Lappeenranta University of Technology
School of Engineering Science
Degree Program in Computational Engineering and Technical Physics / Technomathematics

Veronika Malakhovskaia

MODELLING APPROACHES FOR EPIDEMIOLOGICAL MODELS

Examiners: Professor Heikki Haario
Professor Matti Heiliö

Supervisors: Professor Heikki Haario

ABSTRACT

Lappeenranta University of Technology
School of Engineering Science
Degree Program in Computational Engineering and Technical Physics / Technomathematics

Veronika Malakhovskaia

Modelling approaches for epidemiological models

2017

44 pages, 19 figures, 6 tables.

Examiners: Professor Heikki Haario
Professor Matti Heiliö

Keywords: MCMC, Monte Carlo, epidemiology, Markov chain, SIR-model, agent-based modeling, Gillespie algorithm

Predicting the spread of diseases remains an important aspect for epidemiological problems. For these purposes, such areas as statistics, sociology and biology are used. Nowadays, there is a large number of approaches for modelling epidemiological problems using compartmental models.

The aim of this paper is to obtain solutions for a mathematical model of dynamics of an influenza epidemic using various methods and compare them to identify the optimal method for solving this problem. A number of the algorithms such as Markov Chain Monte Carlo method (MCMC) for the deterministic modelling and Gillespie algorithm and a discrete approach for the agent-based modelling (ABM) were implemented with Matlab, a numerical computing environment. The data for implementing the algorithms is taken from the case of influenza epidemic in British Boarding School in 1978.

PREFACE

First, I would like to express my deep gratitude to my supervisor Heikki Haario for help, support, patient guidance and interesting research area.

Second, I would like to thank Southern Federal University and particularly the Faculty of Mathematics, Mechanics and Computer Sciences in the persons of Mikhail Karyakin and Konstantin Nadolin. They provided me the great opportunity to continue my study in Lappeenranta University of Technology and obtain a double degree. Also I would like to thank my supervisor Elena Shiryaeva and the head of the department of Computational Mathematics and Physics Mikhail Zhukov for helping me during my work on the diploma.

Third, I am infinitely grateful to my parents for the opportunity to study at Lappeenranta University of Technology and their support and care for me throughout my studies.

Finally, I would like to thank all the people who helped and encouraged me during the studies.

Lappeenranta, 2017

Veronika Malakhovskaia

List of Figures

List of Tables

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

PDF	Probability Density Function
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings Algorithm
AM	Adaptive Metropolis algorithm
DR	Delayed rejection method
ODE	Ordinary differential equation
ABM	Agent-based modelling
S	Susceptible individuals
E	Exposed individuals
I	Infected individuals
R	Recovered with immunity

NOTATIONS

n	Number of measurements
θ	Unknown parameters
y	Measurements
ϵ	Measurement error
$f(x;\theta)$	Model with constants x and unknown parameters θ
$\pi(\theta\|y)$	Posterior distribution
$p(\theta\|y)$	Probability density function
$l(y\|\theta)$	Likelihood function
$\pi_{pr}(\theta)$	Prior distribution
$p_Y(y)$	Normalizing factor
$\pi(\hat{y}\|y)$	Predictive distribution
$SS(\theta)$	Sum of squares
θ^0	Arbitrary point
$\beta(\theta^*, \theta^0)$	Acceptance ratio
C_0	Initial covariance
s	Covariance scaling factor
ϵ_r	Regularization parameter
n_0	Non-adapting period

1 INTRODUCTION

1.1 Background

Modeling of diseases is a tool that is used to study the mechanisms of spreading the epidemic, predicting the future course of the disease and evaluating strategies for fighting the epidemic. Since it is important to find out the development of the disease, a science was created that is called "epidemiology".

Epidemiology is a general medical science that studies the patterns of the occurrence and spread of diseases for developing preventive measures. The subject of studying epidemiology is the *incidence* – a set of cases of the disease in a certain area at a certain time among a certain population group.

Major areas of epidemiology include determining the causes of the disease, outbreak investigation and comparisons of treatment effects. Epidemiologists rely on other scientific disciplines, such as biology, statistics and social sciences for the effective use of the data and for obtaining relevant conclusions.

The first attempts to quantify the causes of mortality were made in 1662 by John Graunt. He analyzed the lists of causes of deaths that were periodically published in free access. The analysis of the Graunt is considered as the beginning of the "theory of competing risks". The earliest study of mathematical modeling of the spread of the disease was conducted in 1766 by Daniel Bernoulli. Bernoulli created a mathematical model to defend the practice of inoculating against smallpox [1]. The modern theory of epidemiology began with the study of malaria diseases. Sir Ronald Ross was the first scientist to consider cases of malaria. A simple deterministic model was formulated by A. G. McKendrick and W. O. Kermack in 1927 [2]. Nowadays different epidemic models are widely studied for improving the predictive function of the spread of the various diseases.

There are two main types of the epidemic models: deterministic and stochastic. A *stochastic* model is used for estimating probabilistic distributions of potential outcomes, allowing a random change in one or more input data over time. The spread of the disease is considered as a random process. This type of the models is used when the fluctuations of exposure or disease dynamics are important, especially in case of small group of the individuals.

A *deterministic* model is used in cases of large populations when the fluctuations are not

so significant. In this type, individuals in the population belongs to different compartments. All these subgroups represent a unique stage of the epidemic. To indicate the stages of the disease, a special common terminology is used. In general, a deterministic model can be described by ordinary differential equations or by partial differential equations.

A simple epidemic model, according to Daley & Gani [2], is the case when there are only two compartments of the population: susceptible individuals and infected ones. If an individual become infected, he will not return to the susceptible group any more.

More complex models include such compartments as recovered and exposed individuals. After catching the infection, an individual may be in the exposed subgroup or directly goes to the infected compartment. After being in infected subgroup an individual can go to the group of recovered population. If the epidemic involves a fatal outcome, one can add a group of removed individuals.

1.2 Research questions

In this thesis the problem of implementing different epidemic models and modelling approaches is studied. The main goals can be described as follows:

- understanding of the basic theory of the epidemic modelling and behaviour of the disease;
- implementation basic deterministic model for epidemiological problem based on real data using Markov Chain Monte Carlo methods;
- implementation agent-based model and its comparison with the deterministic one;
- implementation the Gillespie algorithm and its comparison with deterministic one;
- determination of the most optimal approach for the modelling of epidemiological problems

1.3 Outline

This thesis is divided into five chapters. The first chapter contains the introduction in the basic epidemiology and epidemic models. Also it includes the research questions of the work.

Introducing the basic idea of the Bayesian inference and the Bayes' rule is considered in the chapter 2. Furthemore, the basic ideas of the MCMC methods are given in the second chapter.

The third chapter contains the implementation of the stochastic agent-based models to the given epidemiological model based on real data from the case of influenza epidemic in the English Boarding School, 1978. The Gillespie algorithm is considered as one of the possible modelling approaches.

The final part of the work presents results, discussions and conclusions.

2 Bayesian estimation and MCMC methods

2.1 Bayesian inference in parameter estimation

In Bayesian analysis we are interested in estimation of the model parameters, we use the following model that is presented in equation (1). The model consists of n sample measurements $y = (y_1, y_2, ..., y_n)$, known constants or variables $x = (x_1, x_2, ..., x_k)$, the vector of unknown parameters $\theta = (\theta_1, \theta_2, ..., \theta_k)$ and measurement error ϵ:

$$y = f(x; \theta) + \epsilon. \tag{1}$$

The Bayesian solution to the problem of parameter estimation is the posterior distribution of the parameters. In other words, it is a conditional probability distribution of the unknown parameter θ. We are interested in the posterior distribution $\pi(\theta|y)$ with probability density function, where θ, as mentioned above, is the unknown parameter values and y is vector of observations.

To define $\pi(\theta|y)$ we consider a *probability density function* (PDF) $p(\theta; y)$. This function gives the probability for each combinations of the parameters. We can write PDF as:

$$p(\theta; y) = l(y|\theta)\pi_{pr}(\theta), \tag{2}$$

where $\pi_{pr}(\theta)$ is a prior distribution of known parameters. The likelihood function $l(y|\theta)$ gives the probability for obtaining data y with the parameter value θ. If for two possible parameter values θ_1 and θ_2 we see that $l(y|\theta_1) > l(y|\theta_2)$, it means that the observation y is more likely to happen under the parameter θ_1 [3]. The likelihood is a function of the model parameter which can be determined as the product:

$$l(y|\theta) = \prod_{i=1}^{n} l(y_i|\theta), \tag{3}$$

in case the measurement errors ϵ_i are independent.

For obtaining probability density function (PDF) we should normalize the probability. That

means the probabilities sum must be equal to 1. This normalizing factor can be written as $p_Y(y)$. According to the rule of total probability, we can calculate this value using the integration over all values of vector θ [3]:

$$p_Y(y) = \int l(y|\theta)\pi_{pr}(\theta)d\theta. \tag{4}$$

After these notations, the posterior density can be written in the Bayes rule formula:

$$\pi(\theta|y) = \frac{l(y|\theta)\pi_{pr}(\theta)}{p_Y(y)}. \tag{5}$$

Note that in terms of theory of probability this formula can be rewritten as a probability of two random variables A and B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}. \tag{6}$$

The problem of how the prior distribution $\pi_{pr}(\theta)$ will affect the posterior distribution has a significant impact. The reason is that the prior contains the information which is available before obtaining new data.

One of the crucial parts in Bayesian inference is the prediction of future observations $\hat{y}$ given the current ones y. The prediction is based on the posterior distributions of θ. Assuming that y and $\hat{y}$ are conditionally independent given the value of θ, the predictive distribution $\pi(\hat{y}|y)$ can be written as:

$$\pi(\hat{y}|y) = \int l(\hat{y}|\theta)\pi(\theta|y)d\theta. \tag{7}$$

The distance between the model and the observations can be measured with the sum of squares:

$$SS(\theta) = \sum_{i=1}^{N}(y_i - f(x_i, \theta))^2. \tag{8}$$

Some iterative optimization methods can be used for minimizing the sum of squares [4].

2.2 Monte Carlo methods for parameter estimation

In Bayesian analysis for unknown parameters we are interested in forming the posterior distribution for the parameters. Since it is rather difficult to do this analytically, we have to do it numerically.

Monte Carlo methods (MC) are a class of stochastic algorithms for sampling from a probability distribution to investigate a certain problem. MC methods are widely used for calculating numerical approximations of multi-dimensional integrals in different branches of science. According to Laine [3], the MCMC algorithm that is used in the simulation ensures that the chain will take values in the domain of the vector of unknown parameters θ.

For this work a class of *Markov chain Monte Carlo methods* (MCMC) is used for obtaining the samples from the probability distribution which is based on constructing a Markov chain. The state of the chain after a number of steps is basically used as a sample of the desired distribution. The quality of the sample is considered as a function of the number of steps.

Markov chains are used for generating a sequence of random variables $X1, X2, X3, ...$ with the Markov property. This property can be written as "the probability of moving to the next state depends only on the present state and not on the previous states":

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_{n+1} = x | X_n = x_n), \tag{9}$$

if all the probabilities are well-defined, such that $P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) > 0$.

The classical and widely used Monte Carlo methods are the random walk methods. There are several types of this method, for example, Gibbs sampling (requires that all the conditional distributions of the target distribution should be sampled exactly). Also the slice sampling (where one can sample from a distribution by sampling uniformly from the region under the plot of its density function) can be used for creating the random walk. But the most known algorithm is random walk Metropolis-Hastings algorithm (MH).

This method obtains a sequence of random samples from a probability distribution for which direct sampling is difficult. This sequence can be used to approximate the distribution. Gen-

erally, this type of MCMC algorithms is used for sampling from high-dimensional distributions.

The basic steps of the MH algorithm can be represented as following:

1. Choose an arbitrary point θ^0 as the starting sample.
2. New point θ^* is obtained from the proposal distribution. The sample is depend on the previous point of the chain θ.
3. For every iteration step calculate the acceptance ratio as the probability:

$$\beta(\theta^*, \theta^0) = \min\left(1, \frac{\pi(\theta^*)}{\pi(\theta^i)}\right), \tag{10}$$

   where π is a posterior density. If the point is rejected, algorithm should proceed with the previous point of the chain. The procedure is to be repeated until the desired distribution are obtained.

The Metropolis-Hastings algorithm assumes that the target distribution is the stationary distribution of the Markov chain. That means that the values generated will eventually the posterior distribution $\pi(\theta|y)$ [3].

There are several assumptions that make up a significant role in the Metropolis-Hastings algorithm. First, it should be assumed that there is a symmetric proposal distribution. It means that the probability of changing from the current point to the proposed one is the same as the changing in the opposite way. Despite some cases where non-symmetric distribution can be used, the symmetric one remains the most used option.

Second, one of the most crucial points of the algorithm is the way of selecting the distribution. Relative ease in sampling and similarity to the target distribution should be considered. Also it should be mentioned that inappropriate implementation of the algorithm can provide high computational cost. For example, if the distribution is sufficiently small, it will take more time for cover the area of distribution.

It should be noted that in calculations we can remove the normalization constant, since we need only find out the ratios of the posterior densities. Thus, in the multidimensional case of the parameter estimation, the algorithm is computationally feasible [4].

2.3 Adaptive MCMC method

The ability to choose the proposal distribution has an influence on the Metropolis-Hastings algorithm. One of the problem in adapting the proposal distribution, according Laine [3], is when the accepted values become depend on the history of the chain, the standard convergence results do not apply because the process stops to be Markovian. This problem can be solved by discarding those part of the chain for which adaptation has already been used. But it this solution provides the increasing of computational cost.

Haario et al. [5] considered adaptive method where the covariance is adapted with history of chain generated so far [3]. An algorithm of the Adaptive Metropolis method (AM) is presented below:

1. Choose an initial value θ^0, initial covariance $C = C_0$, covariance scaling factor s, non-adapting period n_0 and regularization parameter ϵ_r.
2. Propose a new θ^* using Gaussian distribution centred at the current value $N(\theta^{i-1}, C)$.
3. For every iteration i calculate the acceptance ratio of MH algorithm and decide if the θ^* is accepted or not.
4. After the end of non-adapting period n_0, adapt the proposal covariance matrix

$$C = cov(\theta^0, \ldots, \theta^i)s + I\epsilon_r. \tag{11}$$

5. Repeat steps 2-4 until desired number of values has been generated.

Also the delayed rejection method (DR) can be used for making the distribution. In this method a second move is proposed instead of retaining the same position. The acceptance probability of the second candidate is computed according the preservation of the Markov chain convertibility [3].

The first stage acceptance probability can be computed as the standard MH acceptance ratio given in the formula (10). If the point θ^* is rejected, the second point is obtained with new acceptance ratio:

$$\beta_2(\theta, \theta^*, \theta^{**}) = \min\left(1, \frac{\pi(\theta^{**})[1 - \beta_1(\theta^{**}, \theta^*)]}{\pi(\theta)[1 - \beta_1(\theta, \theta^*)]}\right), \tag{12}$$

where β_1 is the standard acceptance ratio of the Metropolis-Hastings algorithm.

As, according to Laine [3], the property of Markov chain convertibility is preserved, the DR method leads the stationary distribution π as the MH algorithm.

As one can combine AM and DR methods, the new method DRAM is obtained [3]:

1. Select an initial value θ^0, initial covariance $C^{(1)} = C_0$, covariance scaling factor s, non-adapting period n_0 and regularization parameter ϵ_r. Choose the scalings for the higher-stage proposal covariances C^i, $i = 1, \ldots, N_{tr}$. N_{tr} is the amount of tries that is allowed.

2. Until the acceptance of a new point or N_{tr} tries have been made:

   Propose a new θ^* using Gaussian distribution centred at the current value $N(\theta^{i-1}, C^(k))$.

   Accept or reject the point according to the k^{th} stage of acceptance ratio β.

3. Set $\theta^i = \theta^*$ if the point is accepted, set $\theta^i = \theta^{i-1}$ if the point is rejected.

4. After the end of non-adapting period n_0, adapt the covariance matrix

$$C^{(1)} = cov(\theta^0, \ldots, \theta^i)s + I\epsilon_r. \tag{13}$$

5. Repeat steps 2-4 until desired number of values has been generated.

2.4 Case of seal virus

In the end of the 1980-s the seals in the North Sea were infected with the phocine distemper virus. This case is used for providing a simple example of the SIR compartment model with the analysis made by MCMC methods. The seals population is divided into the four groups according to their state in the epidemic. S is susceptible individuals, I is infected seals, H is healed seals and R is removed ones.

The mechanism of the epidemic can be described like this: susceptible individuals interact with the infected seals which are not isolated from the other members of the group. After that susceptible seal became infected. These individuals can either die or heal. If they are healed, they will become protected from the virus and cannot be infected again.

With this notations and assumptions this SIR model can be given in terms of the system of differential equations:

$$\frac{dS}{dt} = -\alpha SI, \tag{14}$$

$$\frac{dI}{dt} = \alpha SI - \beta I, \tag{15}$$

$$\frac{dR}{dt} = (1-\nu)\beta I, \tag{16}$$

$$\frac{dH}{dt} = \nu\beta I, \tag{17}$$

where α is the contact rate, β is the removal rate, ν is the survival rate. These rates and the initial number of the infected population are unknown parameters θ that we need to determine: $\theta = [\alpha, \beta, \nu, I(0)]$. The initial population of susceptible seals is known for simplifying the model.

Table 1 provides the information about the parameters of the given epidemic model (14)–(17).

Table 1. Parameters for the model of the "toy-case".

α	Contact rate	Estimated
β	Removal rate	Estimated
ν	Survival rate	Estimated
S_0	Initial population of the susceptible individuals	3400
I_0	Initial population of the infected individuals	10
R_0	Initial population of the removed individuals	0
H_0	Initial population of the healed individuals	0

The resulting time evolution for the susceptible, infected, removed and healed compartments is shown in Figure 1. As one can see, the model obtained is well fitted to the known data, in particular, the data of the number of removed individuals.

The distribution of parameters and model response which were obtained by using the MCMC are presented in Figures 2 and 3. It can be seen that identifiability of components is not so

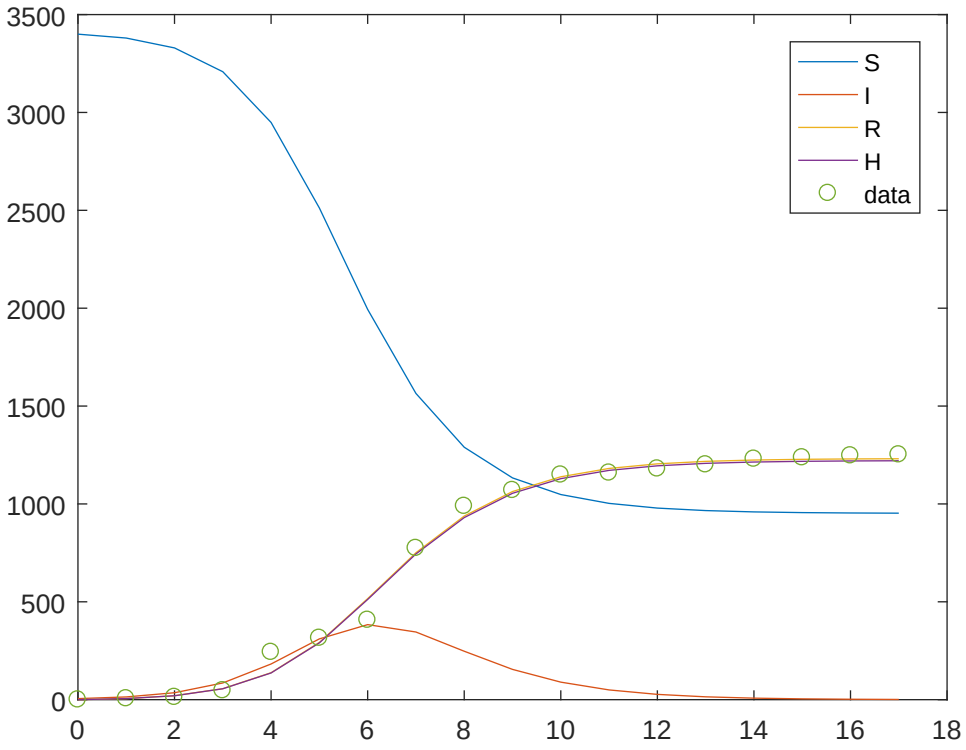

Figure 1. Graph of the compartments.

good, except the component that was measured. The explanation can look like follows: we have no information about the mortality data and how it can be produced. As Solonen et al. [6] claim, there are at least two ways in which the mortality rate can be represented. Without this prior knowledge the parameters of the model cannot be identified in a proper way.

Also, according to the Figure 2, there is strong correlation between the contact rate α and the removal rate β. So, the basic reproduction number r can be estimated as $r = \frac{\alpha}{\beta}$. This parameter has the significant value for the epidemiological problems as it represents the the mean number of cases which are obtained by the infected individual over the course of its infectious period [6]. The identifiability of the basic reproduction number r in the SIR model was considered, for example, in [7]. Moreover, Kypraios [8] provided the concept that the basic reproduction number r in a stochastic SIR model can be estimated from partial observations, and the infection and removal rates can be unidentified.

According this example, Markov Chain Monte Carlo methods can be used for estimating the parameters of the given model when only limited number of observations is available Moreover, this approach can be used to determine what type of observations would be effective for estimating the model parameters [9].

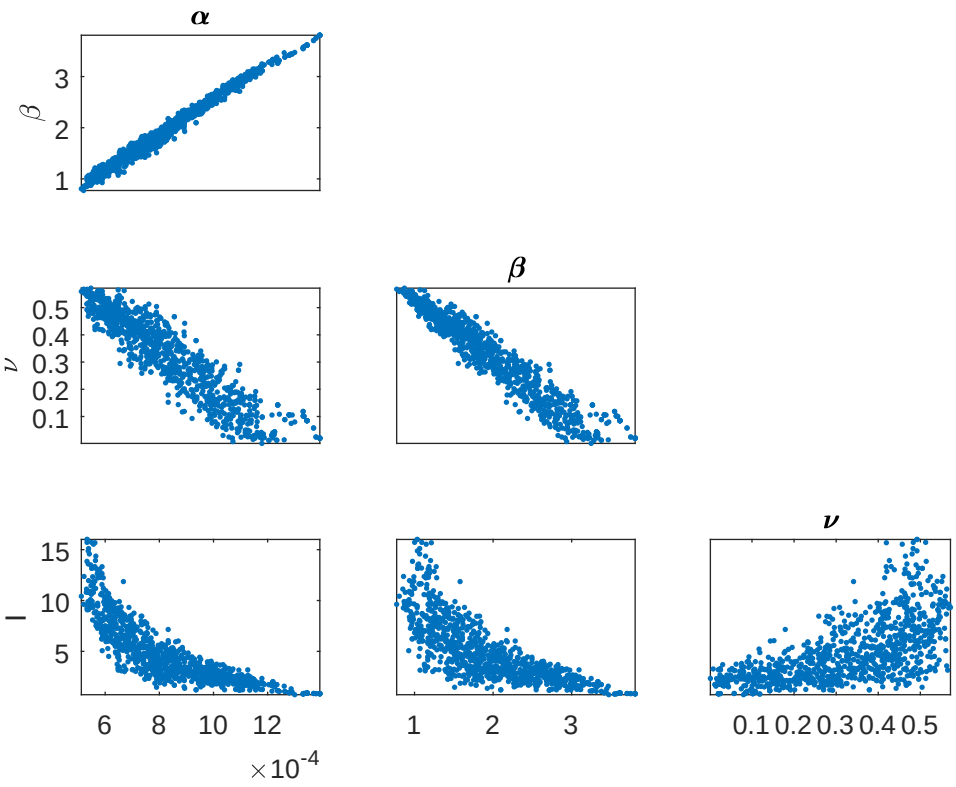

Figure 2. Parameter posterior sample produced by MCMC method.

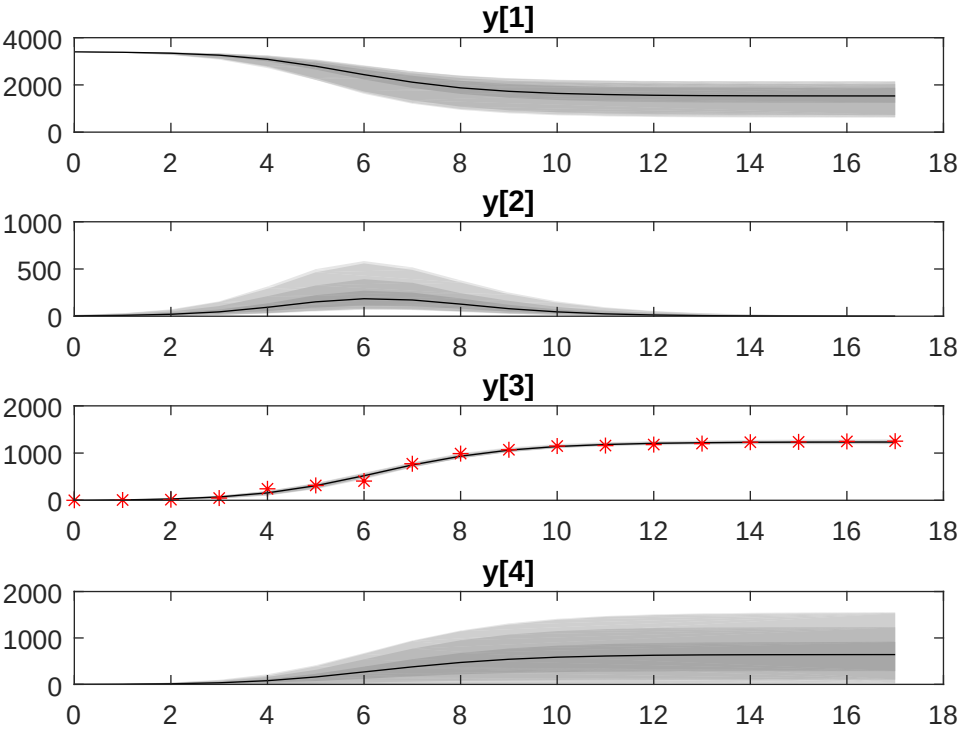

Figure 3. Predictive distributions for the compartments made by MCMC method. Grey regions show the confidence intervals of the response. Black lines shows the mean response.

2.5 Case of influenza epidemic in the British Boarding School

For the thesis an example based on real data was used for estimating parameters of the SIR model. This case was considered as a simple epidemic model by Murray in his book [10].

In 1978 there was an influenza epidemic in the British Boarding School. British medical journal provided the information on development of the epidemic. The total amount of the boys (between the ages of 10 and 18, all except 30 being full boarders [11]) in a school was equal to 763. Of this number, 512 students were infected during the two-week period of the epidemic.

Murray [10] supposed that one infected boy started the disease in the school. Table 2 gives the statistics of the infected individuals from the British Boarding school, 1978.

Table 2. Number of the infected students during the period on the influenza epidemic

Days	Number of infected boys
0	1
1	3
2	7
3	25
4	72
5	222
6	282
7	256
8	233
9	189
10	123
11	70
12	25
13	11
14	4

The compartmental diagram of the given SIR model are represented on the Figure 4.

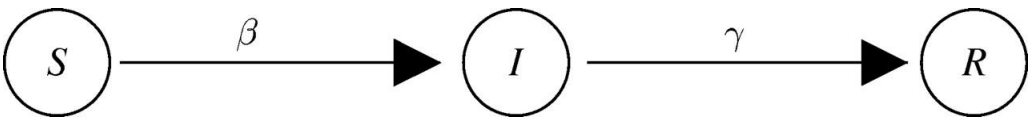

Figure 4. Compartmental diagram of the influenza epidemic.

Here β is a contact rate and γ is the recovery rate.

The model of the disease can be described by the system of three ordinary differential equations (ODE):

$$\frac{dS}{dt} = -\beta SI, \tag{18}$$

$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{19}$$

$$\frac{dR}{dt} = \gamma I \tag{20}$$

where β is a contact rate and γ is the recovery rate. For this case initial conditions are assumed to be known: $S_0 = 762$, $I_0 = 1$, $\beta = 0.002$ and $\gamma = 0.4$.

Figure 5 represents the evolution of epidemic during the two-week period. Figure 6 represents how the data fit the model. The graph shows that the data fit the model quite well. Figure 7 represents the distribution of the parameters α and β. It can be seen, that the MCMC chain obtained is stable and the acceptance rate here is pretty high. Unlike in the previous example, there is no a strong correlation between the parameters in the Figure 8. Figure 9 shows the predictive distributions. It can be seen that in this case the confidence region is not wide. So, there is no problem with the identifiability of parameters unlike in the example of seal population.

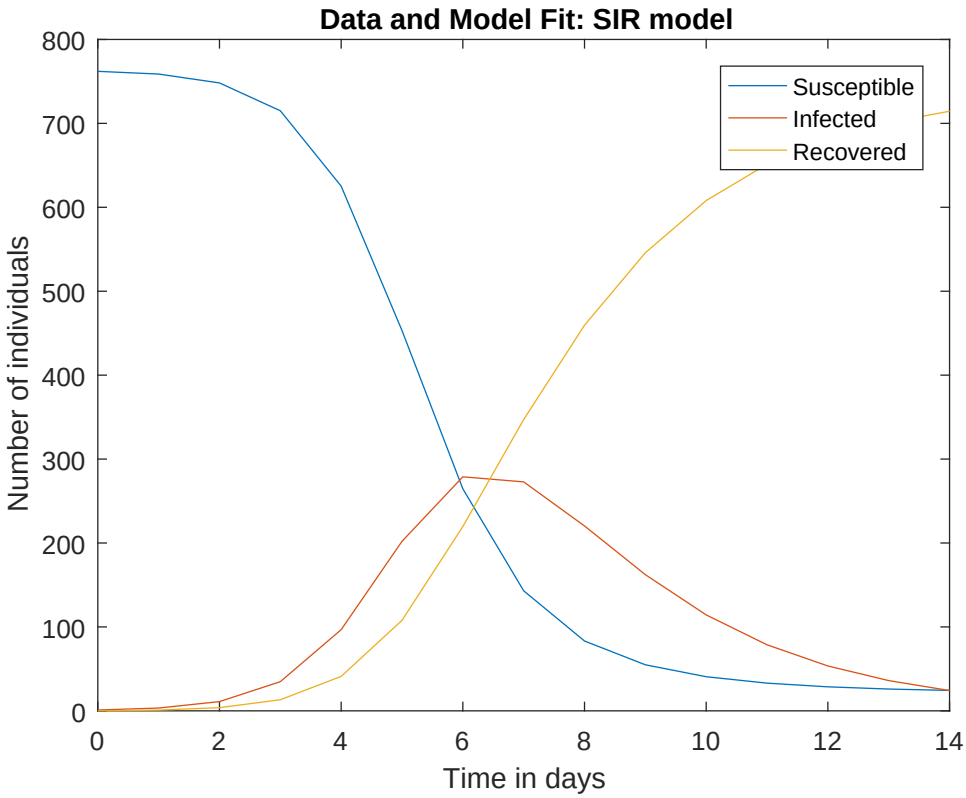

Figure 5. Graph of the compartments.

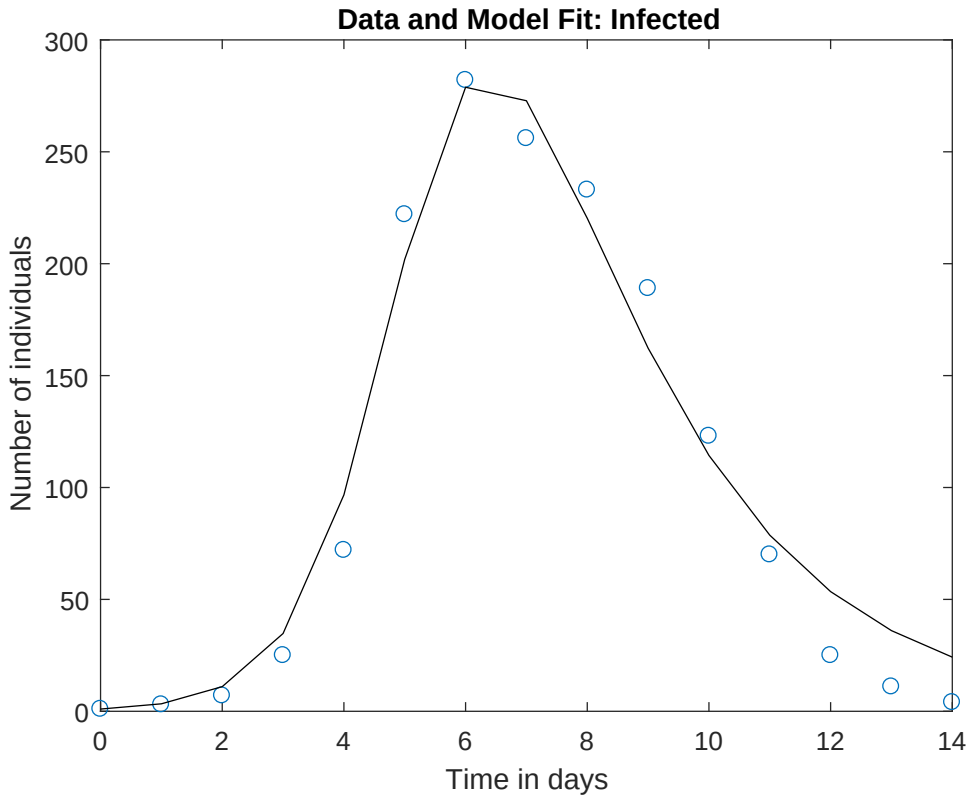

Figure 6. Fitting data to the model.

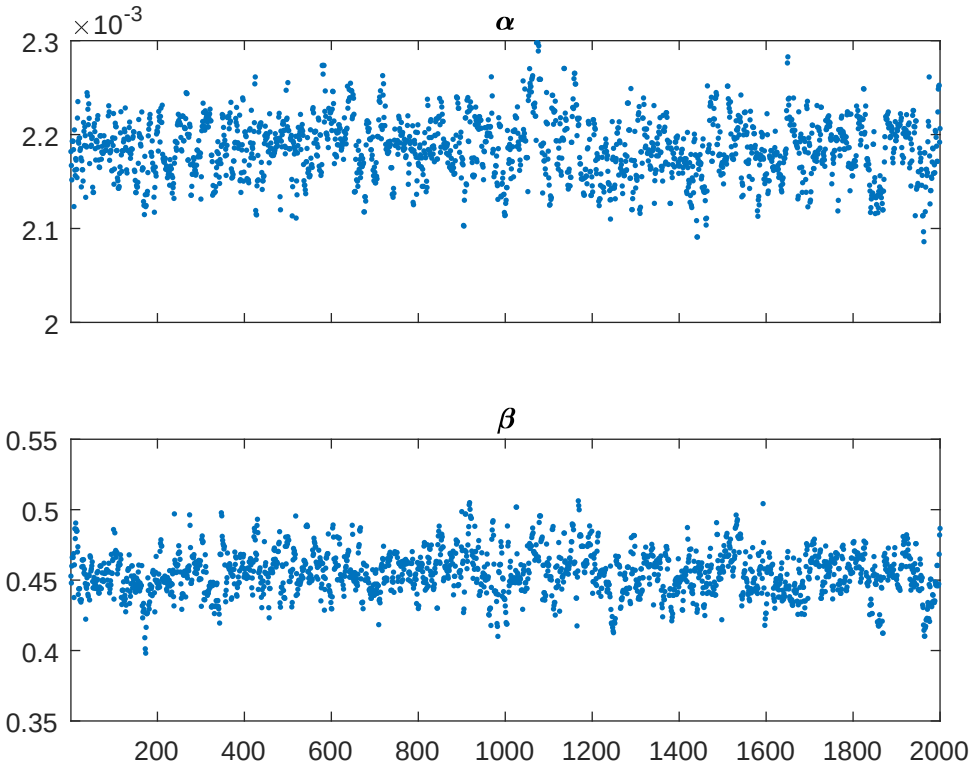

Figure 7. Distribution of the parameters α and β.

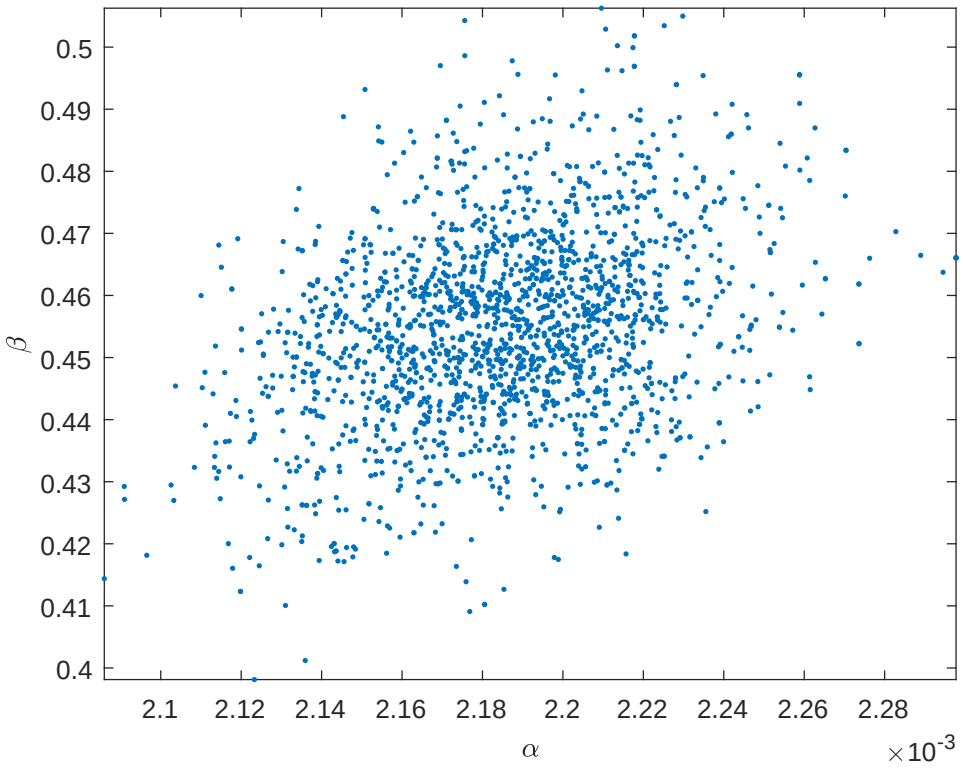

Figure 8. Parameter posterior sample.

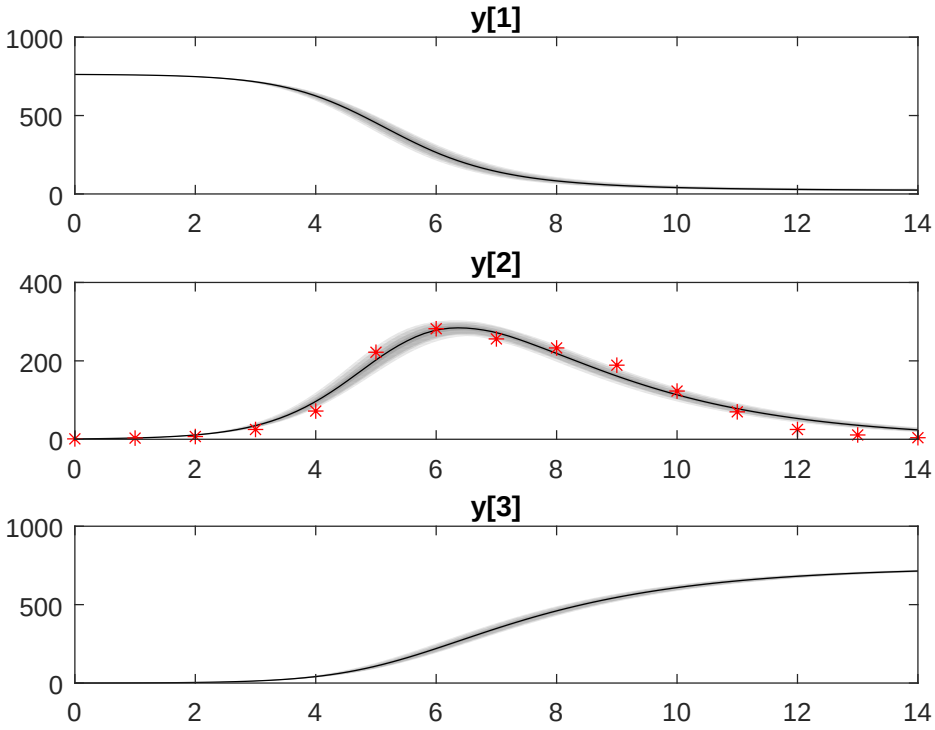

Figure 9. Predictive distributions for the compartments. Grey regions show the confidence intervals of the response. Red stars show the known data for infected individuals.

3 Agent-based approach for solving epidemiological problems

3.1 Agent-based model

Agent-based modeling (ABM) – a computer modeling technique which is used in cases of systems with random variales and irreducible interactions [12]. The main goal of the ABM approach is to build a copy of the real epidemic model and show explicitly its heterogeneous part [13].

In the ABM, there are different compartments that can be called as the "*agents*". According to Gilbert et al. [14], there is no generally agreed definition of this term. So, usually it can be defined as "self-contained programs that can control their own actions based on their perceptions of their operating environment" [14]. Uhrmacher et al. [15] proposed the following definition: an agent can be represented as a process that situated in a certain environment. The goal of agent-based modeling is to create system where the agents will interact with their environment "intelligently".

Agents are a fundamental part of the agent-based modeling approach. They represent entities in the system to be modeled. The behaviour of the agents determines the outcoming result. According to Helbing [16], depending on the given problem, agents can represent such types as individuals, groups, or countries. Agents can have the following properties [16]:

- birth, death, reproduction;
- ability to take part in the competitions and symbiosis;
- emotions;
- memory;
- perception and many other properties.

In this thesis agents are the compartments of the SIR model.

The ABM also represents the environment in which these entities can interact to each other. Each of these modeled entities has a state and an its own explicit behavior [12]. The behavior of the agents should be specified by using rule-based approach. This type of specification

can be represented by logical operations and conditional operator. Due to this assumption, ABM becomes more flexible [16].

For every agent there is a "*space*" or an "*environment*" of their existing. In the simple case of modeling, as Barnes [12] claims, the environment can be described as an empty container. Computational cost and results of the simulations depend on the choice of the space for the agents.

Scientific ABM do not imply plausibility. The system can be simplified in many aspects, for example, one can take into account not all the properties of a certain agent but only a few amount of the necessary things. But despite that, the results of this modeling must correspond to the later empirical observations. Generally, scientific models are focused on obtaining the processes rather than the visual representation [16].

3.2 Benefits and drawbacks of the ABM

As every kind of approaches, ABM have either advantages or disadvantages. First of all, agent-based models provide ample opportunities for observing the dependencies between different groups [17]. This allows to apply this type of modeling to epidemiological problems and, particularly, to SIR models.

The other reasons for using this method are the ability to determine the consequences of different hypotheses and execute the process in parallel form. Latter benefit helps to reduce the computational time [17]. Also, Helbing [16] mentions that agent-based models can be combined with other models, especially in case of using the continious environment. Parker et al. [18] provide an example, where the agent-based system can be used for creating the model of people's evacuation.

But despite the advantages of the ABM, there are some drawbacks of using this approach. The main disadvantage of the ABM approach is its computational cost. As the model depends on the its complexity, it may take a lot of time to simulate the process. Another challenge of using this approach is the dependence on the stochastic processes. It means that after each simulation the result can varies from the previous one. So it is necessary to make a large amount of the simulations for obtaining statistical confidence in the results.

3.3 Aspects of agent-based modeling

According to Helbing [16], there are several aspects and principles that are used for creating an appropriate model. Some of the aspects that are necessary are represented in this list:

- All the empirical and experimental observations should be explained before the creating the model. Also, the simple properties of the system should be described in a proper way.
- One should give an explanation about the purposes of the simulation process. The goals of the study should be known before the simulation.
- The agents of the model should be chosen according the given problem. All unnecessary agents should be excluded from the task.
- There is no central control over individual behavior. As Epstein [19] claims, agents are autonomous.
- The behaviour of the agents and all the states of the model should be described thoroughly for improving the system.
- Agents do not have global information, they are bounded due to their role in the system [19].
- Finally, Helbing [16] proposed the principle that one should choose the model with meaningful parameters instead of the meaningless one.

3.4 Discrete modelling: the Gillespie algorithm

In the theory of probabilities, the Gillespie algorithm generates a statistically correct possible solution of stochastic equations. This approach was presented by Dan Gillespie in 1976. The method was implemented by D.G. Kendall later used by M.S. Bartlett in his studies of epidemics outbreaks [20].

Mathematically, it is a kind of a Monte Carlo method. The Gillespie algorithm allows stochastic simulation of a system with several equations because every process is explicitly simulated [20].

The main idea of the algorithm is to determine when something happens next and what will be the population levels of species after a given period of time. The steps of Gillespie's first reaction algorithm are given below:

1. Initialize the number of susceptible, infected and recovered individuals. Also, parameters of the model should be initialized in this step.

2. Determine which reaction occurs next. For that, one need to calculate the probability that an event μ will occur in the next time interval Δt:

$$P(\tau, \mu)\Delta\tau = P_0(\tau)a_\mu\Delta\tau, \tag{21}$$

   where $P_0(\tau)$ is the probability that no reaction occurs during $(t, t+\tau)$, $a_\mu\Delta t$ is the probability that reaction μ occurs during $(t, t+\tau+\Delta\tau)$. The rate at which any reaction occurs is $R_{total} = \sum_1^m R_\mu$.

   One can obtain $P_0(\tau)$ from the formula:

$$P_0(\tau) = \left[1 - \sum_{i=1}^{M} h_i c_i \epsilon\right]^K = \lim_{K\to\infty}\left[1 - \frac{\sum_{i=1}^{M} h_i c_i \tau}{K}\right]^K = e^{-\sum_{i=1}^{M} h_i c_i \tau}, \tag{22}$$

   where K is the number of subintervals for interval $(t, t+\tau)$ with width ϵ. h_i is the number of reactant combinations, c_i is the average probability that a particular combination of reactants will react in the next time interval $\Delta\tau$. M is the number of reactions.

   After that one can use direct method of deriving $P(\tau, \mu)$:

$$P(\tau, \mu) = h_\mu c_\mu P_0(\tau) = h_\mu c_\mu e^{-\sum_{i=1}^{M} h_i c_i \tau} = a_\mu e^{-a_0\tau} = \left(a_0 e^{-a_0\tau}\right)\left(\frac{a_\mu}{a_0}\right) = P(\tau)\cdot P(\mu|\tau), \tag{23}$$

   where a_μ is the average probability that reaction will occur in the next time interval, $a_0 = \sum_{\mu=1}^{M} a_\mu$.

3. Determine the times to the next event by choosing the minimal of a random value from the exponential distribution with rate parameter to determine the time to the next event.

4. Update the system states and increase time by Δt.

5. Return to the Step 2.

The given SIR model can be represented in the two reactions: infection (24) and recovery (25).

$$S + I \longrightarrow 2I \tag{24}$$

$$I \longrightarrow R \tag{25}$$

Updating of the system states is described in the Table 3.

Table 3. Updating system states for the stochastic SIS model

	System States
Infection	$S_i = S_{i-1} - 1$ $I_i = I_{i-1} + 1$ $R_i = R_{i-1}$
Recovery	$S_i = S_{i-1}$ $I_i = I_{i-1} - 1$ $R_i = R_{i-1} + 1$

Since the algorithm is computationally expensive, there are many modifications of it, for example, the tau-leaping method [20]. Despite that, the original, first reaction algorithm is used as the basic version of agent-based approach.

3.5 A discretization approach for agent-based simulation

Here we present another discretization approach that directly transforms an ODE system into an agent population, and stochastically transforms the agents at every time step [21]. Every differential equation (18)-(20) is presented in the discretized form:

$$\frac{S_{i+1} - S_i}{\Delta t} = -\beta S_i I_i, \tag{26}$$

$$\frac{I_{i+1} - I_i}{\Delta t} = \beta S_i I_i - \gamma I_i, \tag{27}$$

$$\frac{R_{i+1} - R_i}{\Delta t} = \gamma I_i. \tag{28}$$

Making simple transformations, one can obtain following results:

$$\frac{S_{i+1}}{S_i} = 1 - \beta I_i \Delta t, \tag{29}$$

$$\frac{I_{i+1}}{I_i} = 1 + \beta S_i \Delta t - \gamma \Delta t, \tag{30}$$

where $\beta I_i \Delta t$ is a fraction of decrease in the number of susceptible individuals, $\beta S_i \Delta t$ is a fraction of increase in the number of infected individuals, $\gamma \Delta t$ is a fraction of increase in the number of recovered individuals.

Updating of the states is made by following rules:

1. Create r_1 — a vector of randomly selected values from the uniform distribution (values between 0 and 1) for each agent of the compartment S.

   If the value $r_1(i) < \beta I \Delta t$ then $S = S - S_{rem}$, $I = I + S_{rem}$, where S_{rem} is the sum of those values $r_1(i)$ (number of individuals who got the infection).

   Thus, there is a transition from susceptible individuals to infected ones.

2. Create r_2 — a vector of randomly selected values from the uniform distribution (values between 0 and 1) for each agent of the compartment I.

   If $r_2(i) < \gamma \Delta t$ then $I = I - I_{rem}$ and $R = R + I_{rem}$, where I_{rem} is the number of recovered individuals.

   Thus, there is a transition from infected compartment to recovered individuals.

4 RESULTS

MCMC method was applied for deterministic epidemiological model and two ABM approaches were implemented to the stochastic SIR model.

Figure 10 illustrates the comparison of the numerical solution (obtained using standard Matlab solver *ode45*) and solution with discretization approach in case of small amount of individuals. Total number of population is equal to 763. Initial number of the susceptible individuals is equal to 762. The number of the infected people is equal to 1. And the number of recovered ones is 0. The value of the contact rate β is equal to 0.002. The recovery rate γ is 0.4. data is taken from the book written by Murray [10].

Green lines represent ABM approach with 10 simulations. Blue, red and orange lines represent the dynamics of the numbers of susceptible, infected and recovered individuals respectively. Solution is obtained using the standard Matlab *ode45* solver.

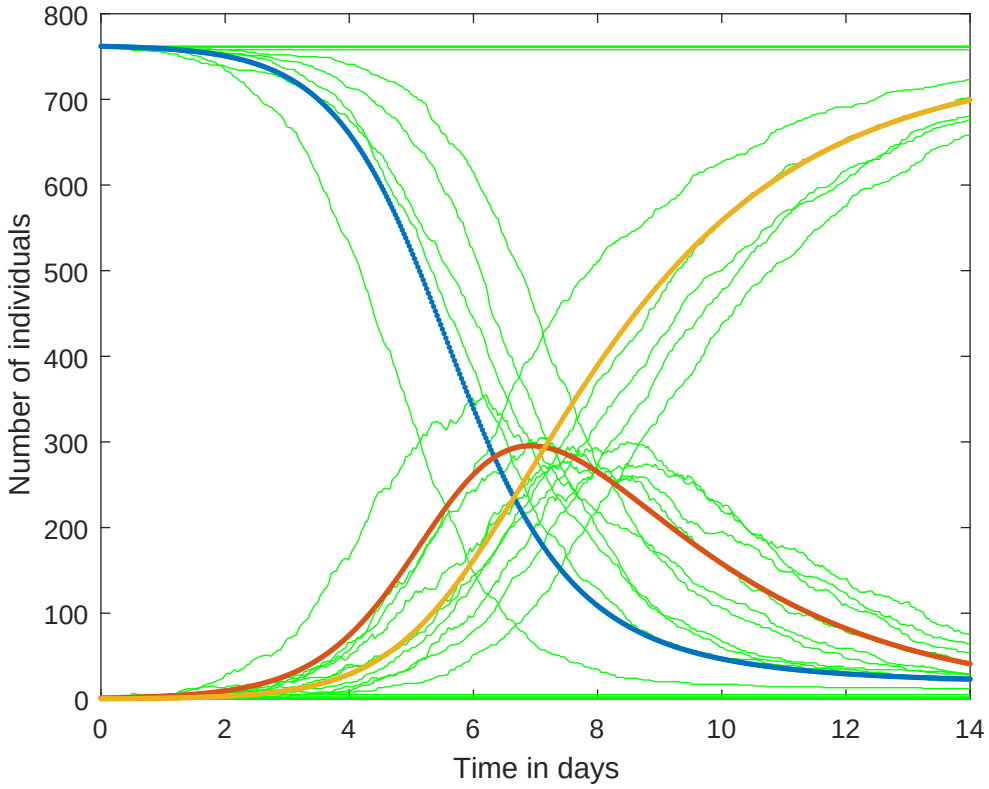

Figure 10. Agent-based model with discretization approach for $S_0 = 762$, $I_0 = 1$, $R_0 = 0$. Green lines show the solution with discretization approach, blue, red and orange lines represent the dynamic of susceptible, infected and recovered individuals obtained with *ode45* solver.

According to the Figure 10, the small values of the individuals in the compartments provide unsatisfactory results. The accuracy of the solution is not high. Due to randomization used to determine the probability of events and small amount of infected individuals in the beginning (only 1 person), a scenario is possible when susceptible individuals do not contact the infected ones at all. It makes sense in the real-based situations, the outbreak of the disease might not start with one infected person.

The situation changes when the number of infected individuals increases. On the Figure 11 this case is considered. When the initial amount of the infected persons is increased to 10 people, the situation with the results of the implemented algorithm becomes more stable. The probability of keeping the number of infected people in the closed space unchanged becomes smaller. Nevertheless, due to the small total number of individuals, the accuracy of the solution remains low.

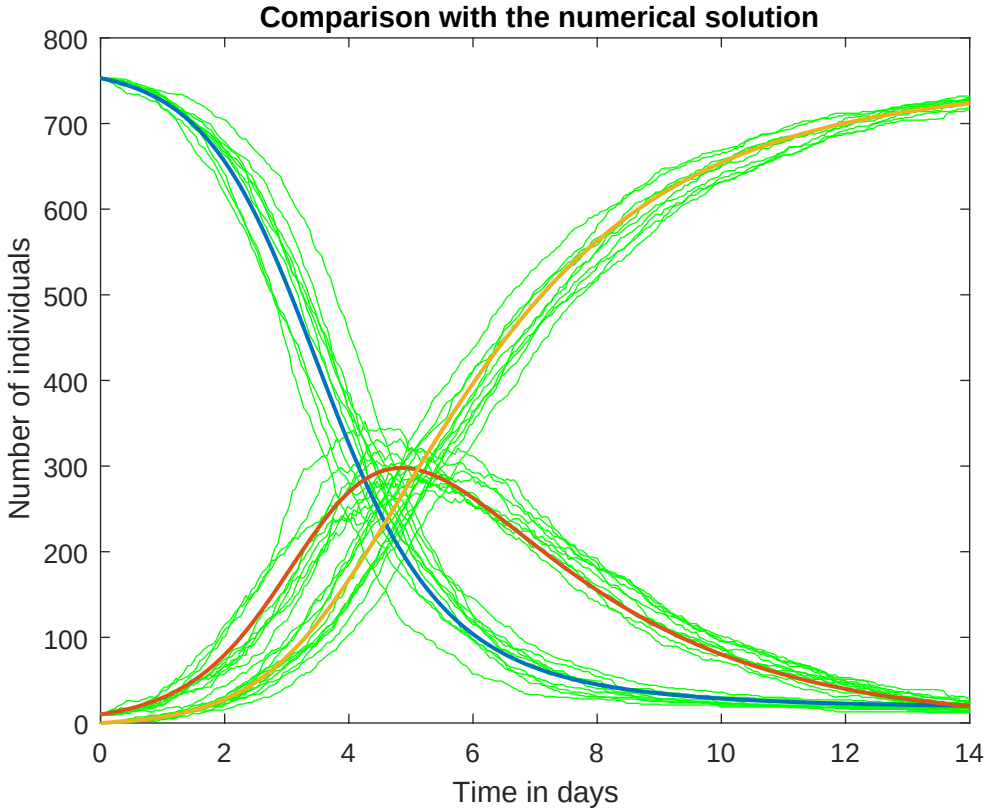

Figure 11. Agent-based model with 10 simulations of the discretization approach for $S_0 = 753$, $I_0 = 10$, $R_0 = 0$. Green lines show the solution with discretization approach. Blue, red and orange lines represent the dynamic of susceptible, infected and recovered individuals obtained with *ode45* solver.

The other situation is presented on the Figure 12, where the number of individuals is increased in 10 times. The solution is close to the numerical one with minor discrepancies. On the Figure 12 green lines show the solution with discretization approach. Blue, red and orange dots represent the Matlab *ode45* solution.

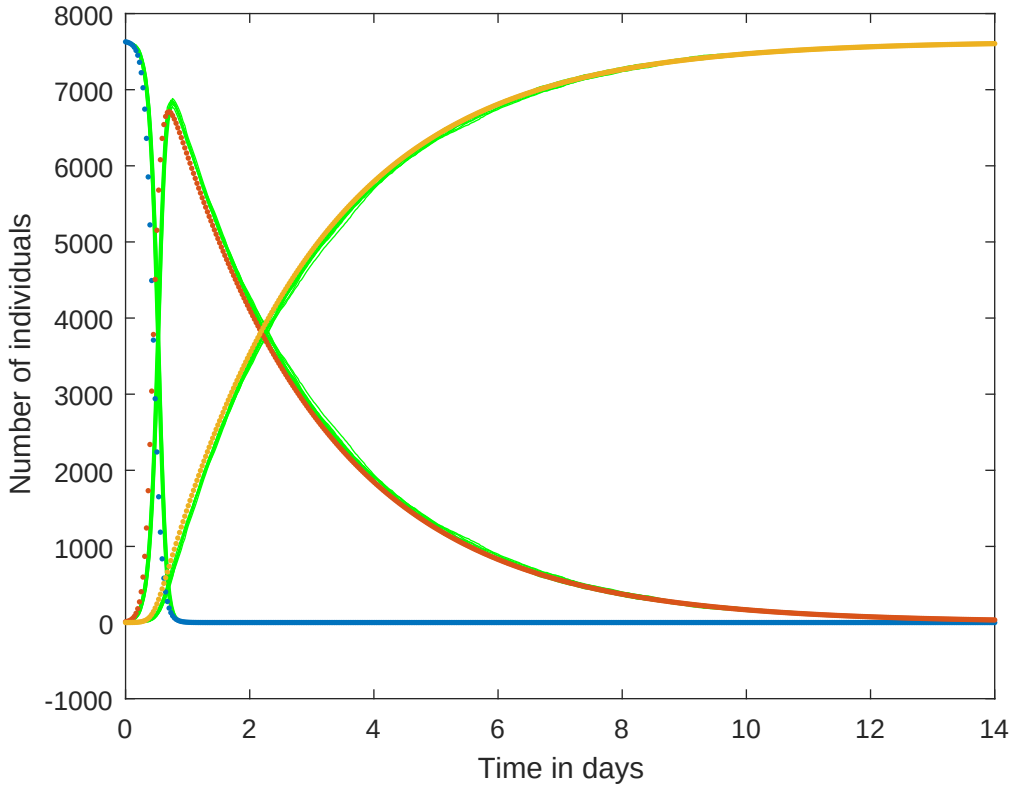

Figure 12. Agent-based model with discretization approach for $S_0 = 7620$, $I_0 = 10$, $R_0 = 0$. Green lines show the solution with discretization approach. Blue, red and orange dots represent the *ode45* solution.

Figure 13 represents the results of the Gillespie algorithm in comparison of the numerical solution. Total number of population is equal to 763. Initial number of the susceptible individuals is equal to 762. The number of the infected people is equal to 1, the number of recovered ones is 0. Initial number of the susceptible individuals is equal to 762. The number of the infected people is equal to 1. And the number of recovered ones is 0. The value of the contact rate β is equal to 0.002. The recovery rate γ is 0.4.

Here, blue lines show the solution obtained with *ode45* solver. Green lines are the solution of the Gillespie algorithm. The small number of individuals provides the visible fluctuations in the solution. The discrepancies with the numerical solution are significant in case of dynamic of the infected individuals.

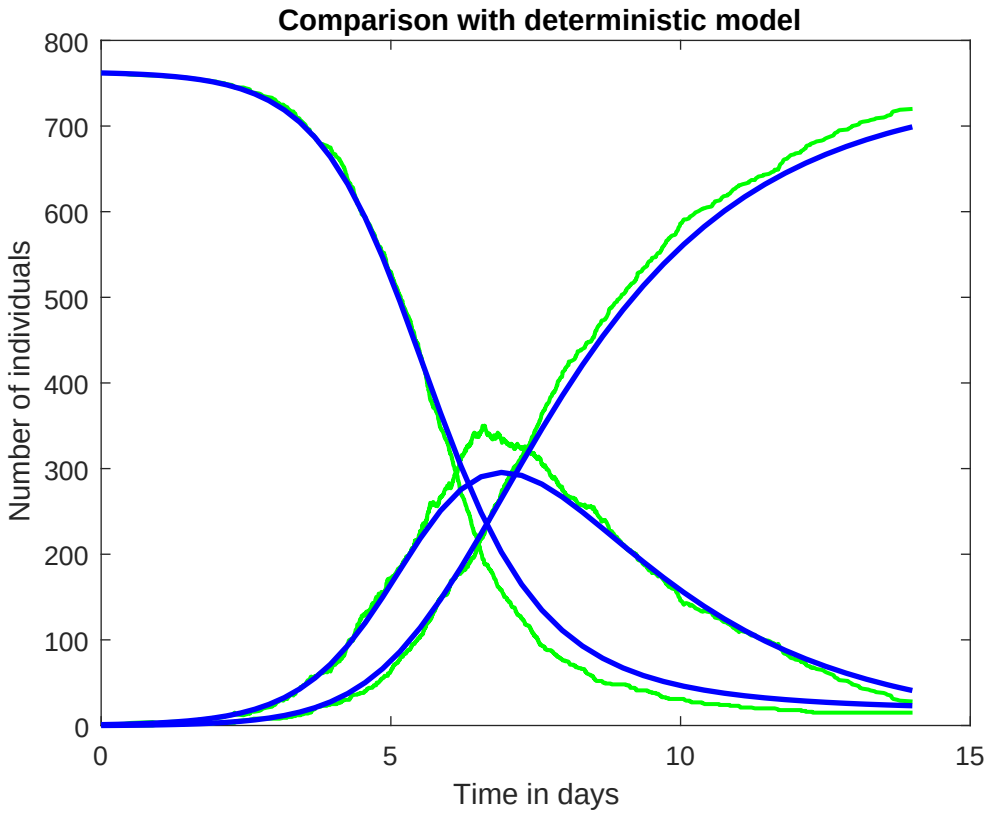

Figure 13. Agent-based modelling with Gillespie algorithm for $S_0 = 762$, $I_0 = 1$, $R_0 = 0$. Blue lines show the solution using *ode45* solver. Green lines are the solution with the Gillespie algorithm.

Figure 14 illustrates the case when the number of the infected individuals remains the same during the whole period of the disease. Just as with the discrete approach to solving the problem (see Figure 10), using the Gillespie algorithm, there is a possibility that because of the very small initial number of infected people, an outbreak of disease will be impossible. Different scenarios of epidemic development with a small initial value of infected individuals are shown on Figure 15. Green lines represent 10 simulations of the Gillespie algorithm.

In case of building the agent-based model that should be close to the real life, this case makes sense as well as the possible situation with the discrete method. While the built-in algorithm solves the problem in "ideal" conditions, under which the spread of the disease is guaranteed, the Gillespie algorithm is more approximate to the realities of life.

Figures 16 and 17 represent the distribution of the time for each agent-based event in case of 10 simulations and for 1 simulation respectively. It can be clearly seen that at a time when the process of spreading the disease is quite active, the time necessary for the next event to occur is not so great. Approximately the first 5-10 days, the Δt value does not exceed 0.08. After the outbreak of the disease goes to a decline and the solution of the system of

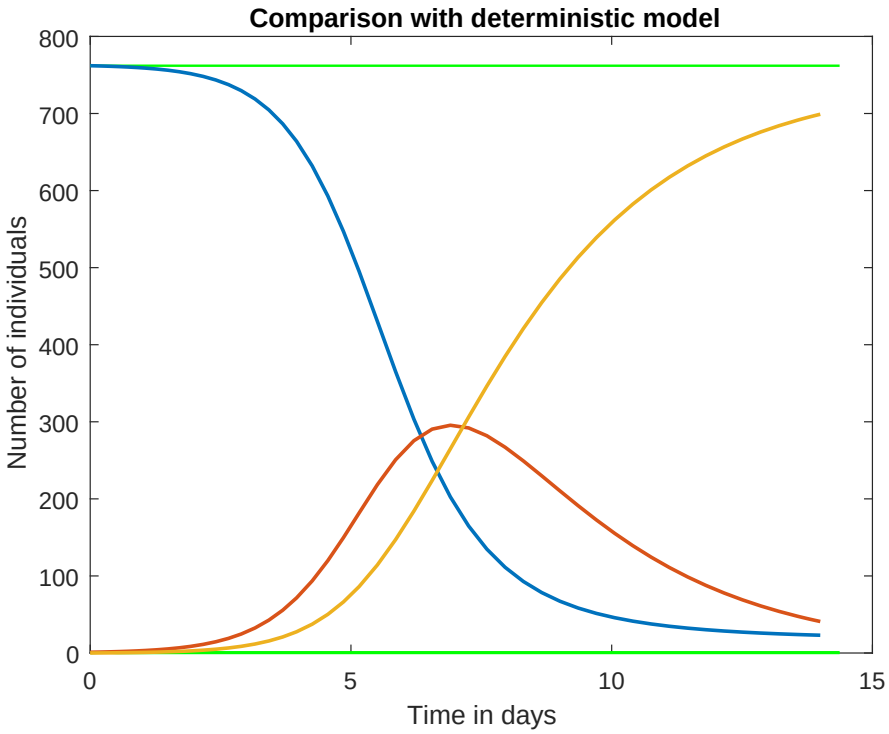

Figure 14. A case when there is no outbreak of the disease. Agent-based modelling with Gillespie algorithm for $S_0 = 762$, $I_0 = 1$, $R_0 = 0$. Blue, red and orange lines show the solution using *ode45* solver. Green lines are the solution with the Gillespie algorithm.

differential equations becomes stable, the time value increases up to 0.6 in some cases. On average, the time doubled.

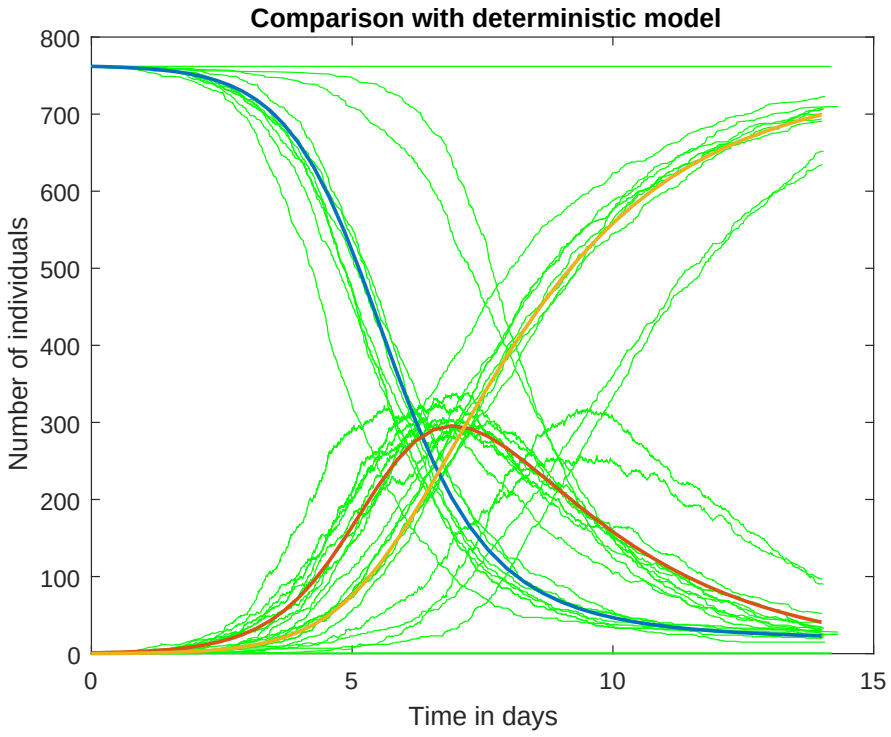

Figure 15. Agent-based modelling with Gillespie algorithm for $S_0 = 762$, $I_0 = 1$, $R_0 = 0$. Blue, red and orange lines show the solution using *ode45* solver. Green lines are the solution with the Gillespie algorithm for 10 simulations.

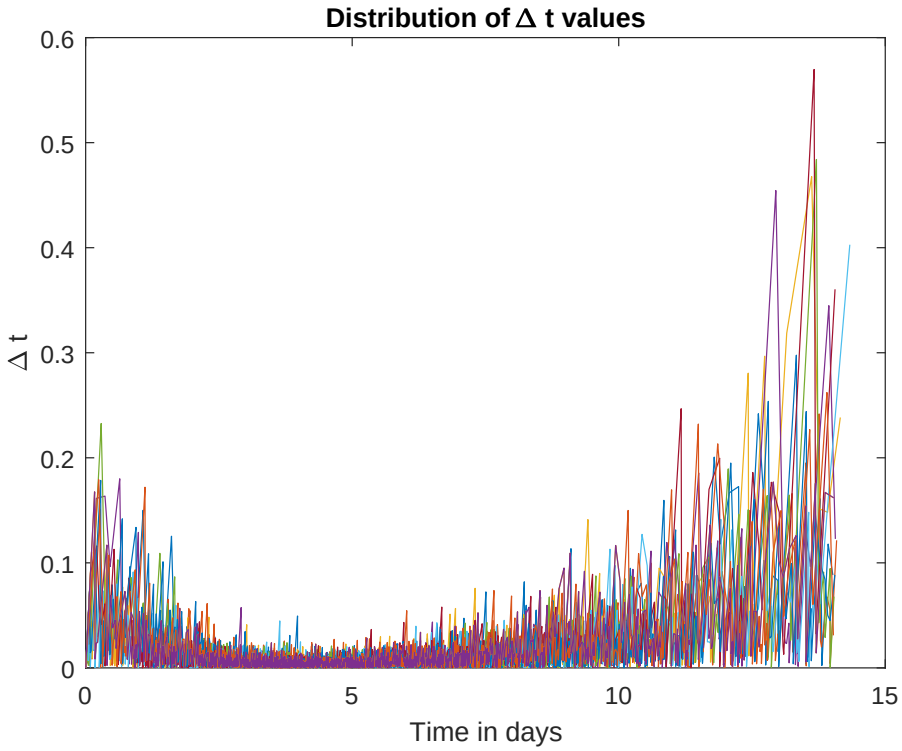

Figure 16. Distribution of the time for each agent-based event in case of 10 simulations.

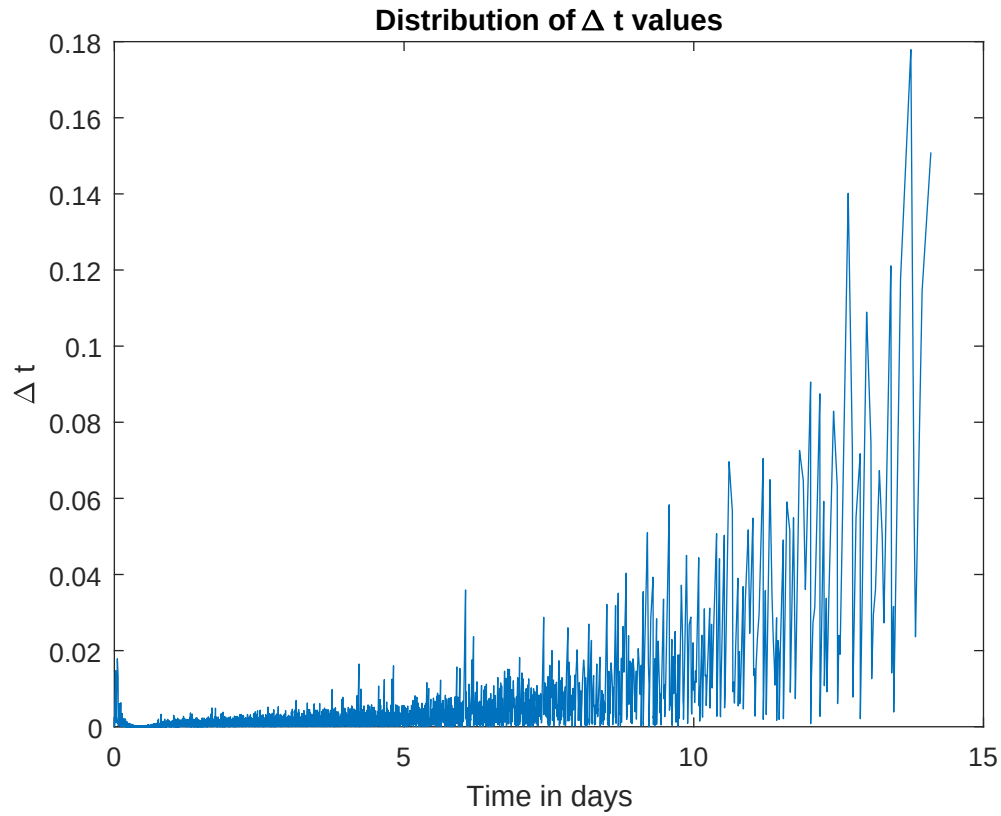

Figure 17. Distribution of the time for each agent-based event in case of 1 simulation. First 5-10 days the distribution is quite stable with the insignificant number of the outbreaks.

With the increase in the number of total population the fluctuations become smaller or disappear. This is clearly confirmed in the Figure 18. Here, green lines represent the Gillespie algorithm and red, blue and orange dots show the numerical solution. Gillespie solution is very close to the numerical one. Due to the meanings of the contact and recovery rates, the rate of spread of the disease is quite high in comparison with the case of a small amount of the population.

Figure 19 shows the distribution of the parameter Δt in case of large number of population. Within first days of the infection spread the meaning of this parameter is small. It does not exceed 0.02. It means that the events of the given agent-based model occur very fast. But by the end of the 2 weeks period the values of the Δt become increase due to stabilization of the solution of the ODE system.

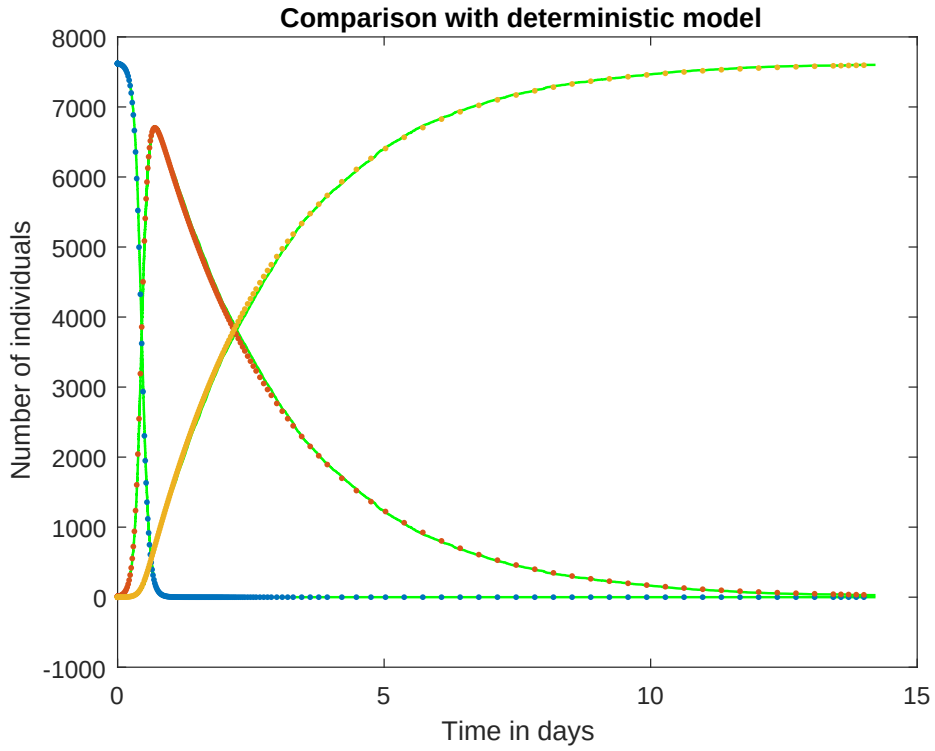

Figure 18. Agent-based modelling with Gillespie algorithm for $S_0 = 7620$, $I_0 = 10$, $R_0 = 0$. Green lines represent the Gillespie algorithm. Red, blue and orange dots show the numerical solution.

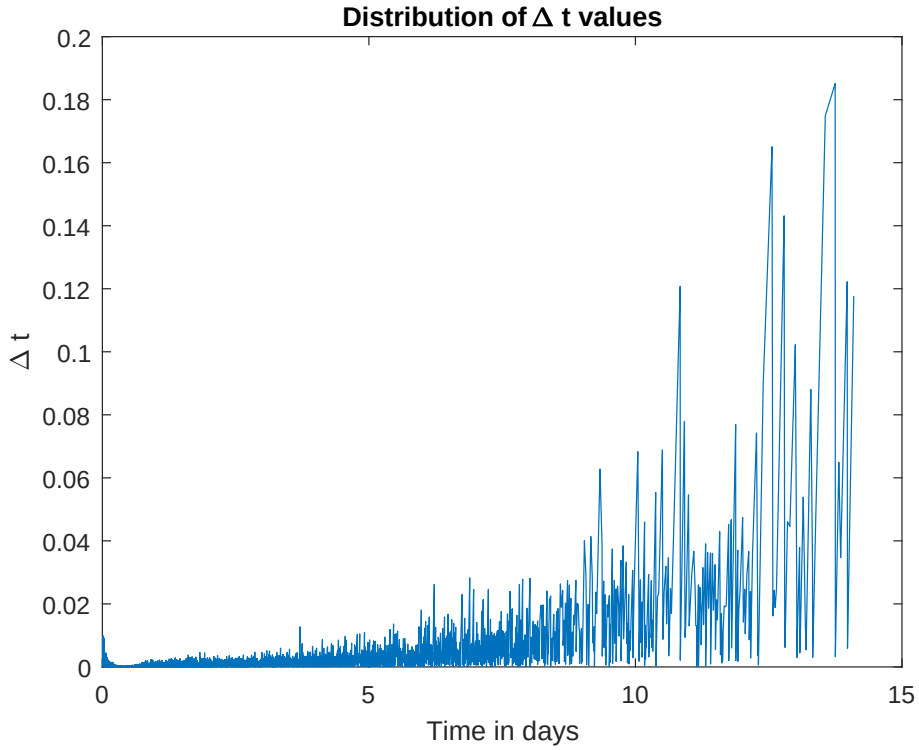

Figure 19. Distribution of the time for each agent-based event in case of 1 simulation. Total number of individuals is equal to 7630.

After the results were obtained, the execution time of the algorithms was compared according to size of the population. Discrete approach of the agent-based modelling and the Gillespie algorithm were executed with different total amount of the individuals: 763, 7630, 76300 and 763000. Results of the comparison are given in the Tables 4 and 5:

Table 4. Time of Gillespie algorithm execution and Matlab ODE solver, Δt is chosen randomly.

Size of the population	Time (sec)	
	Gillespie	ODE solver
N = 763	0.021586	0.163342
N = 7630	0.186073	0.157564
N = 76300	1.140396	0.157092
N = 763000	11.715125	1.046510

Table 5. Time of discrete algorithm execution and Matlab ODE solver, $\Delta t = 0.0281$.

Size of the population	Time (sec)	
	ABM	ODE solver
N = 763	0.081132	0.390010
N = 7630	0.102972	0.405175
N = 76300	0.397446	0.466955
N = 763000	3.771290	0.663510

As it can be seen from the Tables 4 and 5, with the increase in the number of population, the time required to execute the algorithm increases from 0.021586 to 11.715125 for the Gillespie algorithm and from 0.081132 to 3.771290 for discretization approach with the constant Δt. The time taken to solve the system of equations using the built-in Matlab ode solver is changed insignificantly. Also one can notice that the Gillespie algorithm executes slower than the ABM.

Choosing different values for Δt, you can improve the result for a discrete method of solving the problem. In the Table 6 the comparison of discretization approach and built-in Matlab solver for $\Delta t = 0,0468$ is presented:

Table 6. Time of discrete algorithm execution and Matlab ODE solver, $\Delta t = 0,0468$.

Size of the population	Time (sec)	
	ABM	ODE solver
N = 763	0.057326	0.404637
N = 7630	0.091201	0.409515
N = 76300	0.425790	0.476306
N = 763000	3.908590	0.661298

5 CONCLUSIONS

In this thesis, modelling approaches for different types of epidemiological problems were investigated. Three algorithms – Markov Chain Monte Carlo method, Gillespie algorithm and the discretization approach – were implemented to model spreading of the influenza epidemic in a British boarding school in 1978.

The Markov Chain Monte Carlo method was used for obtaining a solution for the deterministic system of ordinary differential equations. It was shown that for a small amount of the population the identifiability of parameters is very good. However, in the literature [6] there does exist opposite examples where identifiability of parameters is not so good due to large number of individuals and uncertainty production of mortality data. The compartmental diagram (see Fig. 1) of the influenza epidemic case shows that using Markov Chain Monte Carlo method provides a very good fitting model to the given data of infected individuals. Moreover, this method provides a strong correlation between the contact and removal rates so the basic reproduction number can be obtained easily.

For solving the stochastic system of differential equations two methods were proposed. First method, Gillespie algorithm, was implemented to the different number of population. Results show that increasing a number of individuals provides smoother graphs and the accuracy of the results become higher. At the same time, this algorithm showed a greater approximation to real life than a standard Matlab algorithm. In particular, the Gillespie algorithm assumes the probability that if initially the number of infected people is very small, then outbreaks may not happen. This shows the effectiveness of this algorithm in comparison with the built-in one.

The second agent-based algorithm was the method of discretization. The rules for this approach were developed during the implementing the algorithm. The results showed practically the same behavior as in the Gillespie algorithm. Large population provided better results which are very close to the numerical solution obtained with the Matlab *ode45* solver. The possibility to have the situation when outbreak might not happen is shown as well as Gillespie algorithm.

In addition, the execution time of the algorithms for a different number of populations was measured using standard Matlab functions. The results showed that, depending on the choice of the number of individuals and the value of parameter Δt, the CPU time can increase in more than 3 times. Nevertheless, in the same time, the discretization approach showed the faster results than Gillespie algorithm. And, in some points, discrete method showed the

better results in comparison with standard Matlab solver.

The implementing of agent-based approaches provides the various ways for the future work. One of the possible direction can be the developing agent-based system that is maximally close to the real life situations. This will help to observe the development of diseases not from the point of view of ideal conditions such as a closed space, in which infection can be inevitable, but in terms of the real behaviour of individuals who may not be in contact with other infected.

REFERENCES

[1] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599 – 653, 2000.

[2] D.J. Daley and J. Gani. *Epidemic Modelling: An Introduction*. Cambridge Studies in Mathematics. Cambridge University Press, 2005.

[3] M. Laine. *Adaptive MCMC methods with applications in environmental and geophysical models*. Contributions: Ilmatieteen Laitos. Finnish Meteorological Inst., 2008.

[4] A. Solonen, H. Haario, and M. Laine. *Statistical analysis in modelling*. 2011.

[5] H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 04 2001.

[6] A.Solonen, H. Haario, J. M. Tchuenche, and H. Rwezaura. Studying the identifiability of epidemiological models using mcmc. *International Journal of Biomathematics*, 6(2):1350008 (18 pages), 2013.

[7] N.G. Becker. *Analysis of Infectious Disease Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.

[8] T. Kypraios. *Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi parametric time series models*. PhD thesis, Lancaster University, 2007. http://eprints.lancs.ac.uk/26392/1/kypraios.pdf.

[9] A. Solonen, H. Haario, and M. Laine. Simulation-based optimal design using a response variance criterion. *Journal of Computational and Graphical Statistics*, 21(1):234–252, 2012.

[10] J.D. Murray. *Mathematical Biology: I. An Introduction*. Interdisciplinary Applied Mathematics. Springer New York, 2011.

[11] News and notes. *British Medical Journall*, 1(6112):586–590, 3 1978.

[12] D. J. Barnes and D. Chu. *Introduction to Modeling for Biosciences*. Springer-Verlag London Limited, 2010.

[13] J.L. Casti. *Would-be worlds: how simulation is changing the frontiers of science*. J. Wiley, 1997.

[14] N. Gilbert and K. Troitzsch. *Simulation For The Social Scientist*. UK Higher Education OUP Humanities & Social Sciences Sociology. McGraw-Hill Education, 2005.

[15] A.M. Uhrmacher and D. Weyns. *Multi-Agent Systems: Simulation and Applications*. Computational Analysis, Synthesis, and Design of Dynamic Systems. CRC Press, 2009.

[16] D. Helbing. *Social Self-Organization, Understanding Complex Systems*. Springer-Verlag Berlin Heidelberg, 2012.

[17] C.S. Taber and R.J. Timpone. *Computational Modeling*. Computational Modeling. SAGE Publications, 1996.

[18] J. Parker and J.M. Epstein. A distributed platform for global-scale agent-based models of disease transmission. *ACM Trans. Model. Comput. Simul.*, 22(1):2:1–2:25, 2011.

[19] J.M. Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton Studies in Complexity. Princeton University Press, 2012.

[20] Gillespie algorithm. [online]. Available: https://en.wikipedia.org/wiki/Gillespie_algorithm, 2017. [Accessed: May 16, 2017].

[21] H. Haario. Oral communication, 2017. [July 23, 2017].