Universiteit Utrecht

Faculty of Science

Phase-Type Distributions: explorations for applications in Epidemiology

Bachelor Thesis

Eco Roman Olff

Mathematics and Applications

Supervisor:

Dr. M.C.J. Bootsma

Mathematical Institute, Utrecht University

Julius Center for Health Sciences and Primary Care, UMC Utrecht

June 11, 2020

Contents

1 Introduction

In epidemiology, "the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems"[2], there are various mathematical models to describe the spread of infectious diseases. One of the best-known models is the compartmental-based SIR-model. This model divides the population, of size N, into three groups: susceptible (S), infected (I) and recovered (or removed, based on if an individual recovers from a disease or dies) (R), which directly means a recovered individual is not susceptible again. In the SIR-model, $S(t), I(t)$ and $R(t)$, therefore, contain the amount of people that are either susceptible, infected or recovered at time t. The model describes the transition of individuals between these compartments. Whenever a susceptible individual is in contact with an infected individual, there is a probability that the susceptible becomes infected: the transmission probability. The rate of infection is most often described by the number of "successful" contacts a susceptible has on average, β, which is defined as

$$\beta = pc \tag{1.1}$$

where c is the number of contacts a susceptible individual has with an infected individual per time-step and p is the probability that a contact between a susceptible individual and an infected individual leads to a new infection. Therefore the total transition from state S to state I occurs $\frac{\beta SI}{N}$ times per time-step. Whenever an individual is infected, there is a rate of recovery γ, which is defined as the proportion of infected recovering per time-step, $1/D$ with D the time that an individual is infected. Therefore the total transition from state I to state R then occurs γI times per time-step. There is no transition from state S to state R since one can obviously not recover when not infected. The dynamics can be described in the system of ordinary differential equations 1.2:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\frac{\beta}{N}SI \\ \frac{dI(t)}{dt} &= \frac{\beta}{N}SI - \gamma I \\ \frac{dR(t)}{dt} &= \gamma I. \end{aligned} \tag{1.2}$$

This can also be written in another way for a process in discrete time (which will be further explained in chapter 2) with step size Δt , which we will use later on. The SIR-model can be given as:

$$\begin{aligned} S_{k+1} &= S_k - (\frac{\beta}{N}S_kI_k)\Delta t \\ I_{k+1} &= I_k + (\frac{\beta}{N}S_kI_k - \gamma I_t)\Delta t \\ R_{k+1} &= R_k + (\gamma I_k)\Delta t \end{aligned} \tag{1.3}$$

The transitions can also be plotted in a state transition diagram as in figure 1.

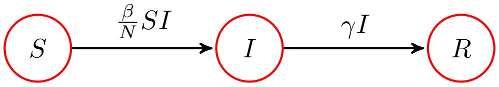

Figure 1: The state transition diagram of the basic SIR-model. The rates given above the arrows are the total transition rates for to the next compartment.

The dynamics of the SIR-model are fully determined by the state at the beginning, also known as the initial state. In this thesis we will be especially focusing on the time that an individual is infected. In the SIR-model, every time step γI individuals recover on average. Therefore for an infected individual the probability that that individual recovers in each time-step is γ and the distribution on the length of the infectious period is an exponential distribution with parameter $\gamma > 0$. However, diseases mostly do not satisfy the condition of having an exponentially distributed recovery function. The assumption of an exponentially distributed infectious period is thus often violated in reality. For example, Influenza typically has an infectious period of 5 to 7 days. If we consider a time-step of 1 day, the rate of recovery would be between $\frac{1}{5}$ to $\frac{1}{7}$. For an exponential distribution with such parameter, actually only 12% of the lengths of the infectious periods is within the 5 to 7 days range. As a possible improvement on this SIR-model, this thesis will discuss phase-type (PH) distributions. First, the definition and characteristics will be reviewed, then multiple examples of PH distributions will be given and finally the possibility of application within epidemiology will be discussed.

2 What are phase-type distributions?

Before introducing phase-type distributions, there are a few topics that need clarification. In this section we first consider discrete-time stochastic processes, Markov-chains, memorylessness and transition matrices before looking at the PH distributions itself. For this section it is assumed that probability density functions (pdf; denoted by f), mass density function (mdf; denoted by F) and exponential distributions are prior knowledge. This section uses the articles of Fackrell (2009)[3] and McClean, Garg, Barton, and Fullerton. (2010)[10].

2.1 Discrete-time stochastic processes

In daily life many processes are a description of quantities changing over time. Some widely known examples of this are:

- number of incoming calls at a helpdesk;
- the number of people waiting in a queue (for example in line at a shop or restaurant);
- stock prices.

Such processes are called stochastic processes.

Definition. *Stochastic Process*
A stochastic process is a collection of random variables $(X_t)_{t\in T}$. There are two types of stochastic processes, namely discrete-time stochastic processes or continuous-time stochastic processes:

- Discrete-time: $T = \mathbb{N} \cup \{0\} = \{0, 1, 2, \dots\}$;
- Continuous-time: $T = \mathbb{R}_+ = [0, \infty]$.

From now on we will only use discrete-time stochastic processes, therefore it will no longer be stated that a stochastic process is a discrete-time stochastic process.

2.2 Markov chains

A special and in particular rather useful type of stochastic processes are Markov chains.

Definition. *Markov chain*
A stochastic process $(X_n)_{n\geq 0}$ is called a Markov chain if:

1. it is a chain, i.e. the X_n all take values in a countable set I. We call I the state space of $(X_n)_{n\geq 0}$ and each $i \in I$ a state.
2. it satisfies the Markov property, i.e.

$$\mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_n + 1 | X_n = i_n)$$

   for all $n \geq 0$ and $i_0, i_1, \dots, i_{n+1} \in I$.

We call a Markov chain (time-)homogeneous if for $i, j \in I$, we have that

$$\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i),$$

i.e. the conditional probability does not depend on n. This transition probability will be denoted by p_{ij}. In this article only homogeneous Markov chains will be considered.

The Markov property implies *memorylessness*, i.e. the transition probability to a next state from the current state does not depend on the path to the current state. Thus the probability for the future state is independent from the path to the current state. This can also be given as:

$$\mathbb{P}(\text{future} \mid \text{present, past}) = \mathbb{P}(\text{future} \mid \text{present}).$$

A Markov chain $(X_n)_{n\geq 0}$ is defined by three characteristics:

- the state space I;
- the initial distribution $\lambda = (\lambda_i)_{i\in I}$; where $\lambda_i = \mathbb{P}(X_0 = i)$;
- the transition matrix $P = (p_{ij})_{i,j\in I}$; where $p_{ij} = \mathbb{P}(X_1 = j|X_0 = i)$.

We can shortly note this as Markov(λ,P). We notice that λ is a probability vector and therefore $\lambda_i \geq 0$ for all $i \in I$ and $\sum_{i\in I} \lambda_i = 1$. Furthermore, we notice that P is a stochastic matrix and therefore that $p_{ij} \geq 0$ for all $i, j \in I$ and that $\sum_{j\in I} p_{ij} = 1$ for all $i \in I$, which means that all columns add up to 1.

Multiplication explanation

We denote the distribution at time t as vector $\mathbf{v}_t$. Whenever the probability matrix is operated on the vector $\mathbf{v}_t$ you get $\mathbf{v}_{t+1}$, thus

$$\mathbf{v}_{t+1} = P \cdot \mathbf{v}_t. \tag{2.1}$$

In particular, we notice that $\mathbf{v}_0 = \lambda$ is the initial distribution. Consequently, the first iteration is initialised by acting the transition matrix P on the given initial distribution λ:

$$\mathbf{v}_1 = P \cdot \mathbf{v}_0 = P \cdot \lambda. \tag{2.2}$$

As an illustration, the following example of a compartmental disease model, in which there is a constant rate at which susceptible individuals become infected, can be interpreted this way; given the initial distribution where 90% of the population is susceptible and 10% of the population is infected, the initial distribution can be given as $\lambda = (0.9, 0.1, 0)^T$ (in which T means the vector is transposed). The transition probabilities are given in figure 2 below. This thus means that a susceptible individual has a 20% chance to get infected and that an infected individual has a 40% chance to recover. The corresponding transition matrix can be given as

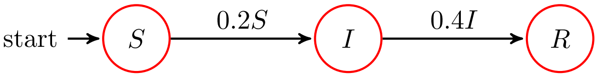

Figure 2: An example of a state transition diagram based on the SIR-model. The transition rates above the arrows are the total transition rates per time-step.

$$P = \left(\begin{array}{c|ccc} & S & I & R \\ \hline S & 0.8 & 0 & 0 \\ I & 0.2 & 0.6 & 0 \\ R & 0 & 0.4 & 1 \end{array}\right).$$

If we now consider the multiplication $P \cdot \lambda$, we see that

$$\mathbf{v}_1 = P \cdot \lambda = \begin{pmatrix} 0.8 & 0 & 0 \\ 0.2 & 0.6 & 0 \\ 0 & 0.4 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0.9 \\ 0.1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.72 \\ 0.24 \\ 0.04 \end{pmatrix}. \tag{2.3}$$

This can thus be interpreted that after 1 timestep, e.g. 1 day, that 72% of the population is susceptible, 24% of the population is infected and 4% is recovered.

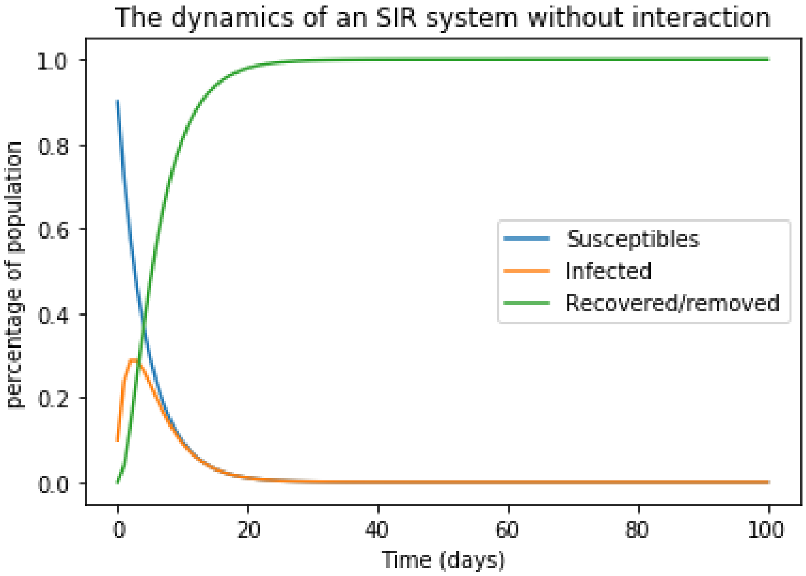

Figure 3: Dynamics of SIR model without interaction with infection rate 0.2 and recovery rate 0.4 starting with 90% susceptible individuals and 10% infected individuals.

This way, operating P repeatedly, the progression of the model can be mapped, i.e. the population distribution after n time steps can be calculated via

$$\mathbf{v}_n = P^n \cdot \lambda. \tag{2.4}$$

Corresponding to this transition matrix, a numeric simulation has been done (see figure 3) to see that with these dynamics everyone in the end has recovered.

Definition. *Absorbing and Transient states*
A state i is called *absorbing* if:

$$\mathbb{P}(X_{t+1} = i | X_t = i) = 1.$$

All states that are not absorbing are called *transient*.

Example.
As an example, the transition diagram of figure 1 can also be given in terms of a transition matrix. To make sure the transition matrix will not be too complicated in terms, we write α as the rate of infection. The corresponding transition matrix can be given as:

$$\left(\begin{array}{c|ccc} & S & I & R \\ \hline S & 1-\alpha S & 0 & 0 \\ I & \alpha S & 1-\gamma I & 0 \\ R & 0 & \gamma I & 1 \end{array}\right)$$

with $\lambda = (1, 0, 0)$. This thus suggests that when in states S and I there is a probability to remain in that state or move to the next state while in state R, the only possibility is to remain in state R. State R is therefore called an absorbing state while states S and I are called transient.

2.3 Phase-type distributions

A phase-type distribution is a distribution on a non-negative random variable. For a phase-type distribution a Markov chain with all but one transient states and one absorbing state, denoted as state 0, is needed. This distribution gives the time of going through the transient states to the absorbing state. In epidemiology this could for example measure the length of stay (LOS) in a hospital. In the example of an SIR-model as used before it gives a distribution on the time spent in the I-component as this is the transient state when translated to a PH distribution.

Definition. *Phase-type (PH) distributions*
PH distributions are defined by 3 characteristics:

- The state space $S = \{0, 1, 2, \dots, k\}$;
- The initial probability distribution $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_k) = (\lambda_0, \boldsymbol{\lambda})$, with $\sum_{i=0}^{k} \lambda_i = 1$;
- An infinitesimal generator $\mathbf{Q}$, with

$$\mathbf{Q} = \begin{pmatrix} 0 & \vec{0} \\ \vec{t} & \mathbf{T} \end{pmatrix}. \tag{2.5}$$

In this matrix $\vec{0}$ is a $1 \times k$-vector of zeros, $\vec{t}$ is a $k \times 1$-vector of absorption rates from the transient states and $\mathbf{T}$ is a $k \times k$-matrix with the transition rates between the transient states. The transition rate from state i to state j is denoted by t_{ij} and thus the absorption rate from state i is given as t_{i0} with $i = 1, 2, \dots, k$. Note that it is needed that $t_{i0} \geq 0$ for all $i = 1, 2, \dots, k$ and $t_{i0} > 0$ for at least one $i \in \{1, 2, \dots, k\}$. For $i, j \neq 0$ and when $i \neq j$ the transition rate satisfies $t_{ij} \geq 0$ and when $i = j$ the transition rate satisfies

$$t_{ii} = - \sum_{j=0, j \neq i}^{k} t_{ij}. \tag{2.6}$$

A PH distribution with k non-absorbing states (order k) and 1 absorbing state is represented by its *representation* $PH(\boldsymbol{\lambda}, \mathbf{T})$ in which $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_k)$ such that $\lambda = (\lambda_0, \boldsymbol{\lambda})$. This thus says that there are k states and the distribution has initial probability $\boldsymbol{\lambda}$ and transition matrix $\mathbf{T}$.

Distribution functions
Since PH distributions are distributions, PH distributions have their own probability density function and cumulative density function attached to it. The cumulative density function is given as:

$$F(t) = \begin{cases} \lambda_0 \text{ if } t = 0 \\ 1 - \lambda e^{Tt} \mathbf{e} \text{ if } t > 0 \end{cases} \tag{2.7}$$

for $t = 0, 1, 2, \dots$ in which $\mathbf{e} = (1, 1, \dots, 1)$ is a 1-vector of length k. The probability density function is given as:

$$f(t) = -\lambda e^{Tt} T \mathbf{e}, \tag{2.8}$$

again for $t = 0, 1, 2, \dots$ with $\mathbf{e} = (1, 1, \dots, 1)$ a 1-vector of length k.

Laplace-Stieltjes Transform
To find the mean and variation of a PH distribution, the moment generating function is of use. In order to find the moment generating function we will use the *Laplace-Stieltjes Transform.*
For every well-defined function $F(t)$ defined for $t \geq 0$, and complex number s, the Laplace-Stieltjes Transform is given as:

$$\phi(s) = \int_0^\infty e^{-st} dF(t). \tag{2.9}$$

As mentioned in equations 2.8 and 2.7 PH distributions have probability density function

$$f(t) = -\lambda e^{Tt} T\mathbf{e}$$

and cumulative density function

$$F(t) = \begin{cases} \lambda_0 \text{ if } t = 0 \\ 1 - \lambda e^{Tt}\mathbf{e} \text{ if } t > 0. \end{cases}$$

The cumulative density function can also be written as:

$$F(t) = \int_{-\infty}^{t} f(x)dx. \tag{2.10}$$

For PH distributions this is the same as:

$$F(t) = \int_{0}^{t} f(x)dx = \int_{0_+}^{t} f(x)dx + \lambda_0, \tag{2.11}$$

in which the integral from 0 to t can also be written as $F(t) = \sum_{x=0}^{t} f(x)$. Since the cumulative density function can be written this way, the Laplace-Stieltjes transform of a PH distribution can be written as:

$$\phi_+(s) = \int_{0_+}^{\infty} e^{-st} dF(t) = \int_{0_+}^{\infty} e^{-st} f(t)dt = \int_{0_+}^{\infty} e^{-st}(-\lambda e^{Tt}\mathbf{e})dt = \int_{0_+}^{\infty} -\lambda e^{(-s+T)t}\mathbf{e}dt. \tag{2.12}$$

Thus the transform is:

$$\phi(s) = -\lambda(s\mathbb{I}_k - T)^{-1}T\mathbf{e} + \lambda_0. \tag{2.13}$$

From this transform one can now derive the moment generating function for the PH distribution as:

$$\frac{d^k\phi(s)}{ds^k} = m_k = (-1)^k k! \lambda T^{-k}\mathbf{e}. \tag{2.14}$$

Since the expected value of a distribution is equal to the first moment and the variance of a distribution is equal to the second moment minus the first moment squared, we get:

$$\mathbb{E}[t] = m_1 = -\lambda T^{-1}\mathbf{e}, \qquad Var(t) = \mathbb{E}[t^2] - \mathbb{E}[t]^2 = m_2 - m_1^2 = 2\lambda T^{-2}\mathbf{e} + (\lambda T^{-1}\mathbf{e})^2 \tag{2.15}$$

3 Different types of phase-type distributions

In this section two types of PH distributions will be discussed and reviewed. The first type is the Coxian PH distribution and after that the mixture of Coxian PH distributions will be analysed. This section uses the articles of Fackrell (2009)[3] and McClean et al. (2010)[10].

3.1 Coxian Phase-Type (CPH) distributions

One of the most widely used types of PH distribution is the Coxian phase-type distribution (CPH distribution). This type of PH distribution is characterised by its composition having all but one transient states and a single absorbing state. Furthermore, all the transient states are sequential which implies that there is a single way to move through the transient states while from every transient state there is a positive probability to go to the absorbing state a. The transient states form a subset of the state space: $\mathcal{I} \subset I$. The initial distribution is $\lambda = (1, 0, 0, 0, \ldots, 0)$.
It can be noted as:

$$p_{ij} = \begin{cases} 0 \text{ if } j > i+1 \\ 0 \text{ if } j < i \\ u \text{ if } j = i+1 \\ v \text{ if } j = i \\ 1-u-v \text{ if } j = a \end{cases} \tag{3.1}$$

for all states $i \in \mathcal{T} = \mathcal{I} \setminus \{a\}$ and

$$p_{aj} = \begin{cases} 0 \text{ if } j \in \mathcal{I} \\ 1 \text{ if } j = a. \end{cases} \tag{3.2}$$

The concept of a CPH distribution can more clearly be illustrated by means of a transition diagram. In figure 4 below a CPH distribution with n transient states is given.

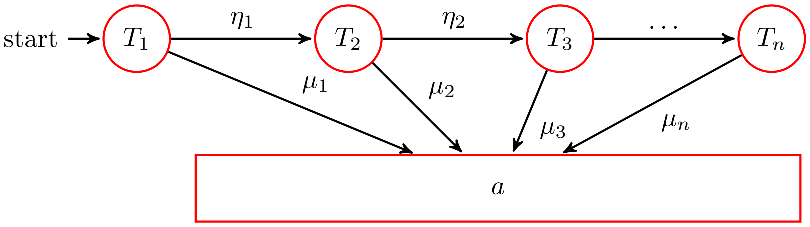

Figure 4: The state transition diagram of a Coxian phase-type distribution. Here T_i are transient states with $i = 1, 2, \ldots, n$ and a is the absorbing state.

In figure 4 we can easily see that from every transient state (T_i) there is a direct path to the absorbing state a and to the next transient state.

The CPH again has state space I, initial distribution λ and infinitesimal generator Q. The probability density function for Coxian PH distributions is given as:

$$f(t) = \mathbf{p}e^{Qt}\mathbf{q}, \tag{3.3}$$

in which $\mathbf{p} = (1, 0, 0, \ldots)$ is the specified initial distribution for CPH distributions,

$$Q = \begin{pmatrix} -(\xi_1+\mu_1) & \xi_1 & 0 & \ldots & 0 \\ 0 & -(\xi_2+\mu_2) & \xi_2 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & -(\xi_{n-1}+\mu_{n-1}) & \xi_{n-1} \\ 0 & \ldots & & 0 & -(\xi_n+\mu_n) \end{pmatrix}$$

is the infinitesimal generator (in which ξ_i is the probability $p_{i,i+1}$ ($\xi_n = 0$) and μ_i is the absorption probability for state i) for a distribution with n transient states and $\mathbf{q} = -Q\mathbf{e} = (\mu_1, \mu_2, \ldots, \mu_n)^T$ is the vector with absorption probabilities from each transient state. Since the (-) from 2.8 is a scalar (namely -1), because of the commutativity property of scalar multiplication this can be moved to $\mathbf{q}$.

To calculate the mean and variance of the CPH distribution, we can use 2.15 since the CPH distribution is a type of PH distribution. Therefore we find that for CPH distributions

$$\mathbb{E}[t] = m_1 = -\mathbf{p}Q^{-1}\mathbf{e}, \qquad Var(t) = \mathbb{E}[t^2] - \mathbb{E}[t]^2 = m_2 - m_1^2 = 2\mathbf{p}Q^{-2}\mathbf{e} + (\mathbf{p}Q^{-1}\mathbf{e})^2. \tag{3.4}$$

3.2 Mixed Coxian Phase-Type distributions

Since in many situations it is not realistic that there is a single absorbing state (think about an example of hospital charges, a person can either die, go back home or go to a residential home), a distribution with multiple different absorbing states is useful. This can be done by combining multiple Coxian phase-type distributions. A Mixed Coxian Phase-Type (MCPH) distribution is a combination of $C \geq 2$ (C of components) CPH distributions. An MCPH describes the probabilities a random variable $\mathcal{T} = (\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_C)$ in which $\mathcal{T}_c$ is a random variable which is described by a CPH distribution with infinitesimal generator Q_c. The probability density function of an MCPH distribution is given as:

$$f(\mathcal{T}) = \sum_{c=1}^{C} a_c f_c(\mathcal{T}_c) = \sum_{c=1}^{C} a_c \mathbf{p}_c e^{Q_c \mathcal{T}_c} \mathbf{q}_c, \tag{3.5}$$

where a_c is the mixing proportion of the component with $\sum_{c=1}^{C} = 1$, $\mathbf{p}_c = (a_c, 0, 0, \ldots)$ is the entry probability vector per compartment and $\mathbf{q}_c$ is the absorption vector for every compartment.

Theorem 3.2.1. *An MCPH distribution has transition probabilities* 0 *between states from different compartments.*

Proof. Let an MCPH distribution be a combination of n compartments. The transition matrix of this distribution is given as:

$$Q = \begin{pmatrix} Q_1 & 0 & 0 & \ldots & 0 \\ 0 & Q_2 & 0 & \ldots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \ldots & 0 & Q_n \end{pmatrix}$$

with

$$Q_c = \begin{pmatrix} -(\lambda_{1,c} + \mu_{1,c}) & \lambda_{1,c} & 0 & 0 & \ldots & 0 \\ 0 & -(\lambda_{2,c} + \mu_{2,c}) & \lambda_{2,c} & 0 & \ldots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -(\lambda_{(n-1),c} + \mu_{(n-1),c}) & \lambda_{(n-1),c} \\ 0 & 0 & \ldots & 0 & 0 & \mu_{n,c} \end{pmatrix}.$$

We thus see that there is no transition between states in different compartments Q_k and Q_l with $k, l \in 1, 2, \ldots, n$ and $k \neq l$. □

An MCPH distribution has $m \in \mathcal{S}$ absorbing states, in which $\mathcal{S} = I \backslash \mathcal{I}$. Therefore the absorption vector from a CPH distribution becomes an absorption matrix $q = (\mathbf{q}_1, \mathbf{q}_2, \ldots \mathbf{q}_C)^T$ (for an MCPH distribution with C compartments), in which $\mathbf{q}_c = (\mu_{c,1}, \mu_{c,2}, \ldots, \mu_{c,m})$. Furthermore we write $p = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_C)$
We can thus write the probability density function of an MCPH distribution as follows:

$$f(\mathcal{T}) = pe^{Q\mathcal{T}}q. \tag{3.6}$$

4 What are phase-type distributions used for in general?

Before we look into the use of phase-type distributions in epidemiology, other fields in which phase-type distributions are used are considered. Since PH distributions give a distribution of the time it takes until absorption, PH distributions can be widely used in a wide variety of fields. In the first part of this section we will discuss some of these fields.

4.1 Applications of phase-type distributions

When searching ”phase-type distribution” as a literal in Web of Science, 478 results appear. These results can be divided into categories (fields). This separation of articles into categories by Web of Science is shown in figure 5. Noticeable is that there are no articles refining the search on ”epidemiology” (as a topic). There are categories ”health policy services” (5 articles), ”Health care sciences services” (3 articles), ”public environmental occupational health” (3 articles) and ”medical informatics” (5 articles). Yet there is only 1 article concerning improving the SIR-model by means of phase-type distributions, found by refining the search on ”SIR” (as a topic).

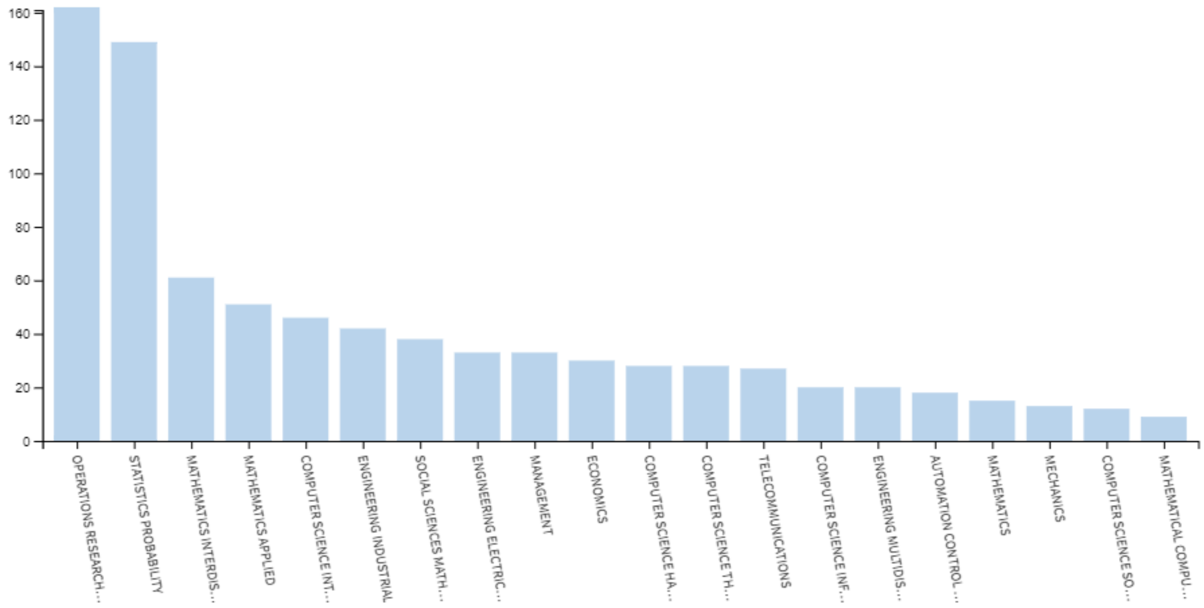

Figure 5: Overview of the use of phase-type distributions in different fields (categorised by Web of Science). The 20 fields with highest usage numbers are shown.

One can see that phase-type distributions are mainly used in operations research management science (161 times, 33.7% of all articles on PH distributions). Within this category, there is a mix of topics. Main topics in the operations research management science category are the Length of Stay (LOS) of (non-infectious disease) patients in hospitals and industrial engineering. In this part we discuss some examples which can be translated to infectious disease modeling.

Length of Stay
In the article of Fackrell (2009)[3], different phase-type distributions had been modeled to fit data about the length of stay of patients in the Royal Melbourne Hospital that had been transferred from other hospitals. These patients were not specifically known for having an infectious disease. In figure 6 below, one can see that PH distributions are capable of fitting to such a distribution.

Fackrell[3] concludes that a general PH distribution of order 6 is more appropriate for constructing such distribution than Coxian PH distributions until order 25. Since the distribution functions for the PH distri-

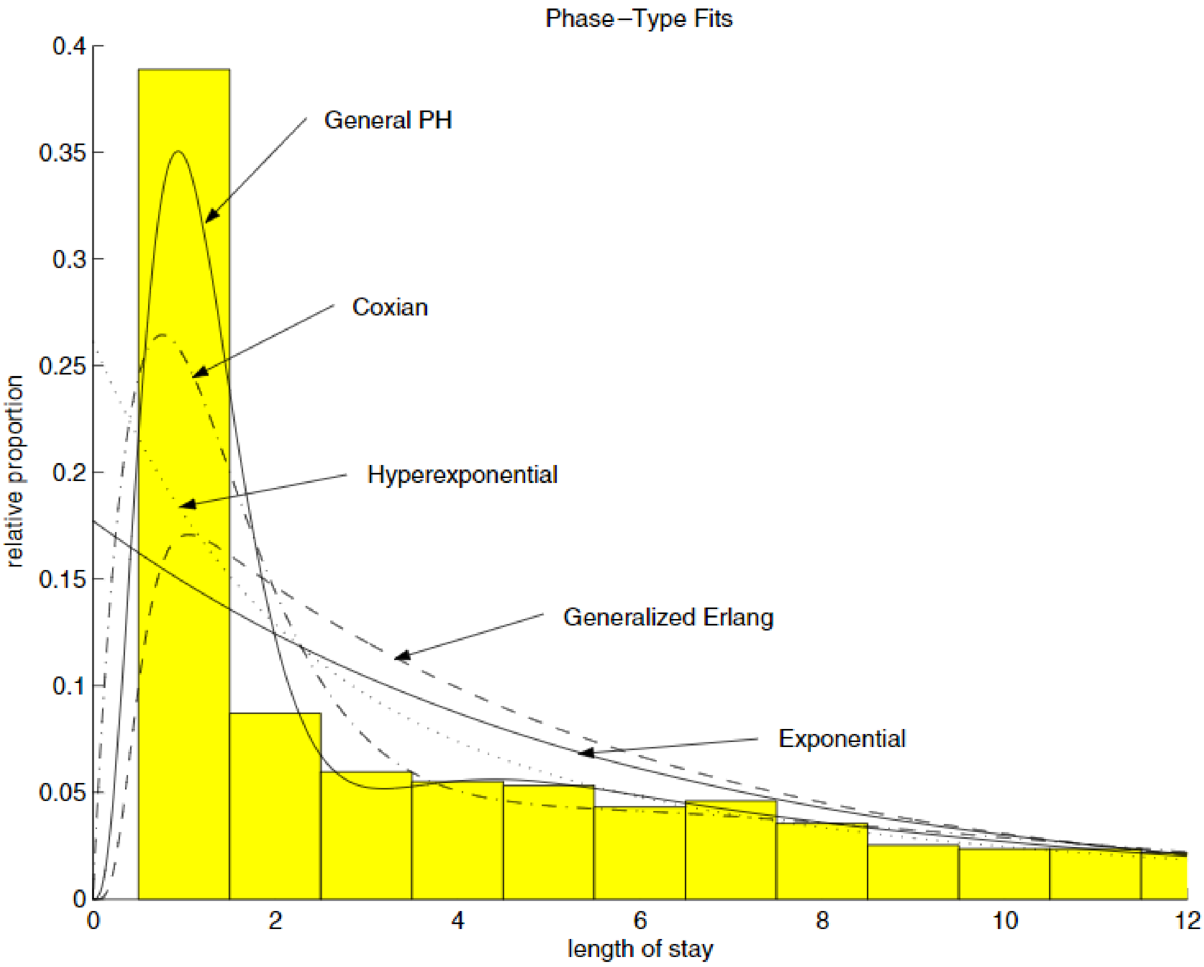

Figure 6: Probability densities of different PH distributions are plotted to data of the Royal Melbourne Hospital. This figure originates from Fackrell (2006)[3].

butions are similar to that of the length of the infection or infectious period, this might indicate that the use of PH distributions might be useful.

4.2 Phase-type distributions in epidemiology

In the article of Zhu and Chen (2020)[12] assumptions were made on the shape of the distribution functions for the incubation period and the infectivity of COVID-19 as shown in figure 7 below.

Figure 7[12] gives an intuition on the distribution function for infectious diseases. This distribution function has a somewhat similar shape to that of the LOS of figure 6. Therefore the use of PH distributions in SIR models to improve the part of the rate of recovery might be beneficial.
Many processes in epidemiology and medical care are processes that happen until arriving in some absorbing state. An example of such a process is the a bacterial outbreak in a hospital.[1] The outbreak time can then be modelled by a phase-type distribution. During the outbreak one can also consider looking at the processes where doctors get infected and recover. Furthermore, interventions taken during the process may have an impact. In the article of Castro, López-García, Lythe, and Molina-París (2018)[1], they consider the interventions of screening all the healthcare workers and when testing positive sending them off-duty (which

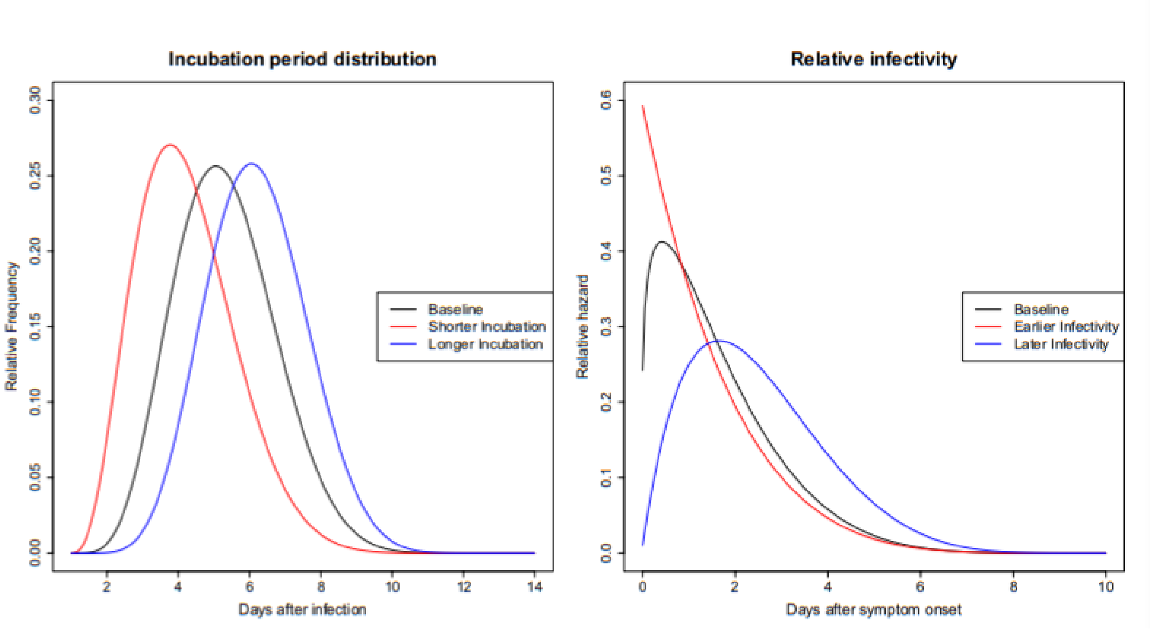

Figure 7: Assumption on the incubation period and infectivity of COVID-19 made by Zhu and Chen (2020)[12].

is different state in the PH distribution) as well as giving them a treatment.

Hospital capacity is another issue which may benefit from the use of PH distributions. To describe the LOS of patients, in many cases Coxian PH distributions are used, while sometimes a mixture of CPH distributions is used. To describe the LOS of elderly patients in a hospital, a CPH distribution with two absorbing states (one for death and one for discharge) has been set up.[6] A survival tree has been made with the population split up using the Akaike information criterion. The CPH distribution has then been used to estimate the LOS for all groups of the population.[6] This same type of distribution on LOS has been used to estimate either the patient survival[9] or the mortality rate in hospitals[11], to estimate the cost of geriatric patients in Northern-Ireland[8] and to cluster patients based on their LOS.[4]

In another study, a mixture of CPH distributions have been used to model the readmission of patients to a hospital.[5] Since elderly patients require an increasing number of hospital beds, a more thorough understanding of the movement of elderly patients between hospital uptakes and discharge to society is needed to better monitor the number of available hospital beds (for non-elderly patients) and with that the waiting times and hospital experience. To match the data about readmissions and discharges of elderly patients, a phase-type distribution with consecutively three stages of care in the initial hospital uptake, two stages of care in the community and three stages in the first readmission has been created. In this model, there is no transition between non-final states of each stage to the next stage. Gordon, Marshall, and Cairns (2016)[5] found that this mixture of CPH distributions fitted the data of these elderly patients better. As a further improvement of the model, in addition to the mixture of CPH distributions, a conditional (Bayesian) segment was added. This segment made it possible to condition the time until an event occurs to be dependent on the time since leaving the earlier stage.

To conclude

In future research the use of Coxian phase-type distributions in epidemiology needs to be examined further in order to use it to its full potential. Having only one result[7] in a search on PH distributions as an improvement of the SIR-model in Web of Science when combining the search terms ”phase-type distribution” and ”infection” and no results when combining ”phase-type distribution” and ”epidemiology”, this is an application of PH distributions that requires more attention. Currently, being governed by the 2019-2020

Covid-19 (SARS-CoV-2) pandemic, this especially is a topic which concerns all of society. By using data from former outbreaks like the SARS outbreak in 2003 or the MERS outbreak in 2012, it might be manageable to create a PH distribution which describes the length of the outbreak, length of infections, et cetera. This thesis aimed to provide insight in the possibility of using PH distributions in epidemiology. Therefore this study provides a starting point for further research on analysing the application of a PH distribution in infection diseases such as the SARS-CoV-2 pandemic. If this is possible, future research on this topic might lead to a different and improved insight in and dealing with future pandemics.

References

[1] Castro, M., López-García, M., Lythe, G., & Molina-París, C. (2018). First passage events in biological systems with non-exponential inter-event times. *Scientific reports*, 8(1), 1-16.

[2] Dicker, R. C., Coronado, F., Koo, D., Parrish, R. G. (2006). *Principles of epidemiology in public health practice; an introduction to applied epidemiology and biostatistics.*

[3] Fackrell, M. (2009). Modelling healthcare systems with phase-type distributions. *Health care management science*, 12(1), 11.

[4] Garg, L., McClean, S., Meenan, B. J., & Millard, P. (2011). Phase-type survival trees and mixed distribution survival trees for clustering patients' hospital length of stay. *Informatica*, 22(1), 57-72.

[5] Gordon, A. S., Marshall, A. H., & Cairns, K. J. (2016). A conditional approach for modelling patient readmissions to hospital using a mixture of Coxian phase-type distributions incorporating Bayes' theorem. *Statistics in medicine*, 35(21), 3810-3826.

[6] Gordon, A. S., Marshall, A. H., & Zenga, M. (2018). Predicting elderly patient length of stay in hospital and community care using a series of conditional Coxian phase-type distributions, further conditioned on a survival tree. *Health care management science*, 21(2), 269-280.

[7] Lefèvre, C., & Simon, M. (2016). SIR epidemics with stages of infection. *Advances in Applied Probability*, 48(3), 768-791.

[8] Marshall, A. H., Shaw, B., & McClean, S. I. (2007). Estimating the costs for a group of geriatric patients using the Coxian phase-type distribution. *Statistics in medicine*, 26(13), 2716-2729.

[9] Marshall, A. H., & Zenga, M. (2009). Simulating Coxian phase-type distributions for patient survival. *International Transactions in Operational Research*, 16(2), 213-226.

[10] McClean, S., Garg, L., Barton, M., & Fullerton, K. (2010, October). Using mixed phase-type distributions to model patient pathways. *In 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 172-177). IEEE.

[11] Zenga, M., Marshall, A. H., Crippa, F., & Mitchell, H. (2016). The coxian phase-type distribution as a contribution to the multilevel model of in-hospital mortality. *Communications in Statistics-Theory and Methods*, 45(6), 1819-1830.

[12] Zhu, Y., & Chen, Y. Q. (2020). On a statistical transmission model in analysis of the early phase of covid-19 outbreak. *Statistics in Biosciences*, 1.

A Appendix

Listing 1: Python code for figure 3

```
import numpy as np
import matplotlib.pyplot as plt

# The SIR-model
def base_sir_model(init_vals, params, t):
    S_0, I_0, R_0 = init_vals
    S, I, R = [S_0], [I_0], [R_0]
    beta, gamma = params
    dt = t[1] - t[0]
    for _ in t[1:]:
        next_S = S[-1] - (beta*S[-1])*dt # Dynamics of susceptible group
        next_I = I[-1] + (beta*S[-1] - gamma*I[-1])*dt # Dynamics of infected group
        next_R = R[-1] + (gamma*I[-1])*dt # Dynamics of recoverd group
        S.append(next_S)
        I.append(next_I)
        R.append(next_R)
    return [S, I, R] # Return the list of values with new time step added


# Define parameters
t_max = 100 # The time to model
dt = 1 # Time step
t = np.linspace(0, t_max, int(t_max/dt) + 1)
N = 100 # Number of individuals within the population
init_vals = 0.9, 0.1, 0 # Initial values

beta = 0.2 # Infection rate
gamma = 0.4 # Rate of recovery
params = beta, gamma
# Run simulation
results = base_sir_model(init_vals, params, t)

max_index = float("-inf")
max_value = float("-inf")

# Define x-axis as the time
time = [i for i in range(len(results[1]))]
time2 = [i for i in range(len(results[1]))]

# line S points
y1 = results[0]
# plotting the line S points
plt.plot(time, y1, label = "Susceptibles")
# line I points
y2 = results[1]
# plotting the line I points
plt.plot(time2, y2, label = "Infected")
# line R points
y3 = results[2]
# plotting the line R points
plt.plot(time2, y3, label = "Recovered/removed")
# Name x-axis as time
plt.xlabel('Time␣(days)')
# Name y-axis as the percentage of population
plt.ylabel('percentage␣of␣population')
# Name the figure
plt.title("The␣dynamics␣of␣an␣SIR␣system␣without␣interaction")
# show a legend on the plot
plt.legend()
# Display the figure.
plt.show()
```