Treball Final del Grau en Enginyeria Física

Development of a simulator to evaluate public health control strategies of tuberculosis in big cities

Author:

Joan Francesc Gilabert Navarro

Codirectors:

Clara Prats
Cristina Montañola

Universitat Politècnica de Catalunya

Barcelona

Juny 2015

Abstract

Tuberculosis is still one of the most important infectious diseases worldwide, killing over 1.5 million people each year. It is estimated that one third of world population has already been infected by *Mycobacterium tuberculosis*. In particular, the Ciutat Vella neighbourhood in Barcelona has a tuberculosis incidence which is comparable with the incidence in countries like Sudan.

In this work we have improved a previous agent-based model, and we used it to develop a simulator of the tuberculosis dynamics in Ciutat Vella, as a first step towards the simulation of tuberculosis dynamics in big cities. The model was implemented in NetLogo, a free and open-source agent-based simulation tool that incorporates a helpful user-friendly interface. NetLogo is time-consuming for large populations (our model has around 100.000 individuals), therefore, some optimization work was required in order to make the experiments feasible.

This simulator has allowed us to perform virtual experiments to asses the efficacy of public health strategies. In particular, we have performed two series of experiments. In the first one, we checked the influence of the distribution of infected times on the number of sick individuals. In the second one, we modelled different types of screenings, in order to determine which was the most effective public-health measure.

As conclusions, we saw that the kind of distribution of infected times used had a great impact on the results. Also, we found that among the several screenings, random variable screening, selective screening and contact tracing, the latter was the most effective. This is good news, since it is a common measure in epidemiological control. Finally, we observed the limitations of NetLogo as a simulation tool. For big populations, as could be the approximately 1.600.000 inhabitants of Barcelona, NetLogo is not practical in times of execution time. Therefore, a more powerful tool is needed. We discussed the possibilities of extending our model to an example of such tools, the YADES simulator, which has the added advantage that is a demographical simulation tool, which would increase the fitness of the model to the target population.

Acknowledgement

First and foremost, I wish to thank my co-directors, Drs. Clara Prats and Cristina Montañola, for their guidance and advice, which have been extremely helpful during these last months.

I am also thankful to inLab FIB, for providing me with an space to work in the middle of a very professional and comfortable environment. This has surely made my work easier.

Moreover, I'm grateful to Drs. Daniel López and Quim Valls from the MOSIMBIO research group in the UPC; Drs. Pere-Joan Cardona and Cristina Vilaplana from the Experimental Tuberculosis Unit in the Institut de Recerca Germans Trias i Pujol – Can Ruti and Dr. Josep Casanovas from InLab and BSC; from all their brilliant insight, comments and suggestions for my work.

Finally, I would like to thank my family and friends for their support and faith in me.

Contents

Chapter 1

Introduction

Tuberculosis (TB) is an infectious disease that has co-evolved with humanity. Its particular strategy of remaining almost hidden and slowly acting has become its success. The Global Tuberculosis Report 2014 from the World Health Organization (WHO) estimates that, in 2013, 9.0 million people developed TB, and 1.5 million died from the disease (WHO, 2014). In Figure 1.1 a map with the estimation of the TB incidence worldwide is shown. This map shows how TB is still an important disease in a great part of the world.

1.1 Theoretical basis

1.1.1 Tuberculosis natural history

Tuberculosis is an airborne infectious disease caused by the *Mycobacterium tuberculosis*. It can affect many organs, but the most usual symptoms are found in the lungs. About 75% of the tuberculosis cases are pulmonary tuberculosis. Some of these symptoms are tissue destruction, chronic cough and fever. Although since the first treatments for tuberculosis were developed the number of cases had been decreasing, especially in developed countries, it is still an important disease. Tuberculosis is the second cause of death from infectious diseases. It is estimated that about one-third of world's population has been infected, with 9 million new infections every year. About 1,3-1,5 million people dies annually because of the disease (WHO, 2013).

Although the natural history of tuberculosis is still not fully understood yet(Cardona, 2010), there are several factors that have been already identified as crucial from the epidemiological point of view. Among others, we can mention that a patient with a cavitation has a higher spreading rate than a non-cavitated individual. There are other risk factors that complicate the eradication of tuberculosis, for example, HIV infection. The immunodeficiency caused by HIV is a crucial risk factor for tuberculosis, and it is found that the two diseases are often present at the same time. Another difficulty is the appearance of drug-resistant strains of the tuberculosis bacterium. Multiple drug- resistant tuberculosis (MDR-TB) can be caused by sick individuals that abandon the treatment after its initial

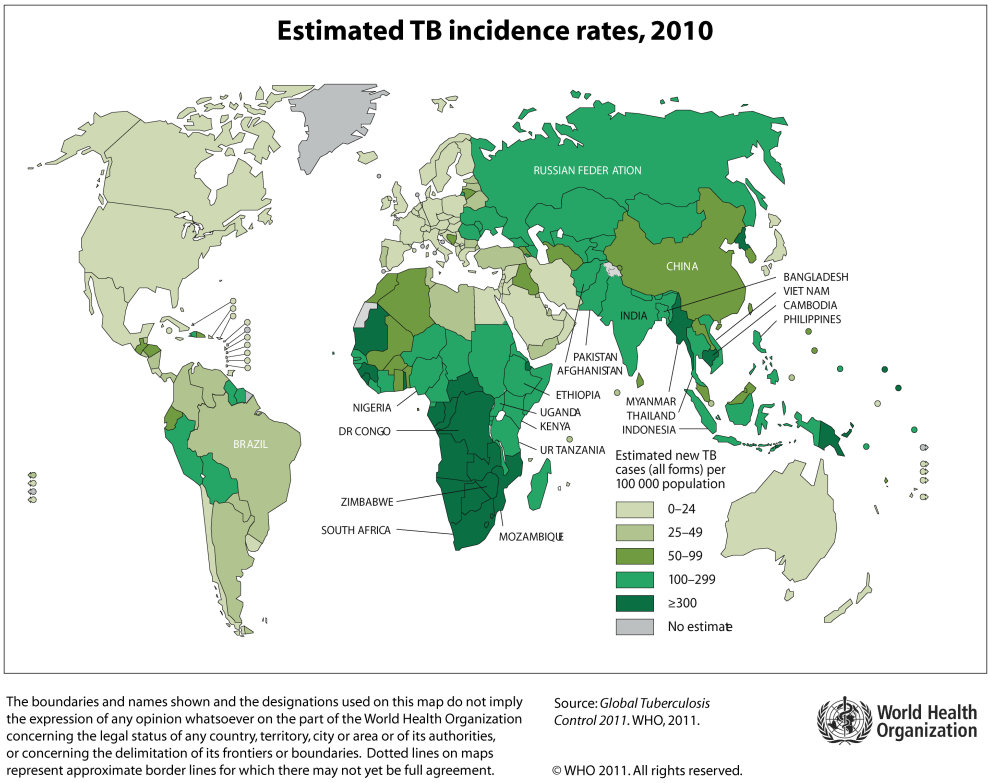

Figure 1.1: Map of the tuberculosis incidence worldwide in 2010, data corresponding to the WHO *Global Tuberculosis Control 2011* (WHO, 2012)

phase. Since the bacterium is not entirely eliminated from their system, it can develop this resistance that hinders the treatment. Even a more worrying strain has been discovered, named as Extensively drug-resistant tuberculosis (XDR-TB), resistant to even more drugs. Its treatment is more difficult and expensive (Ozcaglar et al., 2012).

It is important to note that there is a difference between being infected with tuberculosis (latent tuberculosis infection or LTBI) and actually developing the disease (active tuberculosis or ATB). Only 10% of those infected on average will eventually develop it. To understand how this is possible, we will describe how infection and sickness take place. LTBI occurs when bacteria responsible for the disease enters into a healthy individual via the respiratory system. The bacteria travel to the alveoli, where they are phagocyted by an alveolar macrophage. Inside the macrophage it multiplies and eventually destroys the macrophage. This process is repeated with different macrophages until there is an immune response that encapsulates the bacteria, forming the TB lesions.

Some bacteria may be driven to the superior bronchial space by foamy macrophages in order to be expelled, but once there they can re-enter through the aerosols and cause another lesion. In immunodeficient individuals (HIV-infected or other causes) the immune response can be not enough to stop the cycle and thus the individual would develop TB. However, there is another risk regarding this cycle. The bacteria may reach the superior lobe of the lung, where the conditions are more favourable for the bacteria than those of the alveoli, making it an easier environment to the infecting vector. If this happens in immunocompetent hosts, there would be an inflammatory response that could lead to tissue destruction and forming cavities. This reaction is not as frequent in immunodeficient hosts as it is in immunocompetent ones, since the first do not show such a strong inflammatory response.

1.1.2 Models in epidemiology

Mathematical models in epidemiology are a tool of great importance in the study of the dynamics of a disease. Multiple schemes have been used for this purpose, from the basic models of differential equations with compartments (SEIR, SIR, SIS...) to stochastic models that may focus on an individual scale. Along with these different schemes, some variations of models have allowed obtaining more realistic results by taking into account heterogeneities in the population. For example, variation within age groups, gender, or even the effects of vaccination.

The compartment models consist of different pools of individuals characterized by the status of the disease. In the most common models (SEIR, SIR, SIS) the S stands for susceptible, or persons capable of being infected; E stands for exposed, persons who have been in contact with the disease but are not infectious; I stands for infectious, individuals who might transmit the disease and finally, R for recovered, individuals who have overcome the disease. These categories are the common foundations of this type of models,

but depending on the disease to model the behaviour of these compartments can change or even the compartments used may differ. An in-depth review on this matter can be found in Sumi and Kobayashi (2003).

In the particular case of tuberculosis, a model that accounts for a vast range of heterogeneities might be of great help. For example the incidence between men and women is very different, sometimes even a 50% difference, and so it is for age. Other aspects that might be worth to consider when attempting to model tuberculosis is HIV incidence, which is a crucial risk factor. When trying to build a model of tuberculosis it might be wise to consider the location to simulate, since the situation of the disease in developed countries such as Western Europe is very different from that of some African countries. Specifically for European communities the incidence is so low that considering continuous differential equation model might not make any sense and a method focused on smaller scales may have a better chance of obtaining good and reliable results. An example of one of such method is agent-based modelling (ABM).

1.1.3 Agent-Based Modelling

As an alternative from the more classical mathematical models, TB can be modelled from a bottom-up perspective that allows observing how the dynamics of the global system emerge from the low-level interactions. Agent-Based Modelling (ABM) is an approach for a bottom-up modelling that in recent years has gained popularity as a tool to understand the insights of social complexity. An ABM allows the simulation of the dynamics of a population by controlling the characteristics and behaviour of each individual of the system (Ferrer et al., 2009). Moreover, ABM is particularly useful for projecting a population answering "what if" questions such as the effect of a certain policy on the spread of a disease in a target group. Notably, an ABM of TB spreading would provide the possibility of incorporating the contact tracing (Begun et al., 2013).

Agent-based models are mathematical models constructed using the bottom-up approach, i.e., by modelling the elemental parts of the system and the relationships among them, one can obtain the behaviour of the whole system. The fundamental units of those models are the agents, entities with a set of rules defined which can evolve according to them. The level of depth achieved by this kind of models allows the modeller to obtain phenomena present on the system that other models such as the compartment models may overlook. Of course, the level of detail comes at the price of a bigger computational cost. However with the increasing computational capacities and the possibilities that High-Performance Computing offers (HPC) the drawback that it represents is notoriously compensated by the benefits of the model.

ABM models have been used in multiple fields, from social sciences to markets, economics and even for military purposes (Bonabeau, 2002). As mentioned before, the bottom-up approach would be beneficial for the study of tuberculosis in a small com-

munity since it would describe more precisely the interactions with individuals, and the dynamics arising from them.

1.2 State of the art

Mathematical models have been used for estimating long-term dynamics of tuberculosis epidemics (Zwerling et al., 2015). Most of these models have been classically built with a structured top-down strategy (Ferrer et al., 2009). They divide the population into different classes (e.g., susceptible, exposed, infected and recovered in the case of an SEIR model) and fix specific fluxes between these groups. This strategy is feasible whenever the size of each class satisfies the continuum hypothesis. Nevertheless, the use of differential equations can be questioned when the population under study is not big enough. This is often the case since the incidence of sick people among infected is small (on average, only 10% of infected of tuberculosis develop an active disease).

In the last decade, agent-based modelling has been introduced to the study tuberculosis epidemiology. An agent-based microsimulation model, which is close to an ABM, was published by Murray (Murray, 2002) . In this paper, the author tackled a heterogeneous population at the individual level whose dynamics was governed by fluxes extrapolated from literature's population (compartment) models. This model took into account TB-HIV co-infection as a risk factor, the kind of active tuberculosis developed (either pulmonary or extra-pulmonary), the infectious strain, and the social structure of the population (household and neighbourhood). It was calibrated with literature data from 7 countries and 2 USA prisons. The simulations considered a closed population during a 4-year period and studied the effect of the different factors on the cluster size and distribution of *Mycobacterium tuberculosis* isolates in a particular community.

Espíndola et al. de Espíndola et al. (2011) introduced an ABM to study the drug resistance emergence in relationship with different treatment patterns. In fact, their model is close to a cellular automaton, since all the spatial cells were considered to be occupied by a single individual that may acquire different states (susceptible, latent infected or with an active disease) according to a set of probabilistic rules which are not time-dependent. Two infection routes were considered, either neighbouring (locally) or non-neighbouring (globally) caused. They took into account either drug susceptibility or drug resistance for infected and infectious individuals. Simulations covered a 317 x 317 lattice (100,489 individuals or spatial cells) for a period of up to 300 years, with a time step of 1 day. This model was not fitted to real data but focused on carrying out a mathematical study of the drug resistance emergence patterns and their interaction with the involved model parameters.

Guzzetta et al. (Guzzetta et al., 2011) tested three different approaches to study tuberculosis evolution in Arkansas: (i) an Ordinary Differential Equations (ODE) model

considering a homogeneous population; (ii) an age-structured ABM with homogeneous mixing and closed community; and (iii) an age-structured ABM coupled with a spatially explicit socio-demographic model that drives the transmission dynamics according to the three levels considered (households, schools and workplaces, and general population) in a closed population. The socio-demographic model had previously been calibrated to Arkansas data. The ABM considered the different infection status of individuals, taking into account smear-positive or negative disease states and including many age-dependent factors. It was focused on the transmission dynamics but did not explicitly consider the diagnosis and treatments of individuals with an active disease. Thus, heterogeneity of population was related with the age-structure and socio-demographic-structure. The results showed good agreement with Arkansas data, particularly the third approach which showed better fit at long-term (7 years).

A more recent work can be found in Kasaie, Dowdy, and Kelton (Kasaie et al., 2013). In this paper, an ABM coupled with a 3-level social network to tackle the role of social contacts on the infection and reinfection dynamics of tuberculosis was presented. The model at the agent level was mostly based on fluxes between different states and considered conditions of a TB high-incidence region with mean disease duration of 11 months. The model was fitted to global WHO data (WHO, 2012). A population of about 10,200 individuals (2,000 households, 50 neighbourhoods) was simulated for a period of 150 years (with a transitory of 100 years). At this point, no heterogeneity among the population was considered. In a subsequent paper (Kasaie et al., 2014) fitted the model to 2013 WHO data (WHO, 2013). The authors introduced the control of particular strains to track the transmission patterns, taking into account the molecular epidemiology.

In general, although being ABMs, most of the mentioned models above were conceived from an SEIR (or top-down) perspective. Many of them ((Guzzetta et al., 2011);(Kasaie et al., 2013); (Kasaie et al., 2014);(Murray, 2002)) included sophisticated models for the transmission routes taking into account the socio-demographic structure of the population and the different degrees of contacts between social groups. Only Guzzetta et al. (Guzzetta et al., 2011) deal with a particular reality (Arkansas), while models by Kasaie et al., (Kasaie et al., 2013); (Kasaie et al., 2014) are fitted to global data. None of them considers high heterogeneity within population, including factors like the kind of disease, possible immunodeficiency, smoking/drinking habits or immigrant-native origin simultaneously.

1.3 Tuberculosis in Barcelona

In 2013, Barcelona registered an incidence of 20.4/100,000 habitants, superior to the 11.8/100,000 hab incidence of overall Spain (Orcau i Palau et al., 2014). This situation is common in big cities that receive a substantial flow of migrants because those fluxes increase the incidence of infection. However, further study of the situation of tuberculosis

in Barcelona reveals interesting data.

TB incidence presents vast differences depending on the district, from the lowest rate of Sarrià-Sant Gervasi area of 8,9/100000 hab to 67/100000 hab in Ciutat Vella. In particular, the neighbourhood of El Raval in Ciutat Vella presents a high rate of 119.9/100000 hab. The reasons behind the disparity of occurrence are probably due to the difference in immigrants' distribution (immigrant population accounts for the 43% of Ciutat Vella's inhabitants, and over 80% of TB cases diagnosed in this district on 2012 corresponded to immigrants) and population density across the city, as well as the different life conditions for them (e.g. overcrowding, that facilitates spreading, or the access to sanitary services, which may delay the diagnostic). Figures 1.2 and 1.3 give a good image of the situation of TB both in the whole city of Barcelona and the district of Ciutat Vella.

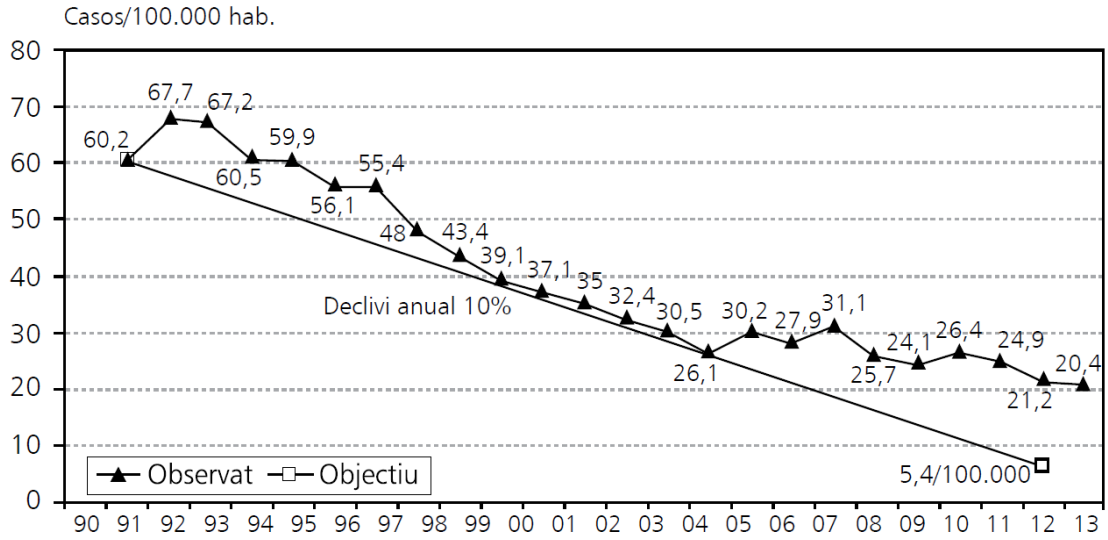

Figure 1.2: Evolution of the TB situation in Barcelona, 1990-2013, along with the objective annual decrease. Extracted from the 2013 TB in Barcelona report (Orcau i Palau et al., 2014)

This situation makes the district of Ciutat Vella the perfect candidate for our model. The detailed reports of the TB situation in Barcelona grant us access to an important quantity of data, which is very helpful when trying to build a model. Also, the higher incidence of Ciutat Vella creates a more reproducible situation in a simulation using ABM. For all these reason in the following sections we will be referring constantly to Ciutat Vella, since our model is focused on the data of that district of Barcelona.

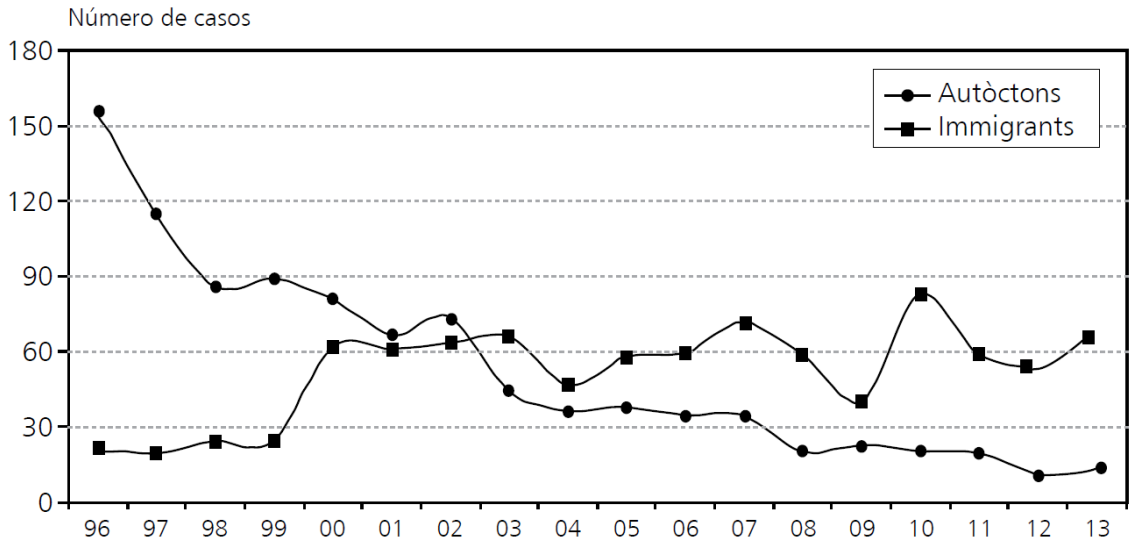

Figure 1.3: Evolution of the TB cases in Ciutat Vella according to the country of origin, 1996-2013. Extracted from the 2013 TB in Barcelona report (Orcau i Palau et al., 2014)

1.4 Objectives and outline

The final objective of this study is to develop a simulator to evaluate public health control strategies of tuberculosis in medium-sized and big cities. In order to progress towards this aim, several specific objectives were defined:

1. To improve a preliminary individual-based model of tuberculosis spreading in a closed community. The model simulates the dynamics of the population and the transmission of the disease among a heterogeneous population.

2. To implement the model in Netlogo, a user-friendly tool that should facilitate its use by non-modellers. The implementation of our model with this tool should facilitate its use by non-modellers such as health services workers or tuberculosis control units.

3. To optimize the implemented model in order that simulations are feasible in NetLogo. This optimization would be a first step towards ensuring the feasibility of the virtual experiments with the simulators, considering that NetLogo is a highly time-consuming tool.

4. To review the tuberculosis epidemiological data of last years in Barcelona and to use them for calibrating the model. Data was obtained from national statistics and the Tuberculosis Investigation Unit in Barcelona.

5. To use the simulator as a test-platform of tuberculosis epidemiology control strategies for decision-making.

6. To discuss the need for more powerful platforms for simulating the disease dynamics in big cities and to analyse the modifications that should be carried out in the model.

The work is organized as follows. Chapter 1 revises the recent works done on agent-based simulation of TB epidemics and explains the basic concepts of biology, epidemiology and modelling needed for understanding the work. Chapter 2 presents the agent-based model for the TB transmission in Barcelona and the optimization improvement implemented in order to overcome the time-consuming limitations of Netlogo. The contents of this chapter can be found in an accepted article for the Proceedings of the 2015 Winter Simulation Conference (Montañola sales et al., 2015). In Chapter 3, it is explained how the parameters of the model have been adjusted. The results of our experiments are presented in Chapter 4 , Chapter 5 includes the concluding remarks and finally Chapter 6 highlights the lines of further work.

Chapter 2

Model

2.1 ODD description of the model

ABM has been used in multiple fields, from simulation in social science to business, and of course biology and epidemiology. As a consequence of this variety, ABM models have had almost as many description procedures as models have been developed. In order to deal with all the different description systems and standardize the description of ABM models the ODD (Overview, Design concepts, and Details) protocol was developed (Grimm et al., 2006, 2010). This protocol consists of three blocks, which are subdivided into seven elements: Purpose, Entities, State variables and scales, Process overview and scheduling, Design concepts, Initialization, Input data and Submodels; although not all of the elements have to be used in the description.

In the following sections we are going to present an ODD description of the ABM developed for studying TB epidemiology in a city.

2.1.1 Overview

Purpose: The objective of this ABM is to analyse the evolution of pulmonary tuberculosis incidence in a closed community. It is fitted to Ciutat Vella neighbourhood, and the possible effects of epidemiology control strategies and public health decisions are checked by means of virtual experiments.

Entities, state variables and scales: The fundamental entities of the model are persons. We consider the persons to be able to go through five infection states: healthy, infected, sick (i.e., with an active TB), under treatment and recovered. Persons in four out of the five states, all but healthy are simulated as individuals. Healthy individuals are not considered simple entities since they have no remarkable properties, the only information useful is their position. Therefore, they are considered as *properties of the space*. The state variables of the persons mainly refer to their status in the tuberculosis infection cycle as well as the time spent in such phases and individual diagnostic time when getting sick.

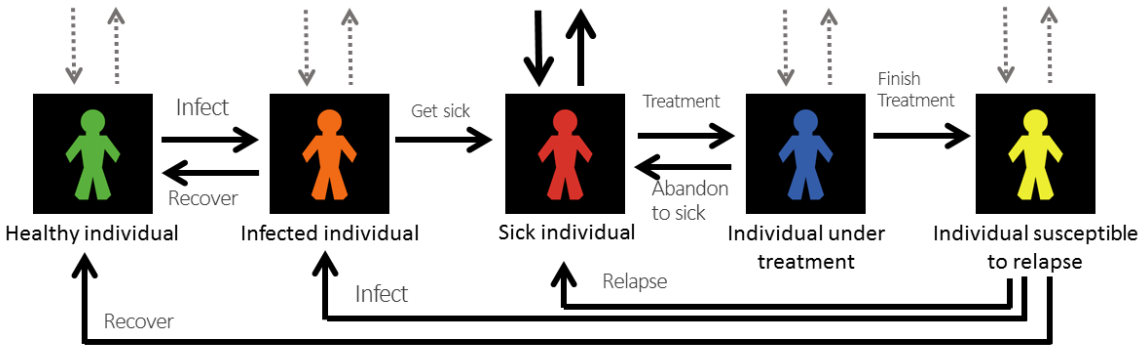

Figure 2.1: State diagram of the whole TB infection process

Other individual state variables and parameters are age, autochthon/foreign, possible risk factors (e.g. smoking), diabetes and possible immunosuppression (mainly HIV infection). Once a person gets infected, the presence (or not) of pulmonary cavitation is also considered. A state diagram of the model is presented in Figure 2.1. The population simulated is 100.000 individuals for the optimization and sensitivity analysis, and 105.123 persons for the calibration and the experiments, this represents roughly the entire population of Ciutat Vella.

The model is partially spatially explicit, i.e., space is considered but it does not mimic the real space of Ciutat Vella. Simulation occurs in a discrete space of 501 x 501 spatial cells. Each spatial cell represents an abstract local space where two persons can meet, and the bacilli can be spread in a day. Time step is set to 1 day, and the simulation may cover up a period of: 1, 2, 5, 10 or 20 years, or run until tuberculosis is eradicated, although this option is not used in this work.

Process overview and scheduling: Our model was built in NetLogo, a popular simulation and modelling tool among social simulation practitioners (Tisue and Wilensky, 2004). NetLogo is well suited for modelling a wide variety of agent-based systems. It has a user-friendly interface that allows non-experts to run simulations and perform virtual experiments (see Figure 2.3 for a screenshot of the implemented program user interface).

The simulation starts with the set-up of the initial configuration, where the population is randomly generated according to the input distributions of parameters and randomly distributed in the 501 x 501 grid. The model assumes discrete time steps of 1 day. Each day, all individuals execute a series of actions, and their variables are updated immediately.

The individual actions may be: to grow, to move, to get infected, to get sick, to be diagnosed and start a treatment, to abandon or finish the treatment, to recover, to die and in some experiments to be screened, i.e. to be identified as an infected individual by a public health campaign, the details of these actions will be provided later. When an individual dies, a new person is introduced with random particular characteristics according

to the initial distribution of individual parameters. At the end of each time, step global variables are updated. Figure 2.2 shows the flow diagram of the computational model.

2.1.2 Design concepts

Basic principles: The model is based on general knowledge about the natural history of tuberculosis. There are two essential characteristics of TB that must be taken into account in any epidemiological model. On the one hand, an infected individual does not necessarily develop an active disease; on average, only a 10% of infected people become sick. Moreover, a person remains infected for a long period and may develop an active tuberculosis after several years, but the probability of developing the disease decreases with time. Infected people are usually not diagnosed. On the other hand, only TB sick can disseminate the infection. The infection rate increases if the patient has a TB with cavitation. Once a TB sick is diagnosed; the pharmaceutical treatment takes six months. Once the treatment is finished, the possibility of getting sick again remains at 1% for 2 years. Also, additionally to the treatment administered to the active cases, in some experiments a second different treatment is incorporated. This treatment is longer than the treatment given to persons with active TB; it lasts 9 months, and is administered to infected persons to prevent the development of an active disease. There is also a probability of relapse to the infected state that is calculated similarly to the first treatment.

Emergence: Emerging phenomena are mainly related to long-term dynamics of the infection at the population level. On the one hand, only non-treated people with an active TB can spread the disease. Therefore, diagnosis time is an essential parameter for the prevalence of the disease. On the other hand, infected persons may develop an active tuberculosis a few years after the infection. Therefore, global consequences of specific conditions at a certain moment may be detected some years later.

Interaction: Local interactions between individuals are explicitly modelled and crucial for the dynamics of the system. They refer to the meeting of two persons propitiated by the spatial proximity between them and the possibility that one of those individuals with an active TB may infect the other person.

Stochasticity: Stochasticity is introduced at all levels of the simulation. The initial distribution of individual properties is randomly executed according to input distributions. Movement is assumed to be random. Each action is associated with a certain probability and thus executed according to a stochastic number.

Collectives: Two collectives may be distinguished, according with the individuals origin: autochthon and foreign. The difference between them is the diagnosis time.

Observation: Output data show the daily evolution of number (or prevalence) of healthy people, infected people, sick people, people under treatment and persons already treated.

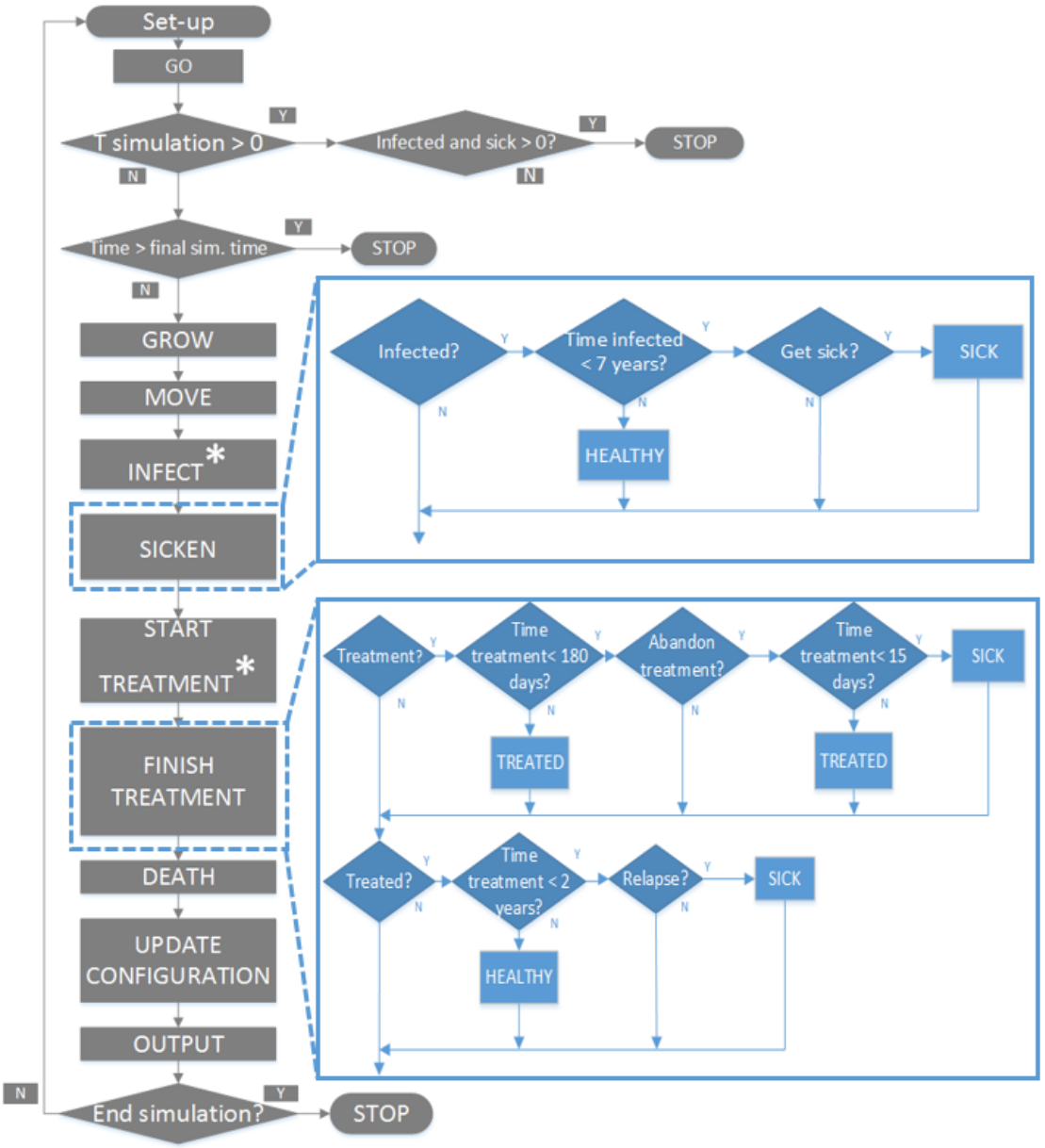

Figure 2.2: Flow diagram of the computational model. The processes marked with an asterisk only affect certain states, i. e., only healthy and treated people can be infected. The processes highlighted are the most complex ones, and are therefore detailed

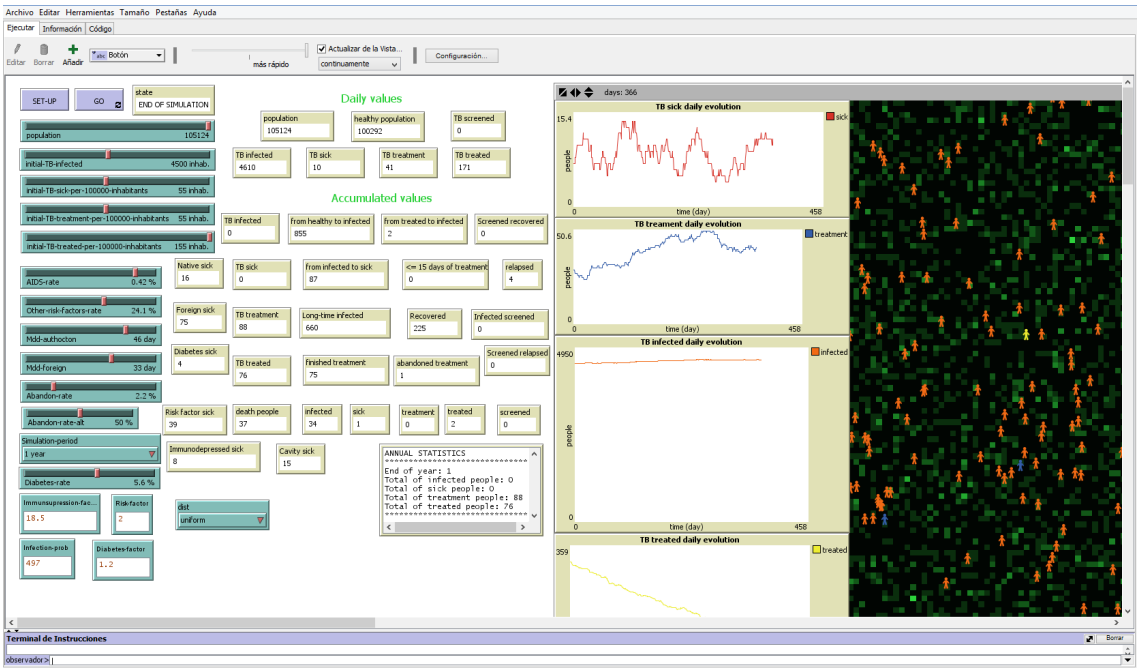

Figure 2.3: Screenshot of the program user interface at the end of a 1-year-long simulation. Green slides (left) are initial values that can be changed. Numerical outputs, as well as the world's representation and the graphical evolution of several states are also seen

These data are exported to a csv file, and an annual report is shown to the user in the interface screen.

2.1.3 Details

Initialization: The user can change some initial conditions at the beginning of the simulation. For this specific study, most of the input parameters were taken from official reports (Ajuntament de Barcelona, 2010; Orcau et al., 2011; Orcau i Palau et al., 2012; Ajuntament de Barcelona, 2012; Orcau i Palau et al., 2014). The initial population was fixed to either 105.123 o 100.000 individuals, depending on which process was going on. All percentages shown in Table 2.1 were used for calculating the configuration of initial population: rates sick, under treatment and recovered individuals per 100.000 inhabitants; mean diagnosis delay (MDD); mean treatment abandon rate; individuals with risk factors and with HIV infection. Some other initial variables are assigned randomly: individuals age (following the percentages shown in Table 2.1), and time spent in the infection state assigned.

Submodels:

- *To grow:* all individuals increase their age in 1 day each time step.

Table 2.1: Official data of Ciutat Vella corresponding to 2010 (Ajuntament de Barcelona, 2010; Orcau et al., 2011) and 2012 (Ajuntament de Barcelona, 2012; Orcau i Palau et al., 2012) used in simulations. All percentages are with respect to the total population of Ciutat Vella, except (*) that are with respect to the total number of TB sick people.

District of Ciutat Vella	Number of indiv. 2010/2012	% 2010/2012
Total population	104.507/105.123	-
Autochthon population	56.611/59.801	54,17/56.89 %
Foreign population	47.896/45.419	45,83/43.21 %
Population <= 10 years old	7.688/7.963	7.36/7.57 %
Population 10 - 65 years old	80.373/81.769	76.91/77.78 %
Population > 65 years old	16.446/15.488	15.89/14.73 %
Detected cases of TB	104/64	-
Rate of sick per 100.000 hab.	100/55	-
Rate of treatment per 100.000 hab.	100/55	-
Rate of treated per 100.000 hab.	99/155	-
Cavitation forms*	-	10.57/22 %
Detected cases of VIH	1304/440	1.25/0.42 %
Risk factors*	-	65.4/24.1 %
Total annual mortality	936/877	0.89/0.83 %
Treatment abandon rate*	-	9.1/2.2 %
Diagnosis delay (days)	48/39	-

- *To move:* all individuals move randomly to a neighbouring cell, once a day.

- *To get infected:* if there is a number of individuals susceptible to TB (healthy and treated) different from zero in the proximity of a sick individual, meaning one of the 4-neighbouring spatial cell, this sick person may infected one of them with a certain probability. The total of susceptible neighbouring individuals is computed and then the infection process is repeated as many times as individuals have been found. The infection probability depends on the type of TB disease that the sick person has, either cavitated or non-cavitated. A cavitation is considered to duplicate the infection probability. These probabilities are fixed in order to reproduce the ratios of 10 infections / cavitated TB sick and 5 infections / non-cavitated TB sick in 60 days. Once infected, the infection time of the individual is set to 0 and starts increasing each time step.

- *To get sick:* once infected, the individual may develop an active TB according to a particular annual probability that decreases with infection time during the 7 years post-infection. It is neglected for the subsequent years (t > 7 years). Since simulation time does not cover periods longer than 10 years, the approximation is good enough. This probability is multiplied by a certain factor for immunodeficient people, and the same happens if there are other risk factors (smoking, alcoholism) or

if the patient has diabetes. The chance of becoming a TB sick individual is evaluated at each time step for all infected persons. Globally, the average of a 10% of infected developing an active disease is satisfied. The possibility of relapse (getting sick again) for recovered patients is also daily evaluated according to the individual relapse probability (see below). Once an person gets sick, the disease time counter starts running until the individual diagnostic time is reached.

- *To be diagnosed and start a treatment:* each individual has a particular diagnostic time that is randomly assigned when getting sick. This individual times are assumed to be distributed following a normal distribution centred around 42 or 46 days for autochtonous people and 33 for immigrant people, which is the median diagnostic time according to the 2012 TB in Barcelona report (Orcau i Palau et al., 2012) and standard deviation 4 (for the optimization process described later in this chapter,the values used are different, 42 days for the delay for autochtonous and 51 for immigrant people, this data is taken from the 2010 TB in Barcelona report (Orcau et al., 2011)) . When the sick time counter reaches this values, the individual is diagnosed. Once diagnosed, it is assumed to start a medical treatment and stop spreading TB. Individual time under treatment is initially fixed to 0 and then updated each time step.

- *To abandon the treatment:* there is a certain probability that an individual abandons the treatment before finishing it. This possibility is daily evaluated for each patient under treatment, according to the input abandon probability. If a person abandons the treatment during the initial 15 days post-diagnosis, it becomes ill again. If it abandons the treatment after 15 to 180 days post-diagnosis ($t_{treatment}$), it is considered to be recovered but with a certain probability of relapse the following 2 years. This probability is considered to decrease linearly from the 100% of a 15-day abandon to the 1% of the 180-day treatment period.

- *To recover:* when a sick individual is diagnosed and treated for 180 days, it becomes recovered and a relapse probability of 1% is assigned to it (this is the probability of getting sick again the following 2 years). After 2 years, the individual is considered to be healthy.

- *To die:* each individual has a certain probability of dying according to its age. These probabilities are fixed using demographic data from Ciutat Vella in 2010 or 2012. Accordingly, the daily dying probabilities are considered to be $6.877 \cdot 10^{-5}$% for individuals under 10, $5.454 \cdot 10^{-4}$% for individuals between 10 and 65, and $1.2249 \cdot 10^{-2}$% for individuals over 65 (as before, a different set of data was used in the optimization, in particular the probabilities were: 0, $1.293 \cdot 10^{-2}$% and $2.192 \cdot 10^{-2}$%, for the 0-10, 10-65, +65 years old individuals, respectively). It is a coarse simplification,

but it is enough for the purposes of this model. Besides, TB sick people have a specific probability of dying from tuberculosis. This probability is daily evaluated for each sick individual, taking into account that the 40% of non-treated TB sick may die in 5 years. Each time an individual dies, a new individual is introduced to the simulation world with the aim of keeping a constant population. Its characteristics are fixed according the distribution of the initial population.

- *To get screened:* this procedure is used only in some of the last experiments. A certain amount of infected is screened, that is, it is identified as infected by medical authorities and undergoes a preventive treatment. Therefore, these infected individuals change to a new class called treatment-screen that has almost equivalent to the treatment class, with the same behaviour and different duration.

2.2 Optimization

NetLogo provides a visual interface that helps non-modellers when creating experiments. The existence of this interface is a crucial feature of this model since its obvious users are public health workers who most probably do not have any experience with building models.

Although the proposed implementation in NetLogo shows good results, with around 100,000 agents the simulation is too slow. Therefore, we proposed an optimization to reduce the execution time. In this section we only present a summary of the optimization process, a detailed description is presented in appendix A. It seems that most of the simulation time was spent updating the world-view since the simulation was moving and updating 100,000 agents continually.

During a simulation, most of the agents are in the healthy state (> 95 %). The agents in this state do not have any particular information apart from the few attributes that are shared by all agents and that are assigned to them initially. They can move, grow in age and die according to a probability, but there are no relevant changes in their variables until they get infected. Besides, the order of magnitude of this subpopulation (10^5) is far from the order of magnitude of infected (10^3) and sick (10^1) subpopulations. Therefore, it should be possible to describe them with global (or local) continuous variables.

Thus, we proposed to transform the healthy agents into properties of the environment. In NetLogo, the spatial cells of the grid are called patches and are described by local variables that may change according to specific rules. Therefore, we define the local number of healthy individuals as a new property of the patches. With this modification, all agents in a healthy state were removed (> 95 % of the agents), but we still consider them from a collective approach. The infection process occurs in a similar way taking into account the amount of healthy individuals in the Von Neumann neighbourhood of a sick individual.

The particular properties of the resulting infected individuals are assigned according to the percentages shown in Table 2.1. In the following section, we show how we obtain similar results with respect to the original (non-optimized) version.

After changing the implementation, we needed to check that the results achieved with the optimized version were equivalent to the previous one. To do so, we slightly modified the initial conditions of the original simulator to get observable (reproducible) tendencies in all subpopulations. Then, we compared the evolutions of the population of each state during one year of simulation obtained by both versions. In Figure 2.4 we can see those evolutions for both the original model and the modified one. The results show the mean of 50 runs for each case. These graphics show a good agreement between the results of the two versions. Therefore, we can consider the optimized model is equivalent to the original one.

Once we established that the comparison of the results was satisfactory, we recalculated the execution time to see how important is the reduction obtained with this new implementation. In Table 2.2 we can see the execution time for both models in a Quad 3.20 GHz Intel Core i5 personal computer. The optimized version reduced time spent by 88%.

Finally, we run some extra test in order to check how the NetLogo model scales as the size of the system (population size) increases. To do so, we decided to run a 1-year-long simulation with a similar population corresponding of Barcelona, 1.600.000 habitants (a medium city) and Bombay, 12.500.000 habitants (a big city). In the Barcelona's case, the simulation took almost 12 hours, whereas in the Bombay's case NetLogo crashed after 3 hours of work.

Table 2.2: Comparison between the original (all states have agents) Netlogo implementation and after the optimization of considering the healthy individuals as a property of the spatial cell.

Model	Original	Optimized
Execution time (min)	126.03	15.03

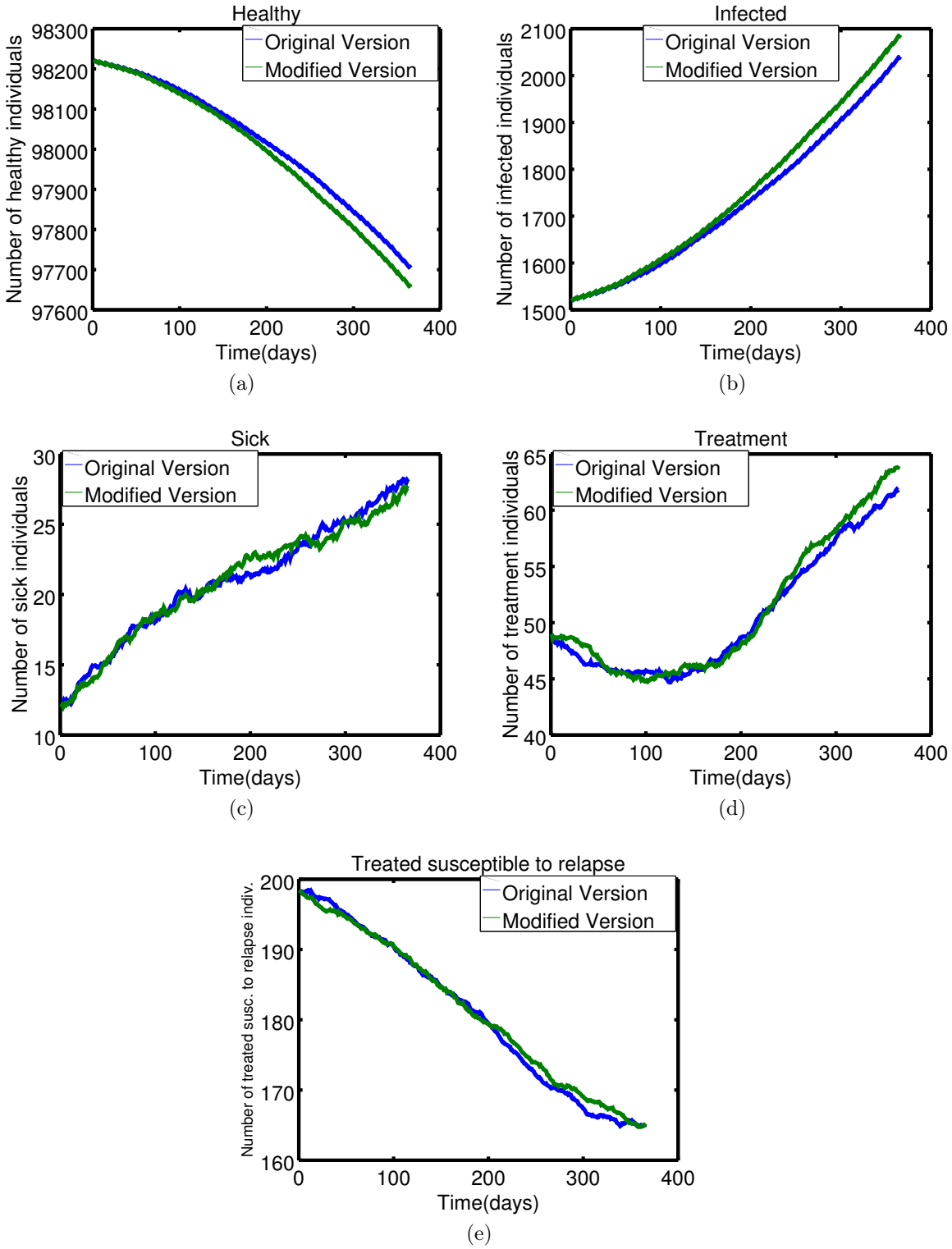

Figure 2.4: Results of the optimized version of the model compared to the original one

Chapter 3

Calibration and sensitivity analysis

Once the model was fully implemented and optimized, the next step was to fit the model to real-world behaviour. Once this was done we would be able to perform virtual experiments that could help public health decision makers to do their jobs. We used the data from the report about TB in Barcelona of 2012 (Orcau i Palau et al., 2012) as initial conditions, and the same report but of 2013 (Orcau i Palau et al., 2014) to check the results after one year of simulation time.

To fit the model we needed to perform a sensitivity analysis and then a calibration process. In the following subsection both of them are explained.

3.1 Sensitivity analysis

The first step in the analysis was to identify the parameters of the model that can change and the results that we should obtain. By doing so, we can identify which of these parameters significantly affect the results and which of them do not, in order to then calibrate the model by modifying their value. We start listing the parameters to change, which are five, listed below:

Parameters to change

- Number of initial infected individuals
- Infection probability
- Immunodeficiency multiplication factor
- Risk-factor multiplication factor
- Diabetes multiplication factor

The first parameter is the number of infected individuals that are created during the set-up. The number of initial infected individuals is included in the list of parameters

because of the lack of data. There is not a reliable estimation of the number of infected individuals and the procedures for the diagnosis of the infection cannot provide with that information either. Therefore, it will be used to fit the model to the Ciutat Vella data.

The second parameter, as its name indicates, is the probability of a healthy or treated individual of getting infected when it is close to a sick person.

The last three parameters are multiplication factors that increase the probability of getting sick and are properties of the agents. The immunodeficiency factor increases the probability of getting sick if the infected individual has some type of immunodeficiency, mainly HIV infection. The risk-factor accounts for the increase in the probability of getting sick if the individual is a heavy smoker or drinker.Lastly, the diabetes factor does the same as the previous with infected individuals that suffer diabetes.

Next we listed the values that we wanted to check against the real data to fit the model, they can be seen below:

Results to check

- Total of sick people during simulation period
- Percentage of HIV/TB co-infection
- Percentage of diabetic TB sick
- Percentage of other risk factors present on TB sick

We performed the sensitivity analysis by varying the elements in the first list up and down by a 10% and then comparing the value of the elements of the second list. This analysis allowed us to evaluate which parameters have a stronger impact on the infection and sickening processes. The results are analysed after 30 executions using a box-plot diagram with the GNU Octave software.

A box-plot diagram is a useful resource to represent groups of numerical data. It is based on quartiles, and it is constructed as follows. The red line inside the box is the second quartile, that is, the median. The limits of the box are the first and third quartiles. The extension of the verticals lines are either the maximum value or 1.5 times the difference between the first and third quartile (Interquartile range or IQR) for the upper bound, or the minimum or 1.5 times the IQR for the lower bound.

The superposition of different box-plots gives us good idea of the stochastic variability of the model, and at the same time, it allows us to determine the effect of the change of the different parameters in the output of the model.

In Figure 3.1 the results of the analysis for the five parameters listed above can be seen. We can see how the infection probability (Figure 3.1e) is the parameter that affects

less the output of the model. Thus, this parameter will be fixed at the current value of the rest of the work. The three multiplication factors (Figures 3.1b-3.1d) also have little effect on the measured output, which is the number of accumulated sick individuals in one year of simulation. However, we will use these three parameters on the calibration process to make sure the HIV/TB co-infection rate is fitted to that from (Orcau i Palau et al., 2014) along with the percentage of risk-factors and diabetes. Finally we have the initial number of infected individuals (Figure 3.1a) whose initial value was an assumption of the model due to the lack of data, and will be fitted in the posterior calibration process. This parameter is found to be the most sensitive parameter for the model since its variation produces a directly proportional change in the output. Therefore, we will consider this parameter key to the calibration.

3.2 Calibration

After identifying the most crucial factors, we started the process of the model calibration. In this process we focused on the parameters of the first list to which the model was more sensitive, that is, for a variation of the parameter the output also presented a substantial variation.

This process was carried out in a similar way to the sensitivity analysis. We then performed 30 executions of the model varying the parameters to calibrate and represent the results using Octave with a box-plot diagram. In this case, we also marked with a horizontal line the value of the output found in the literature (Orcau i Palau et al., 2014) in order to illustrate whether the output of the model fits the real-world data.

The final results of this process are shown in Figure 3.2. The most important is, without any doubt, the initial number of infected individuals. It has a great influence in the number of sick individuals. As can be seen in Figure 3.2a, the TB prevalence results with a value of 4500 infecteds were close to 78, which is the value found in the report.

The multiplication factors showed little effect in the number of sick individuals. However, they had a remarkable influence on the other three results from the second list. In Figure 3.2b we can see the outcome of the calibration of the diabetes multiplication factor. This parameter affecd mainly the percentage of diabetic sick individuals, which according to the literature is 6.99 %. The calibration yielded a value for this parameter of 1.2.

The next parameter was the other risk-factors (alcoholism, heavy smoking) multiplication factor. Similarly to the previous setting, this number did not affect the number of individuals with an active disease significantly, but the percentage of sick with one of these risk-factors, which according to the data is the 40.5 %. In Figure 3.2c it can be seen that the value that fits this behaviour was 2.

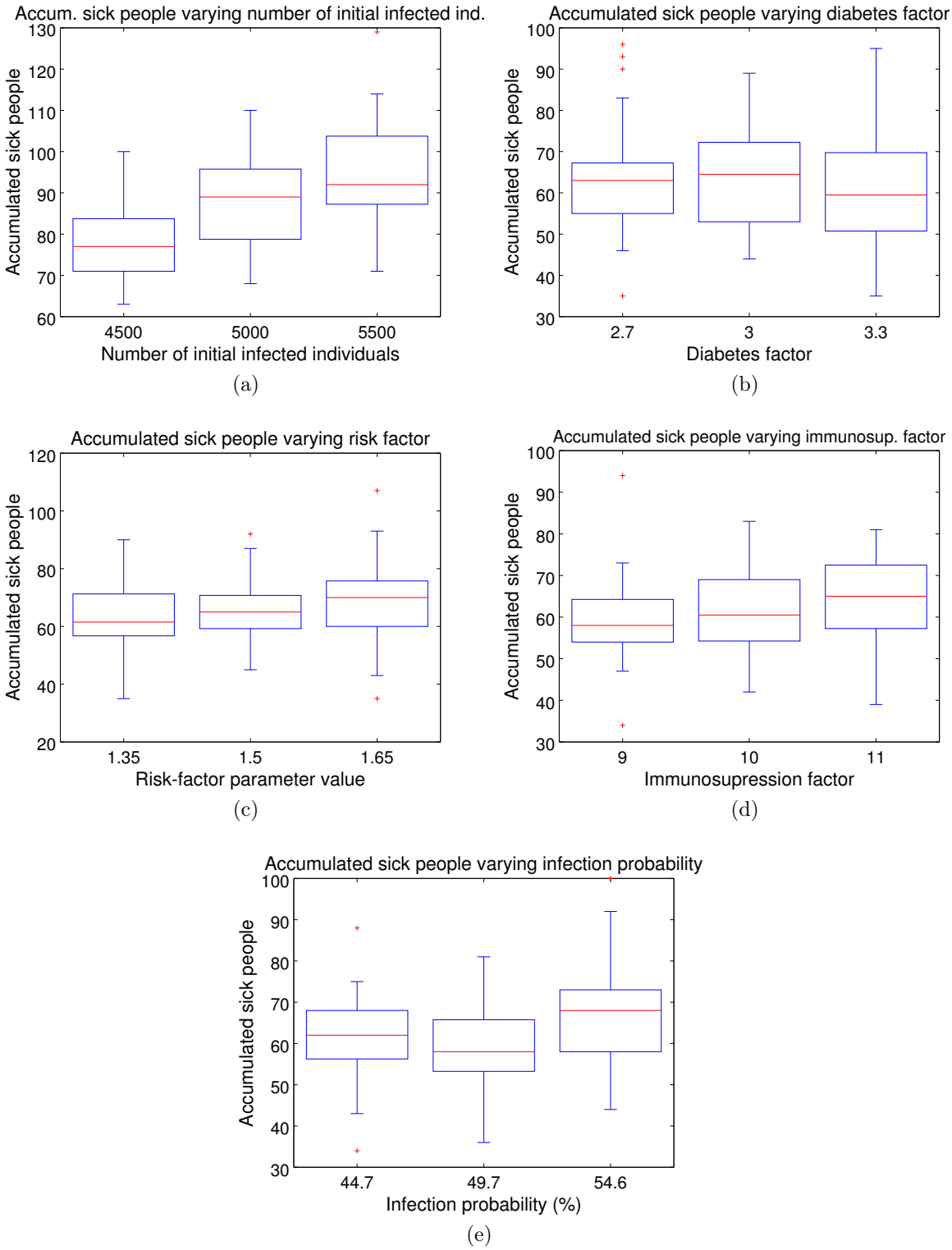

Figure 3.1: Results of the sensitivity analysis: (a) for the number of initial infected individuals, (b) for the diabetes multiplication factor, (c) for the other risk factors multiplication factor, (d) for the VIH multiplication factor and (e) for the infection probability

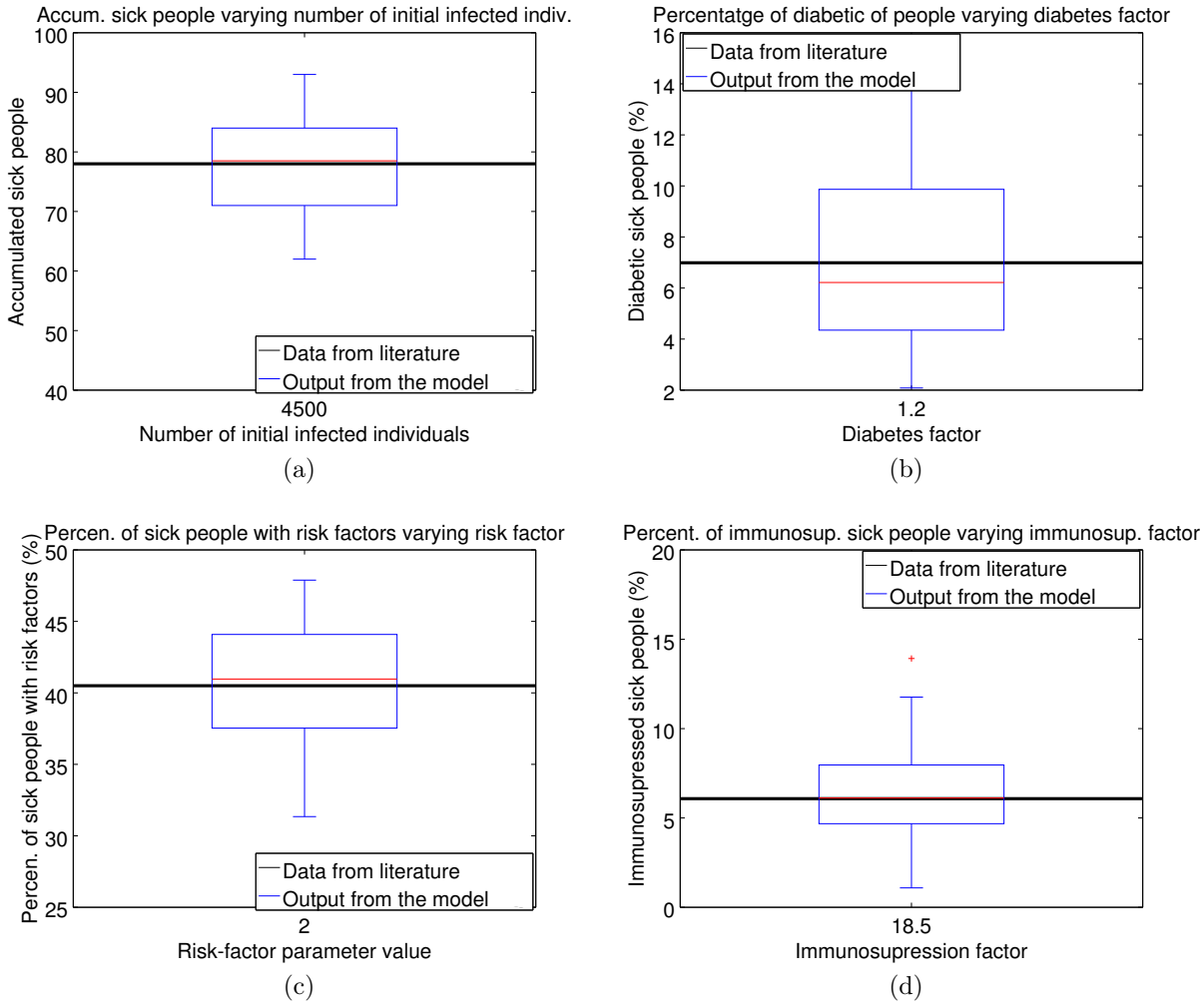

Figure 3.2: Results of the calibration process: (a) for the number of initial infected individuals, (b) for the diabetes multiplication factor, (c) for the other risk factors multiplication factor and (d) for the HIV multiplication factor

Finally the last parameter to calibrate was the immunodeficiency factor (Figure 3.2d). It was found in the literature that the percentage fo HIV/TB coinfection is 6.08 %. This result can be obtained for a value of 18.5 of the HIV multiplication factor. This value is significantly bigger than the other factors, but this could easily be explained by the natural history of the tuberculosis. The immunodeficiency caused by HIV makes it easier for the *M. tuberculosis* to develop in the host.

Chapter 4

Results

Once the model is accurately calibrated, we can start to use it to perform virtual experiments. There are two lines of experimentations in which we are interested. The first one is the effect of the initial distribution of infected times and the second is the impact that a screening campaign might have on the TB situation.

4.1 Distribution of infected time

The infection time is the time elapsed (in days) since the individual was infected. In our model, we assumed that a person can keep infected during a maximum of 7 years. Therefore, we can separate the infected individuals according to which year of infection they are, and call this classification the distribution of infected times.

We do not have any data of what kind of shape this distribution has; thus the initial choice was to assume a uniform distribution. In fact, all the work previously exposed is done with a uniform distribution.

Tuberculosis is a disease with a crucial long-term dynamics, due to, for instance, the fact that the sickening process can occur several years after the infection or the possibility of relapsing or reinfection. This may lead to distributions of infected times very different from a uniform distribution.

For example, if it were the case that in a community with a high incidence the authorities took bold measures against tuberculosis and manage to decrease the incidence remarkably. After a few years, the community would have a great number of individuals with a long infection time and a smaller number of people with a young infection. This situation would correspond to a distribution with the shape of a positive exponential function.

We could think about the opposite case, a community with a low incidence that experiments a sudden TB outbreak. This outbreak could be caused for example by an immigra-

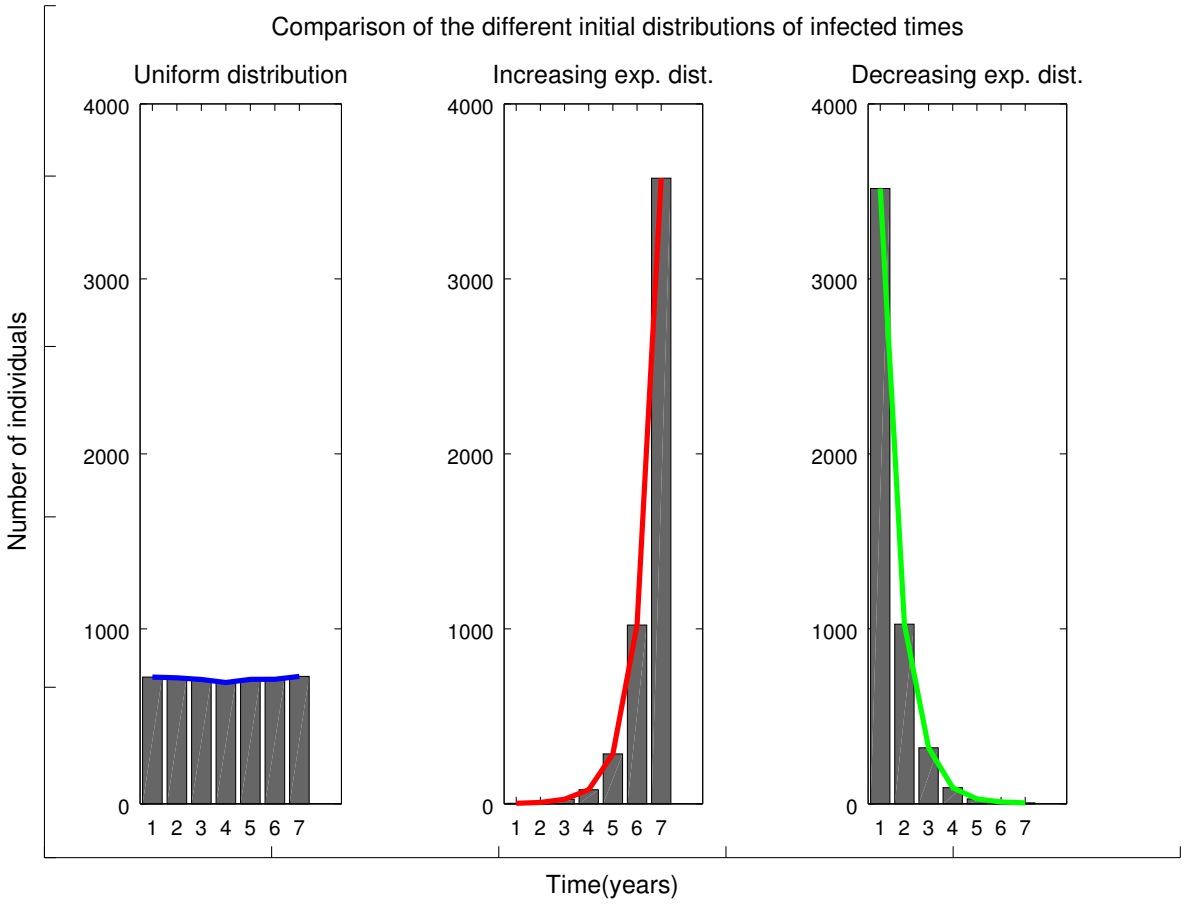

Figure 4.1: Different initial distributions of infected times used in the experiment

tion wave from countries with a high TB incidence, e.g. Pakistan (in 2013 according to the Barcelona City Hall (Ajuntament de Barcelona, 2012) a 7.7% of the population of Ciutat Vella was born in Pakistan, a country with a TB incidence of 270 per 100.000 habitants according to the WHO). In this situation presumably there would be a significant number of new infections and the distribution would present a high number of recently infected individuals with lower levels of older infections. In this case, the distribution would look like a negative exponential function.

In Figure 4.1 we can see the three initial distribution that would correspond to the uniform assumption and the two situations presented above. These three distributions are used as initial conditions in the experiments.

With each of these distributions, we executed 30 runs of a simulation with a duration of 1 year. At the end of this year, we recorded the number of people who had become sick during this period. With all these data, we built a box-plot diagram for each distribution that can be seen in Figure 4.2.

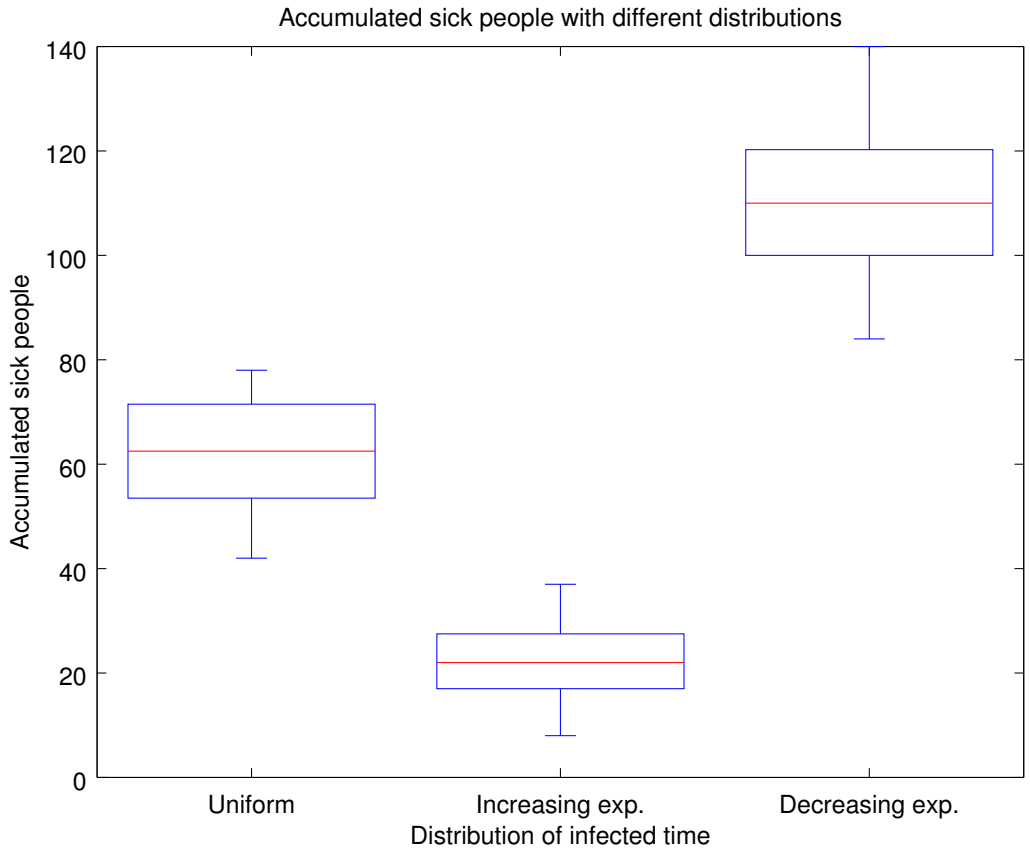

Figure 4.2: Effect of the distribution of initial infection time on the number of sick individuals

We can see in Figure 4.2 how important the chosen initial distribution can be for the results of the simulation. From this outcome, we could think that it is worth to concentrate efforts on tackling new infections since a distribution where most of the infected are old infections shows six times less sickened people during a year. However, this result does not give enough information on its own to draw any conclusions, and we will focus now on results obtained in a longer simulation.

Next we run a simulation during 10 years with each of the initial distributions shown in Figure 4.1 with 10 runs of each and record at the end of every year the distribution of infected time. With all the data, we average the distributions and the standard deviations, which are shown in the error bars of Figure 4.3. With the error bars, we can see again that the model presents a high variability, which increases as the simulation advances. This variability is caused by the randomness that the model has due to the use of pseudo-random numbers multiple times. Also, it is possible that more runs are needed to smooth the results, but even with the optimization the simulations of ten years are very time-consuming.

As for the distributions, we observe how the peaks in the exponential distributions move forward. In the increasing exponential, the peak disappears at the end of the first year since all of the infected in year seven who do not develop the disease become healthy. Once we have reached the seven years of simulation these peaks are gone and the distributions do not show any particular shape. A very different case is found for the simulations with the initial uniform distribution, where the final distribution is apparently decreasing, with a linear decrease. This decrease indicates that the rhythm of infection increases every year with a uniform distribution. This situation is probably caused by the fact that the model assumes a closed population, and therefore for long simulations the infected population tends to grow, causing more individuals to become sick which in its turn means more infected. To assess the magnitude of the decreases at the end of the simulation we perform a linear regression to the three distributions, painted in black in Figure 4.3. The slopes obtained are -0.12 for the uniform distribution, -0.012 for the exponential increasing distribution and -0.03 for the exponential decreasing distribution. Therefore, we can assume that the exponential distributions tend to essentially flat distributions after 10 years, while the uniform finally presents a slight linear decrease.

4.2 Screening

4.2.1 Concept of screening

According to Porta (2008) the concept of screening can be defined as:

> «*The presumptive identification of unrecognised disease or defect by the application of tests, examinations, or other procedures which can be applied rapidly.*

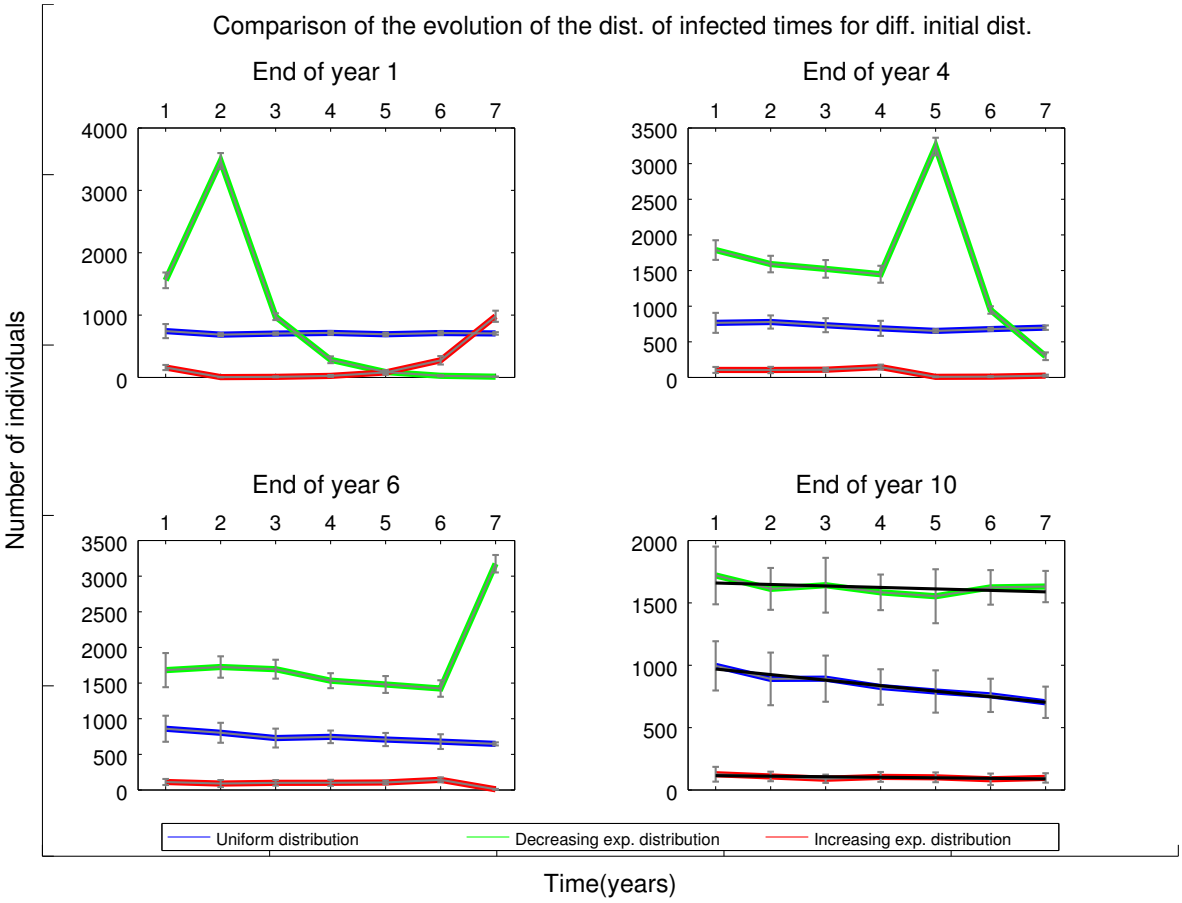

Figure 4.3: Distribution of infected times at the end of years 1, 4, 6 and 10 of simulation of duration 10 years with uniform initial distribution (blue), increasing exponential initial distribution (red) and decreasing exponential initial distribution (green)

> *Screening tests sort out apparently well persons who probably have a disease from those who probably do not. A screening test is not intended to be diagnostic. Persons with positive or suspicious findings must be referred to their physicians for diagnosis and necessary treatment.*
>
> *The initiative for screening usually comes from an agency or organization rather than from a patient with a complaint. Screening is usually concerned with chronic illness and aims to detect disease not yet under medical care. Screening may identify risk factors, genetic predisposition, and precursors, or early evidence of disease.*»

In simpler words, the goal of a screening campaign is to detect diseases before the person affected shows any symptoms. Screening is a established tool in epidemiology. It is particularly fit to be used in TB epidemiology because of the long time that a person may remain in the LTBI phase and, therefore, show no symptoms. These individuals are the best candidates for screening campaigns.

However, for a screening campaign to be effective, it has to be well planned. The public to which it is aimed has to be selected according to the details of the disease, the population on which the tests will be performed and the means available. In epidemiology there are several types of screenings defined, such as: mass screening, acting on the bulk of the population; selective or targeted screening, acting on a particular sector of the population that present risk factors or contact tracing, acting on the contacts of a diagnosed individual that would have the greatest probability of having been infected.

The goal of this experiment is to determine which of the several types of screening that we will test is the most effective one. This experiment will also serve as a demonstration of the capabilities of the developed simulator in public health decision-making. We will examine three types of screening: variable random screening, which is a slight variation of mass screening, selective screening and contact tracing (Porta, 2008). The details of the protocols and the simulations will be provided in the subsequent sections, and finally a comparison between the three types and a reference value, where no screening is applied, will be carried out in the last section.

4.2.2 Variable random screening

The first approximation taken to reproduce the screening process was to randomly identify, in the middle of the first year, a certain percentage of the newly infected individuals (infected less than 1 year ago). This percentage was an input of the experiment, and it is called the degree of the screening.

The particular experiment consisted of varying the degree of the screening with the following values: 0% (no intervention), 25%, 50%, 75% and 100%.The simulation started from a uniform distribution such as the one in Figure 4.1 with the parameters obtained from the calibration process and a duration of 10 years. The rest of input data can be found in Table 2.1, among the data corresponding to 2012. The outcome of the experiment consists of the distribution of infected times and the number of new sick and infected individuals and the count of infected persons. All these information was stored at the end of each simulation year, and the simulation was run ten times with a duration of ten years.

In Figure 4.4 we can see a summary of the evolution of the annually recorded infected times distribution. At the end of the first year, the action of the screening has taken place, and consequently there is a reduction in the number of infected persons with less than two years of infection.

It should be noted that the detection is done by half of the first year. Therefore, after this detection there are still about 180 days during which new infections can occur. Still the number of infected with less than one year is always smaller than the reference value with no intervention.

In the following years there is no other action, the screening process is only performed in the first year of simulation. However, the decrease advances in time mainly for two reasons. Firstly, every year all of those infected individuals with seven years of infection that do not get sick become healthy. Secondly, with less young infection there is less probability of individuals sickening, and, therefore, less new infections.

As the simulation progresses, the difference between the reference values and the outcome of the intervention becomes clearer. There is an significant difference between the reference and the result of the screening at 25%. The same happens between 25 and 50% and between 75 and 100%. Surprisingly, there is no remarkable difference between the 50 and 75% at the end of the simulation.

This distribution allows us to take a first look at the effect of the measure. However, to assess thoroughly this impact we need more information. This information is given by the evolution of the number of sick individuals.

In Figure 4.5 we can see the evolution of the number of persons who have been diagnosed at the end of each year of the simulation of 10 years of duration. Figure 4.4 already showed how the variable screening caused a decrease in the number of infected, and in Figure 4.5 we see a similar behaviour for the number of sick individuals. It is clear that a process of screening also causes a decrease in the number the number of sick persons. This decrease is particularly notable in the last years of the simulation when there is a difference of about 15 individuals for the 25% screening. In the 100% case, this difference is of almost 30 persons less.

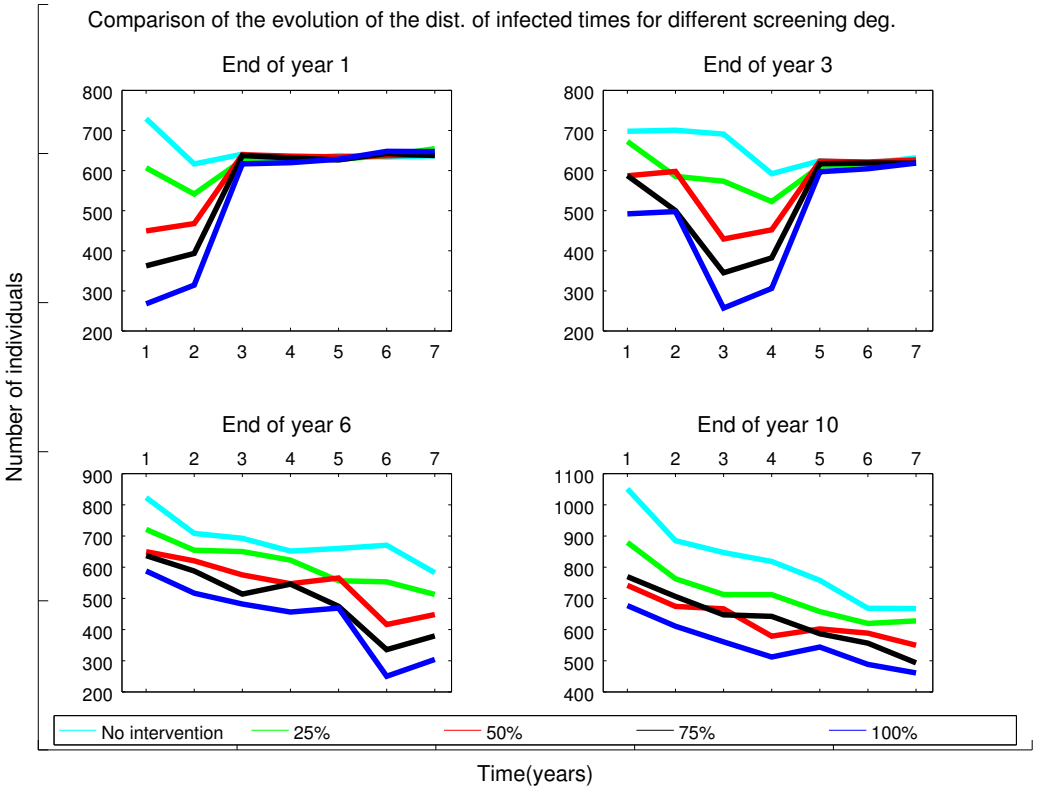

Figure 4.4: Distribution of infected times for different degrees of random variable screening at different moments in a simulation of 10 years. Superior row: left, end of year 1; right, end of year 3. Inferior row: left, end of year 6; right, end of the simulation

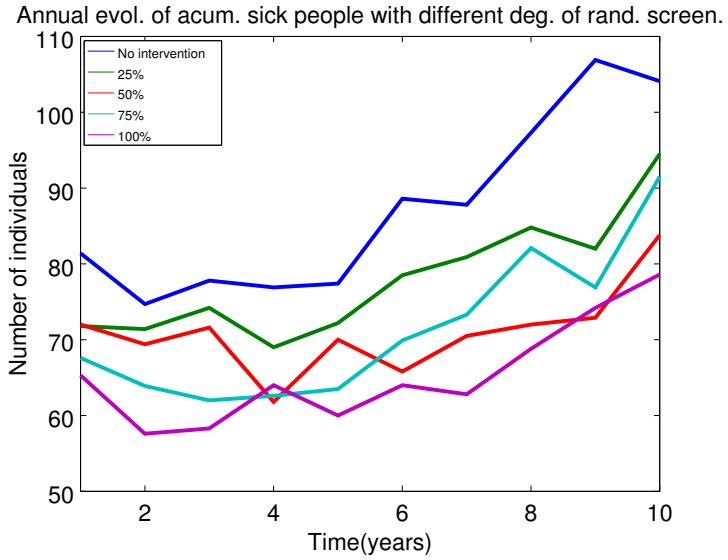

Figure 4.5: Evolution of the number of accumulated sick persons at the end of each year of simulation for different degrees of random variable screening with values 0, 25, 50, 75 and 100%

In the discussion of the evolution of the infected time distribution, we found a striking similarity between the outcome of the 50 and 75% screenings. This outcome is repeated in Figure 4.5, where the evolution corresponding to the 50% screening is not only very close to that of the screening with 75%, but in some points it is lower than the value for the 100% screening. Of course this could be caused by the variability introduced by the randomness of the model, but the analysis of both figures leaves no doubt that there is no added benefit in performing a screening over 75% of the newly infected population compared to that obtained with a screening over the 50%.

Finally it is worth mentioning that the no intervention scenario in a closed population leads to a gradual increase in the infected and sick population. Thus, public health measures are necessary to control the diseases incidence. Also, it is observed that in the fight against TB the outcome to control should probably be the number of sick individuals, which is easier to control and know in real-world situations due to the particularities that TB presents with its latent infection.

In these figures, there are no error bars due to the already high density of information enclosed in the figures. In this case, as well in the comparison between screenings that will come later, we will show the results of the analysis of the standard deviation in a separate figure. In Figure 4.6 we can see the variation, in percentage, of the standard deviation with respect to the average.

The variation has been calculated as follows: from the ten executions of the ten-year-long simulations we calculated the distribution and averaged them. Also, from the bulk of data the standard deviation of the distribution was calculated. We then had ten vectors of standard deviation (one for each year of simulation) per degree value. These vectors have seven elements, one per each year in the distribution. Then we divided the standard deviation by the average value corresponding, took the maximum and minimum for each year of simulation and plotted the evolution of this variation along the simulation, which is what is depicted in Figure 4.6. This figure shows that the variation is almost zeros for the initial conditions and increases as the simulation advances reaching values no greater than a 10% of the average, independently of the screening. Taking into account that we only used 10 executions for the experiments, due to the computational cost, the error is what should be expected considering the stochasticity of the model.

4.2.3 Selective screening

The first type of screening presented interesting results but was artificial and unrealistic since it is not easy to detect TB infections, let alone 75 % or even a 100% of them. Consequently, we need to design a screening that is more realistic because the goal of this simulator is to serve as a tool for public health decisions.

The difficulty of identifying TB infection resides in the fact that the individual does

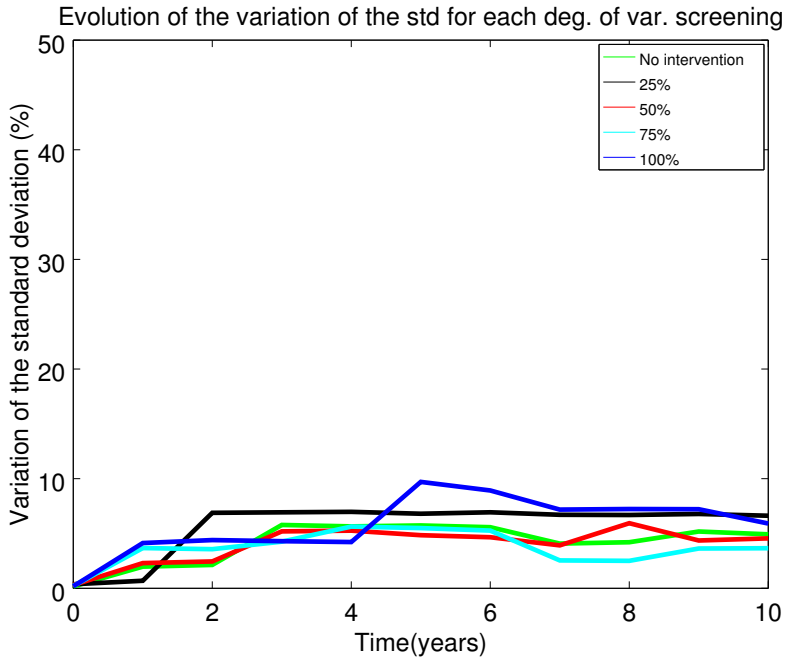

Figure 4.6: Evolution of the variation of the standard deviation for the five degrees of random variable screening used in the experiment (0, 25, 50, 75 and 100%). The values on the y-axis are expressed as a percentage of the average value of the distribution

not show any symptoms until he or she develops the active disease. There are tests to identify TB infected, such as the Tuberculin Skin Test or TST, but administrating it to everybody is not a feasible measure.

A better strategy would be to concentrate on groups that are easier to control. In our model, the best candidates are the HIV-infected persons and the immigrants. The first are included in the group of notifiable diseases in most countries, and this is also the case in Spain.

Generally, not every immigrant is accounted for in the administration registers of population. However, in some cities the immigration services control the arrival of persons from countries with a high incidence and administrate a TST to determine whether this person is infected. However, to think that a large percentage of the immigrants could be detected in a city like Barcelona is not realistic; therefore, when designing the experiment we will only detect a few of them to account for possible errors.

Taking all the previous factors into consideration, the designed selective screening consisted of an identification of 20% of all infected immigrants and 80% of all persons with HIV/TB co-infection. As in the variable screening, this took place once in the middle of the first year. The experiments were performed in a simulation with a duration of 10 years and the same initial conditions that were described in the last section. The outcome

is also the same that was stored in the previous section of the variable screening. These results will be commented later when we compare the different types of screening since in this case there is no variation of parameters.

4.2.4 Contact tracing

In the last section, we took a step towards a more realistic screening protocol with the selective screening. Following the same path our following step was to implement a screening procedure that is well known in epidemiology, the contact tracing procedure. A guide of how this procedure works can be found in National Institue for Health and Care Excellence (2015).

This process consists of identifying the contacts that a person diagnosed with active TB, which is denominated as index case in epidemiology, which have a substantial risk of having been infected. This include household contacts, i.e. persons sharing a bedroom, kitchen, bathroom or sitting room with the index case or possible contacts in the workplace, specially if the type of work includes being with people in a limited space, e.g. flight attendant, or working with children, e.g. school staff member or student.

Of course, after this analysis, a whole set of actions is deployed depending on the details of the case. Our implementation is, however, much simpler. Based on data corresponding to Barcelona from the talks in the *World Tuberculosis day* held in Barcelona (Millet et al., 2015), we know that in the period between 2009-2011 contact tracing of 541 active TB cases allowed to diagnose an extra 43 cases of active TB and 1239 TST positive (i.e., infected with LTBI). From this data, we construct our contact tracing approximation. In this procedure for each active TB case diagnosed, we will have two positive TST, i.e. we will identify 2 infected individuals. Moreover, with a 10% probability and extra active TB case might be diagnosed.

Again, this process was included in a 10-year-long simulation in the same conditions as the previous experiments, and the same outcome was stored. This outcome as in the case of the selective screening will not be analysed here but in the next section, where it will be compared to the results obtained in the no-intervention scenario, the selective screening and the variable screening with a 100% degree.

4.2.5 Comparison between types of screening

We already analysed the results obtained from the variable screening in section 4.2.2. However, in that analysis we focused on the impact of the degree of the screening. Now we are going to concentrate on the efficacy of the different types of screening tested.

The outcome was the same for all the types of screening, the distribution of infected times and the number of sick individuals at the end of each year. In Figure 4.7 we can see a summary of the evolution of the distribution of infected times for the no intervention scenario, the contact tracing, the selective screening and the random variable screening of degree 100%. From the random type, we have only chosen the most effective degree because our goal is to compare the types of screening in order to determine which one is the most effective.

The distribution of infected times at the end of years 1, 3, 6 and 10 can be seen in the Figure 4.7. The behaviour of the random screening has already been analysed in section 4.2.2 so we will focus on the other two screenings.

The actions of the selective screening occur only in the first year, but meanwhile in the random screening we identified new infections, in this case, the infection time plays no role. This difference explains why the minimum of the distribution is in the first year in the variable screening, whereas in the selective screening it is in the second year. As the simulation advances, it can be appreciated how the selective screening does not have a great impact on the infected population since the lines of this procedure, and the no intervention scenario are very close. At the end of the simulation, however, we see that there is a little effect overall, despite the fact that the distributions are identical for the late years of infection.

The distribution corresponding to the contact tracing is entirely different from those seen before. This result in not so surprising since the mechanism involved in this procedure is utterly different to the others screening already exposed. In this case, the distribution follows initially very closely the distribution corresponding to the reference value. The shape of both distributions, contact tracing and reference, is very similar to the whole simulation, with the significant difference of the magnitudes. By the end of the sixth year there is already and offset of 200 individuals between the two distribution, and this gap continues increasing as the simulation advances and reaches a maximum value of almost 300 persons by the end of the simulation. This final state of the contact tracing is nearly identical to that of the variable screening. For most of the simulation the contact tracing shows worse results than the variable screening, however in the long term their behaviour is almost identical.

In Figure 4.8, we can observe an evolution of sick individuals very similar to that observed in Figure 4.7. In the first years, there is an important reduction in the number of sick individuals for the selective and the random screening, particularly for the latter. However as the simulation advances, the evolution of these interventions show an increase, although the results are much better than the reference value. The situation for the contact tracing is, again, utterly different. In the first years, the effect is almost negligible, but it slowly starts to appear by the end of the third year of simulation, and from then on the number of sick individuals does not stop reducing. By the end of the simulation (10 years), the number of sick persons with the contact tracing procedure is the lowest,

with a difference of more than 30 individuals.

With all these data, we are now able to compare properly the different types of screening. Firstly, it is important to note that while the selective screening and the contact tracing are quite realistic and are indeed tools used in epidemiology, the random screening is unrealistic, it was used as a first approximation to the experimentation on the model, although it is based on a established protocol as the mass screening (Porta, 2008). We have to take into account this fact as well as the efficacy if we want to use the simulator as a virtual laboratory. As has been commented on this last section, the least effective method is the selective screening because is the one with the smallest objective population. The other two procedures have shown similar results, with contact tracing being slightly better, especially in at long term. From a decision-making point of view, and considering both the realism and efficacy, the random screening would be eliminated the first, since it can not be implemented, and from the two remaining procedures, contact tracing is definitely more effective than the selective screening. However, not even contact tracing seems to have a shot at eliminating TB, observing the results it can be seen to have better results in controlling the disease.

As we already explained at the end of section 4.2.2, the analysis of the error is carried out differently in this experiment due to the higher amount of information. In the comparison between different types of screenings, the variation of the error is calculated in the same way explained at the end of section 4.2.2. The results of this analysis are shown in Figure 4.9. Again, the variation starts very close to zero and increases as the simulation advances. However, we find an upper bound for the error of 10% again and it does not depend on the experiment performed.

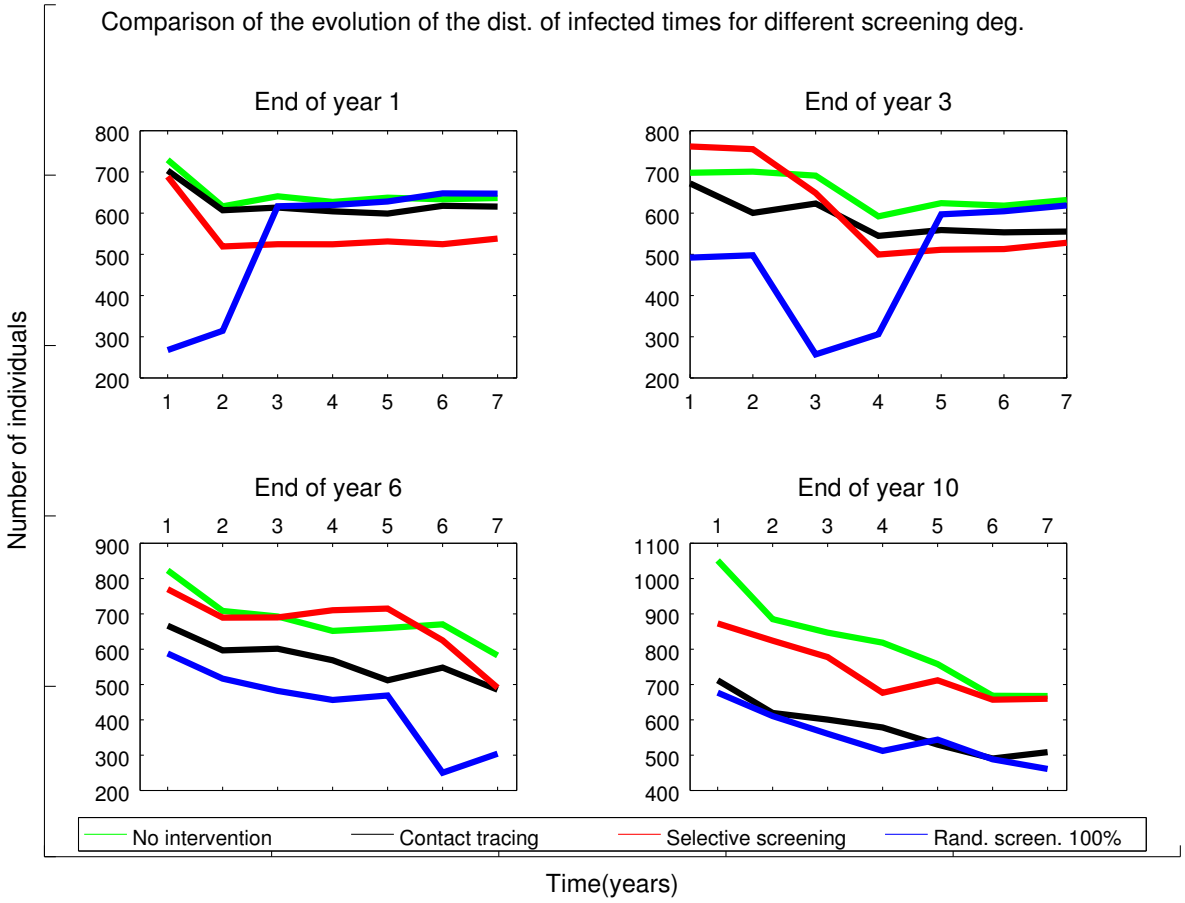

Figure 4.7: Distribution of infected times for different screenings: no intervention, selective screening, contact tracing and random screening at 100%. The plots represent different moments in a simulation of 10 years. Superior row: left, end of year 1; right, end of year 3. Inferior row: left, end of year 5; right, end of the simulation

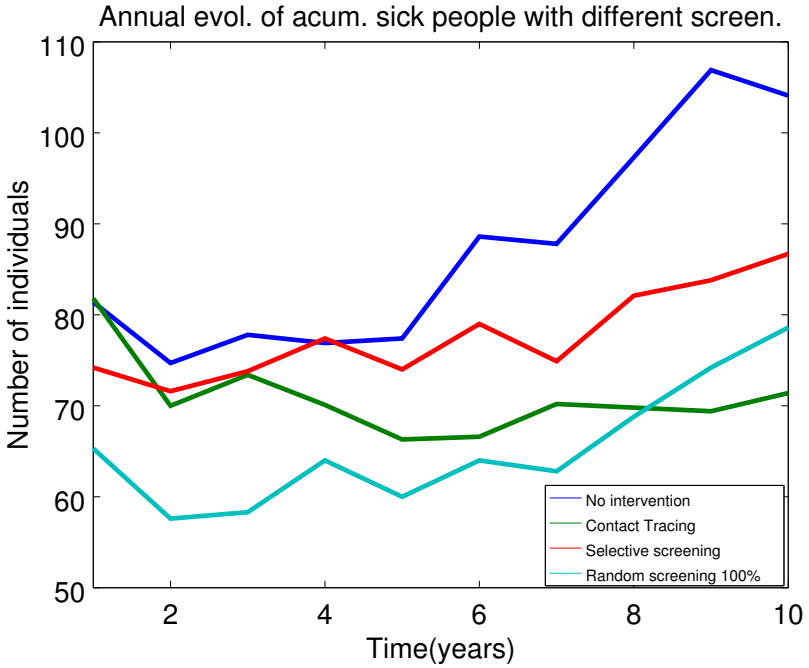

Figure 4.8: Evolution of the number of accumulated sick persons at the end of each year of simulation for different types of screening: No intervention, Contact Tracing, Selective screening and random screening of 100%

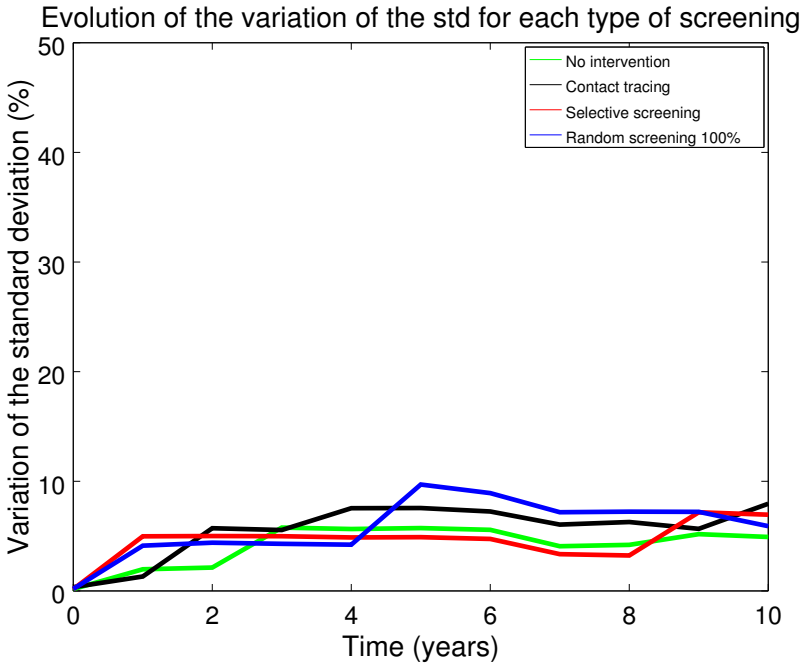

Figure 4.9: Evolution of the variation of the standard deviation for the different types of screening. The values on the y-axis are expressed as a percentage of the average value of the distribution

Chapter 5

Conclusions

The presented work is a first step towards a user-friendly integrated platform for the simulation of tuberculosis spreading in a certain community. The ABM has been built from a mechanistic perspective that accounts for the existing knowledge on the disease natural history and the infected individuals' behavior. It also allows the study in heterogeneous populations, taking into account factors like the autochthon/immigrant origin of individuals, their immunologic capacities and several important risk factors, as well as their age. All the parameters have been calibrated to fit the tuberculosis situation in the Barcelona district of Ciutat Vella, but the model is robust enough to be calibrated to any other region. This process of calibration has taken into account some of these heterogeneities that are relevant to the situation of Ciutat Vella, as well as more general parameters, such as the number of infected, which is always complicated to work with since no data is available.

Its implementation on NetLogo has made evident the computing limitations of this platform, although it has proven to be useful for its use by non-experts. Therefore, we have proposed and proved the success of a solution that overcomes these limitations. We combined the ABM approach of those individuals that are relevant to the course of the disease spreading, with a collective treatment of those individuals that do not require such an explicit control (healthy subpopulation). We have shown the capability of the optimized model to provide equivalent results to those obtained with the original model with a considerable reduction of the computing time. This strategy is an important innovation that may be used in epidemiological studies of other diseases. However, this improvement does not make the NetLogo implementation adequate for the simulation of big cities. We saw in Chapter 2 how for a simulation of 1 year with the entire population of Barcelona the execution time is almost 12 hours, and for really big cities like Bombay the program crashes. This indicates that the NetLogo implementation is not enough in order to simulate big cities and a tool that scales better is needed.

With the model optimized and calibrated, we performed some experiments to prove its utility. We designed two sets of experiments, one related to the distribution of infected time and one related with the screenings. The first experiment consisted of varying the initial infected time distribution and then evaluating how this changes affect the output

of the model. The model showed a high sensitivity to these changes. This result points out how careful one has to be with the assumptions taken at the moment of modelling a system. A piece of the model that a first sight would not seem too relevant produced radically different results when changed. The second experiment consisted of the testing of various types of screenings, which are interventions frequently used in epidemiology. In this experiment, we used three different interventions and compared them to each other and a reference value with no action taken. This experiment allowed us to conclude which of the three is the most effective, and also permitted us to prove the utility of the simulator as an assistance tool for public health decision-making, which was the goal of this work in the first place.

The most crucial limitation of the current model is the absence of a socio-demographic structure of the population that accounts for the social network of individuals and, therefore, allows long-term simulations. In the case of Barcelona, and Ciutat Vella in particular, the migratory fluxes will be of especial interest since there is a significant input of immigrants from high TB incidence regions. This improvement does not seem feasible in a platform like NetLogo; it should be implemented in a High-Computing platform probably with a parallel simulation design that reduces computing time. This platform should be conceived with the final purpose of being used by non-experts and, therefore, should be designed with a user-friendly interface. A few ideas on how to overcome these limitations will be presented in next section.

Chapter 6

Future perspectives

In multidisciplinary projects, NetLogo has the advantage of providing a user-friendly interface which has shown to be useful for including professionals who do not have any modelling experience. This interface facilitates enormously the control of the input values as well as the visualization of the outputs of the model. However, this advantage has a significant drawback: NetLogo designed for (small) sequential simulations. Even with our proposed optimization for the tuberculosis transmission model, a 1-year long simulation takes around 15 minutes whereas we are interested in long-term dynamics, with a duration of 10 years or more. As a result, we have been working with a small population. Therefore, we only considered a district of a big city. However, if we need to model a larger city, like Shanghai or Bombay, Netlogo limitations will become a problem. In this section, we will propose a solution for this problem along with some additional features that could be made in the model to improve it.

One way to leverage simulation costs is to run simulations in a parallel architecture environment (also known as High-Performance Computing, HPC). Parallel simulation consists on the use of multiple processors to partition and therefore speed up simulations execution time. However, this is not an easy task, and it requires some advanced skills on programming and HPC architecture. A possible solution is to use an existing parallel simulator to model TB transmission. Montañola Sales (2015) developed Yades, a parallel simulation framework for studying socio-demographic dynamics. The tool is able to prototype agent-based demographic models to perform simulations of large populations in parallel environments.

Yades provides 5 demographic components that can be specified in their models: fertility, mortality, marital status, economic status and migrations (both emigrations and immigrations). Since the basis of a epidemic model is a demographic model, we propose to implement our TB model in Yades. The advantage of this solution would not only allow us to tackle the limitation of computational costs of Netlogo, but also it will also enrich our model by taking into account the demographic structure. Yades is meant for demographers, anthropologists and other policy makers and uses the most common demographic data: census, survey and ethnographic (qualitative) data. The basic structure

of the agent-based model is a family, that may consist of only a member, or more than one either adults or children, following different familiar patterns to be defined by the modeller. It also divides the simulation space into geographical regions, where families live and interact among them. Moreover, these families can migrate between regions. Immigration arrivals from other countries is also taken into account, which would be crucial in a Tuberculosis model.

In terms of the architecture, Yades is implemented as a parallel discrete-event simulator. This methodology differs from Netlogo in the management of events and time. Netlogo uses a time-step simulation where we have a fixed time step and in every step a series of processes take place. In an event-driven simulation approach, the simulation advances as events occur (see Figure 6.1). From all the possible events that can happen: a migration, a change in marital or economic status, a TB infection, etc., the simulator dispatches the one that will occur in lesser time. Then, the simulation time moves to that event time. The process is repeated until the simulation reaches the end time. This procedure has shown to be fastest by only executing the simulation steps an events necessary.

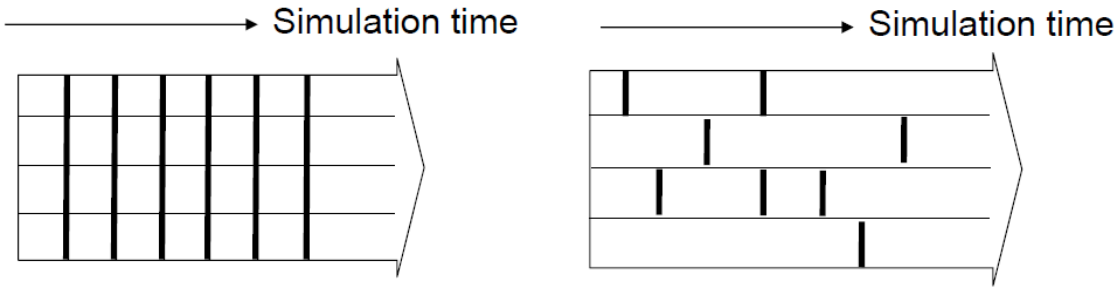

Figure 6.1: Difference between time-step simulation (left) and event-driven simulation (right). Inspired by (Fujimoto, 2000)

Yades parallel architecture is implemented by using a simulation library, μsik, which establishes communication between processors with the standard Message Passing Interface (MPI) (Pacheco, 1997). However, the technical details of Yades implementation are not of interest for this section. An interested reader may found a detailed explanation in (Montañola Sales, 2015). The purpose of this section is to describe a possible implementation of the TB model with Yades, as a line of future work.

As described, Yades works by the means of events, and its model uses an agent-based approach, where agents have a set of properties (attributes). These two facts are key to the implementation of the TB model. As a result, it would be possible to expand Yades architecture to implement the TB transmission process by adding a few extra events and attributes to the individuals. In Figure 6.2 we show a Unified Modelling Language (UML) diagram of how the simulator will look like after incorporating tuberculosis processes to it. This diagram illustrates the object-oriented structure of the system. Yades has 4

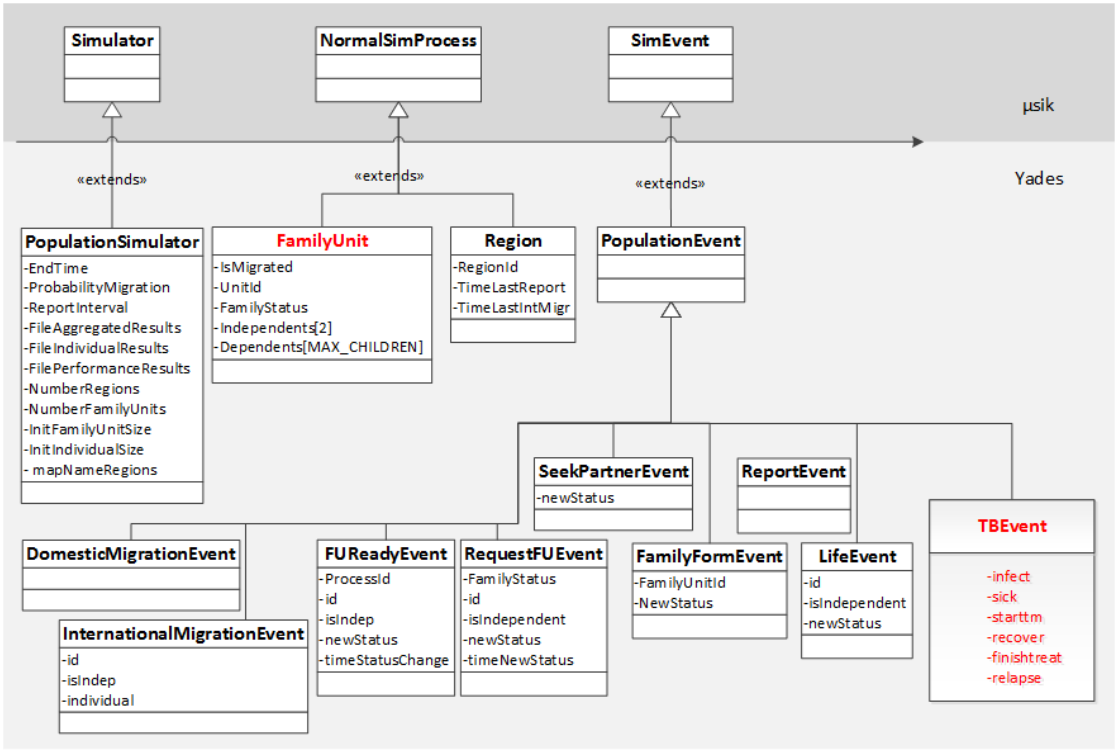

Figure 6.2: UML diagram of the new Yades simulator including the tuberculosis dynamics. In black are marked the events and attributes corresponding to the original Yades, and in red are marked the events and attributes that should modified or added in order to incorporate the TB model

different objects: the PopulationSimulator, the agents FamilyUnit and Region, and the events of type PopulationEvent. In this figure we can see how the different classes inherit from μsik library. In red colour, we show which features should be changed and incorporated in Yades implementation. Elements in black would remain untouched after the implementation of the Tuberculosis component. Comparing the two categories it is clear how the vast majority of the events and attributes of the simulator wouldn't need to be changed at all to incorporate TB dynamics to Yades. As a result, we would need to add a TB status attribute to the FamilyUnit and define a new TB event (TBEvent) wich will handle six processes: infect, sick, start-treatment, finish-treatment, relapse-or-recover.

Thus, it seems that incorporating our TB model in Yades is feasible. Moreover, Yades provides some demographic processes that could be interesting to consider in the TB model and have not been included in the Netlogo model so far (i.e. labor market, relationships, migrations and so on). However, there are aspects that would require some work. Interactions are crucial in the Tuberculosis simulation, and in our model they happen in a homogeneous mixing, where no real space is modelled. In Yades, the spatial structure is simplified since regions are areas where local interactions happen but do not have a grid structure. Furthermore, the individuals are grouped in families and groups of families in regions, but there are no kinship networks to link individuals.

Current literature highlights the importance of personal connections in Tuberculosis spread. Therefore, we would like to incorporate social networks in the simulator to allow establishing the contact between sick and susceptible individuals. However, social network data of the Ciutat Vella it is not likely available. Therefore, we propose to increase the complexity of our model in Yades by incorporating the network of connections through work centres, schools and households. A household might be a nuclear family, several families connected through common members or adult individuals with no relationship that share housing.

Connecting the dots of these networks would require extensive data describing the schools and work centres of Ciutat Vella. A first approximation in Yades would be to distribute uniformly children among the current number of school in the district. In the same way, we could distribute employed adults in work places, the type of whom should be determined in advance (for instance, office workers could be relevant). This proposed solution would permit not only to establish a certain network of close contacts, but also to establish a certain probability of random infection, for casual contacts. However, this approach would have some limitations. For example, it assumes that everybody resides and spends their working day in Ciutat Vella. To overcome this issue, it should be discussed whether it would be worth to expand the simulation scenario by including the whole city of Barcelona. As a result, we could assimilate the districts of the town to the regions of the simulator and run them in parallel.

We are currently working on the implementation of Barcelona districts in Yades. Our future plan includes to first validate the demographic evolution of Barcelona districts and then expand the model with the tuberculosis transmission component to contrast our results with Netlogo implementation. The ideas presented in this section are the result of the discussion with the team members and are subject to future modifications.

As part of the implementation, the first step is to collect all the data necessary for the simulation. We already have all the data corresponding to TB situation in Ciutat Vella, but now we need additional data corresponding to demographic and economic situation, that was not incorporated into the NetLogo model. We look for this data in the microcensus of the Spanish National Institute of Statistics (INE) (INE, 2011). Microcensus are anonymous surveys performed on small but statistically representative sample populations. From this source, we extract information of the population, such as: the percentages of people within certain age ranges, percentages of the type of families, e.g. couple with children, single male with children, percentages of employed individuals per age group and variables of the like. All this information will be needed to build a proper socio-demographic structure for the simulator. Also, the amount of available, and reliable, information is one of the crucial limit factors when building a model.

In overview, this work, although it has proved its validity in terms of public-health evaluation, represents only a small step towards the completion of our final goal, that

is, to develop a simulator to evaluate public health control strategies of tuberculosis in medium-sized and big cities. However, the analysis performed in this last chapter yields truly positive perspectives for this project. With an amount of time similar to that used to do this work, the extension of the simulator to big cities using the tools and procedures described in this last chapter seems feasible.

Bibliography

Ajuntament de Barcelona (2010). Ajuntament de Barcelona. Demographic statistics.

Ajuntament de Barcelona (2012). Ajuntament de Barcelona. Demographic statistics.

Begun, M., Newall, A. T., Marks, G. B., and Wood, J. G. (2013). Contact tracing of tuberculosis: a systematic review of transmission modelling studies. *PloS one*, 8(9):e72470.

Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. *Pnas*, 99(suppl. 3):7280–7287.

Cardona, P.-J. (2010). Revisiting the natural history of tuberculosis. The inclusion of constant reinfection, host tolerance, and damage-response frameworks leads to a better understanding of latent infection and its evolution towards active disease. *Archivum immunologiae et therapiae experimentalis*, 58(1):7–14.

de Espíndola, A. L., Bauch, C. T., Troca Cabella, B. C., and Martinez, A. S. (2011). An agent-based computational model of the spread of tuberculosis. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(05):P05003.

Ferrer, J., Prats, C., López, D., and Vives-Rego, J. (2009). Mathematical modelling methodologies in predictive food microbiology: a SWOT analysis. *International journal of food microbiology*, 134(1-2):2–8.

Fujimoto, R. (2000). *Parallel and distributed simulation systems*. John Wiley & Sons, Hoboken, New Jersey.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jø rgensen, C., Mooij, W. M., Müller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., Robbins, M. M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R. a., Vabø, R., Visser, U., and DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, 198(1-2):115–126.

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., and Railsback, S. F. (2010). The ODD protocol: A review and first update. *Ecological Modelling*, 221(23):2760–2768.

Guzzetta, G., Ajelli, M., Yang, Z., Merler, S., Furlanello, C., and Kirschner, D. (2011). Modeling socio-demography to capture tuberculosis transmission dynamics in a low burden setting. *Journal of Theoretical Biology*, 289(1):197–205.

INE (2011). Instituto Nacional de Estadística.

Kasaie, P., Dowdy, D. W., and Kelton, W. D. (2013). An agent-based simulation of a tuberculosis epidemic: Understanding the timing of transmission. In R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, E. and AN, editors, *Proceedings of the 2013 Winter Simulation Conference*, pages 2227–2238.

Kasaie, P., Dowdy, D. W., and Kelton, W. D. (2014). Estimating the proportion of tuberculosis recent transmission via simulation. In A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, E., editor, *Proceedings of the 2014 Winter Simulation Conference*, pages 1469–1480.

Millet, J. P., Hoff, S. r. T., and Jø rgensen., B. B. (2015). Estudio sobre una nueva tuberculina. Implementación en Barcelona.

Montañola Sales, C. (2015). *Large-scale simulation of population dynamics for socio-demographic analysis*. Thesis dissertation, Universitat Politècnica de Catalunya - BarcelonaTech.

Montañola sales, C., Gilabert-navarro, J. F., Prats, C., López, D., and Cardona, P. J. (2015). Agent-Based Modeling of Tuberculosis in Barcelona. In Yilmaz, L., Chan, W. K. V., Moon, I., Roeder, T. M. K., Macal, C., and Rossetti, M. D., editors, *Proceedings of the 2015 Winter Simulation Conference*, (in press). ACM.

Murray, M. (2002). Determinants of cluster distribution in the molecular epidemiology of tuberculosis. *Proceedings of the National Academy of Sciences*, 99(3):1538–1543.

National Institue for Health and Care Excellence (2015). Contact tracing and screening after a person is diagnosed with active tuberculosis.

Orcau, A., Manzanares, S., García, J. I., García de Olalla, P., and Caylà, J. A. (2011). La Tuberculosi a Barcelona. Informe 2010. Technical report, Agència de Salut Pública, Ajuntament de Barcelona, Barcelona.

Orcau i Palau, A., Arcas i Ferré, M., Caylà i Buqueras, J. A., and García de Olalla i Rizo, P. (2014). La tuberculosis a Barcelona. Informe 2013. Technical report, Agència de Salut Pública. Consorci Sanitari de Barcelona.

Orcau i Palau, A., Caylà i Buqueras, J. A., and García de Olalla i Rizo, P. (2012). Documents La Tuberculosi a Barcelona Informe 2012.

Ozcaglar, C., Shabbeer, A., Vandenberg, S. L., Yener, B., and Bennett, K. P. (2012). Epidemiological models of Mycobacterium tuberculosis complex infections. *Mathematical Biosciences*, 236(2):77–96.

Pacheco, P. S. (1997). *Parallel programming with MPI*. Morgan Kaufmann Publishers Inc., San Francisco.

Porta, M. (2008). *A Dictionary of epidemiology*. Oxford University Press, Oxford [etc.]:.

Sumi, A. and Kobayashi, N. (2003). Dynamics of Infectious Diseases. *Sapporo Medical Journal*, 72(3-6):23–29.

Tisue, S. and Wilensky, U. (2004). NetLogo: A simple environment for modeling complexity. In *Proceedings of the International conference on Complex Systems, Boston, May 2004*, pages 16–21.

WHO, W. H. O. (2012). *Global Tuberculosis Report 2012*. World Health Organization.

WHO, W. H. O. (2013). *Global tuberculosis report 2013*. World Health Organization.

WHO, W. H. O. (2014). *Global tuberculosis report 2014*.

Zwerling, A., Shrestha, S., and Dowdy, D. W. (2015). Mathematical Modelling and Tuberculosis : Advances in Diagnostics and Novel Therapies. 2015.