \right), \qquad N_{k,i,j} \overset{\mathfrak{D}}{=} \text{Borel} - \text{Tanner}\left( \sum_{l=1}^{d} \Psi_l, 1 \right), \tag{2.10}$$

that is, for each $n \geqslant 1$, $\mathbb{P}(N_{k,i,j} = n) = e^{-n\sum_{l=1}^{d}\Psi_l}(n\sum_{l=1}^{d}\Psi_l)^{n-1}(n!)^{-1}$. Consequently, under the convention that an empty product is 1, the probability distribution of $N$ is, for any $n \in \mathbb{N}$,

$$\begin{aligned} \mathbb{P}(N = n) = \sum_{\substack{\{0 \leqslant y_{k,i} \leqslant n, \{1 \leqslant n_{k,i,j} \leqslant n\}_j\}_{i,k}: \\ \sum_{k=1}^{d}\sum_{i=1}^{X_{-k+1}}\sum_{j=1}^{y_{k,i}} n_{k,i,j} = n}} & \prod_{k=1}^{d}\prod_{i=1}^{X_{-k+1}} e^{-\sum_{l=k}^{d}\Psi_l} \frac{\left(\sum_{l=k}^{d}\Psi_l\right)^{y_{k,i}}}{y_{k,i}!} \\ & \times \prod_{j=1}^{y_{k,i}} e^{-n_{k,i,j}\sum_{l=1}^{d}\Psi_l} \frac{\left(n_{k,i,j}\sum_{l=1}^{d}\Psi_l\right)^{n_{k,i,j}-1}}{n_{k,i,j}!}, \end{aligned} \tag{2.11}$$

which may be calculated (resp., estimated), replacing the $\Psi_k$ by their values (resp., estimations). In practice, for any $p \in [0,1[$, (2.11) enables to compute the $p$-quantile $n_p^N$ of the total epidemic size,

$$n_p^N := \min\{n \geqslant 1 : \ \mathbb{P}(N \leqslant n) \geqslant p\}. \tag{2.12}$$

We obtain moreover an explicit formula for the mean value and variance of the size of the epidemic,

$$\mathbb{E}(N) = \frac{\sum_{k=1}^{d} X_{-k+1} \sum_{l=k}^{d} \Psi_l}{1 - \sum_{l=1}^{d} \Psi_l}, \qquad \mathbb{V}\mathrm{ar}(N) = \frac{\sum_{k=1}^{d} X_{-k+1} \sum_{l=k}^{d} \Psi_l}{\left(1 - \sum_{l=1}^{d} \Psi_l\right)^3}. \tag{2.13}$$

### 2.3. Exposed Population

Depending on the disease, it might also be crucial to study and predict the evolution of the incidence of exposed individuals in the population, which is generally unobservable. We assume that this information is given by the process $\{Z_n\}_n$ defined by the conditional distribution $Z_n \mid X_n \overset{\mathcal{D}}{=} \mathrm{Poiss}(\Psi_0 X_n)$, where $\Psi_0$ is the mean number of individuals infected at time $n$ by an infective of this time (see Section 3.1). This property enables on the one hand to reconstruct the whole past epidemic (i.e., the incidence of infectives as well as of exposed individuals) thanks to the observable data. On the other hand, it allows to simulate the evolution of the incidence of exposed individuals in the future, based on predictions of the evolution of the epidemic process $\{X_n\}_n$.

### 2.4. Worst-Case Scenario: Very Late Extinction of the Epidemic

Even in the case when the epidemic dies out almost surely ($\rho \leqslant 1$), and although one can provide the $p$-quantile $n_p^T$ of the extinction time with the probability $p$ as large as wanted (see (2.8)), the epidemic might become extinct after this time with a small but nonnull probability of order $1-p$. This raises the following question: how would the incidences of infectious and exposed individuals evolve in the (unlikely) case of a very late extinction? In terms of risk analysis, this issue appears to be crucial to evaluate the risks associated with this worst-case scenario. The tools developed in the previous subsections allow to evaluate the probability of all possible outcomes. But since the worst ones, typically a very late extinction, have a negligible probability, these tools do not bring any information in these worst cases and in particular do not inform on the evolution at each time step of the spread of the disease (would it decrease extremely slowly, stay at a constant rate for a very long time, present several peaks in its evolution, etc.?). In order to study the propagation of the epidemic in the decay phase, assuming that extinction occurs very late, we are interested in the distribution of the process $\{\mathbf{X}_n\}_n$ conditionally on the event that the epidemic has not become extinct at time $k$, where $k$ is very large. We therefore consider for any $n_1, n_2, \ldots \in \mathbb{N}$ and any $\mathbf{i}_0, \mathbf{i}_1, \mathbf{i}_2, \ldots \in \mathbb{N}^d$ the conditioned probability $\mathbb{P}_{\mathbf{i}_0}(\mathbf{X}_{n_1} = \mathbf{i}_1, \ldots, \mathbf{X}_{n_r} = \mathbf{i}_r \mid \mathbf{X}_k \neq \mathbf{0})$, where the subscript $\mathbf{i}_0$ denotes the initial value. If $k$ is finite this distribution cannot be easily handled due to its time

inhomogeneity. However, when $\rho \leqslant 1$, it is known ([10]) that this conditioned distribution converges, as $k \to \infty$, to the distribution of a $d$-dimensional Markov process $\{\mathbf{X}_n^*\}_n$:

$$\lim_{k\to\infty} \mathbb{P}_{\mathbf{i}_0}(\mathbf{X}_{n_1} = \mathbf{i}_1, \ldots, \mathbf{X}_{n_r} = \mathbf{i}_r \mid \mathbf{X}_k \neq \mathbf{0}) = \mathbb{P}_{\mathbf{i}_0}(\mathbf{X}_{n_1}^* = \mathbf{i}_1, \ldots, \mathbf{X}_{n_r}^* = \mathbf{i}_r). \tag{2.14}$$

We will further discuss in Proposition 2.5 the relevancy of approximating the conditioned probability for $k$ fixed by the limiting object (2.14). The conditioned process $\{\mathbf{X}_n^*\}_n$ defined by (2.14) is known in the literature as the $Q$-process associated with $\{\mathbf{X}_n\}_n$, also described as the process conditioned on "not being extinct in the distant future." It has the following transition probability [10]: for every $n \geqslant 1$, $\mathbf{i}, \mathbf{j} \in \mathbb{N}^d$, $\mathbf{i} \neq \mathbf{0}$,

$$\mathbb{P}(\mathbf{X}_n^* = \mathbf{j} \mid \mathbf{X}_{n-1}^* = \mathbf{i}) = \frac{1}{\rho}\frac{\mathbf{j}\cdot\mathbf{u}}{\mathbf{i}\cdot\mathbf{u}}\mathbb{P}(\mathbf{X}_n = \mathbf{j} \mid \mathbf{X}_{n-1} = \mathbf{i}), \tag{2.15}$$

where $\mathbf{u}$ is the normalized right eigenvector of $\mathbf{M}$ associated to the Perron's root $\rho$ as introduced in Section 2.1, and computed explicitly in (2.5). In the same way as for the process $\{\mathbf{X}_n\}_n$, we define the 1-dimensional process $X_n^* := X_{n,1}^*$. By construction we then have $X_{n,i}^* = X_{n-i+1}^*$, for each $n$ and each $i = 1 \cdots d$.

**Proposition 2.2.** *The stochastic process $\{X_n^*\}_n$ is, conditionally on its past, distributed as the sum of two independent Poisson and Bernoulli random variables:*

$$X_n^* \mid \mathbf{X}_{n-1}^* \overset{\mathfrak{D}}{=} \mathrm{Poiss}(\mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi}) * \mathcal{B}(p(\mathbf{X}_{n-1}^*)), \tag{2.16}$$

*where $\boldsymbol{\Psi} := (\Psi_1, \ldots, \Psi_d)$, $*$ is the convolution product symbol, and*

$$p(\mathbf{X}_{n-1}^*) := \frac{u_1 \mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi}}{u_1 \mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi} + \sum_{k=2}^d X_{n-k+1}^* u_k}. \tag{2.17}$$

*Proof.* Applying (1.1) and (2.15), we obtain that for all $j \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{P}(X_n^* = j \mid \mathbf{X}_{n-1}^*) \\
&\quad = \mathbb{P}\left(\mathbf{X}_n^* = \left(j, X_{n-1}^*, \ldots, X_{n-(d-1)}^*\right) \mid \mathbf{X}_{n-1}^*\right) \\
&\quad = \frac{j u_1 + \sum_{k=2}^d X_{n-k+1}^* u_k}{\rho \mathbf{X}_{n-1}^* \cdot \mathbf{u}} \mathbb{P}\left(\mathbf{X}_n = \left(j, X_{n-1}^*, \ldots, X_{n-(d-1)}^*\right) \mid \mathbf{X}_{n-1} = \mathbf{X}_{n-1}^*\right) \\
&\quad = \frac{j u_1 + \sum_{k=2}^d X_{n-k+1}^* u_k}{\rho \mathbf{X}_{n-1}^* \cdot \mathbf{u}} \frac{\left(\mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi}\right)^j}{j!} e^{-\mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi}} \\
&\quad = \frac{u_1 \mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi}}{\rho \mathbf{X}_{n-1}^* \cdot \mathbf{u}} \frac{\left(\mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi}\right)^{j-1}}{(j-1)!} e^{-\mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi}} + \frac{\sum_{k=2}^d X_{n-k+1}^* u_k}{\rho \mathbf{X}_{n-1}^* \cdot \mathbf{u}} \frac{\left(\mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi}\right)^j}{j!} e^{-\mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi}}.
\end{aligned} \tag{2.18}$$

The equality $\mathbf{M}\mathbf{u}^T = \rho\mathbf{u}^T$ implies that for all $k = 1\cdots d-1$, $\rho u_k = \Psi_k u_1 + u_{k+1}$, and that $\rho u_d = \Psi_d u_1$. Consequently, $\rho\mathbf{X}^*_{n-1}\cdot\mathbf{u} = u_1\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi} + \sum_{k=2}^{d} X^*_{n-k+1}u_k$, and thus

$$\begin{aligned}
&\mathbb{P}\left(X^*_n = j \mid \mathbf{X}^*_{n-1}\right)\\
&\quad= \frac{u_1\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi}}{u_1\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi} + \sum_{k=2}^{d} X^*_{n-k+1}u_k}\frac{\left(\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi}\right)^{j-1}}{(j-1)!}e^{-\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi}}\\
&\qquad+\left(1-\frac{u_1\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi}}{u_1\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi} + \sum_{k=2}^{d} X^*_{n-k+1}u_k}\right)\frac{\left(\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi}\right)^{j}}{j!}e^{-\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi}}\\
&\quad= \left[\mathrm{Poiss}\left(\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi}\right) * \mathcal{B}\left(\frac{u_1\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi}}{u_1\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi} + \sum_{k=2}^{d} X^*_{n-k+1}u_k}\right)\right](j).
\end{aligned} \tag{2.19}$$

□

*Remark 2.3.* Note that if one compares (2.16) with the transition probability of the unconditioned process $X_n \mid \mathbf{X}_{n-1} \overset{\mathfrak{D}}{=} \mathrm{Poiss}(\mathbf{X}_{n-1}\cdot\boldsymbol{\Psi})$, it appears that $\{X^*_n\}_n$ behaves at each time step like $\{X_n\}_n$, according to a Poisson distribution, except that it has the possibility at each time step to add one unit or not, according to a Bernoulli random variable. Moreover, if $X^*_{n-1} = \cdots = X^*_{n-(d-1)} = 0$, then according to (2.17), $p(\mathbf{X}^*_{n-1}) = 1$, which implies that at time $n$, the probability to add one unit is equal to one, thus preventing the extinction of the process.

**Proposition 2.4.** *The process $\{\mathbf{X}^*_n\}_n$ admits a stationary probability measure $\pi$ with finite first- and second-order moments.*

*Proof.* Since the multitype branching process $\{\mathbf{X}_n\}_n$ satisfies property (2.4), it is known [10] that the $Q$-process $\{\mathbf{X}^*_n\}_n$ is positive recurrent with a stationary probability measure $\pi$ given by

$$\pi(\mathbf{i}) := \frac{\mathbf{i}\cdot\mathbf{u}\ \nu(\mathbf{i})}{\sum_{\mathbf{k}\in\mathbb{N}^d}\mathbf{k}\cdot\mathbf{u}\ \nu(\mathbf{k})}, \quad \mathbf{i}\in\mathbb{N}^d, \tag{2.20}$$

where $\nu$ is the Yaglom distribution of the process $\{\mathbf{X}_n\}_n$, uniquely defined by the following property: for all $\mathbf{i},\mathbf{j}\in\mathbb{N}^d\setminus\{\mathbf{0}\}$, $\lim_{n\to\infty}\mathbb{P}(\mathbf{X}_n = \mathbf{i} \mid \mathbf{X}_0 = \mathbf{j}, \mathbf{X}_n \neq \mathbf{0}) = \nu(\mathbf{i})$. In the literature, this stationary measure for the conditioned process $\{\mathbf{X}^*_n\}_n$ is also referred to as the doubly limiting conditional probability. Moreover, by Proposition 2.2,

$$\mathbb{E}(X^*_n) = \mathbb{E}\left[\mathbb{E}\left(X^*_n \mid \mathbf{X}^*_{n-1}\right)\right] = \mathbb{E}\left[\mathbf{X}^*_{n-1}\cdot\boldsymbol{\Psi} + p\left(\mathbf{X}^*_{n-1}\right)\right] \leqslant \sum_{k=1}^{d}\mathbb{E}\left(X^*_{n-k}\right)\Psi_k + 1, \tag{2.21}$$

which implies that $\overline{\lim}_{n\to\infty}\mathbb{E}(X^*_n) \leqslant \left(1-\sum_{k=1}^{d}\Psi_k\right)^{-1} < \infty$. We consequently obtain by means of Fatou's lemma that, for every $i = 1\cdots d$,

$$\sum_{\mathbf{j}\in\mathbb{N}^d} j_i\pi(\mathbf{j}) = \mathbb{E}\left(\lim_{n\to\infty} X^*_{n,i}\right) = \mathbb{E}\left(\lim_{n\to\infty} X^*_{n-i+1}\right) = \mathbb{E}\left(\lim_{n\to\infty} X^*_n\right) \leqslant \underline{\lim}_{n\to\infty}\mathbb{E}(X^*_n) < \infty. \tag{2.22}$$

We similarly prove that $\pi$ has finite second-order moments by writing

$$\mathbb{V}\mathrm{ar}\left(X_n^* \mid \mathbf{X}_{n-1}^*\right) = \mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi} + \frac{u_1 \mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi} \sum_{k=2}^{d} X_{n-k+1}^* u_k}{\left(u_1 \mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi} + \sum_{k=2}^{d} X_{n-k+1}^* u_k\right)^2} \leqslant \mathbf{X}_{n-1}^* \cdot \boldsymbol{\Psi} + \frac{1}{4}. \tag{2.23}$$

□

Let us discuss the relevancy of approximating the epidemic process $\{\mathbf{X}_n\}_n$ conditioned on nonextinction at some finite time $k$, for $k$ large, by the $Q$-process $\{\mathbf{X}_n^*\}_n$ obtained by letting $k \to \infty$. When considering the case of late extinction, one works under an hypothetical assumption based on the unknown future, hence in practice one does not focus on a specific value $k$ for the survival of the disease in the population. We therefore might consider that $k$ is chosen large enough such that the approximation of the process $\{\mathbf{X}_n\}_n$ conditioned on the event $\{\mathbf{X}_k \neq \mathbf{0}\}$ by the process $\{\mathbf{X}_n^*\}_n$ is valid. Of course, the order of magnitude of such $k$ will depend on the rate of convergence of the conditioned process to $\{\mathbf{X}_n^*\}_n$.

**Proposition 2.5.** *Let $n_1 \leqslant \cdots \leqslant n_r \leqslant k$ and $\mathbf{i}_0, \ldots, \mathbf{i}_r \in \mathbb{N}^d \setminus \{\mathbf{0}\}$. Then the difference $|\mathbb{P}_{\mathbf{i}_0}(\mathbf{X}_{n_1} = \mathbf{i}_1, \ldots, \mathbf{X}_{n_r} = \mathbf{i}_r \mid \mathbf{X}_k \neq \mathbf{0}) - \mathbb{P}_{\mathbf{i}_0}(\mathbf{X}_{n_1}^* = \mathbf{i}_1, \ldots, \mathbf{X}_{n_r}^* = \mathbf{i}_r)|$ decreases, as $k \to \infty$, with $\max\{k^{s-1}(|\lambda|\rho^{-1})^{k/2}, \rho^{k/2}\}$, where $\lambda$ is an eigenvalue of $\mathbf{M}$ such that $\rho > |\lambda| \geqslant |\lambda_3| \geqslant |\lambda_4| \geqslant \cdots$, with the $\lambda_i$ being the other eigenvalues of $\mathbf{M}$. In case $|\lambda| = |\lambda_3|$ we stipulate that the multiplicity $s$ of $\lambda$ is at least as great as the multiplicity of $\lambda_3$.*

*Proof.* Thanks to (2.15) and to the Markov property together with the fact that $\mathbb{P}_{\mathbf{i}}(\mathbf{X}_n = \mathbf{0}) = \mathbf{f}_n(\mathbf{0})^{\mathbf{i}}$, we have

$$\begin{aligned} &\left|\mathbb{P}_{\mathbf{i}_0}(\mathbf{X}_{n_1} = \mathbf{i}_1, \ldots, \mathbf{X}_{n_r} = \mathbf{i}_r \mid \mathbf{X}_k \neq \mathbf{0}) - \mathbb{P}_{\mathbf{i}_0}\left(\mathbf{X}_{n_1}^* = \mathbf{i}_1, \ldots, \mathbf{X}_{n_r}^* = \mathbf{i}_r\right)\right| \\ &\quad = \left|\frac{1 - \mathbf{f}_{k-n_r}(\mathbf{0})^{\mathbf{i}_r}}{1 - \mathbf{f}_k(\mathbf{0})^{\mathbf{i}_0}} - \frac{1}{\rho^{n_r}} \frac{\mathbf{i}_r \cdot \mathbf{u}}{\mathbf{i}_0 \cdot \mathbf{u}}\right| \mathbb{P}_{\mathbf{i}_0}(\mathbf{X}_{n_1} = \mathbf{i}_1, \ldots, \mathbf{X}_{n_r} = \mathbf{i}_r). \end{aligned} \tag{2.24}$$

The right term of (2.24) is known to converge to 0, as $k \to \infty$, thanks to the property that $\lim_k \mathbf{a}_k = \gamma \mathbf{u}$, for some $\gamma > 0$, where $\mathbf{a}_k := \rho^{-k}(\mathbf{1} - \mathbf{f}_k(\mathbf{0}))$ (see [8]). This stems from two convergences, namely, $\lim_k b_k = \gamma$, where $b_k := \rho^{-k}(\mathbf{v} \cdot (\mathbf{1} - \mathbf{f}_k(\mathbf{0})))$, and $\lim_k \mathbf{a}_k b_k^{-1} = \mathbf{u}$. Let us write $\mathbf{a}_k = \gamma \mathbf{u} + \varepsilon_k$, where $\lim_k \varepsilon_k = \mathbf{0}$. Since $b_k = \mathbf{v} \cdot \mathbf{a}_k$ and $\mathbf{u} \cdot \mathbf{v} = 1$, it comes

$$\mathbf{a}_k b_k^{-1} = \frac{\gamma \mathbf{u} + \varepsilon_k}{\gamma + \mathbf{v} \cdot \varepsilon_k} \sim_{k \to \infty} \mathbf{u} + \gamma^{-1}[\varepsilon_k - (\mathbf{v} \cdot \varepsilon_k)\mathbf{u}]. \tag{2.25}$$

It thus appears that the rate of convergence of $\mathbf{a}_k$ to $\gamma \mathbf{u}$ is of the same order of magnitude as the one of $\mathbf{a}_k b_k^{-1}$ to $\mathbf{u}$. Let us determine this rate in an accurate way. We use the following inequality produced by Joffe and Spitzer in [8]: for each $k \geqslant n \geqslant 1$,

$$\left\|\mathbf{a}_k b_k^{-1} - \mathbf{u}\right\| = \left\|\frac{\mathbf{1} - \mathbf{f}_k(\mathbf{0})}{\mathbf{v} \cdot (\mathbf{1} - \mathbf{f}_k(\mathbf{0}))} - \mathbf{u}\right\| \leqslant \frac{2\delta_n + \sum_{j=k-n+1}^{k} \alpha_j}{1 - \delta_n - \sum_{j=k-n+1}^{k} \alpha_j}, \tag{2.26}$$

where we are going to replace $\delta_n$ and $\alpha_n$ by some explicit formulae function of $\rho$ and $n$. For this purpose, we use a detailed asymptotic behavior of $\mathbf{M}^k$, as $k \to \infty$, presented for instance in [11]; we have $\mathbf{M}^k = \rho^k \mathbf{R} + O(k^{s-1}|\lambda|^k)$, where $\mathbf{R} = \mathbf{u}^T\mathbf{v}$. For the sake of clarity the symbol $O(\cdot)$ will denote either a scalar or a matrix with all the entries satisfying the associated property. This implies the existence of some constant $a > 0$ such that, for all $k \in \mathbb{N}$, $(1-\delta_k)\mathbf{R} \leqslant \rho^{-k}\mathbf{M}^k \leqslant (1+\delta_k)\mathbf{R}$, where $\delta_k := ak^{s-1}(\rho^{-1}|\lambda|)^k$. Moreover, following [8], let us write, for all $\mathbf{r} \in [0,1]^d$, $\mathbf{1} - \mathbf{f}(\mathbf{r}) = (\mathbf{M} - \mathbf{E}(\mathbf{r}))(\mathbf{1} - \mathbf{r})$, where $\mathbf{0} \leqslant \mathbf{E}(\mathbf{r}) \leqslant \mathbf{M}$, and $\mathbf{E}(\mathbf{r}) = O(\|\mathbf{1} - \mathbf{r}\|)$ as $\mathbf{r} \to \mathbf{1}$. Then $\rho^{-1}\mathbf{E}(\mathbf{f}_{k-1}(\mathbf{0})) = \rho^{-1}O(\|\mathbf{1} - \mathbf{f}_{k-1}(\mathbf{0})\|) = O(\rho^{k-2})$, which implies the existence of some constant $b > 0$ such that, for all $k \in \mathbb{N}$, $\mathbf{0} \leqslant \rho^{-1}\mathbf{E}(\mathbf{f}_{k-1}(\mathbf{0})) \leqslant \alpha_k\mathbf{R}$, with $\alpha_k := b\rho^{k-2}$. We thus have provided an explicit formula for the sequences $(\delta_k)_k$ and $(\alpha_k)_k$ introduced by Joffe and Spitzer in [8]. Finally let us apply (2.26) to $n = \lfloor k/2 \rfloor$ and replace in this inequality $\delta_n$ and $\alpha_n$ by their explicit expressions that we got. We obtain that, for all $k \in \mathbb{N}$,

$$\left\| \frac{\mathbf{1} - \mathbf{f}_k(\mathbf{0})}{\mathbf{v} \cdot (\mathbf{1} - \mathbf{f}_k(\mathbf{0}))} - \mathbf{u} \right\| \leqslant \frac{2^{2-s}ak^{s-1}\left(|\lambda|/\rho\right)^{k/2} + \left(b/\rho(1-\rho)\right)\rho^{k/2}}{1 - 2^{1-s}ak^{s-1}\left(|\lambda|/\rho\right)^{k/2} - \left(b/\rho(1-\rho)\right)\rho^{k/2}}. \tag{2.27}$$

Consequently, the right member of (2.24) will decrease with $\max\{k^{s-1}(|\lambda|\rho^{-1})^{k/2}, b(\rho(1-\rho))^{-1}\rho^{k/2}\}$, as $k \to \infty$. □

Hence the concept of the $Q$-process will have most practical relevance to approximate the very late extinction case if $\rho$ is near to zero and if $|\lambda|$ is small compared with $\rho$. Note however that the very late extinction scenario is more likely to happen if $\rho$ is near to unity because the time to extinction in most realizations will then be long (see Section 2.1).

### 2.5. Estimation of the Infection Parameter

We assume for this subsection that the parameters $\Psi_k$ of the epidemic model (1.1)–(2.1) are not entirely known. More precisely, we assume that the $\Psi_k$ are of the form, for all $k = 1 \cdots d$,

$$\Psi_k(\theta_0) = a_k\theta_0 + b_k, \tag{2.28}$$

where $a_k > 0$ and $b_k \geqslant 0$ are constants, and $\theta_0$ is an unknown real parameter. We will write in what follows $\mathbf{\Psi}(\theta_0) = \mathbf{a}\theta_0 + \mathbf{b}$, where $\mathbf{\Psi}(\theta_0) := (\Psi_1(\theta_0), \ldots, \Psi_d(\theta_0))$ and so forth. This general assumption corresponds in particular to the case where the $\Psi_k$ are of the form (1.2).

We estimate $\theta_0$ by the following WCLSE in model (1.1)–(2.1). This estimator generalizes the well-known Harris estimator [12] in a BGW process. Let $\Theta := ]\theta_1, \theta_2[$, $\theta_2 > \theta_1 > 0$, such that $\theta_0 \in \Theta$. The WCLSE is based on the normalized process $Y_n := X_n/\sqrt{\mathbf{a} \cdot \mathbf{X}_{n-1}}$ and is defined by

$$\widehat{\theta}_{|\mathbf{X}_0|} := \arg\min_{\theta\in\Theta} \sum_{k=1}^{n} (Y_k - \mathbb{E}_\theta(Y_k \mid \mathbf{X}_{k-1}))^2 = \arg\min_{\theta\in\Theta} \sum_{k=1}^{n} \frac{(X_k - \mathbf{\Psi}(\theta) \cdot \mathbf{X}_{k-1})^2}{\mathbf{a} \cdot \mathbf{X}_{k-1}}. \tag{2.29}$$

We easily derive the following explicit form:

$$\widehat{\theta}_{|\mathbf{X}_0|} = \frac{\sum_{k=1}^{n}(X_k - \mathbf{b} \cdot \mathbf{X}_{k-1})}{\sum_{k=1}^{n} \mathbf{a} \cdot \mathbf{X}_{k-1}}. \tag{2.30}$$

The normalization of the process $X_n$ by $\sqrt{\mathbf{a}\cdot\mathbf{X}_{n-1}}$ appears to be the most natural and suitable for the following reasons. First, this normalization generalizes the normalization $X_n/\sqrt{aX_{n-1}}$ in the monotype case, which is the one leading to the Harris estimator $\widehat{m}$ of $m_0 = a\theta_0 + b$ since we have, for $d = 1$, $a\widehat{\theta}_{X_0} + b = \widehat{m}$. It also corresponds, in the linear case $\mathbf{b} = \mathbf{0}$, to the maximum likelihood estimator of $\theta_0$. In addition, defining for any vector $\mathbf{x}$, $\underline{\mathbf{x}} := \min_i x_i$ and $\overline{\mathbf{x}} := \max_i x_i$, we have

$$\theta_0 + \frac{\underline{\mathbf{b}}}{\overline{\mathbf{a}}} \leqslant \mathbb{E}_{\theta_0}\left((Y_k - \mathbb{E}_{\theta_0}(Y_k \mid \mathbf{X}_{k-1}))^2 \mid \mathbf{X}_{k-1}\right) = \theta_0 + \frac{\mathbf{b}\cdot\mathbf{X}_{k-1}}{\mathbf{a}\cdot\mathbf{X}_{k-1}} \leqslant \theta_0 + \frac{\overline{\mathbf{b}}}{\underline{\mathbf{a}}}, \tag{2.31}$$

hence the conditional variance of the error term $Y_k - \mathbb{E}_{\theta_0}(Y_k \mid \mathbf{X}_{k-1})$ in the stochastic regression equation $Y_k = \mathbb{E}_{\theta_0}(Y_k \mid \mathbf{X}_{k-1}) + Y_k - \mathbb{E}_{\theta_0}(Y_k \mid \mathbf{X}_{k-1})$ is invariant under multiplication of the whole process, and bounded respectively to $\{\mathbf{X}_n\}_n$, leading to the quasi-optimality of $\widehat{\theta}_{|\mathbf{X}_0|}$ at finite $|\mathbf{X}_0|$ and $n$, in the sense of [13].

Let us provide asymptotic results for the estimator $\widehat{\theta}_{|\mathbf{X}_0|}$ defined by (2.30), as the initial population size $|\mathbf{X}_0| = X_0 + X_{-1} + \cdots + X_{-d+1}$ tends to infinity. We denote by $m_{ij}^{(k)}(\theta)$ the $(i,j)$th entry in the $k$th power of the matrix $\mathbf{M}(\theta)$ given by (2.3).

**Theorem 2.6.** *Let us assume that, for each $i = 1\cdots d$, there exists some $\alpha_i \in [0,1]$ such that $\lim_{|\mathbf{X}_0|\to\infty} X_{0,i}|\mathbf{X}_0|^{-1} = \alpha_i$. Then $\widehat{\theta}_{|\mathbf{X}_0|}$ is strongly consistent, that is, $\lim_{|\mathbf{X}_0|\to\infty}\widehat{\theta}_{|\mathbf{X}_0|} \overset{a.s.}{=} \theta_0$, and is asymptotically normally distributed:*

$$\lim_{|\mathbf{X}_0|\to\infty}\sqrt{\frac{\sum_{k=1}^{n}\mathbf{a}\cdot\mathbf{X}_{k-1}}{\sigma^2\left(\widehat{\theta}_{|\mathbf{X}_0|}\right)}}\left(\widehat{\theta}_{|\mathbf{X}_0|} - \theta_0\right) \overset{\mathfrak{D}}{=} \mathcal{N}(0,1), \tag{2.32}$$

*where*

$$\sigma^2(\theta) := \theta + \frac{\sum_{k=1}^{n}\sum_{j=1}^{d}\sum_{i=1}^{d}\alpha_j b_i m_{ji}^{(k-1)}(\theta)}{\sum_{k=1}^{n}\sum_{j=1}^{d}\sum_{i=1}^{d}\alpha_j b_i m_{ji}^{(k-1)}(\theta)}. \tag{2.33}$$

*Proof.* Let us first prove that, for each $k = 1\cdots n$ and each $i = 1\cdots d$,

$$\lim_{|\mathbf{X}_0|\to\infty}\frac{X_{k,i}}{|\mathbf{X}_0|} \overset{\text{a.s.}}{=} \sum_{j=1}^{d}\alpha_j m_{ji}^{(k)}(\theta_0). \tag{2.34}$$

Using the branching property of the process $\{\mathbf{X}_n\}_n$ ([6]) we write

$$X_{k,i} = \sum_{j=1}^{X_{0,1}} X_{k,i,j}^{(1)} + \cdots + \sum_{j=1}^{X_{0,d}} X_{k,i,j}^{(d)}, \tag{2.35}$$

where, for all $l = 1\cdots d$ and $j = 1\cdots X_{0,l}$, $X_{k,i,j}^{(l)}$ is the $i$th coordinate of a $d$-type branching process at time $k$ initialized by a single particle of type $l$. For $k$, $i$, and $l$ fixed the random

variables $\{X_{k,i,j}^{(l)}\}_j$ are i.i.d. with mean value $m_{li}^{(k)}(\theta_0)$. According to the strong law of large numbers and under the theorem assumption, we have, for every $l = 1 \cdots d$ such that $X_{0,l} \neq 0$,

$$\lim_{|\mathbf{X}_0| \to \infty} \frac{\sum_{j=1}^{X_{0,l}} X_{k,i,j}^{(l)}}{X_{0,l}} \overset{\text{a.s.}}{=} m_{li}^{(k)}(\theta_0), \tag{2.36}$$

which together with the theorem assumption leads to (2.34).

To prove the consistency of $\widehat{\theta}_{|\mathbf{X}_0|}$ we apply (2.34) to (2.30), using the fact that $X_k = X_{k,1}$ and $X_{k-i} = X_{k-1,i}$, and obtain

$$\lim_{|\mathbf{X}_0| \to \infty} \widehat{\theta}_{|\mathbf{X}_0|} \overset{\text{a.s.}}{=} \frac{\sum_{k=1}^{n} \sum_{j=1}^{d} \alpha_j \left( m_{j1}^{(k)}(\theta_0) - \sum_{i=1}^{d} b_i m_{ji}^{(k-1)}(\theta_0) \right)}{\sum_{k=1}^{n} \sum_{i=1}^{d} \sum_{j=1}^{d} a_i \alpha_j m_{ji}^{(k-1)}(\theta_0)}. \tag{2.37}$$

By definition,

$$m_{j1}^{(k)}(\theta_0) = \sum_{i=1}^{d} m_{ji}^{(k-1)}(\theta_0) m_{i1}(\theta_0) = \sum_{i=1}^{d} m_{ji}^{(k-1)}(\theta_0)(a_i\theta_0 + b_i), \tag{2.38}$$

hence (2.37) immediately leads to the strong consistency.

We are now interested in the asymptotic distribution of $\widehat{\theta}_{|\mathbf{X}_0|} - \theta_0$. We derive from (2.30) that

$$\sqrt{\sum_{k=1}^{n} \mathbf{a} \cdot \mathbf{X}_{k-1}} \left( \widehat{\theta}_{|\mathbf{X}_0|} - \theta_0 \right) = \frac{\sum_{k=1}^{n} (X_k - \boldsymbol{\Psi}(\theta_0) \cdot \mathbf{X}_{k-1})}{\sqrt{\sum_{k=1}^{n} \mathbf{a} \cdot \mathbf{X}_{k-1}}}. \tag{2.39}$$

By (1.1),

$$X_k - \boldsymbol{\Psi}(\theta_0) \cdot \mathbf{X}_{k-1} = \sum_{i=1}^{d} \sum_{j=1}^{X_{k-i}} \left( Y_{k-i,k,j} - \Psi_i(\theta_0) \right) =: \sum_{i=1}^{d} \sum_{j=1}^{X_{k-i}} \overset{\circ}{Y}_{k-i,k,j}, \tag{2.40}$$

where the $\{Y_{k-i,k,j}\}_j$ are i.i.d. given $\mathcal{F}_{k-1}$, following a Poisson distribution with parameter $\Psi_i(\theta_0)$, and the $\{Y_{k-i,k,j}\}_{i,j}$ are independent given $\mathcal{F}_{k-1}$. Renumbering the $\overset{\circ}{Y}_{k-i,k,j}$ we then obtain

$$\sum_{k=1}^{n} (X_k - \boldsymbol{\Psi}(\theta_0) \cdot \mathbf{X}_{k-1}) = \sum_{i=1}^{d} \sum_{j=1}^{\sum_{k=1}^{n} X_{k-i}} \overset{\circ}{Y}_{k-i,k,j}. \tag{2.41}$$

Applying a central limit theorem for the sum of a random number of independent random variables (see, e.g., [14]), we obtain that, for all $i = 1 \cdots d$,

$$\lim_{|\mathbf{X}_0|\to\infty} \frac{\sum_{j=1}^{\sum_{k=1}^{n} X_{k-i}} \mathring{Y}_{k-i,k,j}}{\sqrt{\sum_{k=1}^{n} X_{k-i}}} \overset{\mathcal{D}}{=} \mathcal{N}(0, a_i\theta_0 + b_i). \tag{2.42}$$

We have used the fact that $|\mathbf{X}_0|$ is a real positive sequence growing to infinity, and $\sum_{k=1}^{n} X_{k-i}$ is a sequence of integer-valued random variables such that $\sum_{k=1}^{n} X_{k-i}/|\mathbf{X}_0|$ converges in probability to a finite random variable. In our case the limit is actually deterministic, since we have shown in (2.34) that

$$\lim_{|\mathbf{X}_0|\to\infty} \frac{\sum_{k=1}^{n} X_{k-i}}{|\mathbf{X}_0|} \overset{\text{a.s.}}{=} \sum_{k=1}^{n}\sum_{j=1}^{d} \alpha_j m_{ji}^{(k-1)}(\theta_0). \tag{2.43}$$

Using (2.41) in (2.39), we write

$$\sqrt{\sum_{k=1}^{n} \mathbf{a}\cdot\mathbf{X}_{k-1}}\left(\widehat{\theta}_{|\mathbf{X}_0|} - \theta_0\right) = \sum_{i=1}^{d} \frac{\sum_{j=1}^{\sum_{k=1}^{n} X_{k-i}} \mathring{\xi}_{k-i,k,j}}{\sqrt{\sum_{k=1}^{n} X_{k-i}}} \frac{\sqrt{\sum_{k=1}^{n} X_{k-i}}}{\sqrt{\sum_{k=1}^{n} \mathbf{a}\cdot\mathbf{X}_{k-1}}}. \tag{2.44}$$

Using again (2.34),

$$\lim_{|\mathbf{X}_0|\to\infty} \frac{\sqrt{\sum_{k=1}^{n} \mathbf{a}\cdot\mathbf{X}_{k-i}}}{\sqrt{\sum_{k=1}^{n} \mathbf{a}\cdot\mathbf{X}_{k-1}}} \overset{\text{a.s.}}{=} \sqrt{\frac{\sum_{k=1}^{n}\sum_{j=1}^{d} \alpha_j m_{ji}^{(k-1)}(\theta_0)}{\sum_{k=1}^{n}\sum_{j=1}^{d}\sum_{l=1}^{d} \alpha_j a_l m_{jl}^{(k-1)}(\theta_0)}}, \tag{2.45}$$

which, combined to (2.42) and (2.44), implies by Slutsky's theorem that

$$\lim_{|\mathbf{X}_0|\to\infty} \sqrt{\sum_{k=1}^{n} \mathbf{a}\cdot\mathbf{X}_{k-1}}\left(\widehat{\theta}_{|\mathbf{X}_0|} - \theta_0\right) \overset{\mathcal{D}}{=} \mathcal{N}\left(0, \sigma^2(\theta_0)\right). \tag{2.46}$$

By (2.33) and the strong consistency, $\lim_{|\mathbf{X}_0|} \sqrt{\sigma^2(\theta_0)}/\sqrt{\sigma^2(\widehat{\theta}_{|\mathbf{X}_0|})} \overset{\text{a.s.}}{=} 1$, from which we finally deduce (2.32). □

### 2.6. Estimation of the Infection Parameter in the Worst-Case Scenario

In order to make predictions of the evolution of the epidemic in case of a very late extinction, that is, in order to make predictions of the behavior of the conditioned process $\{X_n^*\}_n$ introduced in Section 2.4, we need to estimate the parameter $\theta_0$ in the setting of this conditioned process. We point out that $\theta_0$ does not play the same role in the conditioned process $\{X_n^*\}_n$ and in the unconditioned process $\{X_n\}_n$, since, as shown in Proposition 2.2, this parameter interferes not only in the Poisson random variable but also in the Bernoulli

one. It would thus be irrelevant to estimate $\theta_0$ with an estimator aimed for the unconditioned process, such as $\widehat{\theta}_{|\mathbf{X}_0|}$. Let us notice that, according to (2.16), the process $\{\mathbf{X}_n^*\}_n$ could be written as a multitype branching process with state dependent immigration. Because of this state-dependence, and since the parameter $\theta_0$ acts in a nonlinear way in the immigration, the methods developed in estimation theory for branching processes with immigration (see, e.g., [15]) cannot be directly applied here. Similarly as in Section 2.5 we consider the WCLSE based on the process $Y_n^* := X_n^*/\sqrt{\mathbf{a}\cdot\mathbf{X}_{n-1}^*}$, namely,

$$\widehat{\theta}_n^* := \arg\min_{\theta\in\Theta} S_n(\theta), \qquad S_n(\theta) := \sum_{k=1}^{n}\left(Y_k^* - f\left(\theta, \mathbf{X}_{k-1}^*\right)\right)^2, \tag{2.47}$$

where $\Theta$ is defined in Section 2.5, and where

$$f\left(\theta_0, \mathbf{X}_{k-1}^*\right) := \mathbb{E}_{\theta_0}\left(Y_k^* \mid \mathbf{X}_{k-1}^*\right) = \frac{\mathbf{X}_{k-1}^*\cdot\boldsymbol{\Psi}(\theta_0) + p\left(\theta_0, \mathbf{X}_{k-1}^*\right)}{\sqrt{\mathbf{a}\cdot\mathbf{X}_{k-1}^*}}. \tag{2.48}$$

Let $\varepsilon_k^* := Y_k^* - f(\theta_0, \mathbf{X}_{k-1}^*)$ be the error term between the normalized process and its conditional expectation. We obtain that

$$g\left(\theta_0, \mathbf{X}_{k-1}^*\right) := \mathbb{E}_{\theta_0}\left(\left(\varepsilon_k^*\right)^2 \mid \mathbf{X}_{k-1}^*\right) = \frac{\mathbf{X}_{k-1}^*\cdot\boldsymbol{\Psi}(\theta_0) + p\left(\theta_0, \mathbf{X}_{k-1}^*\right)\left(1 - p\left(\theta_0, \mathbf{X}_{k-1}^*\right)\right)}{\mathbf{a}\cdot\mathbf{X}_{k-1}^*}, \tag{2.49}$$

which implies

$$\theta_0 + \frac{\underline{\mathbf{b}}}{\overline{\mathbf{a}}} \leqslant \mathbb{E}_{\theta_0}\left(\left(\varepsilon_k^*\right)^2 \mid \mathbf{X}_{k-1}^*\right) \leqslant \theta_0 + \frac{\overline{\mathbf{b}} + 1}{\underline{\mathbf{a}}}. \tag{2.50}$$

In what follows, we denote by $f'$ the derivative of $f$ with respect to $\theta$, and similarly for the other quantities depending on $\theta$.

**Theorem 2.7.** *The estimator $\widehat{\theta}_n^*$ is strongly consistent, that is, $\lim_{n\to\infty}\widehat{\theta}_n^* \overset{a.s.}{=} \theta_0$, and has the following asymptotic distribution:*

$$\lim_{n\to\infty} \frac{\sum_{k=0}^{n} f'\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)^2}{\sqrt{\sum_{k=0}^{n} f'\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)^2 g\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)}}\left(\widehat{\theta}_n^* - \theta_0\right) \overset{\mathfrak{D}}{=} \mathcal{N}(0,1). \tag{2.51}$$

*Remark 2.8.* Note that (2.51) involves the function $f'$ and thus requires the knowledge of the derivative of the function $u_j$ given in (2.5), which is not an explicit function of $\theta$ since $\rho$ is not either. However, $\rho'$ satisfies $\rho' = \sum_{k=1}^{d} a_k\rho^{-k}[\sum_{k=1}^{d} k(a_k\theta + b_k)\rho^{-k-1}]^{-1}$; hence $u_j'$ is known as soon as $\rho$ can be computed. Consequently, denoting $\rho$ by $\rho(\theta)$ when $\theta$ is the parameter of the model, (2.51) can be used as soon as $\rho(\widehat{\theta}_n^{*\mathrm{obs}})$ is known; for this purpose one can for instance numerically approximate the largest solution $\rho$ of $\sum_{k=1}^{d}\Psi_k(\widehat{\theta}_n^{*\mathrm{obs}})\rho^{-k} = 1$.

*Proof.* The proof heavily relies on a strong law of large numbers for homogeneous irreducible positive recurrent Markov chains applied to the conditioned process $\{\mathbf{X}_n^*\}_n$ and its stationary distribution $\pi_{\theta_0}$ (given by Proposition 2.4), which states [16]: for every $\pi_{\theta_0}$-integrable function $h : \mathbb{N}^d \setminus \{\mathbf{0}\} \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} h(\mathbf{X}_k^*) \overset{\text{a.s.}}{=} \sum_{\mathbf{j}\in\mathbb{N}^d} h(\mathbf{j})\pi_{\theta_0}(\mathbf{j}). \tag{2.52}$$

Note that the Perron's root $\rho(\theta)$ of $\mathbf{M}(\theta)$ and the associated right normalized eigenvector $\mathbf{u}(\theta)$ are $\mathcal{C}^\infty$-functions of $\theta$.

In order to prove the strong consistency of $\widehat{\theta}_n^*$ as $n \to \infty$, since $f''(\cdot, \mathbf{X}_{k-1}^*)$ is not linear, we cannot use the general standard method based on the first-order expansion of $S_n'(\widehat{\theta}_n^*)$ at $\theta_0$, and on the strong law of large numbers for martingales applied to $S_n'(\theta_0)$ correctly normalized. We consequently use the following conditions given in [17]:

(i) $f(., \mathbf{X}_{k-1}^*)$ is Lipschitz on $\Theta$; that is, there exists a nonnegative $\mathcal{F}_{k-1}^*$-measurable function $A_k$ (where $\mathcal{F}_{k-1}^* := \sigma(\mathbf{X}_0^*, \ldots, \mathbf{X}_{k-1}^*)$), satisfying, for all $\delta_1$, $\delta_2 \in \Theta$, $|f(\delta_1, \mathbf{X}_{k-1}^*) - f(\delta_2, \mathbf{X}_{k-1}^*)| \leqslant A_k|\delta_1 - \delta_2|$ a.s.,

(ii) $\overline{\lim}_{k\to\infty} \mathbb{E}_{\theta_0}((\varepsilon_k^*)^2 \mid \mathbf{X}_{k-1}^*) \overset{\text{a.s.}}{<} \infty$,

(iii) $\lim_{n\to\infty} \inf_{\substack{\theta\in\Theta \\ |\theta-\theta_0|\geqslant\delta}} \sum_{k=1}^n (f(\theta_0, \mathbf{X}_{k-1}^*) - f(\theta, \mathbf{X}_{k-1}^*))^2 \overset{\text{a.s.}}{=} \infty$.

Let us note that, in the frame of a general model $f$, since $\overline{\Theta}$ is compact, (i) is satisfied as soon as $f(\theta, \mathbf{X}_{k-1}^*)$ has a first derivative in $\theta$ with $\sup_{\theta\in\Theta} f'(\theta, \mathbf{X}_{k-1}^*) < \infty$, (ii) is satisfied for any optimal estimator in the sense of [13] and moreover could be weakened (see [17]), and (iii) is a necessary condition. First, for all $\theta \in \Theta$ and $\mathbf{j} \in \mathbb{N}^d$, $\mathbf{j} \neq \mathbf{0}$, $f'(\theta, \mathbf{j}) = (\mathbf{a}\cdot\mathbf{j} + p'(\theta, \mathbf{j}))(\mathbf{a}\cdot\mathbf{j})^{-1/2}$, where $p'$ denotes the derivative of $p$ with respect to $\theta$, which thanks to (2.17) is bounded on $\Theta$. Condition (i) is thus satisfied. Condition (ii) follows from (2.50). Condition (iii) comes from the fact that, for every $\delta > 0$ and every $\theta \in \Theta$ such that $|\theta - \theta_0| \geqslant \delta$, applying the mean value theorem to the $\mathcal{C}^1$-function $p(\cdot, \mathbf{X}_{k-1}^*)$,

$$\begin{aligned}\sum_{k=1}^n \left(f(\theta_0, \mathbf{X}_{k-1}^*) - f(\theta, \mathbf{X}_{k-1}^*)\right)^2 &= (\theta_0 - \theta)^2 \sum_{k=1}^n \mathbf{a}\cdot\mathbf{X}_{k-1}^* \left(1 + \frac{p(\theta_0, \mathbf{X}_{k-1}^*) - p(\theta, \mathbf{X}_{k-1}^*)}{(\theta_0 - \theta)\mathbf{a}\cdot\mathbf{X}_{k-1}^*}\right)^2 \\ &\geqslant \delta^2 \sum_{k=1}^n \mathbf{a}\cdot\mathbf{X}_{k-1}^* \inf_{\theta\in\overline{\Theta}} \left(1 + \frac{p'(\theta, \mathbf{X}_{k-1}^*)}{\mathbf{a}\cdot\mathbf{X}_{k-1}^*}\right)^2.\end{aligned} \tag{2.53}$$

Let us show that the function $\mathbf{j} \mapsto \mathbf{a}\cdot\mathbf{j}\inf_{\theta\in\overline{\Theta}}(1+p'(\theta,\mathbf{j})/\mathbf{a}\cdot\mathbf{j})^2$ is $\pi_{\theta_0}$-integrable. For every $\theta\in\overline{\Theta}$, $\mathbf{j}\in\mathbb{N}^d$ and $\mathbf{j}\neq\mathbf{0}$, denoting $\Xi := \sup_{i,\theta\in\overline{\Theta}}\{u_i(\theta), \mid u_i'(\theta)\mid\} < \infty$ (by continuity of $u_i'(\theta)$ on $\Theta_\gamma \supset \overline{\Theta}$), and $\underline{\mathbf{u}} := \min_{i,\theta\in\overline{\Theta}} u_i(\theta) > 0$,

$$\begin{aligned}\left|p^{'(\theta,\mathbf{j})}\right| &= \left|\frac{\left(u_i'(\theta)\mathbf{j}\cdot\boldsymbol{\Psi}(\theta)+u_1(\theta)\mathbf{j}\cdot\mathbf{a}\right)\sum_{i=2}^d j_i u_i(\theta) - u_1(\theta)\mathbf{j}\cdot\boldsymbol{\Psi}(\theta)\sum_{i=2}^d j_i u_i'(\theta)}{\left(u_1(\theta)\mathbf{j}\cdot\boldsymbol{\Psi}(\theta)+\sum_{i=2}^d j_i u_i(\theta)\right)^2}\right| \\ &\leqslant \frac{\Xi^2}{\underline{\mathbf{u}}^2}\frac{3\mathbf{j}\cdot\boldsymbol{\Psi}(\theta)\sum_{i=2}^d j_i}{\left(\mathbf{j}\cdot\boldsymbol{\Psi}(\theta)+\sum_{i=2}^d j_i\right)^2} \leqslant \frac{3\Xi^2}{4\underline{\mathbf{u}}^2} =: C_1.\end{aligned} \tag{2.54}$$

Hence, for all $\mathbf{j}\neq\mathbf{0}$, $\mid \mathbf{a}\cdot\mathbf{j}\inf_{\theta\in\overline{\Theta}}(1+p'(\theta,\mathbf{j})/\mathbf{a}\cdot\mathbf{j})^2 \mid \leqslant (1+C_1/\underline{\mathbf{a}})^2\mathbf{a}\cdot\mathbf{j}$, and applying (2.52) together with (2.22) we obtain that

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=1}^n \mathbf{a}\cdot\mathbf{X}_{k-1}^*\inf_{\theta\in\overline{\Theta}}\left(1+\frac{p'(\theta,\mathbf{X}_{k-1}^*)}{\mathbf{a}\cdot\mathbf{X}_{k-1}^*}\right)^2 \overset{\text{a.s.}}{=} \sum_{\mathbf{j}\in\mathbb{N}^d}\mathbf{a}\cdot\mathbf{j}\inf_{\theta\in\overline{\Theta}}\left(1+\frac{p'(\theta,\mathbf{j})}{\mathbf{a}\cdot\mathbf{j}}\right)^2\pi_{\theta_0}(\mathbf{j}). \tag{2.55}$$

Let $\mathbf{j}\neq\mathbf{0}$ fixed. Since, for all $\theta\in\overline{\Theta}$, $p'(\theta,\mathbf{j})\neq -\mathbf{a}\cdot\mathbf{j}$, the extreme value theorem implies that $\inf_{\theta\in\overline{\Theta}}(1+p'(\theta,\mathbf{j})/\mathbf{a}\cdot\mathbf{j})^2 > 0$. Hence the right term in (2.55) is strictly positive, which together with (2.55) leads to (iii).

Let us now consider the asymptotic distribution of $\widehat{\theta}_n^* - \theta_0$. For this purpose, we follow the steps of the proof of Proposition 6.1 in [17]. Writing the Taylor expansion of $S_n'(\widehat{\theta}_n^*)$ at $\theta_0$ we obtain that $\widehat{\theta}_n^* - \theta_0 = -S_n'(\theta_0)/S_n''(\widetilde{\theta}_n)$, for some $\widetilde{\theta}_n = \theta_0 + t_n(\widehat{\theta}_n^* - \theta_0)$, with $t_n\in\,]0,1[$. Since $S_n'(\theta_0) = -2\sum_{k=1}^n \varepsilon_k^* f'(\theta_0,\mathbf{X}_{k-1}^*)$, we can write

$$\sqrt{n}\left(\widehat{\theta}_n^* - \theta_0\right) = \frac{\sum_{k=1}^n \varepsilon_k^* f'(\theta_0,\mathbf{X}_{k-1}^*)}{\sqrt{n}}\left(\frac{F_n}{n}\right)^{-1}\left(\frac{1}{2}\frac{S_n''\left(\widetilde{\theta}_n\right)}{F_n}\right)^{-1}, \tag{2.56}$$

where $F_n := \sum_{k=1}^n \left(f'(\theta_0,\mathbf{X}_{k-1}^*)\right)^2$. Let us first show that

$$\lim_{n\to\infty}\frac{F_n}{n} \overset{\text{a.s.}}{=} \sum_{\mathbf{j}\in\mathbb{N}^d}\left(f'(\theta_0,\mathbf{j})\right)^2\pi_{\theta_0}(\mathbf{j}). \tag{2.57}$$

This is an application of (2.52) and (2.22), since, for all $\mathbf{j}\in\mathbb{N}^d$, $\mathbf{j}\neq\mathbf{0}$,

$$\left(f'(\theta_0,\mathbf{j})\right)^2 \leqslant \mathbf{a}\cdot\mathbf{j} + 2C_1 + \frac{C_1^2}{\underline{\mathbf{a}}}. \tag{2.58}$$

In view of (2.56), we now prove that

$$\lim_{n\to\infty} \frac{S_n''\left(\widetilde{\theta}_n\right)}{F_n} \overset{\text{a.s.}}{=} 2. \tag{2.59}$$

Computing $S_n''$ thanks to the formula $S_n(\theta) = \sum_{k=1}^n (\varepsilon_k^* + f(\theta_0, \mathbf{X}_{k-1}^*) - f(\theta, \mathbf{X}_{k-1}^*))^2$, it appears that (2.59) is true, as soon as the following holds:

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{\left|\sum_{k=1}^n \varepsilon_k^* f''(\theta, \mathbf{X}_{k-1}^*)\right|}{F_n} \overset{\text{a.s.}}{=} 0, \tag{2.60}$$

$$\lim_{n\to\infty} \frac{\sum_{k=1}^n \left(f'\left(\widetilde{\theta}_n, \mathbf{X}_{k-1}^*\right)\right)^2}{F_n} \overset{\text{a.s.}}{=} 1, \tag{2.61}$$

$$\lim_{n\to\infty} \frac{\sum_{k=1}^n \left(f(\theta_0, \mathbf{X}_{k-1}^*) - f\left(\widetilde{\theta}_n, \mathbf{X}_{k-1}^*\right)\right) f''\left(\widetilde{\theta}_n, \mathbf{X}_{k-1}^*\right)}{F_n} \overset{\text{a.s.}}{=} 0. \tag{2.62}$$

Let us prove (2.60)–(2.62). First, (2.60) is given by a strong law of large numbers proved in [17, Proposition 5.1]. The latter can be indeed applied since $\lim_n F_n \overset{\text{a.s.}}{=} \infty$ (as an immediate consequence of the stronger result (2.57)), and since $f''(\cdot, \mathbf{X}_{k-1}^*)$ fulfills the required Lipschitz condition. Indeed we proved earlier that $u_i'''(\theta)$ is continuous on the compact set $\overline{\Theta}$ and is thus bounded on $\overline{\Theta}$, which implies that $f'''(\cdot, \mathbf{X}_{k-1}^*) = p'''(\cdot, \mathbf{X}_{k-1}^*)(\mathbf{a}\cdot\mathbf{X}_{k-1}^*)^{-1/2}$ is bounded by a $\mathcal{F}_{k-1}^*$-measurable function. In view of (2.61), we consider the function $f(\theta, \mathbf{j})^2$ and its derivative $2f'(\theta, \mathbf{j})f''(\theta, \mathbf{j})$. Similarly as for (2.54), one can show that there exists a constant $C_2 > 0$ such that for all $\theta \in \overline{\Theta}$, and all $\mathbf{j} \neq \mathbf{0}$, $|\, p''(\theta, \mathbf{j})\,| \leqslant C_2$. This implies

$$\left|2f'(\theta, \mathbf{j})f''(\theta, \mathbf{j})\right| \leqslant 2\left|\frac{(\mathbf{a}\cdot\mathbf{j} + p'(\theta, \mathbf{j}))p''(\theta, \mathbf{j})}{\mathbf{a}\cdot\mathbf{j}}\right| \leqslant 2C_2\left(1 + \frac{C_1}{\underline{\mathbf{a}}}\right). \tag{2.63}$$

Consequently,

$$\frac{\left|\sum_{k=1}^n \left(f'\left(\widetilde{\theta}_n, \mathbf{X}_{k-1}^*\right)\right)^2 - \left(f'(\theta_0, \mathbf{X}_{k-1}^*)\right)^2\right|}{F_n} \leqslant 2C_2\left(1 + \frac{C_1}{\underline{\mathbf{a}}}\right)\left|\widehat{\theta}_n^* - \theta_0\right|\left(\frac{F_n}{n}\right)^{-1}, \tag{2.64}$$

which by (2.57) and the strong consistency of $\widehat{\theta}_n^*$ almost surely tends to 0. Writing

$$\frac{\sum_{k=1}^n f\left(\widetilde{\theta}_n, \mathbf{X}_{k-1}^*\right)^2}{F_n} = 1 + \frac{\sum_{k=1}^n \left(\left(f'\left(\widetilde{\theta}_n, \mathbf{X}_{k-1}^*\right)\right)^2 - \left(f'(\theta_0, \mathbf{X}_{k-1}^*)\right)^2\right)}{F_n}, \tag{2.65}$$

this implies (2.61). It now remains to prove (2.62). We write

$$\begin{aligned}
&\frac{\left|\sum_{k=1}^{n}\left(f(\theta_0,\mathbf{X}_{k-1}^{*})-f\left(\widetilde{\theta}_n,\mathbf{X}_{k-1}^{*}\right)\right)f''\left(\widetilde{\theta}_n,\mathbf{X}_{k-1}^{*}\right)\right|}{F_n} \\
&\quad\leqslant \frac{1}{F_n}\sum_{k=1}^{n}\frac{\mathbf{a}\cdot\mathbf{X}_{k-1}^{*}\left|\theta_0-\widetilde{\theta}_n\right|+\left|p(\theta_0,\mathbf{X}_{k-1}^{*})-p\left(\widetilde{\theta}_n,\mathbf{X}_{k-1}^{*}\right)\right|}{\mathbf{a}\cdot\mathbf{X}_{k-1}^{*}}\left|p''\left(\widetilde{\theta}_n,\mathbf{X}_{k-1}^{*}\right)\right| \\
&\quad\leqslant \frac{1}{F_n}\sum_{k=1}^{n}\left(\left|\theta_0-\widetilde{\theta}_n\right|+\frac{C_1\left|\theta_0-\widetilde{\theta}_n\right|}{\underline{a}}\right)C_2\leqslant\left|\theta_0-\widehat{\theta}_n^{*}\right|C_2\left(1+\frac{C_1}{\underline{a}}\right)\left(\frac{F_n}{n}\right)^{-1},
\end{aligned}\tag{2.66}$$

which thanks to (2.57) and the strong consistency of $\widehat{\theta}_n^{*}$ implies (2.62). In view of (2.56), we finally want to prove that $\sum_{k=1}^{n}\varepsilon_k^{*}f'(\theta_0,\mathbf{X}_{k-1}^{*})/\sqrt{n}$ converges in distribution. For this purpose we make use of a central limit theorem for martingale difference arrays [18, 19]; if $\{M_k^{(n)},\mathcal{F}_k^{(n)},1\leqslant k\leqslant n\}$, $n\geqslant 1$ is a sequence of square integrable martingales with associated Meyer process $\langle M\rangle^{(n)}=(\langle M\rangle_k^{(n)})_{1\leqslant k\leqslant n}$ satisfying $\lim_{n\to\infty}\langle M_n\rangle^{(n)}\overset{P}{=}c^2$ for some constant $c$, and such that, for all $\varepsilon>0$,

$$\lim_{n\to\infty}\sum_{k=1}^{n}\mathbb{E}\left[\left|M_k^{(n)}-M_{k-1}^{(n)}\right|^2\mathbf{1}_{\{|M_k^{(n)}-M_{k-1}^{(n)}|\geqslant\varepsilon\}}\mid\mathcal{F}_{k-1}^{(n)}\right]\overset{P}{=}0,\tag{2.67}$$

then $\lim_{n\to\infty}M_n^{(n)}\overset{\mathcal{D}}{=}\mathcal{N}(0,c^2)$. Let us define $M_k^{(n)}:=\sum_{l=1}^{k}\varepsilon_l^{*}f'(\theta_0,\mathbf{X}_{l-1}^{*})/\sqrt{n}$, for every $k\leqslant n$. First, for any $k\leqslant n$, $\mathbb{E}_{\theta_0}(\varepsilon_k^{*}f'(\theta_0,\mathbf{X}_{k-1}^{*})/\sqrt{n}\mid\mathbf{X}_{k-1}^{*})=0$. Second,

$$\mathbb{E}_{\theta_0}\left(\left(\frac{\varepsilon_k^{*}f'(\theta_0,\mathbf{X}_{k-1}^{*})}{\sqrt{n}}\right)^2\mid\mathbf{X}_{k-1}^{*}\right)=\frac{\left(f'(\theta_0,\mathbf{X}_{k-1}^{*})\right)^2g(\theta_0,\mathbf{X}_{k-1}^{*})}{n},\tag{2.68}$$

hence $\{M_k^{(n)}\}_{k\leqslant n}$ is a sequence of square integrable martingales. Moreover, using inequalities (2.50) and (2.58), we obtain, by (2.22),

$$\sum_{\mathbf{j}\in\mathbb{N}^d}\left(f'(\theta_0,\mathbf{j})\right)^2g(\theta_0,\mathbf{j})\pi_{\theta_0}(\mathbf{j})\leqslant\left(\theta_0+\frac{\overline{\mathbf{b}}+1}{\underline{\mathbf{a}}}\right)\left(\sum_{\mathbf{j}\in\mathbb{N}^d}\mathbf{a}\cdot\mathbf{j}\pi_{\theta_0}(\mathbf{j})+2C_1+\frac{C_1^2}{\underline{\mathbf{a}}}\right)<\infty.\tag{2.69}$$

So, by means of (2.52),

$$\begin{aligned}
\lim_{n\to\infty}\langle M_n\rangle^{(n)}&=\lim_{n\to\infty}\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left(\left(\frac{\varepsilon_k f'(\theta_0,\mathbf{X}_{k-1}^{*})}{\sqrt{n}}\right)^2\mid\mathbf{X}_{k-1}^{*}\right) \\
&\overset{\text{a.s.}}{=}\sum_{\mathbf{j}\in\mathbb{N}^d}\left(f'(\theta_0,\mathbf{j})\right)^2g(\theta_0,\mathbf{j})\pi_{\theta_0}(\mathbf{j}).
\end{aligned}\tag{2.70}$$

Third, using Cauchy-Schwarz and Bienaymé-Chebyshev inequalities,

$$\begin{aligned}
&\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\left|\frac{\varepsilon_k^* f'(\theta_0,\mathbf{X}_{k-1}^*)}{\sqrt{n}}\right|^2 \mathbf{1}_{\{|\varepsilon_k^* f'(\theta_0,\mathbf{X}_{k-1}^*)/\sqrt{n}|\geqslant\varepsilon\}} \mid \mathbf{X}_{k-1}^*\right] \\
&\quad\leqslant \sum_{k=1}^{n}\left(\mathbb{E}_{\theta_0}\left[\left|\frac{\varepsilon_k^* f'(\theta_0,\mathbf{X}_{k-1}^*)}{\sqrt{n}}\right|^4 \mid \mathbf{X}_{k-1}^*\right]\right)^{1/2}\left(\mathbb{P}_{\theta_0}\left[\left|\frac{\varepsilon_k^* f'(\theta_0,\mathbf{X}_{k-1}^*)}{\sqrt{n}}\right| \geqslant \varepsilon \mid \mathbf{X}_{k-1}^*\right]\right)^{1/2} \\
&\quad\leqslant \frac{1}{n^{3/2}\varepsilon}\sum_{k=1}^{n}\left|f'(\theta_0,\mathbf{X}_{k-1}^*)\right|^3\left(\mathbb{E}_{\theta_0}\left[(\varepsilon_k^*)^4 \mid \mathbf{X}_{k-1}^*\right]\right)^{1/2}\left(\mathbb{E}_{\theta_0}\left[(\varepsilon_k^*)^2 \mid \mathbf{X}_{k-1}^*\right]\right)^{1/2}.
\end{aligned} \tag{2.71}$$

Let us compute $\mathbb{E}_{\theta_0}((\varepsilon_k^*)^4 \mid \mathbf{X}_{k-1}^*)$. We can show that the 4th central moment of the independent sum of a Poisson and a Bernoulli random variables equals $\mu_4 + 6\mu_2\gamma_2 + \gamma_4$, where $\mu_i$ and $\gamma_i$ denote the $i$th central moment of the Poisson and of the Bernoulli variable. If these variables have parameter $\lambda$ and $p$, respectively, then $\mu_4 = \lambda(1+3\lambda)$, $\mu_2 = \lambda$, $\gamma_4 = p(1-p)(3p^2-3p+1) \in [0,1]$, and $\gamma_2 = p(1-p) \in [0,1]$. We thus obtain

$$\left|\mathbb{E}_{\theta_0}\left((\varepsilon_k^*)^4 \mid \mathbf{X}_{k-1}^*\right)\right| \leqslant \frac{\boldsymbol{\Psi}(\theta_0)\cdot\mathbf{X}_{k-1}^*\left(7+3\boldsymbol{\Psi}(\theta_0)\cdot\mathbf{X}_{k-1}^*\right)+1}{\left(\mathbf{a}\cdot\mathbf{X}_{k-1}^*\right)^2}. \tag{2.72}$$

Hence

$$\begin{aligned}
&\left|f^{*'}(\theta_0,\mathbf{X}_{k-1}^*)\right|^3\left(\mathbb{E}_{\theta_0}\left[\varepsilon_k^4 \mid \mathbf{X}_{k-1}^*\right]\right)^{1/2}\left(\mathbb{E}_{\theta_0}\left[\varepsilon_k^2 \mid \mathbf{X}_{k-1}^*\right]\right)^{1/2} \\
&\quad\leqslant \left(\sqrt{\mathbf{a}\cdot\mathbf{X}_{k-1}^*}+\frac{C_1}{\sqrt{\underline{\mathbf{a}}}}\right)^3 \frac{\sqrt{\boldsymbol{\Psi}(\theta_0)\cdot\mathbf{X}_{k-1}^*\left(7+3\boldsymbol{\Psi}(\theta_0)\cdot\mathbf{X}_{k-1}^*\right)+1}}{\mathbf{a}\cdot\mathbf{X}_{k-1}^*}\left(\theta_0+\frac{\overline{\mathbf{b}}+1}{\underline{\mathbf{a}}}\right)^{1/2}.
\end{aligned} \tag{2.73}$$

Since the highest power of $X_n^*$ involved in (2.73) is $3/2$, and since by Proposition 2.4 the stationary distribution $\pi_{\theta_0}$ has finite second moments, we can apply (2.52) to (2.71) and obtain that

$$\lim_{n\to\infty}\sum_{k=1}^{n}\mathbb{E}_{\theta_0}\left[\left|\frac{\varepsilon_k^* f'(\theta_0,\mathbf{X}_{k-1}^*)}{\sqrt{n}}\right|^2 \mathbf{1}_{\{|\varepsilon_k^* f'(\theta_0,\mathbf{X}_{k-1}^*)/\sqrt{n}|\geqslant\varepsilon\}} \mid \mathbf{X}_{k-1}^*\right] \overset{\text{a.s.}}{=} 0. \tag{2.74}$$

It then ensues from the central limit theorem mentioned above that

$$\lim_{n\to\infty}\frac{\sum_{k=1}^{n}\varepsilon_k^* f'(\theta_0,\mathbf{X}_{k-1}^*)}{\sqrt{n}} \overset{\mathcal{D}}{=} \mathcal{N}\left(0, \sum_{\mathbf{j}\in\mathbb{N}^d}\left(f'(\theta_0,\mathbf{j})\right)^2 g(\theta_0,\mathbf{j})\pi_{\theta_0}(\mathbf{j})\right). \tag{2.75}$$

Finally, (2.56), together with (2.57), (2.59), (2.75), and Slutsky's theorem, implies that

$$\lim_{n\to\infty}\sqrt{n}\left(\widehat{\theta}_n^* - \theta_0\right) \overset{\mathcal{D}}{=} \mathcal{N}\left(0, \frac{\sum_{\mathbf{j}\in\mathbb{N}^d}\left(f'(\theta_0,\mathbf{j})\right)^2 g(\theta_0,\mathbf{j})\pi_{\theta_0}(\mathbf{j})}{\left(\sum_{\mathbf{j}\in\mathbb{N}^d}\left(f'(\theta_0,\mathbf{j})\right)^2\pi_{\theta_0}(\mathbf{j})\right)^2}\right), \tag{2.76}$$

which leads to

$$\lim_{n\to\infty}\sqrt{n\frac{\left(\sum_{\mathbf{j}\in\mathbb{N}^d} f'(\theta_0,\mathbf{j})^2\pi_{\theta_0}(\mathbf{j})\right)^2}{\sum_{\mathbf{j}\in\mathbb{N}^d} f'(\theta_0,\mathbf{j})^2 g(\theta_0,\mathbf{j})\pi_{\theta_0}(\mathbf{j})}}\left(\widehat{\theta}_n^* - \theta_0\right) \overset{\mathcal{D}}{=} \mathcal{N}(0,1). \tag{2.77}$$

It now remains to prove the asymptotic distribution (2.51). Thanks to what precedes, this result is immediate as soon as we prove that

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n}\left(f'\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)\right)^2 g\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right) \overset{\text{a.s.}}{=} \sum_{\mathbf{j}\in\mathbb{N}^d}\left(f'(\theta_0,\mathbf{j})\right)^2 g(\theta_0,\mathbf{j})\pi_{\theta_0}(\mathbf{j}), \tag{2.78}$$

as well as the equivalent result for the numerator. For this purpose, we write

$$\begin{aligned}\sum_{k=0}^{n}\left(f'\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)\right)^2 g\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right) &= \sum_{k=0}^{n}\left(f'(\theta_0, \mathbf{X}_k^*)\right)^2 g(\theta_0, \mathbf{X}_k^*) \\ &\quad + \sum_{k=0}^{n}\left[\left(f'\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)\right)^2 g\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right) - \left(f'(\theta_0, \mathbf{X}_k^*)\right)^2 g(\theta_0, \mathbf{X}_k^*)\right],\end{aligned} \tag{2.79}$$

and show that $(f'(\cdot,\mathbf{j}))^2 g(\cdot,\mathbf{j})$ has a bounded derivative and is thus Lipschitz. We have indeed $|\, g'(\theta,\mathbf{j}) \,| = |\, \mathbf{a}\cdot\mathbf{j} + p'(\theta,\mathbf{j}) \,|\, (\mathbf{a}\cdot\mathbf{j})^{-1} \leqslant 1 + C_1\underline{\mathbf{a}}^{-1}$, hence

$$\begin{aligned}&\left|2f''(\theta,\mathbf{j})f'(\theta,\mathbf{j})g(\theta,\mathbf{j}) + \left(f'(\theta,\mathbf{j})\right)^2 g'(\theta,\mathbf{j})\right| \\ &\quad\leqslant 2C_2\left(1 + C_1\underline{\mathbf{a}}^{-1}\right)\underline{\mathbf{a}}^{-1}(\boldsymbol{\Psi}(\theta)\cdot\mathbf{j} + 1) + \left(1 + C_1\underline{\mathbf{a}}^{-1}\right)^2(\mathbf{a}\cdot\mathbf{j} + C_1) \leqslant C_3\boldsymbol{\Psi}(\theta_2)\cdot\mathbf{j},\end{aligned} \tag{2.80}$$

for some constant $C_3 > 0$. This enables us to write

$$\begin{aligned}&\frac{1}{n+1}\sum_{k=0}^{n}\left|\left(f'\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)\right)^2 g\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right) - \left(f'(\theta_0, \mathbf{X}_k^*)\right)^2 g(\theta_0, \mathbf{X}_k^*)\right| \\ &\quad\leqslant \left|\widehat{\theta}_n^* - \theta_0\right| C_3\frac{1}{n+1}\sum_{k=0}^{n}\boldsymbol{\Psi}(\theta_2)\cdot\mathbf{X}_k^*.\end{aligned} \tag{2.81}$$

By the strong consistency of $\widehat{\theta}_n^*$ together with (2.52) and (2.22), (2.81) almost surely tends to zero. Combined with (2.52) and (2.69) in (2.79), this implies (2.78). □

## 3. Application to Real Epidemics

### 3.1. Explicit Epidemic Model

Although the epidemic model (1.1) has a clear interpretation in terms of the disease propagation, it can be also proved [9] that it is the limit process, as the initial population size tends to infinity, of the incidence of infectives described by a thorough process in a large branching population, taking into account all the health states of the disease and describing in detail the infection and latent processes for each individual in the population. This detailed process is a multitype branching process with age and population size dependence, taking into account the variability of many individual factors such as the reproduction, the survival, the transmission of the disease, and the latent time. We assume a disease of the general SEIR type. We also assume a random latent period and, for the sake of simplicity, a duration of the state $I$ of one time unit only. Hence the incidence of infectives exactly corresponds to the incidence of cases.

The main assumptions for the construction of the limit model $\{X_n\}_n$ are, concerning the disease as follows: (i) at the initial time the disease is rare and the total population size is large; (ii) the infection *via* horizontal route is of Reed-Frost type, with the probability for a susceptible individual to become infected by a given dose of pathogens being inversely proportional to the total population size; (iii) the individual survival law is the same for $E$ and $S$ individuals and is independent of the population and of the time; (iv) the latent time law given the individual survival is independent of the time, the individual age, and the infectives population during the latent period. This last property is possible if we assume that overinfection during the incubation has a negligible effect on the latent time. We moreover assume that the whole healthy population size is relatively stable over time.

The limit process (1.1) is then obtained in an inductive way as the limit in distribution of $\{I_n\}_n$ as the initial total population size $|\mathbf{N}_0| \to \infty$, where $I_n$ denotes the number of new infectives at time $n$. More precisely, denoting by $E_n$ the incidence of exposed individuals at time $n$ and by $a_m$ the largest individual survival age, we obtain that $\{X_n, Z_n\}_n \overset{\mathfrak{D}}{=} \lim_{N_0 \to \infty} \{I_n, E_n\}_n$, with

$$X_n \mid (X_{n-1}, \ldots, X_{n-a_m+1}) \overset{\mathfrak{D}}{=} \text{Poiss}\left( \sum_{k=1}^{a_m-1} \Psi_k X_{n-k} \right), \tag{3.1}$$

$$Z_n \mid X_n \overset{\mathfrak{D}}{=} \text{Poiss}(\Psi_0 X_n), \tag{3.2}$$

and where for all $k = 1 \cdots a_m - 1$, $\Psi_k$ satisfies (1.2) in Section 1, with the latent period distribution given survival, $P_{\text{inc},a}$, assumed independent of $a$, being denoted by $P_{\text{inc}}$. Thus

$$\Psi_k = \theta_0 P_{\text{inc}}(k) \sum_{a=k+1}^{a_m} P_{\text{age}}(a) + p_{\text{mat}} P_{\text{inc}}(k) P_{\text{age}}(k+1), \tag{3.3}$$

and similarly,

$$\Psi_0 = \theta_0 + p_{\text{mat}} P_{\text{age}}(1). \tag{3.4}$$

Moreover, when assuming a relatively stable population size and denoting by $S(a)$ the individual survival probability until age $a$ at least, then $P_{\text{age}}(a)$ is equal to $S(a)\left(\sum_{a'=1}^{a_m} S(a')\right)^{-1}$.

### 3.2. Illustration: The BSE Epidemic in Great Britain

In this section we provide an illustration of the methodology developed in Section 2 by studying the decay phase of the BSE epidemic in Great Britain, based on the observations of the yearly diagnosed cases from 1989 until 2011, with 1988 being the date of the main control measure. The disease that was first officially identified in 1986 [20] reached its peak in 1992 (36682 cases) and is obviously now in its decay phase. Since only a very few cases were recently reported [5], namely, 11 in 2010 and 5 in 2011, the spread of the disease should *a priori* "soon" come to an end. It has been accepted that the epidemic is fading out, with a very low level of risk for cattle and humans. However it might be interesting to have a more precise idea of the extinction phase of the disease, for instance, of its length and of its intensity.

#### 3.2.1. Choice of the Epidemic Model

BSE is a transmissible disease through ingestion of prions (horizontal route) and through maternal route. It is commonly accepted (see, e.g., [21]) that the infectious phase (including the clinical state) lasts at most one year, while the latent state can be of several years. If one compares the epidemic data with the total population of around 9 million cattle in Great Britain, it is reasonable to assume that, even at the peak time, BSE may be considered as a rare disease in a large population. We moreover assume that the probability for an animal to become infected by a horizontal route follows a Reed-Frost type model as in (ii) of Section 3.1, and that the probability for a calf to become infected by its dam is nonnull and constant over time. Moreover, for the sake of simplicity and parsimony, we do not take into account potential heterogeneity factors such as the different regulations from 1989 (the main regulation was the feed ban of July 1988 and was shown to be quite efficient [22]), the different types of breeding and of races, the age for animals older than one year, and the evolution of the surveillance system and of diagnosis tests. So the process $\{X_n\}_n$ defined by (3.1)–(3.4) appears relevant to model the spread of this fatal disease after the 1988 ban, where the infectious state $I$ corresponds to the end of the incubation period and the clinical state, and $R$ corresponds to the death (assumed to be either by routine slaughtering for $E$ and $S$ cattle, and control slaughtering for $I$ cattle). Choosing a time step of one year, $X_n$ then represents the yearly incidence of notified cases, which corresponds to the available data provided by the World Organisation for Animal Health [5].

#### 3.2.2. Choice of the Model Parameters

In order to predict the future epidemic evolution, we need to evaluate all the parameters involved in (3.3). These parameters were already studied in previous works [22–25] using different models, amounts of observations, and kinds of estimators. Since it might have a crucial role in the decay phase of the disease, we focus here on $\theta_0$ the infection parameter from 1989 *via* a horizontal route of transmission (i.e., the mean number per infective and per year of newly infected). Until the 1988 feed ban regulation, the main routes of transmission of BSE were horizontal *via* protein supplements (Meat and Bone Meal, milk replacers), and maternal from a dam to its calf. Since a previous statistical study [22] concluded to the full efficiency of the 1988 ban and since most of cattle are slaughtered before the age of 10 years,

the fact that cases of BSE are still observed more than 20 years later could suggest the existence of a remaining source of infection, either *via* a maternal transmission route, or *via* a horizontal one, for example, *via* the ingestion of excreted prions from alive infected animals, or from prions left in the environment. The prediction of the future disease spread thus strongly relies on the intensity of this infection, quantified by the parameter $\theta_0$. We provide an estimation of $\theta_0$ together with a confidence interval, which will be used for prediction estimations. We provide in addition a sensitivity analysis to the other parameters of the model, which are chosen as follows. First, the maximal age of the individuals in the population $a_m$ is set to $a_m = 10$, which corresponds to the largest reported survival age with a nonnegligible probability. Next, we set $p_{\mathrm{mat}} = 0.1$, which is the largest order of magnitude commonly accepted for the maternal infection probability [23, 25]. Following a previous work, [22], we assume here a discretized Weibull distribution with parameters $\alpha, \beta$ for the distribution of the latent period; that is to say, for each $k \geqslant 1$, $P_{\mathrm{inc}}(k) := e^{-((\alpha-1)/\alpha\beta^{\alpha})(k-1)^{\alpha}} - e^{-((\alpha-1)/\alpha\beta^{\alpha})k^{\alpha}}$, where $\alpha$ is a shape parameter, and $\beta$ is the mode of the probability density of the corresponding continuous Weibull distribution. The parameters $\alpha$ and $\beta$ have a very bad identifiability with the infection parameters on a sole monotonous phase and are consequently estimated in [22] on the whole epidemic series (growth and decay), by Bayesian maximum *a posteriori* (MAP) estimations. We set $\alpha = \widehat{\alpha}^{\mathrm{MAP}} = 3.84$ and $\beta = \widehat{\beta}^{\mathrm{MAP}} = 7.46$. Since the death of susceptible and exposed animals is mostly due to routine slaughtering, the survival probability $S(k)$ (probability for an apparently healthy animal to survive at least until $k$ years) corresponds to the probability of being slaughtered after the age of $k$ years. It is derived from existing literature [24].

### 3.2.3. Estimation of the Infection Parameter

Our study is based on the yearly number of cases of BSE reported in Great Britain from 1989 until 2011 (Figure 1) provided by the World Organisation for Animal Health [5], denoted by $X_n^{\mathrm{obs}}$ for each year $n$. We set $d = a_m - 1 = 9$ and denote by $\mathbf{X}_n^{\mathrm{obs}}$ the $d$-dimensional vector $(X_n^{\mathrm{obs}}, \ldots, X_{n-d+1}^{\mathrm{obs}})$. We choose the first time $n = 1$ of the epidemic model such that the model with its initial values covers the period starting from 1989, which we consider here as a time-homogeneous period. Hence $X_{-d+1} = X_{1989}^{\mathrm{obs}}$, and $n = 1$ corresponds to the year $1989 + d = 1998$. We estimate $\theta_0$ by the WCLSE (2.30) in model (1.1)–(3.3), where $\{p_{\mathrm{mat}}, \alpha, \beta, \{S(k)\}_k\}$ are given by the previous values, and where $\mathbf{X}_0 = \mathbf{X}_{1997}^{\mathrm{obs}}$. According to [5], $|\mathbf{X}_0| = |\mathbf{X}_{1997}^{\mathrm{obs}}| = 167977$. We are thus close to the asymptotic $|\mathbf{X}_0| \to \infty$. The number of observations is $n = 14$. The estimator (2.30) provides the estimation $\widehat{\theta}_{|\mathbf{X}_0|}^{\mathrm{obs}} = 2.4324$. We point out that this estimation is of the same order of magnitude as the maximum *a posteriori* Bayesian estimation $\widehat{\theta}^{\mathrm{MAP}} = 2.43$ based on the whole epidemic until 2007, assuming a uniform prior probability [22]. Using (2.32) we obtain the following confidence interval $[\widehat{\theta}_{\min}, \widehat{\theta}_{\max}]$ with asymptotic probability 95%, where $\widehat{\theta}_{\min} := \widehat{\theta}_{|\mathbf{X}_0|} - 1.96\widehat{c}_1^{-1}$, $\widehat{\theta}_{\max} := \widehat{\theta}_{|\mathbf{X}_0|} + 1.96\widehat{c}_1^{-1}$, and $\widehat{c}_1 := (\sum_{k=1}^{n} \mathbf{a} \cdot \mathbf{X}_{k-1}/\sigma^2(\widehat{\theta}_{|\mathbf{X}_0|}))^{1/2}$. Assuming $\alpha_i = X_{1997-i+1}^{\mathrm{obs}} |\mathbf{X}_{1997}^{\mathrm{obs}}|^{-1}$, we get $\widehat{c}_1^{\mathrm{obs}} = 40.7343$ (observed value of $\widehat{c}_1$), and

$$\mathbb{P}\left(\theta_0 \in \left[\widehat{\theta}_{\min}, \widehat{\theta}_{\max}\right]\right) \simeq 95\%, \qquad \widehat{\theta}_{\min}^{\mathrm{obs}} = 2.3842, \qquad \widehat{\theta}_{\max}^{\mathrm{obs}} = 2.4805. \tag{3.5}$$

Although this confidence interval is an asymptotic one, as $|\mathbf{X}_0| \to \infty$, it is a very good approximation of the true confidence interval for a finite $|\mathbf{X}_0|$, since $|\mathbf{X}_0|$ is here very large. Since the estimation of $\theta_0$ relies on the values given to the other model parameters $\{p_{\mathrm{mat}}, \alpha, \beta\}$, we evaluate in addition its sensitivity to the values of these parameters. For this purpose,

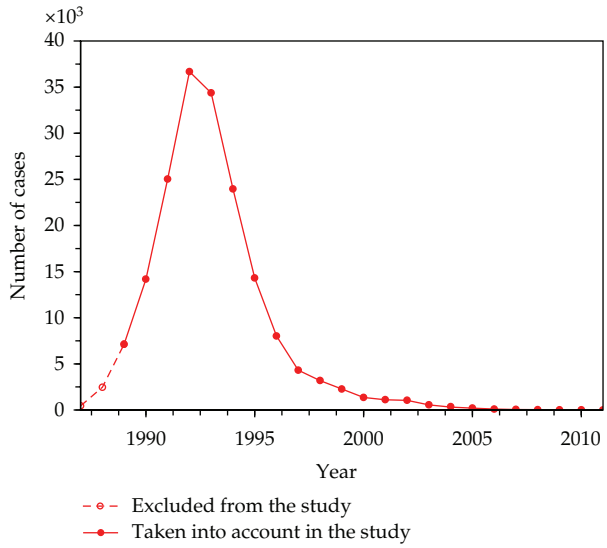


**Figure 1:** Yearly number of cases of BSE reported in Great Britain from 1987 to 2011 [5]. Our study only takes into account the data from 1989, after the 1988 feed ban law.

**Table 1:** Sensitivity analysis. Estimation of the infection parameter $\theta_0$, and its confidence interval $[\widehat{\theta}^{\text{obs}}_{\min}, \widehat{\theta}^{\text{obs}}_{\max}]$ with asymptotic probability 95%, for different values of the maternal infection parameter $p_{\text{mat}}$ and of the latency parameters $(\alpha, \beta)$. The values $(3.84, 7.46)$ correspond to the Bayesian MAP estimations $(\widehat{\alpha}^{\text{MAP}}, \widehat{\beta}^{\text{MAP}})$, and 0.1 to the largest commonly accepted order of magnitude for $p_{\text{mat}}$. The estimations of $\theta_0$ are based on the observed data over the years 1989–2011.

| $p_{\text{mat}}$ | $\alpha$ | $\beta$ | $\widehat{\theta}^{\text{obs}}_{\lvert \mathbf{X}_0 \rvert}$ | $[\widehat{\theta}^{\text{obs}}_{\min}, \widehat{\theta}^{\text{obs}}_{\max}]$ |
|---|---|---|---|---|
| 0.1 | 3.84 | 7.46 | 2.4324 | [2.3842, 2.4805] |
| 0 | 3.84 | 7.46 | 2.4860 | [2.4379, 2.5342] |
| 1 | 3.84 | 7.46 | 1.9492 | [1.9014, 1.9970] |
| 0.1 | 2 | 7.46 | 2.7835 | [2.7287, 2.8382] |
| 0.1 | 20 | 7.46 | 4.0186 | [3.9395, 4.0977] |
| 0.1 | 3.84 | 1 | 1.0127 | [0.9925, 1.0329] |
| 0.1 | 3.84 | 10 | 6.2128 | [6.0914, 6.3341] |
| 0.1 | 3 | 6 | 1.5402 | [1.5095, 1.5710] |
| 0.1 | 4 | 5 | 1.0227 | [1.0020, 1.0434] |

we compute the estimation of $\theta_0$ and the associated confidence interval with asymptotic probability 95%, for different values of $(p_{\text{mat}}, \alpha, \beta)$ (Table 1). The first line of the table corresponds to the parameters chosen for the model. In each of the four following lines, we fix two coordinates and choose an extremal (unrealistic) value for the third one. It appears that the estimation of $\theta_0$ is almost independent of the value of the maternal infection parameter. However, the estimation seems more strongly dependent on the parameters of the latent period distribution. Nevertheless, even for very unrealistic values $(\alpha, \beta)$, all the estimations of $\theta_0$ remain in the same order of magnitude of several units. This is really small compared to estimations obtained for the infection *via* Meat and Bone Meal or lactoreplacers (before 1989) which are of the order of 1000 [22]. However, although these estimations are all very small,

$\theta_0$ seems nonnull. This could suggest the existence of a minor but nonnull infection source which is not of maternal type.

### 3.2.4. Extinction of the Epidemic

We know thanks to Proposition 2.1 that $\{X_n\}_n$ becomes extinct almost surely if and only if $R_0 = \sum_{k=1}^{d} \Psi_k(\theta_0) \leqslant 1$. The estimated basic reproduction is here $R_0(\widehat{\theta}^{\text{obs}}_{|\mathbf{X}_0|}) = 0.1072$. Moreover, solving with a computing program the equation $\sum_{k=1}^{d} \Psi_k(\widehat{\theta}^{\text{obs}}_{|\mathbf{X}_0|})\rho^{-k} = 1$, we obtain the following value for the Perron's root $\rho(\widehat{\theta}^{\text{obs}}_{|\mathbf{X}_0|}) = 0.6665$, which provides the speed of decay of the expected yearly incidence of cases (see (2.6)); from a certain time, the expected number of new cases will decrease from around 33% every year.

### 3.2.5. Prediction of the Incidences of Cases and Incidences of Infected Cattle

Let us predict the spread of the disease from 2012 by means of simulations of $\{X_n\}_n$, where $\theta_0$ is replaced by its previous estimation $\widehat{\theta}^{\text{obs}}_{|\mathbf{X}_0|} = 2.4324$, and where the initial time of the model is 2011, that is, $\mathbf{X}_0 = \mathbf{X}^{\text{obs}}_{2011}$. The simulations are done recursively using the transition law (3.1). We point out that the model initialized by $\mathbf{X}_0 = \mathbf{X}^{\text{obs}}_{1997}$ provides quite realistic simulations on the period 1998–2011 compared to the real observations on the same period, as illustrated in Figure 2(a). The epidemic process $\{X_n\}_n$ thus seems to provide a satisfying prediction of the overall evolution of the real epidemic. In order to predict the incidences of futures cases, we simulate 1000 trajectories of $\{X_n\}_n$ initialized by the observed values $\mathbf{X}^{\text{obs}}_{2011}$, with the estimated infection parameter $\widehat{\theta}^{\text{obs}}_{|\mathbf{X}_0|} = 2.4324$. We illustrate in Figure 2(c), for each year from 2012, the maximum, minimum, median, 2.5%, and 97.5% quantiles associated with these 1000 realizations. It is also relevant to study and predict the evolution of the incidence of infected cattle in the population, which represents the hidden face of the epidemic. The incidence $Z_n$ of infected cattle at time $n$, conditionally on the number $X_n$ of cases at that time, is given by the Poisson distribution (3.2). For every $n \geqslant 2012$ and for each of the 1000 previously simulated values $X_n$, we generate one realization of $Z_n$. We then illustrate in Figure 2(d), the yearly maximum, minimum, median, 2.5%, and 97.5% quantiles associated with the 1000 realizations.

### 3.2.6. Prediction of the Year of Extinction

Let $T := 2011 + \inf\{n \geqslant 1, \mathbf{X}_n = \mathbf{0}\}$ denote the extinction year of the epidemic process $\{X_n\}_n$. According to (2.7) and denoting by $\mathbf{f}_{\theta_0}$ the offspring generating function of $\{\mathbf{X}_n\}_n$ defined in (2.2) (from now on we let the dependence in $\theta_0$ appear in the notation), we have $\mathbb{P}_{\theta_0}(T \leqslant 2011 + n) = \mathbf{f}_{\theta_0,n}(\mathbf{0})^{\mathbf{X}^{\text{obs}}_{2011}}$, for every $n \geqslant 1$, which by iterating $\mathbf{f}_{\theta_0}$ can be computed explicitly. We obtain in particular, for the estimated value $\widehat{\theta}^{\text{obs}}_{|\mathbf{X}_0|} = 2.4324$, the following $p$-quantiles for the extinction time (see (2.8)): $n^T_{0.5} = 2028$, $n^T_{0.95} = 2035$ and $n^T_{0.99} = 2039$. Keeping in mind that $T$ corresponds to the complete extinction of the epidemic (i.e., $d = 9$ consecutive years without any case), these results actually mean that for the infection parameter $\theta_0 = 2.4324$, with probability larger than 50% (resp., 95% and 99%), no case will arise in the population from year 2020 (resp., 2027 and 2031). Moreover, in order to take into account the uncertainty around the estimation $\widehat{\theta}^{\text{obs}}_{|\mathbf{X}_0|}$ of the infection parameter $\theta_0$, we make use of the asymptotic

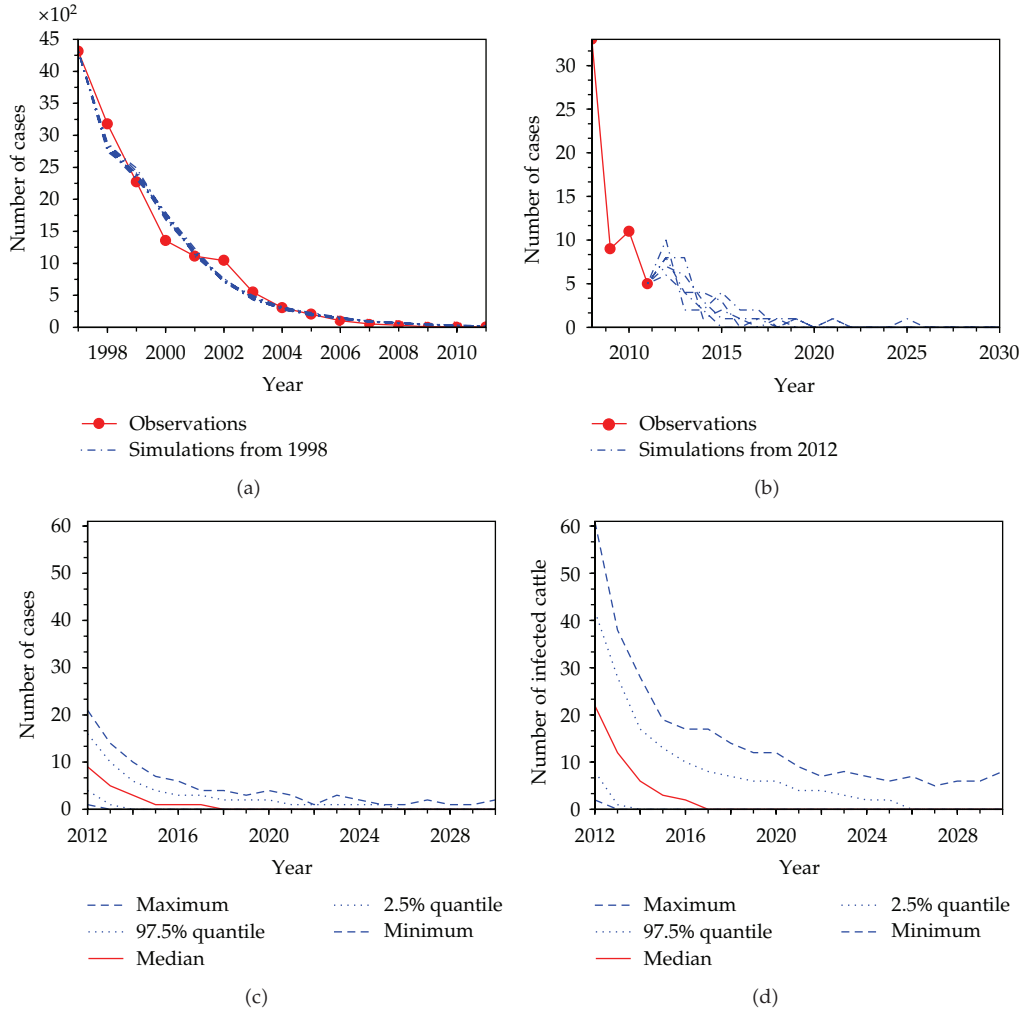


**Figure 2:** Figure 2(a) 10 simulations of $\{X_n\}_n$ initialized by $\mathbf{X}^{\text{obs}}_{1997}$, and comparison with the observations on the period 1998–2011. Figure 2(b) 5 simulations of $\{X_n\}_n$ initialized by $\mathbf{X}^{\text{obs}}_{2011}$. Figures 2(c) and 2(d) prediction, based on 1000 simulations of the process of the yearly incidences of cases (resp., infected cattle) from 2012. 95% of the trajectories remain in the band delimited by the blue dotted lines. All the simulations are done with the infection parameter $\widehat{\theta}^{\text{obs}}_{|\mathbf{X}_0|} = 2.4324$.

confidence interval (3.5) of $\theta_0$ and of the fact that $\theta \mapsto \mathbb{P}_\theta(T \leqslant n)$ is a decreasing function of $\theta$, which implies that for every $n \geqslant 2011$,

$$\mathbb{P}\left(\mathbb{P}_{\theta_0}(T \leqslant n) \in \left[\mathbb{P}_{\widehat{\theta}_{\max}}(T \leqslant n), \mathbb{P}_{\widehat{\theta}_{\min}}(T \leqslant \mathrm{n})\right]\right) \simeq 95\%. \tag{3.6}$$

We collect in Table 2 the observed interval $[\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\max}}(T \leqslant n), \mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\min}}(T \leqslant n)]$, for each $n \geqslant 2020$ (if $n < 2019$ we have, conditionally on the initial value $\mathbf{X}^{\text{obs}}_{2011}$, $\mathbb{P}(\mathbf{X}_n = \mathbf{0}) = 0$ because of the memory which is not equal to 0). Note that these intervals are very narrow, leading to an accurate estimation of $\mathbb{P}_{\theta_0}(T \leqslant n)$.

**Table 2:** Cumulative distribution function of the year of extinction computed with the infection parameters $\widehat{\theta}^{\text{obs}}_{\min} = 2.3842$ and $\widehat{\theta}^{\text{obs}}_{\max} = 2.4805$ defined by (3.5). The values in bold character correspond to the $p$-quantiles $n^T_p$ for $p = 0.5, 0.95$, and 0.99, and to the asymptotic confidence intervals of $\mathbb{P}_{\theta_0}(T \leqslant n^T_p)$ based on (3.6).

| $n$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\max}}(T \leqslant n)$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\min}}(\cdots)$ | $n$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\max}}(\cdots)$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\min}}(\cdots)$ | $n$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\max}}(\cdots)$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\min}}(\cdots)$ |
|---|---|---|---|---|---|---|---|---|
| 2020 | 0.0000 | 0.0000 | 2030 | 0.7610 | 0.7807 | 2040 | 0.9950 | 0.9957 |
| 2021 | 0.0000 | 0.0000 | 2031 | 0.8310 | 0.8465 | 2041 | 0.9967 | 0.9972 |
| 2022 | 0.0010 | 0.0014 | 2032 | 0.8816 | 0.8934 | 2042 | 0.9978 | 0.9981 |
| 2023 | 0.0121 | 0.0152 | 2033 | 0.9186 | 0.9273 | 2043 | 0.9985 | 0.9988 |
| 2024 | 0.0496 | 0.0592 | 2034 | 0.9451 | 0.9513 | 2044 | 0.9990 | 0.9992 |
| 2025 | 0.1211 | 0.1390 | **2035** | **0.9633** | **0.9677** | 2045 | 0.9995 | 0.9994 |
| 2026 | 0.2325 | 0.2579 | 2036 | 0.9755 | 0.9786 | 2046 | 0.9996 | 0.9996 |
| 2027 | 0.3756 | 0.4047 | 2037 | 0.9835 | 0.9857 | 2047 | 0.9997 | 0.9998 |
| **2028** | **0.5303** | **0.5584** | 2038 | 0.9889 | 0.9904 | 2048 | 0.9998 | 0.9998 |
| 2029 | 0.6619 | 0.6862 | **2039** | **0.9925** | **0.9936** | 2049 | 0.9999 | 0.9999 |

**Table 3:** Cumulative distribution function of the total size of the epidemic for the infection parameters $\widehat{\theta}^{\text{obs}}_{\min} = 2.3842$ and $\widehat{\theta}^{\text{obs}}_{\max} = 2.4805$ defined by (3.5). The values in bold character correspond to the $p$-quantiles $n^N_p$ for $p = 0.5, 0.95$, and 0.99, and to the asymptotic confidence intervals of $\mathbb{P}_{\theta_0}(N \leqslant n^N_p)$.

| $n$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\max}}(N \leqslant n)$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\min}}(N \leqslant n)$ | $n$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\max}}(\cdots)$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\min}}(\cdots)$ | $n$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\max}}(\cdots)$ | $\mathbb{P}_{\widehat{\theta}^{\text{obs}}_{\min}}(\cdots)$ |
|---|---|---|---|---|---|---|---|---|
| $\cdots$ | $\cdots$ | $\cdots$ | 16 | 0.1293 | 0.1669 | 28 | 0.8702 | 0.9048 |
| 5 | 0.0001 | 0.0000 | 17 | 0.1795 | 0.2263 | 29 | 0.8942 | 0.9302 |
| 6 | 0.0002 | 0.0003 | 18 | 0.2373 | 0.2915 | 30 | 0.9283 | 0.9495 |
| 7 | 0.0005 | 0.0008 | 19 | 0.3025 | 0.3637 | **31** | **0.9480** | **0.9646** |
| 8 | 0.0013 | 0.0020 | 20 | 0.3741 | 0.4409 | 32 | 0.9635 | 0.9759 |
| 9 | 0.0027 | 0.0046 | 21 | 0.4486 | 0.5189 | 33 | 0.9742 | 0.9838 |
| 10 | 0.0060 | 0.0094 | **22** | **0.5231** | **0.5934** | 34 | 0.9824 | 0.9892 |
| 11 | 0.0119 | 0.0178 | 23 | 0.5975 | 0.6639 | **35** | **0.9881** | **0.9928** |
| 12 | 0.0214 | 0.0310 | 24 | 0.6655 | 0.7287 | 36 | 0.9917 | 0.9952 |
| 13 | 0.0365 | 0.0518 | 25 | 0.7282 | 0.7840 | 37 | 0.9945 | 0.9968 |
| 14 | 0.0585 | 0.0801 | 26 | 0.7829 | 0.8323 | 38 | 0.9964 | 0.9980 |
| 15 | 0.0891 | 0.1181 | 27 | 0.8304 | 0.8726 | 39 | 0.9977 | 0.9989 |

### *3.2.7. Prediction of the Epidemic Size*

Let $N := \sum_{n=1}^{T-2011} X_n$ be the total size of the future epidemic from 2012 (total number of cases from 2012 until the extinction of the epidemic). We compute the distribution of $N$ using (2.11), conditionally on the event $\{\mathbf{X}_0 = \mathbf{X}^{\text{obs}}_{2011}\}$. We obtain in particular, for the estimated value $\widehat{\theta}^{\text{obs}}_{|\mathbf{X}_0|} = 2.4324$, the following $p$-quantiles for the total epidemic size (see (2.12)): $n^N_{0.5} = 22$, $n^N_{0.95} = 31$ and $n^N_{0.99} = 35$. From (2.13) we deduce that the mean value and variance of $N$ for the parameter $\widehat{\theta}^{\text{obs}}_{\min}$ (resp., $\widehat{\theta}^{\text{obs}}_{\max}$) are 21 and 27 (resp., 22 and 28). Moreover, we obtain similarly as for the extinction time and thanks to (3.5) a confidence interval $[\mathbb{P}_{\widehat{\theta}_{\max}}(N \leqslant n), \mathbb{P}_{\widehat{\theta}_{\min}}(N \leqslant n)]$ of $\mathbb{P}_{\theta_0}(N \leqslant n)$, for every $n \in \mathbb{N}$, with asymptotic probability 95%, collected in Table 3.

*3.2.8. Study of the Very Late Extinction Case*

In order to predict the behavior of the "most dangerous" evolution of the epidemic, we first need to compute the estimation $\widehat{\theta}_n^{*\mathrm{obs}}$ (2.47) based on the data in Great Britain. The number of available observations is only $n = 14$, so we are far from the asymptotic setting $n \to \infty$ of Theorem 2.7. However, the large value of $|\mathbf{X}_0|$ can make us hope for a good accuracy. We point out that, by making use of the estimator $\widehat{\theta}_n^*$ on the real data, we make an unverifiable assumption on the future of the epidemic; we consider the data as if they were the beginning of a trajectory with very late extinction. This should have the following consequence: the estimation provided by $\widehat{\theta}_n^{*\mathrm{obs}}$ should *a priori* be a bit smaller than the value 2.4324 provided by $\widehat{\theta}_{|\mathbf{X}_0|}^{\mathrm{obs}}$. Indeed we obtain $\widehat{\theta}_n^{*\mathrm{obs}} = 2.4305$. Using this value, we deduce from (2.51) the confidence interval $[\widehat{\theta}_{\min}^*, \widehat{\theta}_{\max}^*]$ of $\theta_0$ with asymptotic probability 95%, where $\widehat{\theta}_{\min}^* := \widehat{\theta}_n^* - 1.96\widehat{c}_2^{-1}$, $\widehat{\theta}_{\max}^* := \widehat{\theta}_n^* + 1.96\widehat{c}_2^{-1}$, and

$$\widehat{c}_2 := \sum_{k=0}^{n} \left( f'\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)\right)^2 \left[ \sum_{k=0}^{n} \left( f'\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)\right)^2 g\left(\widehat{\theta}_n^*, \mathbf{X}_k^*\right)\right]^{-1/2}. \tag{3.7}$$

Therefore, $\widehat{c}_2^{\mathrm{obs}} = 40.6988$, and

$$\mathbb{P}\left(\theta_0 \in \left[\widehat{\theta}_{\min}^*, \widehat{\theta}_{\max}^*\right]\right) \simeq 95\%, \qquad \widehat{\theta}_{\min}^{*\mathrm{obs}} = 2.3823, \qquad \widehat{\theta}_{\max}^{*\mathrm{obs}} = 2.4787, \tag{3.8}$$

which is of the same magnitude order as the confidence interval [2.3842, 2.4805] obtained with the unconditioned process (see (3.5)). Let us predict the most dangerous evolution thanks to the transition law of the conditioned process given by Proposition 2.2. Note that if one computes the eigenvalues $\rho$ and $\lambda$ introduced in Proposition 2.5, one obtains $\rho(\widehat{\theta}^{*\mathrm{obs}}) = 0.6664$ and $|\lambda(\widehat{\theta}^{*\mathrm{obs}})| = 0.5570$. It thus appears that the convergence of the epidemic process conditioned on nonextinction at time $k$ to the $Q$-process as $k \to \infty$ is not very fast. As a consequence the study of the $Q$-process only provides information about the behavior of the disease spread in the case of an extremely late extinction. First, we see thanks to Figure 3(a) that the simulations provided by the conditioned process initialized by $\mathbf{X}_0 = \mathbf{X}_{1997}^{\mathrm{obs}}$, and where $\theta_0$ is estimated by $\widehat{\theta}_n^{*\mathrm{obs}} = 2.4305$, are true to the real observations on the period 1999–2011. Figure 3(b) is an example of one simulation in the period 2012–2040 of the conditioned process, for $\mathbf{X}_0^* = \mathbf{X}_{2011}^{\mathrm{obs}}$. It appears that the values of this simulated trajectory are rapidly very small and of course are never equal to 0 for $d = 9$ consecutive times. For a finer prediction, we simulate 1000 realizations of this process from 2012 until 2050, with $\mathbf{X}_0^* = \mathbf{X}_{2011}^{\mathrm{obs}}$. Moreover, for every $n \geqslant 2012$ and for each of the 1000 simulated values $X_n^*$, we make one realization of the incidence $Z_n$ of infected cattle at time $n$, according to the law given by (3.2). Figures 3(c) and 3(d) represent the yearly maximum, minimum, median, 2.5%, and 97.5% quantiles associated with the 1000 realizations of, respectively, the incidence of cases and infected cattle, in case of an extreme late extinction. It appears thanks to Figures 3(c) and 3(d) that the supposedly "most dangerous" trajectories nevertheless do not reach high values and do not present a new peak of epidemic.

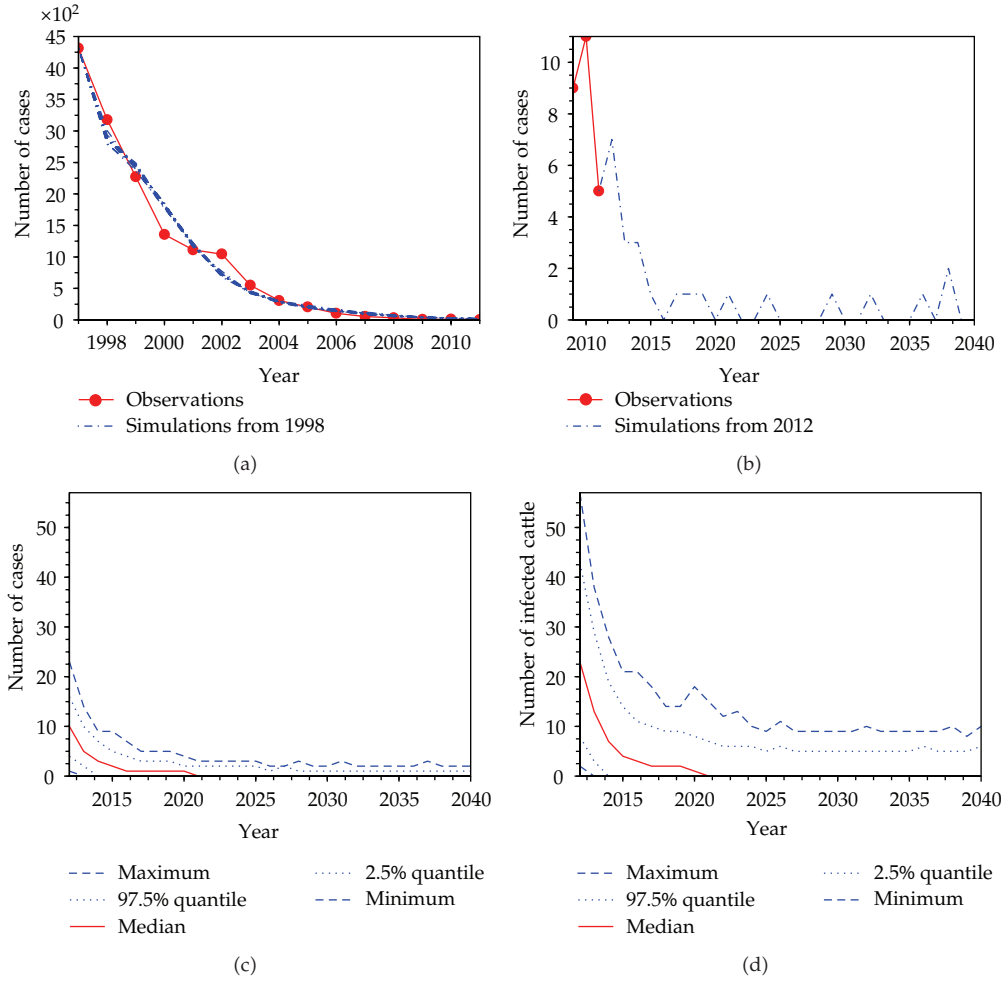


**Figure 3:** Figure 3(a) 10 simulations of $\{X_n^*\}_n$ initialized by $\mathbf{X}_{1997}^{\text{obs}}$, and comparison with the observations on the period 1998–2011. Figure 3(b) one simulation of $\{X_n^*\}_n$ initialized by $\mathbf{X}_{2011}^{\text{obs}}$. Figures 3(c) and 3(d) prediction, based on 1000 simulations of the conditioned process of the yearly incidences of cases (resp., infected cattle) from 2012. 95% of the trajectories remain in the band delimited by the blue dotted lines. All the simulations are done with the infection parameter $\widehat{\theta}_{|\mathbf{X}_0|}^{*\text{obs}} = 2.4305$.

## 4. Conclusion

The methodology developed in Section 2 thus applies well here and enables not only to confirm mathematically what is commonly accepted, namely, that BSE is fading out, but also to predict, with a very large probability, that the last BSE case will occur before 2027, and that until the complete extinction of the epidemic in the population, there will be less than 31 cases to come. We obtain moreover from Figure 2(d) the order of magnitude of the number of infected cattle in the population. In addition, the estimation of the infection parameter concludes to the possible existence of a minor but nonnull infection source which is not of maternal type, and which is very small (only around 3 newly infected animal per year and per infective) compared to the main source of horizontal infection until 1988 due to protein

supplements. Finally, the study of the worst-case scenario shows that even in the case of an extreme late extinction of the disease in the population, the incidence of cases will decrease quite rapidly to 0, with afterwards only 1 or 2 yearly cases occurring regularly but sparsely, with no appearance of a new peak of epidemic. We have shown with this example that the methodology developed in Section 2 provides accurate tools to study the decay phase of an epidemic under the current sanitary measures, which would help to make new policy decisions. This evaluation is all the more relevant since it is obtained not by simply computing what should most probably happen, but also by taking into account the variability of many factors (infection, incubation, and survival), and by studying the potentially most dangerous evolution.

## References


[1] V. M. Panaretos, "Partially observed branching processes for stochastic epidemics," *Journal of Mathematical Biology*, vol. 54, no. 5, pp. 645–668, 2007.

[2] P. Trapman, R. Meester, and H. Heesterbeek, "A branching model for the spread of infectious animal diseases in varying environments," *Journal of Mathematical Biology*, vol. 49, no. 6, pp. 553–576, 2004.

[3] F. Ball, "The threshold behaviour of epidemic models," *Journal of Applied Probability*, vol. 20, no. 2, pp. 227–241, 1983.

[4] F. Ball and P. Donnelly, "Strong approximations for epidemic models," *Stochastic Processes and their Applications*, vol. 55, no. 1, pp. 1–21, 1995.

[5] OIE (World Organisation for Animal Health), "Number of cases of bovine spongiform encephalopathy (BSE) reported in the United Kingdom," 2012, http://www.oie.int/en/animal-health-in-the-world/bse-specific-data/.

[6] K. B. Athreya and P. E. Ney, *Branching Processes*, Springer, New York, NY, USA, 1972.

[7] P. Haccou, P. Jagers, and V. A. Vatutin, *Branching Processes: Variation, Growth, and Extinction of Populations*, Cambridge Studies in Adaptive Dynamics, Cambridge University Press, Cambridge, UK, 2007.

[8] A. Joffe and F. Spitzer, "On multitype branching processes with $\rho \leq 1$," *Journal of Mathematical Analysis and Applications*, vol. 19, pp. 409–430, 1967.

[9] C. Jacob and S. Pénisson, "Limit models for branching processes with memory and population dependence in large populations," Technical Report UR341, INRA Jouy-en-Josas, 2012.

[10] S. Dallaporta and A. Joffe, "The $Q$-process in a multitype branching processes," *International Journal of Pure and Applied Mathematics*, vol. 42, no. 2, pp. 235–240, 2008.

[11] J. N. Darroch and E. Seneta, "On quasi-stationary distributions in absorbing discrete-time finite Markov chains," *Journal of Applied Probability*, vol. 2, pp. 88–100, 1965.

[12] T. E. Harris, "Branching processes," *Annals of Mathematical Statistics*, vol. 19, pp. 474–494, 1948.

[13] V. P. Godambe and C. C. Heyde, "Quasi-likelihood and optimal estimation," *International Statistical Review*, vol. 55, no. 3, pp. 231–244, 1987.

[14] J. R. Blum, D. L. Hanson, and J. I. Rosenblatt, "On the central limit theorem for the sum of a random number of independent random variables," *Probability Theory and Related Fields*, vol. 1, pp. 389–393, 1962-1963.

[15] M. P. Quine and P. Durham, "Estimation for multitype branching processes," *Journal of Applied Probability*, vol. 14, no. 4, pp. 829–835, 1977.

[16] P. Billingsley, *Statistical Inference for Markov Processes*, Statistical Research Monographs, Vol. II, The University of Chicago Press, Chicago, Ill, USA, 1961.

[17] C. Jacob, "Conditional least squares estimation in nonstationary nonlinear stochastic regression models," *The Annals of Statistics*, vol. 38, no. 1, pp. 566–597, 2010.

[18] R. Rebolledo, "Central limit theorems for local martingales," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 51, no. 3, pp. 269–286, 1980.

[19] B. L. S. Prakasa Rao, *Semimartingales and Their Statistical Inference*, vol. 83 of *Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC, 1999.

[20] G. A. Wells, A. C. Scott, C. T. Johnson et al., "A novel progressive spongiform encephalopathy in cattle," *Veterinary Record*, vol. 121, no. 18, pp. 419–420, 1987.

[21] M. Kaatz, C. Fast, U. Ziegler et al., "Spread of classic BSE prions from the gut via the peripheral nervous system to the brain," *The American Journal of Pathology*, vol. 181, no. 2, pp. 515–524, 2012.

[22] C. Jacob, L. Maillard-Teyssier, J.-B. Denis, and C. Bidot, "A branching process approach for the propagation of the Bovine Spongiform Encephalopathy in Great-Britain," in *Workshop on Branching Processes and their Applications*, vol. 197 of *Lecture Notes in Statistics*, pp. 225–240, Springer, Berlin, Germany, 2010.

[23] J. Bradbury, "Maternal transmission of BSE demonstrated in cattle," *The Lancet*, vol. 348, no. 9024, p. 393, 1996.

[24] C. A. Donnelly, N. M. Ferguson, A. C. Ghani, M. E. J. Woolhouse, C. J. Watt, and R. M. Anderson, "The epidemiology of BSE in cattle herds in Great Britain. I. Epidemiological processes, demography of cattle and approaches to control by culling," *Philosophical Transactions of the Royal Society B*, vol. 352, no. 1355, pp. 781–801, 1997.

[25] C. A. Donnelly, N. M. Ferguson, A. C. Ghani, J. W. Wilesmith, and R. M. Anderson, "Analysis of dam-calf pairs of BSE cases: confirmation of a maternal risk enhancement," *Proceedings of the Royal Society B*, vol. 264, no. 1388, pp. 1647–1656, 1997.