Universiteit Antwerpen
Faculty of Medicine and Health Science
Faculty of Science

universiteit hasselt
Faculty of Science

Agent-Based Models For Infectious Disease Transmission

Exploration, Estimation & Computational Efficiency

Dissertation submitted for the joint-degree of
doctor of medical science at the University of Antwerp and
doctor of science at Hasselt University
to be defended by

LANDER WILLEM

Promotoren
Prof. Dr. Philippe Beutels
Prof. Dr. Jan Broeckhove
Prof. Dr. Niel Hens

Begeleider
Dr. Ekaterina Vladislavleva

Antwerp, 2015

PROMOTOREN
Prof. Dr. Philippe Beutels, Universiteit Antwerpen
Prof. Dr. Jan Broeckhove, Universiteit Antwerpen
Prof. Dr. Niel Hens, Universiteit Hasselt en Universiteit Antwerpen

BEGELEIDER
Dr. Ekaterina Vladislavleva, Evolved Analytics Europe

INTERNE JURY
Prof. Dr. Greet Ieven, Universiteit Antwerpen
Dr. Jori Liesenborgs, Universiteit Hasselt

EXTERNE JURY
Prof. Dr. Niko Speybroeck, Université Catolique de Louvain
Prof. Stefano Merler, Fondazione Bruno Kessler

This PhD is supported by an interdisciplinary grant of the Special Research Fund of the University of Antwerp [IDBOF-5175].

ISBN 9789057284915
(D/2015/12.293/18)

Universiteit Antwerpen
Faculteit Geneeskunde en Gezondheidswetenschappen
Faculteit Wetenschappen

universiteit hasselt
Faculteit Wetenschappen

Individu-Gebaseerde Modellen Voor De Verspreiding Van Infectieziektes

Exploratie, Schattingen & Computationele Efficiëntie

Proefschrift voorgelegd tot het behalen van de graad van
doctor in de medische wetenschappen aan de Universiteit Antwerpen en
doctor in de wetenschappen aan de Universiteit Hasselt
te verdedigen door

LANDER WILLEM

Promotoren
Prof. Dr. Philippe Beutels
Prof. Dr. Jan Broeckhove
Prof. Dr. Niel Hens

Begeleider
Dr. Ekaterina Vladislavleva

Antwerpen, 2015

Contents

Abbreviations

ABM	Agent-based model
AH	Absolute humidity
AR	Attack rate
C	General-purpose, low-level, imperative computer programming language, requires compilation
C++	Object-oriented programming language derived from the C programming language
CI	Confidence interval
CPU	Central processing unit
Eurostat	Online database providing European statistics on demographics, economy, policy,... etc
FluTE	Open source agent-based model for pandemic influenza from Chao *et al*. (2010)
FLUTE	ABM implementation in Chapter 5, based on FluTE
FRED	Open source agent-based model for pandemic influenza from Grefenstette *et al*. (2013): Framework for Reconstructing Epidemic Dynamics. Also the name for an ABM implementation in Chapter 5, based on the published model
GNU	Free software, mass collaboration project
GP	General practitioner
GPU	Graphics processing unit
ILI	Influenza-like illness
KCE	Belgian health care knowledge centre
LAIV	Live attenuated influenza vaccine
LCPI	Local cycles per instruction
LHD	Latin hypercube design
LLC	Last level cache
MALAB	Multi-paradigm numerical computing environment and fourth-generation programming language developed by MathWorks
MCMC	Markov Chain Monte Carlo: a sampling technique that uses a probability distribution based on previous steps in the Markov chain
NOAA	National Oceanic and Atmospheric Administration
OpenMP API	Open multi-processing application programming interface

PerfExpert	Profiling tool for high-performance computing applications
POLYMOD	Large European population-based social contact survey
PRNG	Pseudo-random number generator
QALY	Quality adjusted life years
$\mathbf{R_0}$	Basic reproduction number: the expected number of secondary infections caused by a typical primary infection in a totally susceptible population
$\overline{\mathbf{R_0}}$	Average basic reproduction number
R	Software environment and programming language for statistical computing and data analysis
RAS	Realistic age structure
RH	Relative humidity
RSV	Respiratory syncytial virus
SARS	Severe acute respiratory syndrome
SED	Sequential experimental design
SEIR(S)	Susceptible-exposed-infectious-recovered(-susceptible)
SID	ABM implementation in Chapter 5 without specific data layout
SIR	Susceptible-infected-recovered
SR	Symbolic regression
TIV	Trivalent influenza vaccine
VE	Vaccine efficacy
$\mathbf{VE}_S$	Vaccine efficacy for susceptibility
$\mathbf{VE}_{SP}$	Vaccine efficacy for infectiousness to disease (proxy for VE_S)
$\mathbf{VE}_I$	Vaccine efficacy for infectiousness
VZV	Varicella zoster virus
WAIFW matrix	”Who acquires infection from whom” matrix: summarizes age-dependent transmission parameters
WLS	Weighted least squares

Agent-based vs. Individual-based

Agent-based models are microscale simulations in which a complex system is re-created by simulating the unique behavior of multiple agents simultaneously (both individuals or collective entities). In ecology and biology, agent-based models are also called individual-based models to stress the difference with population models (=compartmental models). According Grimm & Railsback (2005), individuals in individual-based models are simpler than fully autonomous agents in agent-based models. This dissertations targets simulation models with autonomous individuals and therefore "agent-based" is used in the title and text.

CHAPTER 1

Introduction

1.1 Infectious Diseases

Increased urbanization in combination with more intensive travel and trade patterns facilitate the rapid spread of pathogens within and between countries. The 2003 severe acute respiratory syndrome (SARS) outbreak, the 2009 pandemic of influenza A(H1N1)v (swine flu) and the very recent Ebola outbreak received massive attention and emphasized the potential threat infectious diseases may pose on public health worldwide. Other currently emerging infectious diseases include avian influenza virus as well as bacteria that have become resistant to an increasing range of antibiotics in established and expanding market economies. These emerging infectious diseases are not only heeded for their potentially devastating impact on public health [1], but also for their macroeconomic impact, which can be as high as 10% of the Gross Domestic Product of affected countries [2–5].

The 1918-1919 Spanish influenza pandemic (= worldwide epidemic) is still the most fatal event in human history with an estimated mortality of 50 million people [1]. Influenza is a respiratory virus causing mostly mild disease in healthy adults but is responsible for significant morbidity and mortality among elderly people, children and patients with underlying health conditions each year [6–8]. Next to seasonal epidemics, pandemics were registered in 1957-1958, 1968-1969 and 2009 and have been followed by periods of high respiratory disease activity associated with large outbreaks and high mortality over multiple years. Influenza type A is most common and targets many warm-blooded species including birds, pigs, horses and humans. High mutation rates and genetic exchange with strains from other species allow the virus to evade pre-existing immunity and cause infection [9].

1.2 Modeling Infectious Diseases

For many health care interventions, pre-introduction clinical trials are unfeasible due to budget or ethical reasons. Therefore, mathematical models are used as pragmatic tools to inform policy [10]. This is particularly the case for large-scale infectious disease interventions such as mass vaccination and school closures. Static (or steady-state) models assume a system to be in equilibrium and are commonly used for health economic evaluations and sometimes inappropriately so for infectious diseases. Dynamic (or time-dependent) models represent the transmission of evolutionary systems and contributed to our understanding of biological mechanisms and the spread of infectious diseases. The level of computational complexity and memory requirements vary substantially between the two most widely used types of dynamic models: deterministic compartmental models and stochastic agent-based models.

Until now, deterministic compartmental models are mostly used to simulate infectious disease transmission in human populations [10]. These models partition the population in compartments typically according to health state (e.g., susceptible, infectious and recovered) and simulate transitions between these compartments over time. Health related compartments could be further subdivided according to age and/or other characteristics. In the context of infectious diseases, we define a model deterministic if there is no randomness in the calculations of the acquisition of infection. Transition rates between compartments are pre-defined based on aggregated population data and do not account for individual stochastic events. Deterministic models typically ignore household compositions, locality and social mobility.

Individual heterogeneity and random events dominate the (non-)propagation of an infection in the initial stage of an emerging infectious disease, or in the final stage of eliminating infectious disease transmission. One can assume that average rates adequately mimic stochastic processes if transition probabilities are homogeneous in a relatively large population. Most existing models ignore chance and individual heterogeneity because of computational difficulties. However, increasing hardware performance makes large-scale stochastic simulations more practically feasible to analyze infectious diseases [11–15].

Individual-based or agent-based models (ABM), which track each individual in the population separately, allow for heterogeneous behavior in relation to social mixing by defining an extensive set of relevant characteristics for each individual (e.g., age, gender, household composition, locality, immunity status, overall health status,... etc.). The need to scale simulations from thousands

to millions of individuals requires the application of advanced and parallel programming techniques to fully exploit the computational power introduced by the recent multi-core processor architectures.

A central metric in epidemiology is the basic reproduction number R_0, defined as the expected number of secondary infections caused by a typical primary infection in a fully susceptible population [16]. Several factors affect R_0 including social contact mixing, the natural disease history of the pathogen and the applied reactive strategy. Generally, the infection will extinct if $R_0<1$ whereas the epidemic will continue if $R_0>1$. Some baseline examples:
R_0(influenza) = 2-3, R_0(rubella) = 5-7, R_0(measles) = 12-18.

1.3 Motivation And Aim

Agent-based models offer endless possibilities to explore complex heterogeneous behavior and stochastic processes with limited expenses but need to be well informed and struggle with the trade-off between model complexity and performance. Current assumptions and performance-bounded limitations have unavoidably consequences on the development and usefulness of present models. The overarching aim of this PhD thesis was to explore and improve agent-based transmission models for infectious diseases and elaborate on five key aspects: model-guided exploration, parameter estimation, social contact patterns, computational efficiency and synthetic populations (Figure 1.1). All these elements are fundamental to ABMs and further improvement could be based on an iterative process, though not necessarily in this order.

1.4 Introducing The Chapters And Their Objectives

Agent-based modeling involves many aspects. Figure 1.1 presents an overview of the topics that are covered in-depth in the chapters. The coherence of the chapters is reviewed in Chapter 7 (Discussion) together with their contribution to the field of agent-based modeling.

Model Exploration (Chapter 2)

Independent of model structure or research question, a thorough exploration of model assumptions and parameters is required to improve confidence in complex models and to obtain system understanding. Considerable efforts have been made to build realistic simulation models, but few have undergone sufficient testing across a range of settings and situations to be fully validated. Parameter uncertainty is routinely explored with static health models but far

less frequently in the context of dynamic models due to the computational complexity. We investigated the usefulness of an iterative and active learning approach to systematically analyze both common and edge manifestations of complex models.

Parameter Estimation (Chapter 3)

Herd immunity, also called community immunity, is central to dynamic transmission models. Infectious diseases are transmitted from individual to individual and recovered (or immune) individuals disrupt this chain. Herd immunity is of special interest to protect vulnerable fractions of the population who cannot be immunized directly because of age or medial issues. Our objective was to develop a dynamic transmission model for seasonal influenza to assess the impact of vaccinating specific age groups on the incidence of infection, disease and mortality. To focus on herd effects over many years and to explore seasonal heterogeneity within a realistic timeframe, we opted for a compartmental modeling framework to estimate unknown parameters through fitting to data. The methodology, opportunities and pitfalls regarding parameter estimation and seasonal behavior we applied to this model are also useful in the ABM context.

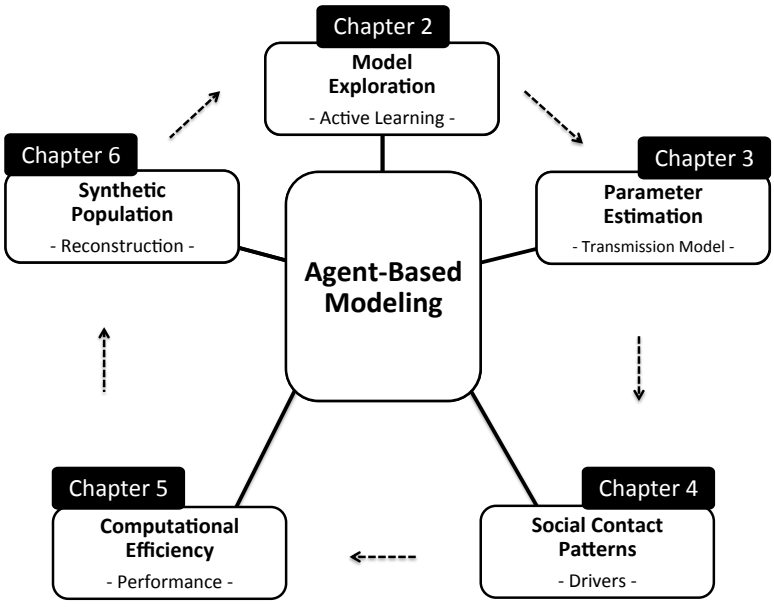


Figure 1.1: Overview.

Social Contact Patterns (Chapter 4)

Social contacts are the driving force of infectious disease transmission. Therefore, transmission models need to be well informed on social contact behavior and infectious disease dynamics to obtain trustworthy results. Social contact patterns are prone to change given autonomous decisions of the individual but also by contextual changes. Seasonal fluctuations in social contact patterns may not be large but could give diseases like influenza, in conjunction with other seasonal effects, a greater opportunity to spread during winters in temperate climate regions. To elaborate on contextual circumstances that reduce or enhance disease transmission, we analyzed social contact patterns with respect to weather conditions.

Computational Efficiency (Chapter 5)

ABMs provide endless opportunities to explore heterogeneous problems since each individual has a unique set of characteristics and health status. The drawback is that each individual needs to be stored and processed separately, which imply a large computational cost. Increasing hardware performance makes large-scale ABMs more frequently used but little attention is given in the literature to computational efficiency. In this chapter, we focus on single- and multi-core performance of discrete-time ABM simulators implemented in C++ to simulate infectious disease transmission.

Synthetic Population (Chapter 6)

Social contact clusters in an ABM need to represent the social behavior in the study area and determine the synthetic population. To simulate epidemic outbreaks, the structure of households is of main interest since they strongly affect the transmission potential of a disease and the final epidemic size. Synthetic reconstruction of households has been described in the literature with household survey data or entirely from aggregated census data. Both methods provide opportunities and contain limitations. We elaborate on the methodology to reconstruct synthetic households with survey data, applied to Belgium (Flanders).

CHAPTER 2

Model Exploration

This chapter is based on published work: "Willem L, Stijven S*, Vladislavleva E, Broeckhove J, Beutels P, Hens N. (2014). Active learning to understand infectious disease models and improve policy making. PLoS Computational Biology 10(4): e1003563" [17].*

Summary

Modeling plays a major role in policy making, especially for infectious disease interventions but such models can be complex and computationally intensive. A more systematic exploration is needed to gain thorough system understanding. We present an active learning approach based on machine learning techniques such as iterative surrogate modeling and model-guided experimentation to systematically analyze both common and edge manifestations of complex model runs. Symbolic regression is used for nonlinear response surface modeling with automatic feature selection. First, we illustrate our approach using an agent-based model for influenza vaccination. After optimizing the parameter space, we observe an inverse relationship between vaccination coverage and cumulative attack rate reinforced by herd immunity. Second, we demonstrate the use of surrogate modeling techniques on input-response data from a deterministic dynamic model, which was designed to explore the cost-effectiveness of varicella-zoster virus vaccination. We use symbolic regression to handle high dimensionality and correlated inputs to identify the most influential variables. Provided insight is used to focus research, reduce dimensionality and decrease decision uncertainty. We conclude that active learning is needed to fully understand complex system behavior. Surrogate models can be readily explored at no computational expense, and can also be used as emulator to improve rapid policy making in various settings.

*Both authors contributed equally.

2.1 Introduction

For many health care interventions, pre-introduction clinical trials are unfeasible for budget or ethical reasons and mathematical models are used as pragmatic tools to inform policy [10]. This is particularly the case for large-scale infectious disease interventions. Simple static fixed risk models are commonly used for health economic evaluation, and sometimes inappropriately so for infectious diseases. Dynamic models representing transmission or evolutionary dynamic systems contributed to our understanding of biological mechanisms and the spread of infections. For instance, in view of its global public health importance, influenza has been the subject of many simulation studies [14, 18–25]. The levels of computational complexity and data capacity needs vary substantially between deterministic compartmental models and stochastic agent-based models, the two most widely used types of dynamic models. Such models are developed through an iterative process of designing, coding and validating with empirical data but few have undergone sufficient testing across a range of settings and situations to be fully validated [10]. In order to improve confidence in model-based conclusions, it is necessary to gain a thorough understanding of the system and assess how model assumptions and parameters alter the results and policy decisions [24].

Agent-based models are computationally expensive and can be too complex to fully explore and understand a systems behavior [10]. Different scenarios and parameter values may be explored to account for methodological, structural and parameter uncertainty [19–21, 26]. Parameter values can be drawn from a distribution or changed at random over a plausible range. Parameter uncertainty using linear regression in a Latin hypercube design (LHD) is now routinely explored with static health economic models. Unfortunately, these techniques are far less used in the context of dynamic models due to the computational complexity and lack of knowledge on some of the fundamental parameter values and their distributions [27–29]. Nonetheless changes in a limited set of parameters or the full set should be explored. Clearly, independent of which method is chosen, it should be transparent and justified in the context of the model [10].

Symbolic regression can be used to explore parameter influence, which enables nonlinear response surface modeling with automatic feature selection. It aims to capture input-response behavior with algebraic expressions without a priori assumptions of model structure [30, 31]. Many variants of this method exist [32–34], but here we apply the Pareto-aware symbolic regression (SR) that uses multiple selection objectives [30, 31, 35]. The algebraic expressions are

surrogate models for the original computationally intensive simulation model. The model responses can be instantaneously predicted for a set of inputs using the algebraic expressions. These expressions provide information on the relationships between inputs and responses. Ensemble-based SR uses a collection of surrogate models as a final solution and the accordance of the models defines a measure for the prediction uncertainty. High model divergence indicates that more simulation samples are needed from the corresponding input region.

The goal of this chapter is to present an iterative modeling approach with a model-guided experimentation process to systematically analyze both common and edge manifestations of model runs. Figure 2.1 presents the methodology we recommend to explore simulation models. First, inputs are sampled using a maximin LHD design where the minimum distance between all points is maximized. Second, each point of the input space is used to initialize the simulation model. Next, input-response data from the simulations are modeled with SR to create surrogate models for response predictions, feature selection and to identify conditions with large prediction uncertainty. Finally, these insights are used to enhance the experimental design of subsequent simulations by adapting the sampling strategy or reducing dimensionality.

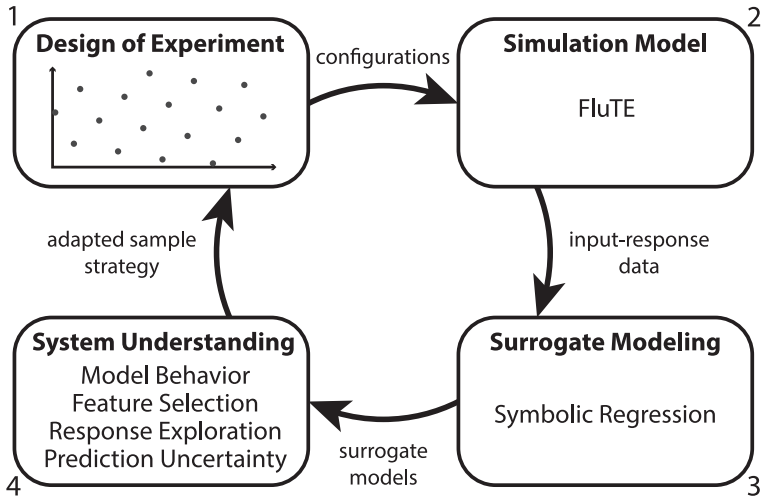


Figure 2.1: Iterative active learning approach with a simulation model. (1) A Latin hypercube design is used to make configuration files. (2) These configurations are used for the simulation model. (3) All input-response data are modeled with symbolic regression. (4) The surrogate models obtained with symbolic regression are used to achieve system understanding. The response prediction uncertainty can be used to adapt the experimental design (1) for the following modeling cycle.

While the general goal is to use a combination of previously described methods from machine learning, known as sequential experimental design [36–38] or active learning [39–41], we want to emphasize that our approach with SR is an improvement in the field of infectious disease modeling. An iterative modeling approach from [28] was based on step-wise linear regression to estimate response hypersurfaces and was limited to polynomials of the third order. Okais *et al*. [29] presented a framework to perform a preliminary sensitivity analysis considering logic and scientific relevance before conducting multivariate sensitivity analysis. However, consistent reproducible methods for the latter were not described. Longini *et al*. [18] calculated confidence intervals for disease burden estimates based on prior distributions instead of fixed parameter values, but they did not describe variable importance or system exploration analyses. Van Hoek *et al*. [42] performed variable importance analyses without feedback to the experimental design.

We apply our methodology to the open-source model FluTE [14], which is a stochastic agent-based model for pandemic influenza. We illustrate a stepwise system exploration and sensitivity analysis of a computationally expensive simulation model. The model simulates a population with realistic social contact networks and transmission based on the natural history of influenza. Several parameters can be varied to modify the spread of influenza, with the basic reproduction number R_0 being typically described as the most essential. R_0 is defined as the expected number of secondary infections caused by a primary infection in a fully susceptible population. For the purpose of our study, we focus on two main outcomes of this model: the clinical attack rate (AR) and the day at which the influenza epidemic peaks.

High dimensionality and correlated inputs cause problems in exploring system behavior, which can be resolved through iterative modeling with SR. The approach we present is relevant for many public health problems and we illustrate this through an example of a previously published dynamic model-based economic evaluation of varicella zoster virus (VZV) vaccination [42]. Varicella (chickenpox) is a typical childhood infection caused by VZV and after recovery from chickenpox, the virus may reactivate later in life to cause herpes zoster (shingles). The probability to experience herpes zoster increases with time since the primary VZV infection, but is reduced by natural re-exposure to VZV (e.g., typically parents are re-exposed when their child has chickenpox). Although infant VZV vaccination was shown to dramatically reduce chickenpox morbidity and mortality, there are lingering concerns about its adverse impact on shingles as it reduces VZV re-exposure opportunities [43]. Many modeling and economic studies aimed to tackle this problem (reviewed in [44–46]) but

the van Hoek *et al*. model [47] is of special interest because it includes empirical observations on social mixing patterns and combined childhood and adult vaccination strategies. This age-structured dynamic transmission model was used to perform cost-effectiveness analysis for the United Kingdom [42] with 185 input parameters, including 100 correlated transmission rates, to calculate the incremental gain of Quality Adjusted Life Years (QALY) and costs. Parameter uncertainty was incorporated and conclusions were based on the results of 1000 runs. We analyzed this input-response data with SR to identify driving parameters and compared our findings with a linear regression analysis [48]. Reducing the dimensionality may improve uncertainty analysis since the most influential parameters can be sampled more intensively.

2.2 Methods

The method section follows the approach presented in Figure 2.1. A guided step-by-step example with a basic simulation model is provided in Appendix A.

Design Of Experiments

We used space-filling LHD to create parameter sets for the FluTE simulations. In the general case, a sample value from the first interval of the first input parameter is matched at random with sample values from intervals chosen for the other input parameters [28]. Then the second interval of the first input parameter is matched at random with sample values from previously unused intervals of the other features. Each interval of every input parameter will be sampled once and only once. LHD has the advantage that the number of samples is independent of the number of dimensions of the input space but can be determined based on the computational budget, the input dimensions and the complexity of the simulation. Computing a space-filling LHD can be an onerous task and therefore we used the maximin designs of spacefillingdesigns.nl [49,50]. We extended the designs using the Intersite-projected distance method of the Sequential Experimental Design (SED) toolbox to increase the number of sample points [36,51,52].

Influenza Model

We made use of an open-source agent-based model for influenza epidemics written in C++, called FluTE [14]. All individuals in the model are members of different social mixing groups. Influenza transmission within each group is based on random mixing. The geographical distribution, employment rates and commuting behavior of the population are based on census data from 2000

for Seattle (500 000 people) and the Los Angeles County (11 million people), distributed together with the source code of the model. The simulation runs in 12-hour time steps, representing daytime (work, school and community contacts) and nighttime (home and community contacts). Contact probabilities were tuned such that the final age-specific clinical attack rates were similar to past influenza pandemics and observed household attack rates. The model can simulate several intervention strategies based on changes in susceptibility and infectivity due to vaccination or antivirals and on changes in contact probabilities between individuals due to social distancing measures.

VZV Model

The economic evaluation of VZV vaccination was based on a deterministic dynamic compartmental model with 185 inputs, including 100 correlated transmission rates between 10 age groups [42, 47]. Underlying contact rates were estimated from a survey of social mixing patterns and bootstrapping the original sample specified uncertainty. The model was adapted and calibrated to data from the UK. Source code was not available but we made use of a dataset with 1000 runs, previously subjected to linear regression analysis [48].

Surrogate Modeling Using Symbolic Regression

Symbolic regression (SR) captures input-response behavior by efficiently exploring hundreds of thousands of algebraic expressions of the input variables [30, 31]. Aside from choosing the modeling primitives, no assumptions or restrictions are made on the model structure and genetic programming is used to optimize the search process. SR is a biologically inspired method that imitates Darwin's evolution theory by applying genetic variation and natural selection on the modeling ensemble [53]. We used the SR implementation from the DataModeler package in Mathematica [54]. The result of a SR run is an ensemble of tree-based regression models that give a good approximation of the response variable. The algorithm consists of the following steps [55]:

Model Initialization. In the first step of the algorithm, a population of SR models is generated randomly and the algebraic expressions of the models are represented by parse trees. Every model is a potential solution that explains the response behavior using a subset of the input variables. Each parse tree consists of terminals of input variable or a constant and primitive functions. We used the following set of primitive functions: $\{+, -, *, \div, \bullet^{-1}, \bullet^2, \bullet^x, \sqrt{}, log, exp\}$ with arbitrary arity of summation and

multiplication. Primitive functions can be chosen according to the problem domain.

Model Evaluation. The model fitness is determined by minimizing two objective functions: model complexity and prediction error $1 - R^2$, with R the Pearson correlation coefficient between the observed and predicted response.

We define model complexity as the sum of the number of nodes in all possible subtrees of a given tree, which is equivalent to the visitation length, i.e. the total number of links traversed starting from the root node to each of the terminal nodes of the parse tree. The complexity objective is used to avoid excessive growth of the model expressions. Because of the complexity objective, the presence of a variable in a sufficiently evolved population indicates that the variable is relevant for describing the response [35].

Model Archiving And Elitism. After the evaluation of all models, a fixed-size archive of the best achieving models is maintained. This is an elitism strategy that ensures that the best achieving models are never lost after recombination. The archive is populated by a selection of the least-dominated models from both the population and the archive of the previous generation. Dominance is defined as performing better on at least one objective and not worse on the other objectives. A model is Pareto optimal if no other model dominates it. This way, we can define the Pareto front of a model set.

Model Evolution. A new generation of SR models is generated every step of the algorithm. Models are created with genetic operators like crossover (90%) and mutation (10%). Crossover is the process of combining parent models into child models by using subtrees of both parents (Figure 2.2A). Parent models for crossover are selected from the previous generation and the model archive based on their dominance in a Pareto tournament. Mutation of a model introduces random alterations in its expression tree (Figure 2.2B). Every 10 generations the population is re-initialized with random models to ensure diversity in the population and to counteract inbreeding.

This evolutionary process is repeated over many generations. A maximum number of generations, a time budget or a model accuracy threshold can be used as criteria to stop the process. We used time budgets based on the size and dimensionality of the data sets. Timings are listed in Table 2.1 and an example with different time budgets for RUN 3 is described in Appendix A.

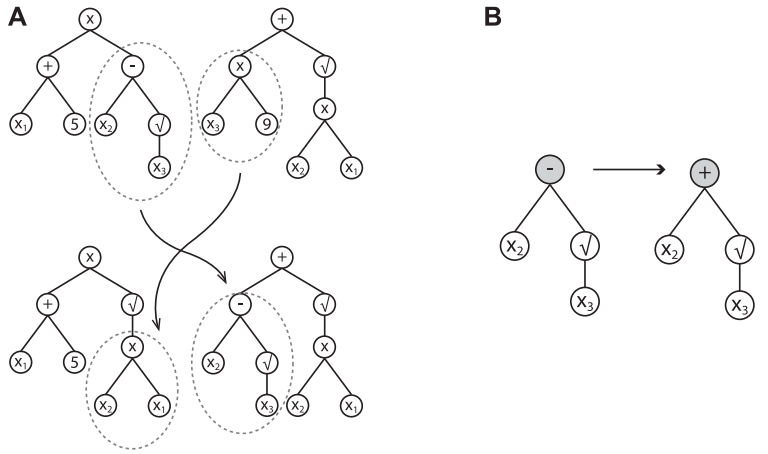


Figure 2.2: Example of algebraic expression recombination and modification. (A) One point crossover: two individuals swap subtrees resulting in two new expressions. (B) One point mutation: the operator minus is replaced by another operator of the same arity.

Table 2.1: Symbolic regression settings.

Name	Value
Population size	1000
Archive size	100
Crossover rate	0.9
Mutation rate	0.1
Population tournaments	5
Primitive functions	$+, -, *, \div, \bullet^{-1}, \bullet^2, \bullet^x, \sqrt{}, log, exp$
Time budget FluTE RUN 1,2,3	1000s
Time budget FluTE RUN 4	7200s
Time budget FluTE RUN 5	3600s
Time budget QALY	3000s
Independent evolutions FluTE	8
Independent evolutions QALY	12

System Understanding

Conclusions are based on a model selection from the knee of the Pareto front and we perform nonlinear optimization of the constants within these models. A model ensemble of high-quality and minimal complexity obtained through an effective SR algorithm can facilitate system understanding and focus research. Variable presence in the final ensemble (taken over several independent SR runs) provides a robust indication of the importance of input variables. Only

inputs significantly related to the response can survive a harsh evolutionary pressure and get into the final ensembles [35]. Besides variable importance, final ensembles also provide dimensionality trade-offs in model complexity and accuracy. Also contributing to system understanding are automatically generated meta-variables through the observation of low order transformations of driver inputs. These transformations can potentially linearize final models and enable further application of powerful linear and regularized linear learning. Next, sensitivity analysis of constructed ensembles highlights facilitated system understanding. The prediction divergence of the model ensemble is a measure for the prediction uncertainty. Conditions that are hard to predict might be missing from the design and require more attention in following modeling iterations.

2.3 Results

Transmission

We performed a stepwise exploration of the US-tailored simulation model for pandemic influenza (FluTE), applied to Seattle and Los Angeles county [14]. We first simulated epidemics in the Seattle population using four basic model parameters: R_0, whether individuals can travel, the number of infected individuals seeded into the population and whether this seeding occurs only once (static) or on a daily basis (dynamic). Table 2.2 summarizes the parameter ranges.

The surrogate models for the AR were of good quality (error < 0.001). Although each configuration was executed 20 times, almost no stochastic fadeout was observed. The dichotomous variable indicating whether people can travel was absent in most surrogate models. Given the inherent feature selection of the SR algorithm, this parameter appears to be unimportant to predict the AR [35]. The response plot for the AR (Figure 2.3A) shows that the number of infected people seeded into the population had almost no impact when seeding once. Only very low numbers of seeded individuals resulted in a different AR. The impact of the seeding number on the AR increased with daily seeding. We observed a correlation of 60% between the AR and the seeding number and frequency though we expected a major role for R_0.

The day of the epidemic peak advanced logarithmically with an increasing number of infected seeds, although small numbers of seeds could give rise to no or very late peaks (Figure 2.3B). There is no consensus in the literature on pandemic influenza models about how and to which extent infectious individ-

uals should be seeded. Some studies [14, 20, 56, 57] have been published with static seeding of 1, 10 and 100 individuals while others used dynamic seeding. No concerns are made about the potential impact of these different seeding approaches, as only a shift of the epidemic curve due to seeding has been reported [21]. We explored a wide range of seeding values using both static and dynamic approaches, and observed that the seeding approach has impact on the results. The surrogate model divergence for small seeding values was very large so these conditions needed to be sampled more intensively. Model specifications and examples of surrogate models are given in Appendix A.

Stochasticity

With the aim to include more irregular manifestations of the FluTE model, we adapted the initial design by oversampling small numbers of infected seeds with successive powers of two until 1024 (Table 2.2, RUN 2). This resulted in more stochastic effects, with substantially different output for each parameter set. The current standard is to use average [18, 28, 56, 57] or median [20] results from several realizations of each scenario. Average responses can be very misleading due to stochastic fadeout (AR = 0). Figure 2.4 presents the median, minimum and maximum AR and epidemic peak day for all configurations. We opted to use all responses for the surrogate modeling to minimize the loss of valuable information. Conditions with stochastic effects increase

Table 2.2: Parameter design for all modeling iterations with FluTE and obtained variable importance for the attack rate.

Parameter	RUN 1	RUN 2	RUN 3	RUN 4	RUN 5
Region	Seattle	Seattle	LA County	Seattle	LA County
Travel allowed?	yes/no (-)	yes/no (-)	yes/no (-)	yes	yes/no (-)
R_0	1.1–2.4 (++)	1.1–2.4 (++)	1.1–2.4 (++)	1.1–2.4 (++)	1.1–2.4 (++)
Infected seeds	0–5000 (+*)	0–1024 (+*)	0–1024 (+*)	0–1024 (+*)	0–1024 (+*)
Seeded daily?	yes/no (++)	yes/no (+)	no	yes/no (-)	no
Ascertainment				0–90% (+)	80%
Ascertainment delay				1–5d (-)	1d
Response threshold				0–5% (+)	instant
Response delay				0–30d (-)	instant
Vaccination coverage				0–90% (-)	0–90% (++)
$VE_{susceptibility}$				0–66% (-)	0–66% (+)
$VE_{infectiousness}$				0–66% (-)	0–66% (-)
$VE_{symptoms}$				0–66% (-)	0–66% (-)
Scenarios	200	200	50	800	200
Repetitions	20	20	10	20	20

Legend: ++ very important, + important, - almost no impact, * only small values, VE: Vaccine efficacy

model divergence. In published pandemic influenza models, the number of repetitions for each scenario ranged from 1 to 1000 [14, 19, 20]. Small numbers have been justified by observing that independent realizations with a given set of parameters lead to very similar epidemic curves [21]. Nevertheless, we observed stochastic fadeout for scenario's resulting in different median AR and especially the day of the epidemic peak seemed sensitive to stochasticity.

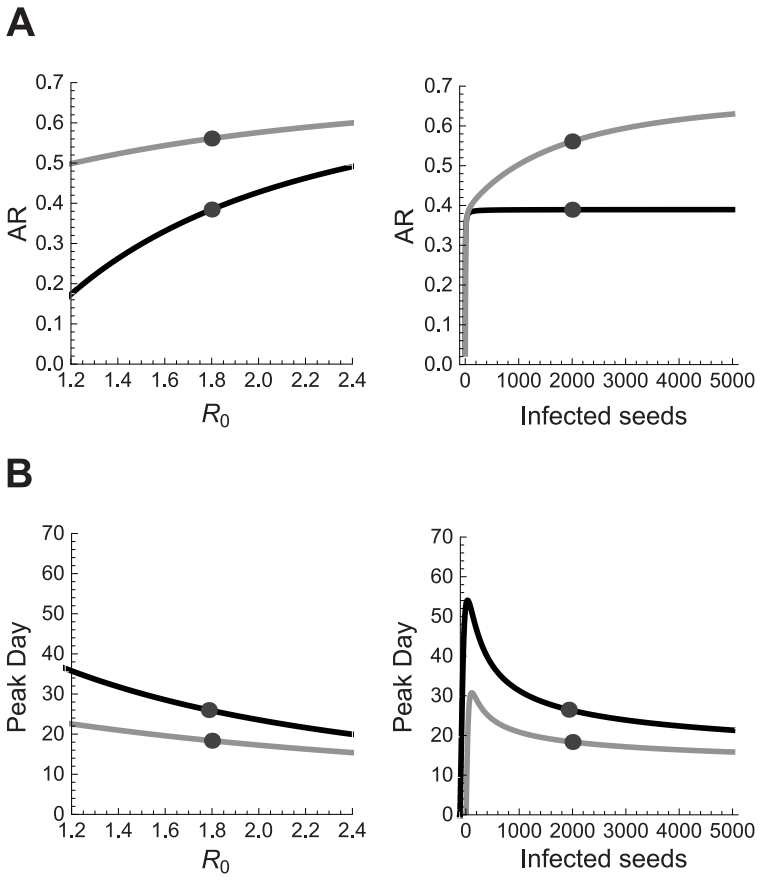


Figure 2.3: Response prediction plots of the high-quality surrogate models obtained with symbolic regression. Response prediction plot for the attack rate (A) and the day on which the epidemic reaches its peak (B) when seeding occurs only once (black) or on a daily basis (gray). Predictions for R_0 assume a fixed number of infected seeds, indicated by the dot in the plot on the right, and vice versa.

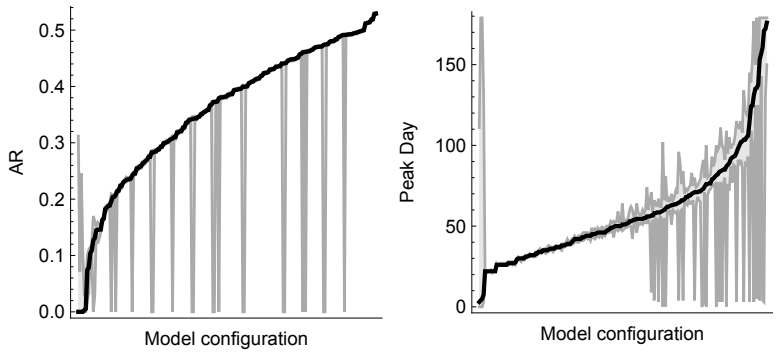


Figure 2.4: Variability of the model output due to stochastic effects. Multiple executions of the stochastic FluTE model with identical epidemiological configuration lead to different outcomes. The variability is shown for the attack rate (left) and day of the epidemic peak (right) for each combination of the R_0, travel and seeding parameters by the median (black), minimum and maximum (grey). The configurations are sorted on their median response value.

Population

In order to assess the effect of population size, we compared FluTE simulations for Seattle (0.5 million people) and LA County (11 million people). We used a single design with four transmission parameters for both populations (Table 2.2, RUN 3) and compared the surrogate models of each dataset. We observed similar response predictions for the AR (Figure 2.5A), indicating that this outcome is insensitive to population size when population size is already substantial (i.e. 0.5m). The travel parameter was absent in most surrogate models for both populations, indicating that this is inherent to the simulation model. The main difference for the enlarged population was the timing of the epidemic (Figure 2.5B). For example, a pandemic with $R_0 = 1.8$ and 100 infected seeds would result in an AR of 0.38 for both populations, but the epidemic peak day in LA County is predicted to be 15 days later compared to Seattle. The similar AR and postponed peak for the larger population are in line with results of previous studies [13,14]. We did not compare urban and rural regions due to lack of data although this may have a large impact [20]. Model ensemble divergence for low seeding numbers was less for LA County, which suggests that large populations absorb stochastic effects.

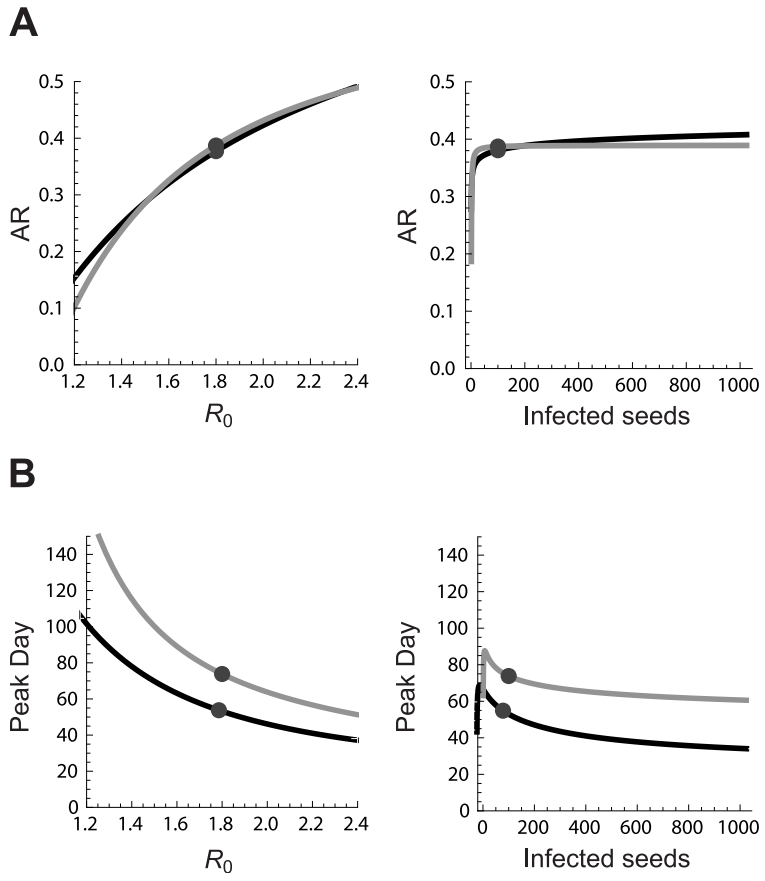


Figure 2.5: Response prediction plots for Seattle and LA County. Response prediction plots of the high-quality surrogate models obtained with symbolic regression for the attack rate (A) and the day of the epidemic peak (B) in Seattle (black) or LA county (gray). Predictions for R_0 assume a fixed number of infected seeds, indicated by the dot in the plot on the right, and vice versa.

Vaccination

After adjusting the transmission settings, seven parameters for reactive vaccination strategies were added to the design (Table 2.2, RUN 4). The computational burden to simulate Seattle was much lower compared to the LA County. Therefore, we used the Seattle population for the initial exploration with vaccination parameters. Based on the resulting input-response data, surrogate modeling showed that mainly the response threshold and ascertainment fraction were important to predict the AR. The importance of R_0 and the vaccination coverage increased when the response threshold and ascertainment parameters were set to mimic instant reactive measures, immediately after emergence.

Emulation

After subsequent simulation and modeling iterations, we obtained surrogate models for LA County that can be used to explore reactive vaccination policies on the outcome of ongoing pandemics. Figure 2.6 shows a basic interface to visualize the response behavior by changing the surrogate model parameters. When vaccination coverage is set to zero, the results from the second design emerge again (Figure 2.3). Further exploration of the surrogate models revealed a saturation effect of the vaccination coverage on the AR. The predicted AR with a vaccination coverage of 60% is almost zero for $R_0 = 1.8$ and vaccine efficacies of 0.5. The protection of the general population by vaccination of a subset is known as herd immunity [58]. The clear visualization of herd immunity with the surrogate models emphasizes the usefulness of our approach since it is hard to observe this effect directly from the numerous individual simulation results. An interactive version of this plot is available at *www.idm.uantwerpen.be* (more info in Appendix A).

Feature Selection

Timely and effective identification and easy exploration of important variables enhances our understanding of the underlying system. SR can be used for feature selection and is capable to handle high dimensionality and correlated variables [35]. Here, we focus on the surrogate modeling step of our approach with a dynamic transmission model to explore the cost-effectiveness of infant and adult VZV vaccination options. We analyzed the results from an economic evaluation described in van Hoek *et al*. [42] with 185 inputs, 100 of which are correlated transmission rates. Parameter uncertainty was explored by using 1000 different configurations. One of the essential findings based on this model [42, 59] is that the incremental Quality Adjusted Life Years (QALYs) might become negative, suggesting infant VZV vaccination might in some cases do more harm than good. Hence the pivotal issue for policy making is to identify and explore the variables that determine the incremental QALYs. Therefore, we performed a SR analysis with all 185 variables to model the vaccination benefits, expressed as incremental QALYs. Figure 2.7 presents the model error and complexity of the surrogate models with the most abundant variable combinations. We obtained surrogate models with only five variables with a model error of 16% (Figure 2.7A), which indicates strong correlation between these inputs and the response. With the sixth variable in Figure 2.7B, a small reduction in model error can be achieved. Figure 2.7C shows models with decreased model error but substantially increased complexity due to the presence of additional variables. Some of these variables are important

to predict the response, but others may mostly increase complexity without explaining the response. Especially with high dimensionality and model complexity, it is possible that some less important variables remain captured in the surrogate models. To tackle this problem, we performed a second SR analysis with the twelve variables from Figure 2.7C. This way, we reduced the number of drivers for the incremental QALYs to eight and ended up with surrogate models with an error of only 10%, which is an improvement compared to the models from Figure 2.7B. Surrogate model specifications and examples are listed in Appendix A tighter with a response plot interface.

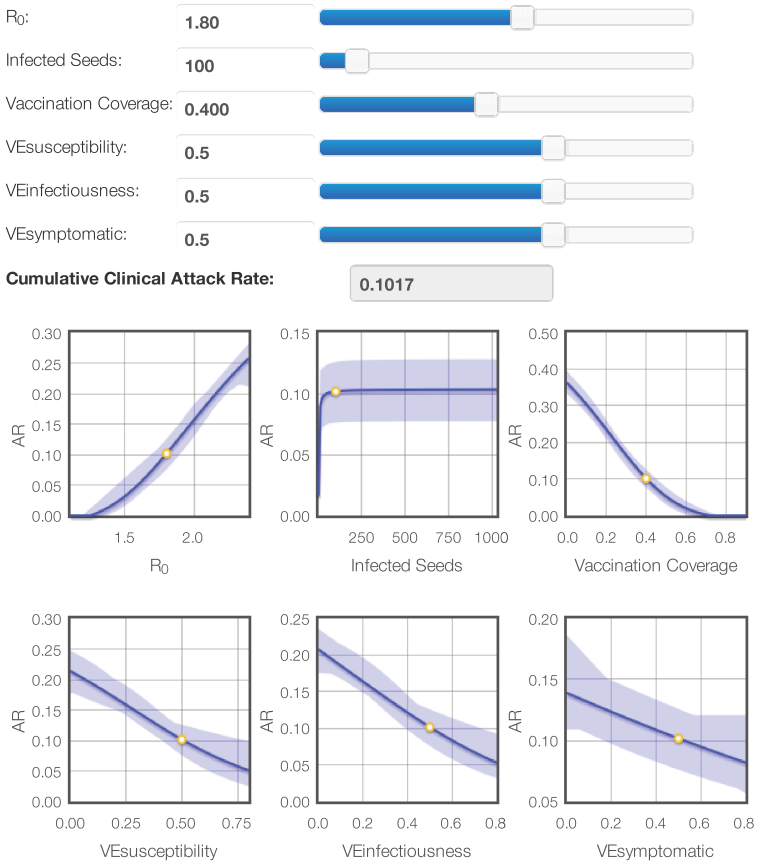
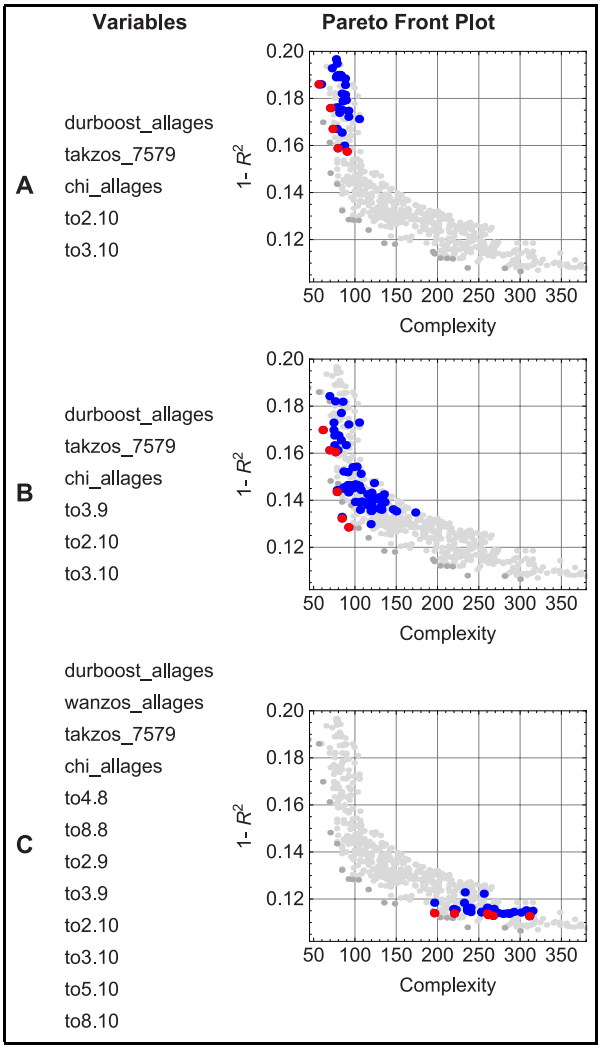


Figure 2.6: Response plot explorer for the attack rate. An interactive version of this plot is available at *www.idm.uantwerpen.be* (see Appendix A).

Marginal contributions were estimated with SR by inspecting the decrease in model error by adding variables. We observed a correlation of 47% between the incremental QALY and combinations of the zoster immunity duration and the change in reactivation due to vaccination. The zoster vaccine uptake for ages 75-79y was responsible for another 11% increase in correlation. Using linear regression, Bilcke *et al*. [48] obtained similar results. They estimated the marginal contribution of all transmission rates at 29%, but they were unable to select particular age-specific transmission rates due to their strong interdependency. Using SR on the same data set, we found that this 29% contribution was explained by just three of the 100 transmission rates.

High variable importance came from rates of transmission from adults to children, despite the fact that mainly children would be in the infectious state pre-vaccination. The transmission rates are based on symmetric contact rates, implying only age-specific differences in susceptibility and infectiousness account for this observation. To study the age component of the transmission rates we estimated the incremental QALYs once with the southwest corner of the transmission matrix and once with the northeast corner. Transmission rates between similar age groups were selected in both experiments but the models with the adult-to-child transmission had a slightly lower model error. It is not exactly clear why this is the case. These variables may capture adult-child interactions, which are relevant for re-exposure to VZV, as well as the susceptibility of children in the presence of universal vaccination. However, this does not imply that transmission would occur more often from adults to children than the other way around.



durboost_allages	Zoster immunity duration after varicella infection
wanzos_allages	Zoster immunity waning
takzos_7579	Zoster vaccine uptake for ages 75y-79y
chi_allages	Change in zoster reactivation after varicella vaccination
to*j.i*	VZV transmission rate from a person of age class *i* to a person of age class *j*. Age classes: {<1, 1-4, 5-14, 25-44, 45-64, 65-74, 75-95, +95}

Figure 2.7: Model error (1-R^2) and complexity of high-quality models for the incremental QALY. Models with the variable combinations presented on the left are shown in color (red = Pareto front). Only the most abundant combinations are given: (A) Surrogate models with five variables, (B) Surrogate models with extra variable (to3.9) and decreased model error compared to (A), (C) Surrogate models consisting of twelve variables with small model error but high model complexity.

2.4 Discussion

We present an iterative process of active learning with SR for the systematic exploration of simulation models. Our initial experimental setup for the pandemic influenza model showed only a subset of the system behavior but provided insights leading to an improved design. We explored common and edge manifestations and ended up with an ensemble of surrogate models for the complex simulation model. The surrogate models for reactive immunization strategies revealed effects like herd immunity and can be useful to instantly evaluate reactive strategies for specific R_0 values based on plausible estimates for vaccination coverage and efficacy.

Although we demonstrated these methods on two vaccination programs, based on two distinctly different dynamic models (stochastic agent-based and deterministic compartmental), we are certain that these methods are relevant to address a wide range of public health problems that are informed by modeling. Surrogate modeling with SR identified the most important variables. A decrease in the uncertainty of these parameters would improve the robustness of the simulation results. Also, the feature selection can be useful during the development of the simulator. E.g., the travel parameter does not seem important in the current FluTE implementation although other studies have stressed the role of travel restrictions on epidemics [13, 21, 22, 60]. A revision of the travel implementation may be required.

Considerable efforts have been made to build realistic simulation models of high quality, but most of these are not fully explored. Ideally, each model should be analyzed systematically to understand system behavior and to assess the impact of model assumptions and parameters on the results. The availability of simple surrogate models based on complex simulation models can be used to understand and emulate the original complex model. Policy makers can easily use an interactive interface, such as the ones we present in this chapter, to mimic the context in which their decisions take place (e.g., transfer model outcomes between broadly comparable countries) and predict the effectiveness or cost-effectiveness of health interventions. In that sense, the use of surrogate models as emulators provides a great opportunity to enhance both the understanding of these models and improve the reliability and speed of policy making based on existing elaborate model structures. With the emulator for FluTE, we could instantly formulate some insights (Figure 2.6) in clear language for policy makers. First, we predict that without reactive measures 36% of the population will be infected. Second, only a few imported cases are enough to start the epidemic hence partial isolation may

only delay the epidemic. Third, 30% vaccination coverage (percentage of the population vaccinated) may result in a 55% reduction in the number of cases and 60% coverage in a 95% reduction due to indirect protection because of the interruption of transmission pathways in a partial immune population.

Future research is required to automate the iterative surrogate modeling approach in order to speed up the process and make it more accessible. The number of realizations should be analyzed more into detail. We acquired already substantial insights on the transmission and vaccination dynamics implemented in the FluTE model with five iterations. However this system exploration could be further expanded, for instance by considering other interventions (e.g. social distancing) separately or in combination with vaccination. While the presented epidemiological results are acquired using the previous generation of simulators, we argue that our approach is applicable to all simulators and should be used for testing and validation when new simulators are developed, and for the emulation to aid policy making across settings after that.

2.5 Acknowledgments

We would like to thank Albert-Jan van Hoek for making the VZV dataset available and Dennis Chao for making the FluTE model available. We also thank the audience of the April 2012 SIMID workshop for constructive feedback on our presentation “Symbolic regression for modeling epidemiological systems” (www.simid.be).

CHAPTER 3

Parameter Estimation

This chapter is based on published work: "Goeyvaerts N, Willem L*, Van Kerckhove K, Vandendijck Y, Hanquet G, Beutels P and Hens N. (2015). Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season specific influenza-like illness incidence. Epidemics, 13, p. 1-9." [61]*

Summary

Dynamic transmission models are essential to design and evaluate control strategies for airborne infections. Our objective was to develop a dynamic transmission model for seasonal influenza allowing to evaluate the impact of vaccinating specific age groups on the incidence of infection, disease and mortality. Projections based on such models heavily rely on assumed 'input' parameter values. In previous seasonal influenza models, these parameter values were commonly chosen *ad hoc*, ignoring between-season variability and without formal model validation or sensitivity analyses. We propose to directly estimate the parameters by fitting the model to age-specific influenza-like illness (ILI) incidence data over multiple influenza seasons. We used a weighted least squares criterion to assess model fit and applied our method to Belgian ILI data over six influenza seasons. After exploring parameter importance using symbolic regression, we evaluated a set of candidate models of differing complexity according to the number of season specific parameters. Transmission parameters (average R_0, seasonal amplitude and timing of the seasonal peak), waning rates and the scale factor used for weighted least squares optimization, had most influence on the fit to the observed ILI incidence. Our results demonstrate the importance of between-season variability in influenza transmission and our estimates are in line with the classification of influenza seasons according to intensity and vaccine matching.

*Both authors contributed equally.

3.1 Introduction

Influenza presents as a mild disease in most healthy adults but is responsible for significant morbidity and mortality among vulnerable groups such as the elderly, patients with underlying health conditions and children. Recently, several countries have introduced routine vaccination of children for two main reasons. First, the rate of influenza hospitalizations in young children is as high as in elderly. Second, children play an important role in influenza virus transmission, so childhood vaccination would provide indirect protective effects for the community.

We developed a dynamic transmission model for seasonal influenza with the aim to enable projecting the effectiveness and cost-effectiveness of mass vaccination strategies [62]. In this chapter, we focus on the dynamic transmission model and estimate key parameters by fitting the model to an age-specific time series of Influenza-Like Illness (ILI) incidence. The model relies on data from Belgium, such as ILI incidence, vaccination coverage and demographic data, though the concepts and methods are generally applicable. In Belgium, influenza vaccines are currently recommended for people over 50 years (with priority for those over 65 years), people with underlying chronic illness, pregnant women and health care workers [62]. In order to gain insights into the existing modeling approaches, we conducted a thorough literature review which yielded 25 articles presenting dynamic transmission models for seasonal influenza. We classified these articles in four main groups based on the modeling approach used.

First, there are standard mathematical models such as SIR (susceptible-infectious-recovered) and compartmental extensions thereof that are mainly designed to capture single epidemics (e.g. [63]). Second, another group of models extends the previous class by including adaptive parameters for seasonality such as seasonally forced transmission rates (e.g. [23, 64]). Influenza occurs in annual epidemics during the winter period, which has been related to many factors e.g. temperature, humidity, viral production and contact patterns [65–67]. Third, some narrative reviews focused on the comparison of various dynamic models (e.g. [68]). Fourth, a final group focused on multi-strain models to evaluate the impact of cross-immunity between different influenza strains by means of theoretical derivations or simulations (e.g. [69, 70]).

We used the dynamic model of Vynnycky *et al*. [23] as a basis for our model because it is an age-stratified model with seasonally forced transmission rates, including annual vaccination. As influenza is mainly spread from person to person through respiratory droplets, transmission depends directly on age-specific

rates of making social contact. Over the last decade, important advances were made in the collection of social contact data to parameterize infectious disease transmission models, such as the large population-based survey conducted in eight European countries as part of the POLYMOD project [71]. The use of empirical observations to inform the 'who acquires infection from whom' matrix has been successfully applied to model the transmission of different airborne infections [16, 72–74]. Vynnycky *et al*. [23] were the first to use POLYMOD contact data to parameterize transmission rates for seasonal influenza.

In this chapter, we propose to directly estimate the dynamic model parameters by fitting the model to multi-season ILI incidence. In many countries, ILI incidence is monitored via surveillance systems such as sentinel networks or online surveys [75]. We capture between-season heterogeneity by including season specific parameters, such that the associated uncertainty can be propagated in predictions for future epidemics and the evaluation of vaccination strategies. By contrasting model predictions against relevant incidence data, our approach improves upon the general practice of imputing pre-specified values to uncertain parameters, such as transmission or waning rates, without formal model validation or sensitivity analyses. Parameter values for dynamic models of seasonal influenza were commonly chosen *ad hoc* or based on inadequate data, e.g. related to historical observations or pandemic influenza.

There are few examples of dynamic transmission models for seasonal influenza that were actually fitted to incidence data. Hsieh [76] estimated age-specific transmission probabilities by fitting to cumulative pneumonia and influenza mortality data from a single low-intensity season, however, the fit to the crude mortality data seems dubious. Finkenstädt *et al*. [64] developed a stochastic model to estimate the rate of antigenic drift from multi-season ILI incidence data. Their results supported the presence of immunity loss, which we will account for by assuming continuous waning after infection or vaccination. Their dynamic model as such was less relevant to our setting as it ignored annual vaccination and age-dependent transmission. The same was true for the deterministic model by Poletti *et al*. [77], which was fitted to ILI incidence data from a single season using a least-squares approach and designed to estimate the impact of behavioral changes during the 2009 H1N1 pandemic. We elaborate on their least-squares method when fitting our model to multi-season ILI incidence data, including age-specific post-stratification weights.

Pitman *et al*. [78, 79] used a model similar to Vynnycky *et al*. [23] to evaluate childhood vaccination in England and Wales. In these studies, parameter values were imputed rather than estimated by fitting the model to epidemi-

ological data. However, during preparation of the current chapter, Baguelin *et al*. [80] presented another model with the same purpose, using an MCMC-based method to estimate model parameters from ILI incidence data. We briefly discuss the main differences between our approaches at the end of this chapter.

3.2 Methods

In this section, we first describe the ILI incidence data after which the dynamic transmission model is introduced, outlining its general structure, the interpretation of the parameters and the model assumptions. At the end of this section, we introduce the statistical methods used to estimate the parameters in the dynamic transmission model.

ILI Incidence Data

The ILI incidence data were collected from a sentinel network of general practitioners (GPs) in Belgium coordinated by the Scientific Institute of Public health. Each week, the GPs reported the total number of ILI consultations by four age groups, 0-4, 5-14, 15-64 and $\geq$65 years, using the following case definition: sudden onset of symptoms, high fever, respiratory (i.e. cough, sore throat) and systemic symptoms (i.e. headache and muscular pain). Weekly estimates for the denominators were available as well, reflecting the population covered by the sentinel network. We analyzed data from the last six pre-pandemic influenza seasons, from October 2003 to August 2009. The ILI incidence data are described in more detail by Hanquet *et al*. [81] and Bollaerts *et al*. [82].

Additionally, a fraction of ILI patients were swabbed by the GPs in the sentinel network during the influenza activity period and tested for influenza A and B. The swabbing of ILI cases was carried out *ad hoc* using quota for each sentinel GP, but without any objective or consistent criteria, and was therefore likely age biased. Due to the non-systematic nature of the swabbing, the weekly laboratory-confirmed influenza case data were incomplete, especially in young children and individuals over 65 years of age. Further, there were few cases of influenza B in Belgium during the study period. We therefore fitted the dynamic model to ILI incidence data rather than to influenza-confirmed ILI incidence data, and calibrated model-based outcomes for the economic evaluation [62].

We thus assumed one generic influenza virus, which should be interpreted as an average of past influenza A and B strains. Therefore, waning of immunity

in our model did capture actual waning of acquired immunity as well as lack of cross protection. The model partly captured season dominance of A and B strains and transmission heterogeneity by incorporating season specific parameters.

Dynamic Transmission Model

General Structure Of The model

We elaborated on the model of Vynnycky *et al*. [23] and use the same notation. This model is an age-stratified SEIRS model with vaccination, classifying the population into compartments of susceptible $S_a(t)$, exposed $E_a(t)$ (infected but not yet infectious), infectious $I_a(t)$, recovered $R_a(t)$ and vaccinated $V_a(t)$ individuals, as displayed in Figure 3.1. Both recovered and vaccinated individuals were assumed fully protected after infection and vaccination, respectively, until their immunity wanes. The population was stratified into age classes of 1 year: [0;1[, ..., [99;100[and we assumed demographic equilibrium. Belgian demographic data from 2009 obtained from Eurostat [83] were used to determine the initial age-specific population distribution N_a and to estimate an age-specific daily mortality rate m_a. We used time steps of 1 day to ensure high precision while maintaining computational feasibility. Short definitions of the parameters in Figure 3.1 are provided in Table 3.1 and further explained below.

We used three time points of transition in the model. We assumed a realistic age-structured (RAS) model in which all individuals move to the next age group on August 31 of each year [84]. Individuals in the final age group (99 years of age) were removed from the population, and as many newborns as there were deaths in the preceding year, were introduced. Therefore, the total population size remained constant. Further, at time point t_{vacc}, any proportion of any age group may receive influenza vaccination, irrespective of their disease or vaccination history. Finally, each year at time point t_{seed}, a fraction p_{seed} of the susceptible population within age group $[a_{1,\text{seed}}; a_{2,\text{seed}}]$ were seeded into the population as newly infectious individuals, to establish a new influenza epidemic. Note that Vynnycky *et al*. [23] assumed the above transition points in relation to aging, vaccination and seeding, all occur at the same time on August 31.

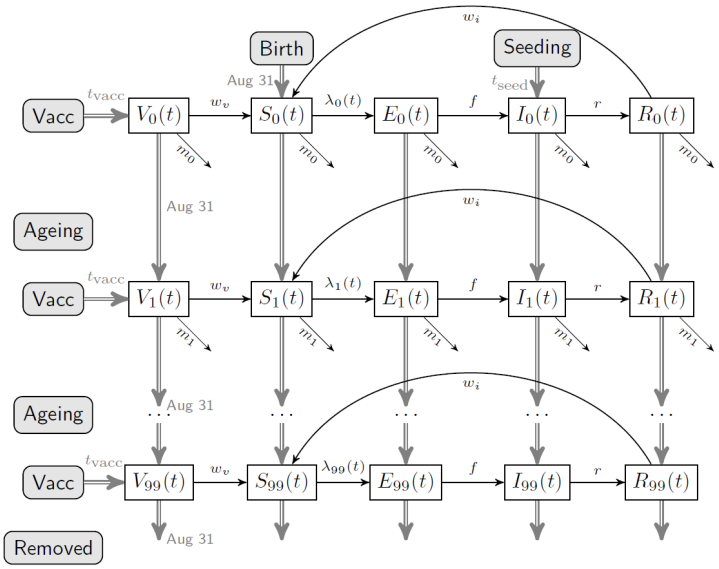


Figure 3.1: Age-stratified SEIRS model with vaccination. Single black arrows indicate time-continuous transitions. Double gray arrows with time point (e.g., t_{seed}) indicate discrete transitions. Detailed transitions for vaccination and seeding are suppressed from display.

The mass action principle relates the force of infection, $\lambda_a(t)$, to seasonally forced age-specific transmission rates:

$$\lambda_a(t) = z(t) \sum_{a'} \beta_{a,a'} I_{a'}(t), \tag{3.1}$$

where $\beta_{a,a'}$ denotes the average daily per capita rate at which an individual of age a' makes effective contact with a person of age a, and $z(t)$ denotes a sinusoidal seasonality function [23]:

$$z(t) = 1 + \delta sin\left(\frac{2\pi(t - t_0)}{365}\right), \tag{3.2}$$

reflecting the relative change of the basic reproduction number at time t, $R_0(t)$, from the average basic reproduction number, $\overline{R_0}$, measured at reference time t_0. The seasonal peak of transmission occurs three months after the reference time t_0. The amplitude parameter δ is bounded $0 \geq \delta \geq 1$ to ensure that $z(t) \geq 0, \forall\, t$. $\overline{R_0}$ is calculated as the dominant eigenvalue of the next generation matrix with elements $\frac{N_a \beta_{a,a'}}{r}$ and $R_0(t) = \overline{R_0} z(t)$.

Social Contact Hypothesis

We assumed that age-specific transmission rates are directly proportional to age-specific rates of making social contact $\beta_{a,a'} = qc_{a,a'}$, where q is a proportionality constant directly related to the value of the average basic reproduction number, $\overline{R_0}$. This is the so-called 'social contact hypothesis' introduced by Wallinga *et al*. [16], who were the first to augment seroprevalence data with data on conversational contacts to estimate transmission rates for airborne infections. We estimated the age-specific contact rates $c_{a,a'}$ from the Belgian POLYMOD contact survey conducted in 2006 [71], assuming that contacts involving physical skin-to-skin touching and taking longer than 15 minutes are a good proxy for those events through which influenza transmission may

Table 3.1: Parameter definitions of the age-stratified SEIRS model with vaccination.

Parameter	Definition
$\overline{R_0}$	Average basic reproduction number measured at reference time t_0
δ	Amplitude of the seasonality function $z(t)$ defined in (3.2). δ determines the peak value of the basic reproduction number ($0 \geq \delta \geq 1$)
t_0	Reference time for the seasonality function $z(t)$ defined in (3.2), at which the basic reproduction number equals $\overline{R_0}$. The seasonal peak of transmission is 3 months later
t_{vac}	Time point of vaccination
t_{seed}	Time point at which infectious individuals are seeded into the population
$a_{1,seed}$	Starting age targeted for seeding infectious individuals into the population
$a_{2,seed}$	Final age targeted for seeding infectious individuals into the population
p_{seed}	Proportion of susceptible individuals seeded into the population as infectious individuals
f	Daily rate at which exposed individuals become infectious
r	Daily rate at which infectious individuals recover and become immune
w_i	Yearly rate at which naturally infected individuals lose immunity
w_v	Yearly rate at which vaccinated individuals lose immunity (assumed equal to w_i)
α	Scale factor to calibrate model-based infection incidence to observed ILI incidence

occur. Previous modeling work revealed that this type of contact fits well the observed seroprevalence profiles for airborne infections such as varicella zoster virus and parvovirus B19 [72, 74, 85, 86].

The age-specific per capita contact rates $c_{a,a'}$ were estimated using the smooth-then-constrain approach described in Goeyvaerts *et al.* [72]. We fitted a negative binomial generalized additive model to the number of reported contacts with age classes of length 1 year, using thin plate regression splines to model the mean as a 2-dimensional flexible function of age. Note that this entailed a continuous contact surface and that the spline basis dimension (taken to be 11) did not correspond to the contact matrix dimension. Subsequently, using age-specific population sizes obtained from demographic data, the estimated contact surface was constrained to account for reciprocity [16]. To reduce boundary effects due to contact data sparseness for the elderly, the contact rates for individuals of age $\geq$ 86 years were based on those of age 85 years. The estimated contact rates are displayed in Figure 3.2, revealing a highly assortative mixing pattern.

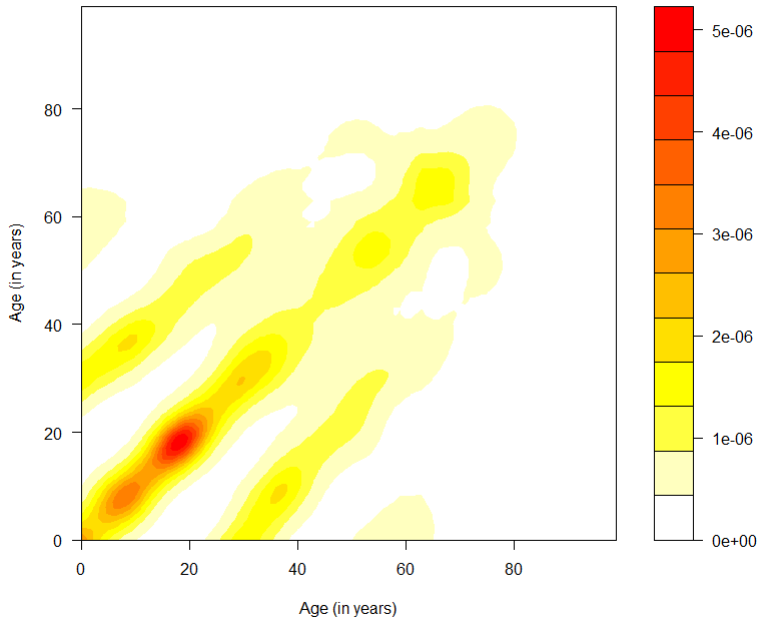


Figure 3.2: Age-specific per capita rates of physical contact $>$ 15 minutes. Estimated from the Belgian POLYMOD data using a smooth-then-constrain approach [72].

Waning Immunity

Genetic variation of the virus produces antigenic novel strains at such a high rate that most people who have had influenza or were vaccinated, were susceptible to a new circulating strain within a few years or even months after infection or vaccination [87,88]. In our model, this process was partially captured by allowing for waning immunity after natural infection or vaccination. The two waning rates, w_i and w_v, were assumed to be equal and age-independent, in the absence of consistent empirical evidence. For example, the literature review by Skowronski *et al*. [89] does not support the historic concern that vaccine-induced antibodies wane more quickly in the elderly compared to the young.

Vaccination Coverage And Vaccine Efficacy

In Belgium, seasonal influenza vaccines are currently recommended for people over 50 years and for high risk groups [62]. We used the following age-stratified vaccination coverage estimates obtained from the Belgian Health Interview Survey of 2008 conducted by the Scientific Institute of Public Health: 0.066% for 6 months - 17 years (arising from a 1% coverage in children with co-morbidities), 11% for 18-49 years, 28% for 50-64 years, 50% for 65-74 years, and 71% for $\geq$75 years [62,81].

Two main types of influenza vaccines are currently registered in Europe: the inactivated trivalent influenza vaccine (TIV), which is injectable, and the live attenuated influenza vaccine (LAIV), which is given as a nasal spray. LAIV vaccine was not authorized in the European Union until 2011 and was not on the Belgian market during the study time period [90]. Therefore, for the purpose of the current chapter, only TIV vaccine efficacy estimates were relevant.

Our model assumes an all-or-none vaccine effect, which means that the vaccine effectively protects a fixed proportion of vaccinated persons, i.e. providing complete immunity against infection, while it completely fails in the remaining part. The 'effective vaccination coverage' was consequently the product of the vaccination coverage and the vaccine efficacy for susceptibility (VE_S) and determined the fraction of the population that moves to the vaccinated state each year. Vaccine efficacy for infectiousness (VE_I) for TIV was found to be non-significant based on experimental challenge studies in seronegative adults [91]. Thus, there is no evidence that vaccinated infected individuals are less infectious compared with unvaccinated infected individuals. In view of these findings, we prefered applying an all-or-none vaccine instead of a leaky vaccine model.

VE_S estimates for TIV were only available from experimental challenge studies in seronegative adults [91]. Therefore, we used estimates of vaccine efficacy for influenza-confirmed ILI (VE for susceptibility to disease, denoted VE_{SP} [92]) obtained from randomized controlled trials and observational studies, as a proxy for VE_S. These studies covered a broader age range and allowed us to stratify VE estimates by both age and type of season, the latter according to influenza intensity (medium-high versus low intensity) and degree of matching between the vaccine and circulating viral strains (relative-good versus poor match). We used the VE_{SP} values for TIV shown in Table 3.2, that were literature-based estimates obtained from influenza-confirmed ILI cases, i.e. based on culture and/or polymerase chain reaction.

Parameter Estimation

Weighted Least Squares

Vynnycky *et al*. [23] imputed pre-specified values, either literature-based or chosen *ad hoc*, to all model parameters presented in Table 3.1 (note that they did not consider α, in the absence of model fitting). For example, their base-case value of $\overline{R_0}$ was 1.8, the best-fitting estimate for the 1957 influenza pandemic in the UK [93]. The parameters presented in Table 3.1 are uncertain and are likely to vary between populations and over time. Therefore, we proposed to estimate these parameters from reported numbers of ILI cases. Other parameters that

Table 3.2: Literature-based vaccine efficacy for infectiousness to disease (VE_{SP}) estimates for TIV per age group, season intensity and vaccine match. Classification of the influenza seasons is based on data from Belgium [81].

Intensity	Match	Season	Age Group	VE_{SP}
medium-high	relative-good	04-05, 06-07, 08-09	6m - 17y	65%
			18y-64y	65%
			≥65y	60%
medium-high	poor	03-04	6m - 17y	48%
			18y-64y	60%
			≥65y	55%
low	relative-good	07-08	6m - 17y	30%
			18y-64y	45%
			≥65y	42%
low	poor	05-06	6m - 17y	16%
			18y-64y	22%
			≥65y	20%

were well recorded, such as demographic and vaccination parameters, were included in the model as fixed values.

We estimated the model parameters using a weighted least squares (WLS) approach. Let $C_a(w_k)$ denote the number of reported ILI cases of age a in calendar week w_k, and let $P_a(w_k)$ denote the corresponding denominator, i.e. the number of individuals of age a covered by the sentinel network in calendar week k. The observed age-specific ILI incidence rate in calendar week k is then calculated as follows: $Y_a(w_k) = C_a(w_k)/P_a(w_k)$. To simplify notation, we suppress the dependency of the model outcome on the input parameters. Let $I_a^*(t)$ denote the number of newly infectious individuals of age a at time t, and let $N_a(t)$ denote the total number of individuals of age a at time t, as predicted by the model. The model-based age-specific incidence rate in calendar week k then equals:

$$Z_a(w_k) = \frac{\sum_{t \in w_k} I_a^*(t)}{\frac{1}{7} \sum_{t \in w_k} N_a(t)}, \tag{3.3}$$

We estimated the model parameters by minimizing the weighted sum of squared differences between the observed ILI incidence rate and the scaled model-based incidence rate:

$$\sum_{i=1}^{4} \sum_{k} v_{a_i}(w_k)[Y_{a_i}(w_k) - \alpha Z_{a_i}(w_k)]^2, \tag{3.4}$$

where the weighted sum was taken over all weekly ILI observations, from week 40 in 2003 to week 35 in 2009, per age group a_i: 0-4, 5-14, 15-64 and $\geq$65 years. The weights $v_{a_i}(w_k)$ are proportional to the corresponding denominator $P_{a_i}(w_k)$ and account for the unequal population sizes represented by the different age groups. The scale factor α calibrates the model-based incidence rate to the observed ILI incidence rate. This factor may absorb several effects, such as the probability for an infected individual to show symptoms, the GP consultation rate, i.e. the probability for a symptomatic infected to consult a GP, and the ILI reporting rate, i.e. the probability for a GP to report a symptomatic influenza case as ILI.

The WLS was a direct measure of goodness-of-fit, with smaller values indicating a better fit to the ILI incidence data. The score was penalized such that models predicting more than 10 new cases per week during at least 10 weeks outside the influenza season from calendar week 40 to 20 were discarded.

Variable Selection And Optimization

We implemented the dynamic transmission model in MATLAB. Starting from a completely susceptible population, the model was pre-run over a burn-in period of five influenza seasons to generate background immunity due to historical infection or vaccination. Our analysis consisted of two stages. First, we explored the parameter space and identified influential parameters by sampling from a Latin hypercube design and by performing feature selection using Pareto-aware symbolic regression [17]. Second, based on the results from the feature selection, we defined a set of nested candidate models with decreasing parameter complexity and obtained parameter estimates using an optimization algorithm.

Using a Latin hypercube design, we sampled parameter combinations from a 12-dimensional parameter space, assuming all parameters were constant across seasons (Table 3.1) Each parameter combination was then used to run the dynamic model. Using Pareto-aware symbolic regression, we analyzed the relationship between the parameter values and the WLS obtained for the observed ILI incidence data.

Symbolic regression aims to capture input-response behavior with algebraic expressions, without a priori assumptions of model structure [30, 31]. First, input variables are randomly combined into expressions using a predefined set of mathematical operators. Second, the resulting expressions are scored by their complexity (i.e. number of parameters and type of operators involved) and prediction error $1 - R^2$, with R the Pearson correlation coefficient between the observed and predicted response (here: WLS). Both objectives are minimized and only the best scoring expressions are retained. Third, the remaining expressions are combined or adapted randomly. Next, the expressions are again scored, selected and so on. This evolutionary process is repeated over many generations to obtain an ensemble of expressions with low prediction error. The complexity objective avoids excessive growth of the expressions and the presence of a parameter in a sufficiently evolved population indicates the variable importance [35]. We used the symbolic regression algorithm from the DataModeler package in Mathematica [54].

Based on the feature selection, we defined a set of nested models with decreasing parameter complexity. For each candidate model, we used the GlobalSearch algorithm from the MATLAB Optimization Toolbox to estimate the parameters. GlobalSearch initiates a gradient-based local solver ("$fmincon$") from multiple starting points. Parameter constraints were taken into account when generating initial points and running the optimization algorithm. We repeated this

optimization process for 1000 different random number streams and selected the best set of parameter estimates i.e. corresponding to the lowest WLS value.

3.3 Results

Symbolic Regression And Variable Selection

We generated 50,000 parameter combinations using a Latin hypercube design and selected the 30% lowest values of WLS (a trade-off between information and noise) for further analysis and feature selection using symbolic regression. Inspection of the parameter presence in the model ensemble from ten independent evolutions, showed that five parameters were driving the goodness-of-fit to the ILI incidence data: $\overline{R_0}$, α and w_i (present in >80% of expressions) followed by δ and t_0 (presence = ≈40%). The following parameters were found less important in explaining the WLS: t_{vacc}, p_{seed}, $a_{1,seed}$, $a_{2,seed}$, r and f (presence ≤30%) and their values were fixed in the model as outlined below. The seeding time point t_{seed} also did not substantially influence the WLS (presence of 12%), however, because the timing of the epidemic peak substantially differs between seasons (Figure 3.3) we retained both the reference and seeding time points (t_0 and t_{seed}) in the model as potential season specific parameters.

We assumed that vaccination took place on October 10 each year, as influenza vaccines in Belgium are generally administered in the second or third week of October. We seeded 200 individuals in each age class of 5-50 years at time point t_{seed}, following Vynnycky *et al*. [23]. Finally, values for the average latent and infectious period were chosen based on the source references of the studies selected in our literature review of dynamic transmission models [62]. There were two eligible source references: a review of experimental challenge studies measuring viral shedding as a proxy for infectiousness [94] and a longitudinal study in households [95]. These studies suggested an average latent period of $1/f$ = 1.0 days [63] and an average infectious period of $1/r$ = 3.8 days [63, 95]. Even though it has been suggested that children on average have a longer infectious period than adults, there is no actual data to support this assumption.

Model Fit And Parameter Estimates

In a first model (Model 1), we allowed the parameters $\overline{R_0}$, t_0 and t_{seed}, to be season specific in order to capture between-season variability, e.g. due to the underlying circulation of different influenza strains. The intensity of influenza transmission, the timing of the seasonal peak of transmission and the timing of

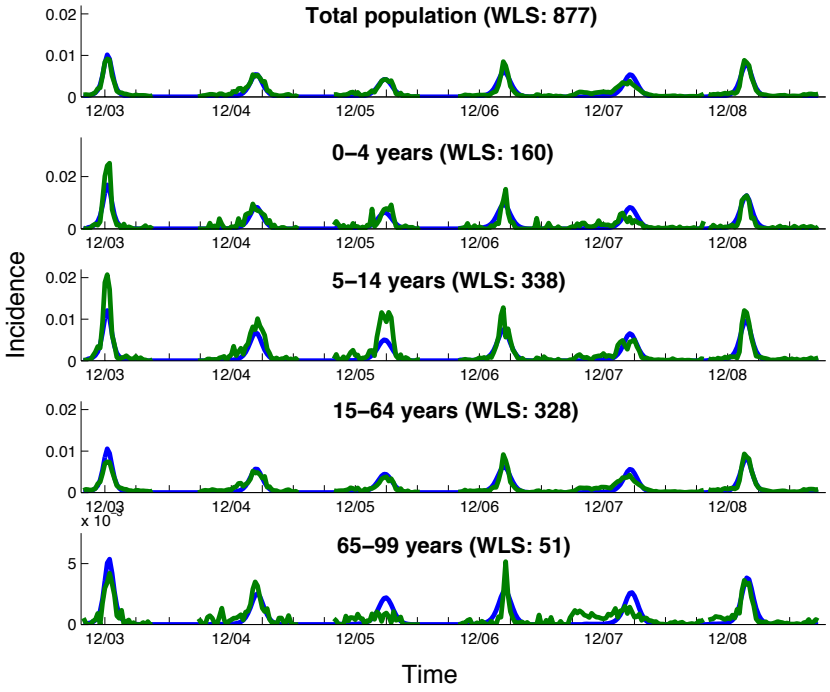


Figure 3.3: Observed ILI incidence rates in Belgium 2003-2009 (green) and corresponding model-based estimates (blue) for Model 3 from Table 3.3. Total population (top) and stratified by age group (four plots below). The weighted least square (WLS) score indicates the fitting accuracy of the model vs the observed incidence.

seeding were thus allowed to vary between seasons. Meanwhile we kept the amplitude δ, the waning rates and the scale factor α constant across all seasons to maintain parsimony and ensure identifiability. We then considered three sub-models of decreasing complexity: a model with constant t_{seed} (Model 2), a model with constant t_{seed} and t_0 (Model 3), and a model without season specific parameters (Model 4).

The parameter estimates obtained with the GlobalSearch algorithm are presented in Table 3.3, while Figure 3.3 shows the fit of Model 3 to the age-stratified ILI incidence data. Allowing the average basic reproduction number $\overline{R_0}$ to be season specific (Model 4 versus Model 3) greatly improves the fit to the ILI incidence reported through the Belgian surveillance system: the WLS decreases from 2150 to 877. Additionally allowing t_0 and t_{seed} to vary by season seems to have a modest impact on model fit.

The model-based and observed total number of ILI cases are similar for Models 1, 2 and 3.The season specific ratios are fairly close to 1, though the dynamic models tend to underestimate the total number of reported ILI cases per season, with the largest differences observed for seasons 04-06 and 07-08. However, early case reports during these seasons are likely confounded by other pathogens causing ILI and circulating before the actual influenza epidemic. The quality of the fit does not differ substantially between age groups. The total number of ILI cases per season is best approximated for the age group of 15-64 years, whereas the models tend to underestimate the total incidence for children aged 5-14 years.

Parameter estimates can only be interpreted conditionally on the model structure and the values of the other parameters. Nevertheless, we observe some

Table 3.3: Parameter estimates for dynamic transmission models obtained by fitting to ILI incidence data from Belgium (2003-2009) using weighted least squares (WLS). See Table 3.1 for parameter definitions.

ID	WLS	Parameter	03-09	03-04	04-05	05-06	06-07	07-08	08-09
		$\overline{R_0}$		3.81	2.91	2.50	2.70	2.64	3.05
		t_0		1/9	30/10	22/10	24/10	20/10	3/10
1	831	t_{seed}		11/11	22/12	1/12	28/12	29/11	29/12
		δ	0.14						
		$w_i = w_v$	0.44						
		α	0.18						
		$\overline{R_0}$		5.06	3.83	3.29	3.71	3.71	4.09
		t_0		8/9	10/10	22/10	2/11	5/9	4/12
2	830	t_{seed}	30/12						
		δ	0.14						
		$w_i = w_v$	0.32						
		α	0.21						
		$\overline{R_0}$		3.99	2.65	2.22	2.19	2.18	2.44
		t_0	22/10						
3	877	t_{seed}	10/10						
		δ	0.52						
		$w_i = w_v$	0.38						
		α	0.21						
		$\overline{R_0}$	4.60						
		t_0	12/9						
4	2150	t_{seed}	12/1						
		δ	0.09						
		$w_i = w_v$	0.25						
		α	0.24						

common features across different models. For example, $\overline{R_0}$ is estimated to be highest in 03-04 and lowest in 05-06, 06-07 and 07-08 (Table 3.3 and Figure 3.4, top), which corresponds well with the seasonal classification in Table 3.2. The 03-04 season was dominated by the new A/Fujian/411/2002 strain, which was not matched with the influenza vaccine of that year, causing a severe epidemic [96]. While values of $\overline{R_0}$ are generally higher for Models 2 and 4, this seems to be compensated by a smaller waning rate. Analyzing the 10% best results obtained with the GlobalSearch algorithm indeed reveals a strong correlation between $\overline{R_0}$, the waning rates and the scale factor α.

Models 1, 3 and 4 have reference time estimates in September-October, which means that the seasonal peak of transmission would occur in December-January. On the other hand, large variability in t_0 is observed for Model 2 with transmission peaks occurring in December-March. Estimates of the seeding time vary substantially between models, possibly indicating a lack of identifiability. Estimates of the average duration of vaccine and naturally induced immunity range from 2.3 to 4.0 years.

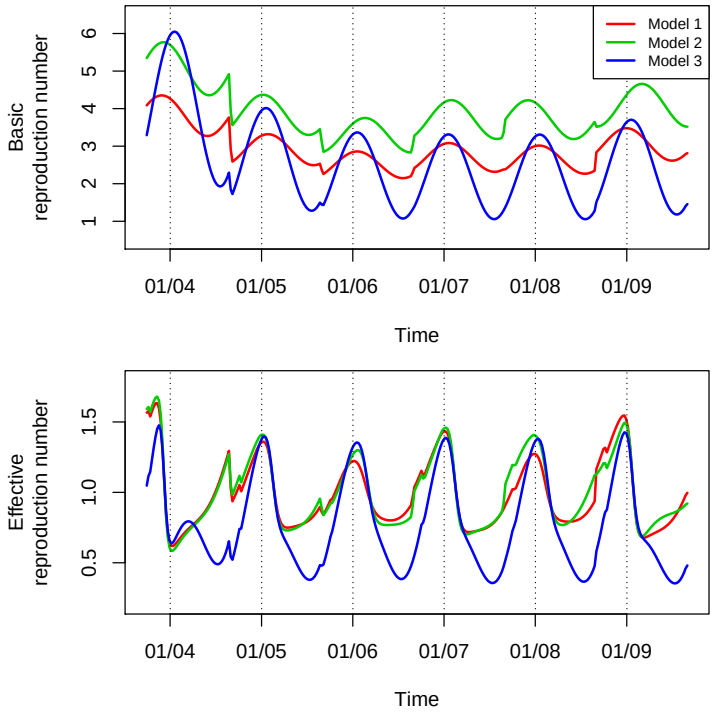


Figure 3.4: Evolution of the average basic reproduction number $\overline{R_0}$ (top) and effective reproduction number R_e (bottom) estimated by Model 1, 2 and 3 (Table 3.3).

3.4 Discussion

This chapter presents a dynamic transmission model for seasonal influenza, designed to evaluate age-specific vaccination strategies, with parameters directly estimated by fitting the model to relevant data of influenza-like illness. This approach greatly improves upon the existing practice of imputing prespecified values to the parameters, without adequately validating the dynamic model outcome. Studies have already indicated the importance of parameter estimation and accounting for uncertainty when using dynamic transmission model outcomes as input for health economic evaluation [26, 97]. With this work we illustrate another important dimension of uncertainty related to the parameterization of the dynamic model.

It is difficult to compare our parameter estimates directly to literature-based ones, since the interpretation of parameters depends on model parameterization and assumptions as well as the population under study, and moreover, we have found that some parameters are highly correlated. Nevertheless, we can derive an estimate of the effective reproduction number over time, $R_e(t)$, which is an important model summary statistic reflecting transmissibility as well as susceptibility (Figure 3.4). Chowell *et al*. [98] estimated R_e from influenza-related mortality data during one month prior to the epidemic peak across three decades of influenza in the US, France and Australia. Figure 3.4 shows that our results match quite well with their estimated average of 1.3 (95% CI: 1.2, 1.4) [98]. Note that small differences in model parameterization entail substantial differences between the estimated evolutions of $\overline{R_0}$ whereas the estimates of R_e are still fairly similar (Figure 3.4). Our results are also in line with other studies showing substantial between-season transmission variability for influenza, partly explained by season dominance of influenza A and B strains [98, 99].

The fit of the models were not indicative of a systematic age bias, though underestimation of the total incidence was more pronounced in children of age 5-14 years. This could be due to age-specific heterogeneity that is not represented by our dynamic model e.g. related to inherent differences in susceptibility or infectivity, potential transmission events not captured by the social contact survey, or over-reporting of ILI in children by the sentinel GPs. We indeed assumed that the GP reporting rate was constant over time and age, i.e. that the reported ILI incidence are representative of the true ILI incidence. In the absence of auxiliary data, it is difficult to disentangle these effects. Furthermore, we assumed that ILI incidence correlates well with true influenza incidence whereas ILI might also reflect infections from other

pathogens such as RSV. Comparing age-stratified ILI incidence to influenza-confirmed ILI incidence shows that the seasonality and the peaks coincide well [62]. We also assumed that physical contacts with a total duration of more than 15 minutes are a good proxy for influenza transmission and that these contact rates are constant throughout the year. Note that our model partly captures time-varying contacts rates through the sinusoidal seasonality function. Since VE_{SP} ignores asymptomatic cases, the proportion of effectively vaccinated might be overestimated. Experimental challenge studies show that the proportion of symptomatic illness in the vaccinated infected is smaller than in the unvaccinated infected, and thus $VE_S < VE_{SP}$ [91].

Estimating the parameters in the dynamic transmission model was a non-trivial computationally intensive task, even for the model without season specific parameters (Model 4). The GlobalSearch algorithm turned out to be sensitive to the initial values and was only able to identify local optima. It was therefore impossible to assess the variance of parameter estimates and to account for the uncertainty originating from the contact data. This precluded the assessment of model identifiability and a simulation study to test the performance of our estimation approach. As future research, likelihood-based and Bayesian approaches using Markov Chain Monte Carlo (MCMC) techniques need to be compared. Recently, advances have been made in using Bayesian approaches to estimate dynamic transmission model parameters, mainly in the context of rotavirus infection [80,100,101]. Note that the approach presented here does not account for serial correlation in the ILI case reports, however, a methodological extension would not be straightforward.

This work differs from Baguelin *et al*. [80] who used MCMC to fit their model to ILI incidence data combined with virological confirmation of ILI cases and auxiliary serological data for one season. For Belgium, serological data were not available. Baguelin *et al*. [80] considered a strain-specific model and stratified the population by 7 age groups. While their Bayesian estimation approach allowed quantifying uncertainty, also with respect to the contact data, the method is highly data-driven and each season and each strain circulating within that season was modeled in isolation. In contrast, assumed one generic influenza virus and modeled influenza dynamics across multiple seasons including waning and boosting of immunity, facilitating future projections of vaccination.

Though the assumption of a generic influenza virus is a strong one, there are other seasonal influenza modeling studies that made the same assumption [64,102]. Extending our model to account for the diversity of circulating strains

which differ from year to year would require many additional assumptions and would likely be a poorer approximation of reality. The advantage of our approach is that it relies on a format of ILI incidence data, as they prevail in many countries using a variety of surveillance systems. Our method could thus prove useful to countries with smaller sized populations and can be used to model ILI incidence from participatory (syndromic) surveillance as well (e.g. Influenzanet [103]).

Axelsen *et al*. [102] very recently published a modeling approach similar to the one described in this chapter. They fitted a SIRS model to multi-season ILI data from Tel Aviv, Israel, also assuming that ILI incidence is representative of true influenza incidence and modeling disease dynamics neutrally for all strains. Axelsen *et al*. [102] also allowed for continuous waning of immunity to capture antigenic drift and included season specific parameters to model large antigenic jumps. They did not consider the impact of seasonal vaccination nor age-related risk, since their goal was not to use the model to project vaccination scenarios, but to understand the relative contribution of various seasonal drivers. Axelsen *et al*. [102] showed that climate variables (temperature, humidity) are important covariates for the seasonality function. Though it is very useful to explore the impact of climate-related dynamics, the use of this model for long-term predictions is not yet possible.

In order to appropriately account for herd immunity when making projections of the effectiveness and cost-effectiveness of different large-scale options for seasonal influenza vaccination, a dynamic transmission model should be incorporated with health economic evaluation [104]. The structure and parameter assumptions in dynamic transmission models are extremely influential for such projections, since they extend on the age-specific incidence of ILI and influenza infections. The method we proposed here for parameter estimation instead of imputation is a step forward to improving such projections. The dynamic transmission models resulting from this study were further used to evaluate the impact of various age-stratified vaccination scenarios in Belgium [62].

3.5 Acknowledgments

This study was co-financed by the Health Care Knowledge Centre (KCE) of the Belgian Federal government and benefited from discussions held as part of the KCE's expert committee and with Joke Bilcke, Adriaan Blommaert and Pieter Neels. We thank Françoise Wuillaume, Viviane van Casteren and Isabelle Thomas (Scientific Institute for Public Health) for collecting and providing sentinel data on ILI and influenza, and Nancy Thiry for useful discussions.

The computational resources and services used in this work were provided by the Hercules Foundation and the Flemish Government - department EWI. We thank Geert Jan Bex for support with using the VSC cluster.

CHAPTER 4

Social Contact Patterns

This chapter is based on published work: "Willem L, Van Kerckhove K, Chao DL, Hens N and Beutels P. (2012). A Nice Day for an Infection? Weather Conditions and Social Contact Patterns Relevant to Influenza Transmission, PLoS ONE 7(11):e48695" [67].

Summary

Although there is no doubt that significant morbidity and mortality occur during annual influenza epidemics, the role of contextual circumstances, which catalyze seasonal influenza transmission remains unclear. Weather conditions are believed to affect virus survival, efficiency of transmission and host immunity, but seasonality may also be driven by a tendency of people to congregate indoors during periods of bad weather. To test this hypothesis, we combined data from a social contact survey in Belgium with local weather data. In the absence of a previous in-depth weather impact analysis of social contact patterns, we explored the possibilities and identified pitfalls. We found general dominance of day-type (weekend, holiday, working day) over weather conditions, but nonetheless observed an increase in long duration contacts (>1 hour) on regular workdays with low temperatures, almost no precipitation and low absolute humidity of the air. Interestingly, these conditions are often assumed to be beneficial for virus survival and transmission. Further research is needed to establish the impact of the weather on social contacts. We recommend that future studies sample over a broad spectrum of weather conditions and day types and include a sufficiently large proportion of holiday periods and weekends.

4.1 Introduction

Influenza is a respiratory virus that causes significant morbidity and mortality during annual epidemics and occasional pandemics [6–8, 105–107]. Seasonal influenza is widespread in temperate climate regions during wintertime and large efforts have been made to study its seasonality. The effect of weather conditions on influenza epidemics received special attention and three main factors have been singled out: efficiency of transmission, host susceptibility and virus survival [65, 108–113]. Temperature and humidity exhibit a distinct seasonality in temperate climates and absolute humidity has been hypothesized to drive influenza seasonality through modulating airborne survival and transmission [66, 114, 115]. A more complete understanding of the virus existence in ambient environment is required to capture the environmental effects on virus infectivity. A latitudinal shift of influenza mortality incidence was observed in Brazil from the northern, tropical regions in summer to the southern, temperate regions of the country in winter, which suggests that the virus migrates from the tropics to temperate regions in both hemispheres during winter [116]. Large respiratory particles shed by infected hosts partially evaporate when the air is dry, become smaller and are more likely to stay airborne [109]. Breathing cold, dry air reduces mucociliary clearance and phagocytic activity of the nasal passage, which normally filter pathogens from the upper respiratory tract [117]. Changes in photoperiod and sunlight exposure alter vitamin D levels and low levels have been shown to impair the body's immune response regulation [118]. Viral stability is also critical for airborne transmission and appears related to environmental factors since the lipid envelope encasing the virus remains longer intact with cold and dry air [109, 119]. Given the sensitivity of airborne transmission to climatological factors, it is believed to be the dominant mode of transmission in temperate regions [109].

Social contact behavior is also known to be important for infectious disease dynamics [16, 71, 120–124]. Social contact patterns are prone to change given contextual changes. For instance, social mixing has been analyzed with respect to day-type (e.g., working days, weekend days and holidays) and generally lower contact rates with more intergenerational mixing are observed on weekends and holiday periods compared to working days [122, 125]. Transmission among school children tends to play a large role in influenza dynamics and both school opening and closing events have been associated with changes in influenza transmission [121, 126, 127]. Also health status interacts on social activity since an increased number of community contacts were observed during influenza seasons in adults without influenza-like illness [128]. In transmission models, mixing patterns are often parameterized based on social contact

surveys, time use studies or social network analyses [16,72,74,129,130], and represented in 'who acquires infection from whom' (WAIFW) matrices. The WAIFW matrix summarizes age-dependent transmission parameters and is a determinant of the basic reproduction number R_0, defined as the expected number of secondary infections caused by a typical primary infection in a fully susceptible population [16].

While most literature on influenza transmission aims to reveal the biological and physical mechanisms associated with different weather conditions, the influence of weather on social contact patterns remains unclear. Mikolajczyk *et al*. touched on this subject by reporting that school children reduced their contacts on rainy days by 16% [120]. Nonetheless, it is hypothesized that the seasonality of many respiratory diseases is driven in part by the tendency of people to congregate indoors when the weather is bad [108]. Analysis of human activity data showed that in addition to day-type, temperature and precipitation also affect daily time usage [131]. In general, people spend 1–2 hours longer indoors during cold weather and about 0.5 hour longer during rainy days [131,132]. However, these changes are small relative to the 21–22 hours individuals usually spend indoors. Furthermore, crowding may occur the year round at public gatherings like shopping malls, festivals, sporting events and conferences. Therefore, seasonal fluctuations in social contact patterns may not be large but could give influenza, in conjunction with other seasonal adjustments, a greater opportunity to spread during winters [108].

The aim of this chapter is to explore modified social contact patterns with respect to weather in order to give more substance to the potential relationship between climatological changes, social contact patterns and influenza seasonality. In the absence of previous in-depth weather impact analysis of social contact patterns, we explored the possibilities and identified limitations of using existing datasets. We partitioned the social contact data according the weather on the day of survey and estimated for each weather condition the mean number of contacts and R_0. Next, relative changes were estimated by the ratio of the mean number of contacts and the R_0's for different weather conditions.

4.2 Methods

Social contact data

A social contact survey was conducted in the Flemish geographic region of Belgium from September 2010 until February 2011. Participants were recruited by random digit dialing on fixed and mobile telephone lines and sampling was performed in order to achieve a representative geographical spread. One person per household was recruited to take part. Sampling was undertaken to obtain an age distribution of 12.5% in the ages 0–8 and 9–24 year, 45% in 25-54 year, 20% in 55-79 year, 4% in 80–89 year and 2.4% in 90–99 year. All participants were asked to fill in a paper diary recording their contacts during one randomly assigned day without changing their usual behavior. No physical samples were collected as part of this study and the ethical committee of the Antwerp University Hospital approved the study protocol. People who agreed verbally to participate received a questionnaire and diary by postal services. They were able to refuse participation even after verbal agreement by not filling in the questionnaire and diary, and/or by not sending it back. The first page of the questionnaire explained that their answers would be used anonymously for noncommercial scientific research purposes. We obtained similar verbal consent with implicit written confirmation from the next of kin, caretakers or guardians on behalf of dependent participants (e.g. children).

Two types of contacts were defined: (1) a two-way conversation with at least 3 words with a person maximum 3 meters away and (2) skin-to-skin touching either with or without conversation. Sampling days were nearly uniformly distributed between all days of the week. Information recorded in the diary included the gender and exact or estimated age of each contacted person, the duration of contacts per person over the entire day as well as the frequency (habitual nature) of contacts with that person. Furthermore, the location or circumstance of each contact was recorded. Three types of diaries were used, adapted to the ages of the participants. For example, the diaries for children (0–12 year) were designed to be filled by a proxy, and included school contacts, which included separate instructions for schoolteachers, whereas those for elderly could also be filled in by a proxy. The diaries were sent and collected by mail. Participants were reminded by phone to fill in the diary one day in advance and followed up the day after. Data were single entered in a computer database and checked manually.

We collected and analyzed social contact data from 1752 participants. The survey was designed to sample 413 participants during a school holiday period (1 - 7 November, 11 November, 27 December - 9 January) or during a weekend

adjacent to a holiday period. Contact data were collected of 1046 participants on regular workdays, 293 on regular weekend days, 286 on workdays during public holidays and 127 on weekend days during or adjacent to public holidays.

Weather Data

We used public data on daily precipitation and temperature from the National Oceanic and Atmospheric Administration (NOAA) [133]. There are 20 weather stations in Belgium that report mean daily temperatures (air and dew-point) based on hourly measurements. The precipitation data was of inferior quality because 631 reports were missing or inconsistent from September 2010 until February 2011. According to the NOAA documentation, some stations report "missing" on days without precipitation. Since weather conditions can vary by locality, we did not use provincial means but matched all participants to the nearest weather station with data on the day they completed the survey. We mapped the participants and the weather stations on 19781 Belgian census tracts (Statistics Belgium, 2001) and used the Pythagorean theorem with the tract coordinates to calculate distances. This way, for every participant a temperature and precipitation measurement was obtained at a median distance of their home of 16 and 19 km respectively.

Many measures of absolute humidity are used in the literature. We used the vapor pressure to be consistent with recent findings [66], which can be calculated from the mean air and dew point temperature. First we calculated the relative humidity (RH) for moist air (assumption: RH $> 50\%$) from the dew point depression [134]:

$$RH = 100 - 5 * (T - T_d), \tag{4.1}$$

with T the air temperature and T_d the dew point temperature. The average RH in Belgium throughout the year published by the Royal Meteorological Institute of Belgium (*http://www.meteo.be/*) is at least 72%, which validates the moist air assumption. Next, the saturation vapor pressure was derived from the temperature using the Clausius-Clapeyron relation [134]:

$$e_s(T) = e_s(T_0) * \exp\left(\frac{L}{R_v}\left(\frac{1}{T_0}\frac{1}{T}\right)\right), \tag{4.2}$$

with $e_s(T)$ the saturation vapor pressure at temperature T, $e_s(T_0)$ the saturation vapor pressure at the reference temperature = 6.11 mbar, T_0 the reference temperature = 273.13 K, L the latent heat of evaporation for water = 2260 kJ/kg and R_v the gas constant for water vapor = 461.5 J/(kg*K). Finally, the vapor

pressure or absolute humidity was calculated by:

$$e(T) = e_s(T) * \left(\frac{RH}{100}\right). \tag{4.3}$$

Diary Weights

The diaries were divided in subpopulations according to the weather and/or the day it was filled in. We adopted the median of all daily temperatures and precipitation levels as threshold to distinguish between days with high and low temperatures and between days with high and low precipitation, respectively. Second, the contribution of each diary was weighted to account for different sampling probabilities in each subpopulation. Information about the participants' age and household size and the day of the week the contacts were collected was used to calculate these weights [135]. Census data from 2001 for Belgium published by Eurostat [136] was used as a reference and weights were constrained to a maximum of 3 to limit the influence of single participants.

Mean Number Of Contacts

We multiplied the total number of contacts from each participant with its diary weight and calculated the mean. We repeated this for contacts of specific types (home, work, school, other) and durations (<15min, >15min, >1h, >4h). We furthermore looked at contacts involving skin-to-skin touching. Next, we measured relative changes in the number of contacts by dividing the weighted means from two conditions.

R_0 Ratio

The contact data were used to calculate transmission rates for the sample populations [16,72,74]. The elements of the social contact matrix $\hat{m}_{ij}$, representing the mean number of contacts in age class j during one day reported by a respondent in age class i, can be estimated by the following expression:

$$\hat{m}_{ij} = \frac{\sum_{t=1}^{T_i} w_t^d y_{ijt}}{\sum_{t=1}^{T_i} w_t^d}, \tag{4.4}$$

where T_i is the number of participants in age class i, w_t^d the diary weight for participant t and y_{ijt} the reported number of contacts made by participant t of age class i with someone of age class j. We used 7 age classes based in part on the Belgian schooling system to reduce sparse data cells in the contact matrix: 0–5 years, 6–12 years, 13–18 years, 19–25 years, 26–45 years, 46–65

years and over 65 years of age. For each sample population we calculated the mean number of contacts between all age groups using all contacts, contacts involving skin-to-skin touching or specific contact types and durations. The reciprocal nature of contacts requires $m_{ij}N_i$ to equal $m_{ji}N_j$, so we defined the elements of the social contact matrix C, representing the per capita daily contact rate between age classes, as:

$$c_{ij} = \frac{m_{ij}N_i + m_{ji}N_j}{2N_iN_j}, \tag{4.5}$$

with N_i the population size in age class i, obtained from demographic data [137]. This reciprocal behavior is not assumed for the specific contact types because for instance contacts for a salesman are most likely not contacts at work for the customers. Therefore, we defined the elements of C for these specific contact types as $c_{ij} = \frac{m_{ij}}{N_j}$.

The next generation matrix G with elements g_{ij} indicates the average number of secondary infections in age class i through the introduction of a single infectious individual of age class j into a fully susceptible population [138]. The next generation matrix is defined by:

$$G = \frac{ND}{L}Cq, \tag{4.6}$$

with N the population size, D the mean duration of infectiousness, L the life expectancy, C the contact matrix and q the proportionality factor. The basic reproduction number R_0 can be calculated as the dominant eigenvalue of the next generation matrix. To estimate the relative change in R_0, we used the R_0 ratio:

$$\frac{R_{0a}}{R_{0b}} = \frac{MaxEigenValue(\frac{ND}{L}C_aq)}{MaxEigenValue(\frac{ND}{L}C_bq)}, \tag{4.7}$$

where indices a and b refer to the contacts registered during different weather conditions. The R_0 ratio can be estimated using only social contact rates when assuming q to be constant since the normalizing constants cancel. Albeit that this q might vary by weather condition, assuming a constant q allows us to focus on the effect of modified social contact behavior.

Bootstrapping

We used a nonparametric bootstrap on the social contact data by resampling participants to calculate 95% confidence intervals. Stratification by participant age was used in the bootstrap to maintain the original age distribution. To estimate the required number of bootstraps to obtain stable results, we calculated R_0 ratio confidence intervals with different bootstrap sizes.

4.3 Results

Social contact data were obtained in Flanders between September 2010 and February 2011 and covered two school holiday periods. Figure 4.1 summarizes the study design by presenting the mean daily temperature, the total daily precipitation, the number of diaries and the mean number of daily contacts over time. It appears that warm weather coincided predominantly with a holiday period in early November and we observed a Spearman correlation of 0.530 between the daily temperature and holiday variables whereas the correlation with precipitation was only 0.005. Therefore, a link between daily temperatures and holiday periods had to be taken into account. Also important is the distribution of participants over time: 70% of the participants filled in their diary on 20% of the survey period (October–November).

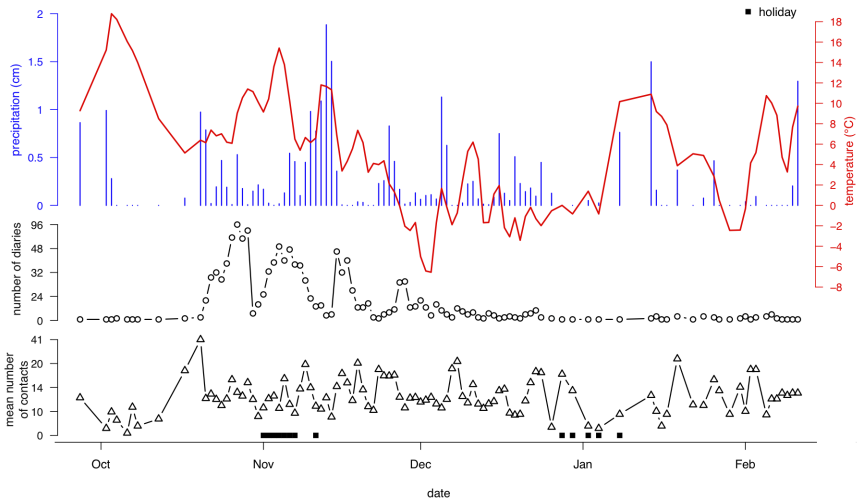


Figure 4.1: Overview of the weather and the social contact data from October 2010 until February 2011. Top: the mean daily temperature (line) and the total daily precipitation (bars). Middle: the daily number of diaries. Bottom: the mean number of social contacts per day. Holiday periods are marked at the bottom with squares.

We partitioned the dataset by the median of the mean daily temperatures (6.83°C) and calculated the ratio of the weighted mean number of contacts for each subpopulation to obtain relative changes. We used a non-parametric bootstrap to generate 95% confidence intervals (CI) and observed a moderate though non-significant increase in the number of contacts on days with low temperatures compared to days with high temperatures (mean 1.130 and CI [0.993;1.283]). The number of school contacts increased significantly (2.004

[1.363;3.027]) with cold temperatures together with contacts of at least 15 minutes (1.190 [1.039;1.341]). A stratification based on the median precipitation value (0.05 cm/day) did not yield significant results.

We validated our methods and dataset with an analysis distinguishing day-types. The boxes in Figure 4.2 present CI for the mean number of contacts ratio for different contact types and durations. We observed a significant increased number of contacts comparing workdays to weekends for all contacts (1.553 [1.397;1.739]) and contacts involving skin-to-skin touching (1.184 [1.049;1.342]). We found similar effects, smaller though significant, for the comparison between regular and holiday periods (Figure 4.2).

Given the effect of holidays on social contact behavior and the correlation between weather and holiday periods, we needed to study these periods separately to eliminate possible confounding. Stratification by working and weekend days did not suffice to overcome the problem with relatively warm holiday periods and an initial partitioning by regular and holiday periods resulted in large uncertainty without significant results. Therefore, it was necessary to partition based on workdays and weekends during regular and holiday periods. For weekends and holiday periods, at least one age group was represented by only 3 to 5 participants, which was too limited. The age distributions of the sample populations for the temperature analysis can be found in Figure 4.3. The analysis for regular weekdays contains at least

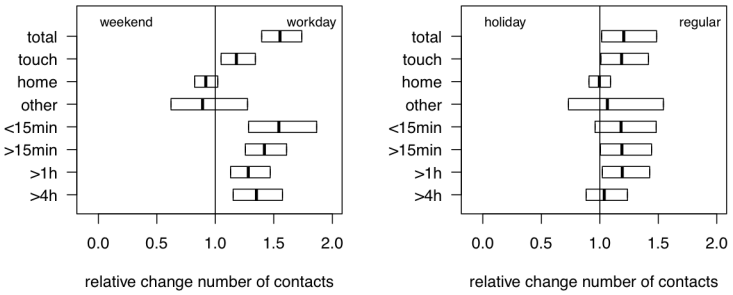


Figure 4.2: Relative change in the number of contacts for weekends vs workdays (left) and holiday vs regular periods (right). The boxes present the 95% confidence interval for different contact types and durations. The relative changes are estimated by the ratio of the mean number of contact for the condition at the right side by the mean number of contact for the condition at the left side.

17 participants in each age group hence we focused on this partition. We encountered similar problems of data sparseness for weekend and holiday periods with a stratification based on day-type and precipitation. A combined analysis with temperature and precipitation was not possible, not even for regular workdays.

We found a significant increase in the mean number of long duration contacts ($>$1hour) on regular weekdays with low temperatures (1.188 [1.025;1.361]) which was found to be a general tendency for all contacts longer than 15 minutes (1.143 [0.995;1.307]). The number of contacts at non-specific locations (henceforth: "non-specific contact type" during leisure, transport, family visits and other activities) decreased significantly (0.461 [0.276;0.737]) on regular workdays with low temperatures.

When we analyzed social contact patterns on regular weekdays with respect to precipitation, we observed a significant increase (1.183 [1.032;1.342]) in the number of contacts longer than 15 minutes when precipitation was low (Figure 4.4). Also the number of school contacts showed a negative association with precipitation (1.502 [1.085;2.041]). We did not observe a change in the number of work or home contacts related to precipitation.

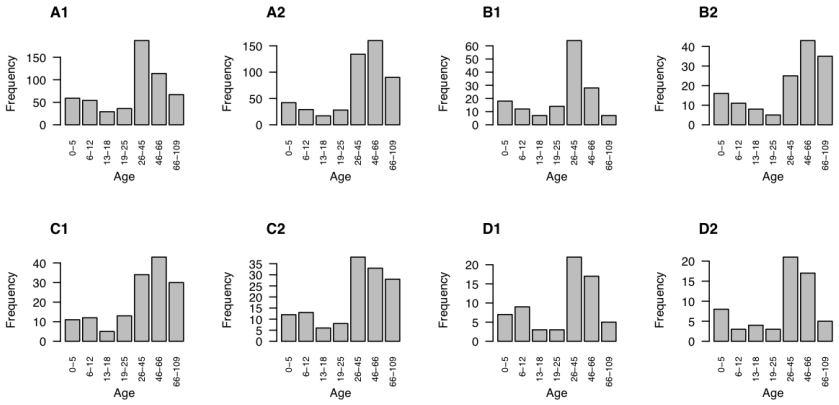


Figure 4.3: Age distribution of the subpopulations after partitioning for day-type and daily temperature. Top: regular workdays with low (A1) and high (A2) temperatures, regular weekend days with low (B1) and high (B2) temperatures. Bottom: workdays during official holiday periods with low (C1) and high (C2) temperatures, weekend days during official holiday periods with low (D1) and high (D2) temperatures. Low: $\leq$median.

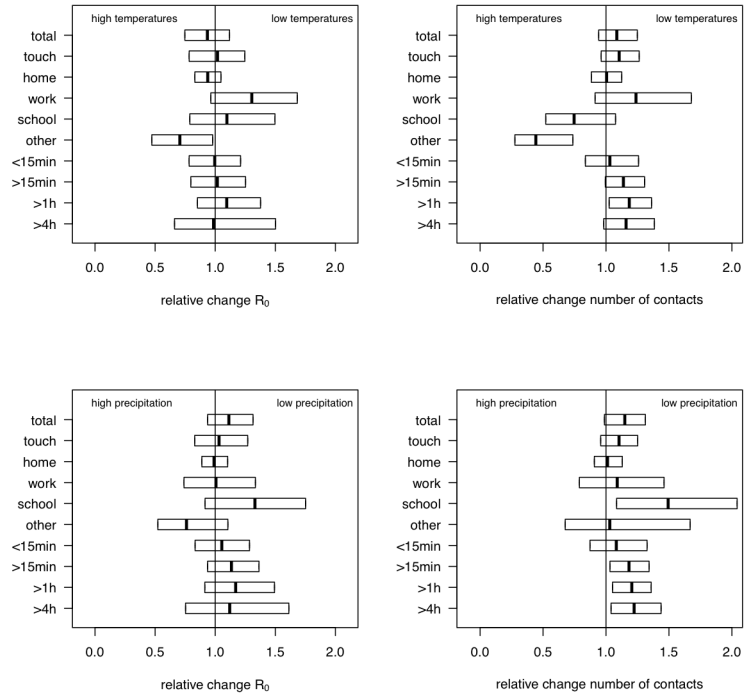


Figure 4.4: Relative change in R_0 and the mean number of contacts considering daily temperatures and precipitation. The boxes present the 95% confidence interval for the R_0 ratios (left) and the mean number of contacts (right) during regular workdays for different contact types and durations. Top: comparison between days with high and low mean daily temperature. Bottom: comparison between days with high and low precipitation.

We investigated the association between absolute humidity (Figure 4.5) and contact patterns and observed a significant increase (1.162 [1.004;1.323]) for contacts longer than one hour with dry air if we stratified by the median absolute humidity (8.41 mbar). This atmospheric condition was also associated with a significant decrease in the number of non-specific contacts (0.445 [0.279;0.732]).

Next to contact frequency, we also focused on transmission potential by including population mixing. We calculated R_0 for each condition and used R_0 ratios to estimate relative changes. The R_0 ratios for the day-types are in line with previous results, i.e. transmission decreases during weekends and holiday periods. The results for the weather conditions showed similar trends as the relative change in the number of contacts, albeit non-significant. Based on sensitivity analysis (Figure 4.6), we used 1000 bootstrap samples to represent uncertainty on our results.

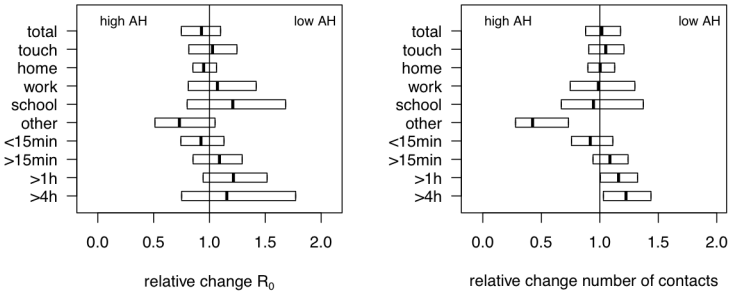


Figure 4.5: Relative change in R_0 and the mean number of contacts considering absolute humidity (AH). The boxes present the 95% confidence interval for the R_0 ratios (left) and the mean number of contacts (right) during regular workdays for different contact types and durations.

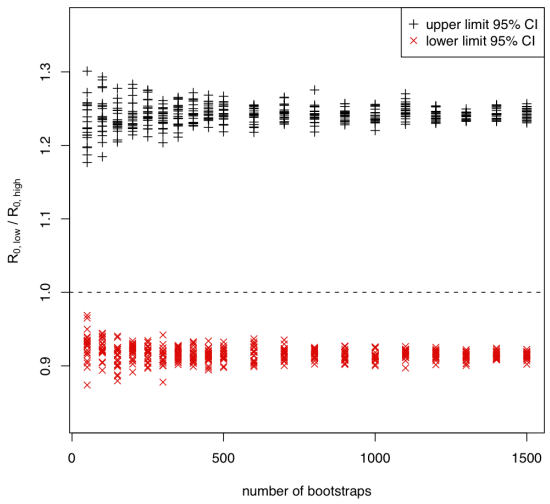


Figure 4.6: Limits of the 95% confidence intervals (CI) for the ratio of the estimated R_0 for regular workdays with high and low precipitation with respect to bootstrap size.

4.4 Discussion

The non-weather related findings we presented were consistent with published results: people have more contacts on regular workdays compared to weekends and holidays [122]. We observed an increased number of school contacts with high temperatures and low precipitation, which is in line with a previous study [120]. Additionally, the number of long duration contacts (>1 hour) increased on regular weekdays with low temperatures and almost no precipitation. Apart from temperature and precipitation, we observed an increased number of prolonged contacts when absolute humidity of the air was low. We also looked at transmission dynamics by calculating R_0 ratios for different weather conditions and observed similar trends for long duration contacts on regular weekdays with low temperature, precipitation and humidity.

The impact of weather conditions on biological mechanisms of influenza transmission has been studied, but the link with social contact patterns remains unclear [65, 108–112, 115]. We presented evidence that weather may influence social mixing patterns. Although one may expect people to leave their home more often on a fine day, this does not necessarily imply that they will make more contacts. Based on the current data, we did not find a relationship between the number of contacts and the weather but the contact duration seemed to depend on the weather. One possible explanation might be that people congregate longer inside when the weather is bad, a phenomenon which has also been observed in time use studies [131, 132]. Prolonged contacts are of particular importance for infectious disease transmission since they tend to be more intensive and more often involve closer interactions [139]. Fluctuations in social contact durations together with other seasonal adjustments might give influenza a greater opportunity to spread during wintertime [65]. In order to focus on the effect of modified social contact behavior, we assumed constant transmission parameters for different weather conditions. However, temperature and humidity are likely to have some impact on virus transmission and host immunity [66, 109, 112]. Especially low absolute humidity is believed to be beneficial for influenza virus survival and transmission [66].

Proper estimation of the impact of weather conditions on social contact patterns requires considering the effect of weekend days and holiday periods. However, we experienced difficulties related to data sparseness when the data was partitioned by day-type. The survey was designed to estimate the effect of day-type without extra partitioning based on weather condition. Each subpopulation has to be representative for the general population. We used weights to account for different age distributions but this was insufficient if the exposure

to a particular weather condition on a particular day-type occurred for only 5 people of a single age group. Therefore, the results presented here are limited to regular workdays. The number of unique weather conditions in the dataset was limited since the majority of the participants were recruited at the end of October and during November.

Future research in this area should carefully consider aspects of data sparseness in study design. Since weather conditions cannot be predetermined, we recommend to sample on many different days to get a broad spectrum of weather conditions. Also a sufficiently large number of respondents should participate on weekends and holiday periods to gain predictive power for these periods. Additional data on the location of the contacts, indoors or outdoors, can also lead to new insights in this challenging research field. Although we found associations between the weather and contact duration, this is no guarantee for a clear-cut causal relation with influenza seasonality. More studies are needed assessing the role of climatic conditions and contact patterns on observed epidemiological transmission patterns to provide adequate information to plan and evaluate mitigating strategies [114, 127].

4.5 Acknowledgments

We thank Patrick Deboosere and Didier Willaert (Interface Demography, Vrije Universiteit Brussel) who kindly provided data on the Belgian census tracts. We thank both referees of the paper for their helpful and constructive remarks.

CHAPTER 5

Computational Efficiency

This chapter is based on published work "Willem L, Stijven S, Tijskens E, Beutels P, Hens N and Broeckhove J. (2015). Optimizing agent-based transmission models for infectious diseases. BMC Bioinformatics, 16:183" [140].

Summary

Infectious disease modeling and computational power have evolved such that large-scale agent-based models (ABMs) have become feasible. However, the increasing hardware complexity requires adapted software designs to achieve the full potential of current high-performance workstations. We have found large performance differences with a discrete-time ABM for close-contact disease transmission due to data locality. Sorting the population according to the social contact clusters reduced simulation time by a factor of two. Data locality and model performance can also be improved by storing person attributes separately instead of using person objects. Next, decreasing the number of operations by sorting people by health status before processing disease transmission has also a large impact on model performance. Depending of the clinical attack rate, target population and computer hardware, the introduction of the sort phase decreased the run time from 26% up to more than 70%. We have investigated the application of parallel programming techniques and found that the speedup is significant but it drops quickly with the number of cores. We observed that the effect of scheduling and workload chunk size is model specific and can make a large difference. Investment in performance optimization of ABM simulator code can lead to significant run time reductions. The key steps are straightforward: the data structure for the population and sorting people on health status before effecting disease propagation. We believe these conclusions to be valid for a wide range of infectious disease ABMs. We recommend that future studies evaluate the impact of data management, algorithmic procedures and parallelization on model performance.

5.1 Introduction

Agent-based models (ABMs) offer endless possibilities to explore heterogeneous problems and spatial patterns but come with a large computational burden. ABMs are increasingly used to model infectious disease transmission, but little attention is given in the literature to model implementation and performance, e.g., in [14, 15, 19–21, 141–145]. Usually the simulation time on large clusters is mentioned, but it is not clear whether computational resources are optimally used. However, computational performance is a significant aspect of a simulators' usefulness. Especially model exploration and sensitivity analysis, which require bulk calculations, benefit from efficient algorithms [17, 56]. Furthermore, improving model performance facilitates model development and testing on workstation systems.

Performance is implementation specific and therefore we compared different close-contact infectious disease simulators starting from two published ABMs for pandemic influenza: FluTE from Chao *et al*. [14] and FRED (a Framework for Reconstructing Epidemic Dynamics) from Grefenstette *et al*. [15]. Both simulators are written in C++ and are free, open source software under the GNU General Public License and the BSD 3-Clause, respectively. The FluTE population model consists of census tracts with communities of 2000 residents on average. The simulation runs in discrete time steps of 12-hours representing daytime with work, school and day community contacts and nighttime with household and home community contacts. All children go to school in the home community and adults are assigned to workplaces based on employment rates and commuting data. The community is the central unit in FluTE and one person is assigned to only one community per time step. The implementation of FRED is based on specific places for social contacts. Different places are used ranging from small households and classrooms to large schools and communities. All members of one place can have social contacts and one person might be assigned to multiple places per time step.

Individual behavior, social contact structures and population setup are very important to simulate infectious diseases. ABMs are suited to model these features because each person can be represented and stored separately. Inherent to these models are many checks and data transfers compared to the number of floating point calculations. For many years, hardware developers have been able to increase the central processing unit (CPU) performance [146]. Mass storage and memory subsystems have improved more slowly for cost reasons, which has introduced a performance gap between processing and accessing data. To reduce this imbalance, a hierarchy of small high-speed cache

memories has been added to the CPU. Instead of fetching data multiple times from the main memory, it is loaded into cache and re-used [147]. The processor loads data into the cache in chunks called cache lines, which leads to efficient processing if in addition to one memory location also the nearby locations are referenced in the near future. This memory characteristic is important for the data layout of software [148]. For example, if person data is stored jointly in a person object ("Array of Structs") and next to a person's age also their gender and zip-code are checked, it will already be available in the high-speed cache. On the other hand, if person attributes are stored in separate containers ("Struct of Arrays") and only the ages are checked, many more ages are available in one cache line and less slow memory accesses are required.

High-speed memory and other advances in CPU technology have enabled performance improvements for sequential software with about a factor of two for every eighteen months during a few decades [147]. Unfortunately, these improvements have now encountered physical limits and processor manufacturers have turned to multi-core and hyper-threading architectures to increase the accumulated peak performance [149]. These novel architectures require adaptations of existing software and new programming approaches to fully exploit the performance potential. Extra attention is needed for shared resources [150] like population data or random numbers.

Random numbers are a key resource of Monte Carlo methods and the more randomness they exhibit, the better [151]. Computer algorithms are by definition deterministic procedures. They can only approximate randomness by generating a stream of so-called pseudo-random numbers. The only true randomness in a sequence of pseudo-random numbers is the "seed" value that gets the series started. The complexity increases even more with parallel simulations. Some good pseudo-random number generators (PRNG) lose their efficiency or quality, or even both, when they are parallelized [152]. In parallel applications, independent streams of random numbers are required for each thread to prevent latency. Different parallelization techniques are used in practice. In "random seeding", all processes use the same PRNG but with a different seed with the hope that they will generate non-overlapping series. More robust and versatile is the "leapfrog" method where one PRNG sequence is distributed (see Methods).

In this chapter, we focus on single- and multi-core performance of discrete-time ABM simulators implemented in C++ to simulate infectious disease transmission. We used a limited close-contact disease simulator as case study. However, the features that we look into are also applicable to more extensive models or

other types of ABMs. We investigate data management, algorithmic procedures and parallelization. We illustrate good-practice of a PRNG in a parallel context. The goal of this chapter is to formulate recommendations for ABM simulators that are straightforward to realize and significantly benefit the performance.

The chapter is structured as follows: First, we describe the methods starting with three different implementations of the population based on a general data structure. Second, we define an extension by adding a sorting algorithm. Third, we specify methods to run simulations in parallel with a shared-memory approach. Fourth, we describe the input data, run parameters and the work environment we used. Next, the results and discussion section presents all findings. Finally, we end with concluding remarks and avenues for further research.

5.2 Methods

Model Structure And Implementation

We have opted for a model structure consisting of households, schools, workplaces and districts similar to published studies [14, 15]. Figure 5.1 shows a schematic overview of the locations, which represent a group of people we define as a "cluster". Social contacts can only be made within a cluster. During nighttime, people can have social contacts with members of their household and home district. During daytime, people stay at home or go to a workplace or school, depending on their age, which also determines their day district. Contact between infectious and susceptible people may lead to disease transmission, which is a stochastic process based on social contact rates, infectiousness and susceptibility.

Figure 5.2 presents the model implementation with a general class diagram. We use a *Simulator* to organize the activities from the people in the *Area*. The *Area* has a *Population*, different *Cluster* objects and a *Contact Handler*. The *Contact Handler* performs Bernoulli trials based on the age of the contacts and random numbers. We included a 2x2 social contact matrix, based on literature [67, 71, 122], in which the transmission rate is doubled for contacts between children (<18y). Each *Cluster* contains links to its members. The *Population* stores all person data (id, age, household, home district, day cluster, day district and health related parameters) within or without *Person* objects but we elaborate further on this issue in the next paragraph. An infection is assumed to follow a temporal pattern of susceptible-exposed-infectious-recovered (SEIR) states similar to an influenza-like disease [14, 15]. After

infection, people need 2 days of latency (infected but not infectious) before becoming infectious and 6 days to recover and acquire immunity against future infections.

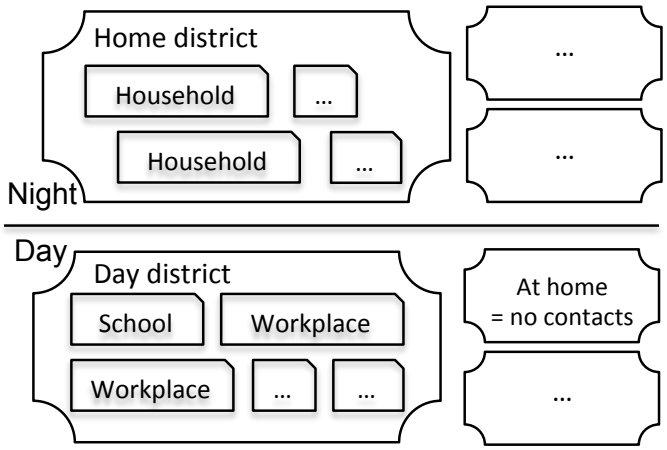


Figure 5.1: Social contact structure. People are member of a household cluster and the corresponding home district at night. During daytime, people can stay at home or go to a school or workplace in a day district.

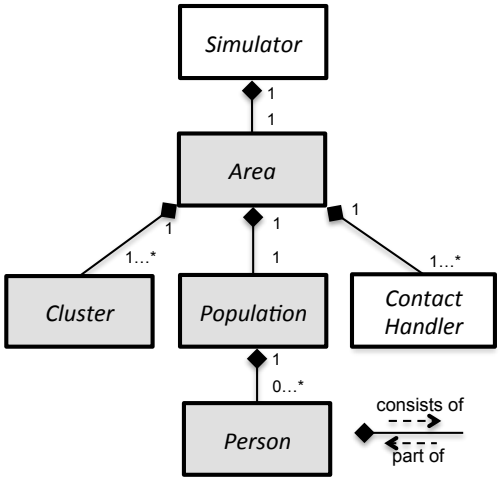


Figure 5.2: Model design: classes and compositions. The digits represent the number of links that are possible. E.g., the *Area* can have 1 or many (*) *Cluster* objects, but a *Cluster* can only be part of 1 *Area*. The FLUTE, FRED and SID model differ in the implementation of the grey classes: FLUTE has less *Cluster* types in *Area* and the *Population* in SID does not contain *Person* objects.

We have constructed three implementations for the previous described transmission model: "FLUTE" and "FRED" are based on the corresponding open source models and "SID" has a novel data layout. The *Area* in FLUTE contains only home and day district *Clusters*. Membership to smaller sub-clusters like households, schools and workplaces can be retrieved from stored cluster IDs in *Person*. People in a district that are also member of the same sub-cluster have two opportunities for social contact and transmission. Therefore, during the processing of social contacts in the district, sub-cluster IDs need to be checked. If two people from a district are also member of the same sub-cluster, we used an aggregated transmission probability instead of performing two random draws. The *Area* of FRED and SID has also separate households and day *Clusters* (= workplaces and schools). We illustrate the difference with the following pseudo-code for home districts and households with transmission probability P_{tr} and P_{tr*} for one or two social contacts respectively.

Transmission algorithm for home districts in FLUTE:

1. **Loop** over all members $\{x\}$
2. **If** x is infectious
3. **Loop** over all members $\{y\}$
4. **If** y is susceptible
5. **If** x and y have equal household ID
6. **If** random number $< P_{tr*}(age_x, age_y)$
7. **Start** disease in y
8. **Else**
9. **If** random number $< P_{tr}(age_x, age_y)$
10. **Start** disease in y

Transmission algorithm for households and home districts in FRED and SID:

1. **Loop** over all members $\{x\}$
2. **If** x is infectious
3. **Loop** over all members $\{y\}$
4. **If** y is susceptible
5. **If** random number $< P_{tr}(age_x, age_y)$
6. **Start** disease in y

Population Structure

Data for an individual is stored as a *Person* object in FLUTE and FRED and the *Population* is a container of *Person* entities, stored consecutively in memory. In SID, the *Population* has a different container for each person attribute and the data of one person is always located at the same index in each of those different containers. For example, to access the age of person *i* in FLUTE or FRED we use "`population[i].age`" while in SID we use "`population.age[i]`".

Population data have been extracted from the 2010 U.S. Synthetic Population Database (Version 1) from RTI International [153,154] for Brooklyn and Nassau County, New York. Every county or state from this database can be used to obtain individual age, household, school and workplace data. People of 16 to 18 years of age with a school and work ID in the original database were assigned to the school to guarantee that people were assigned to only one day cluster. To compare different model implementations, we needed an extra social contact layer (Figure 5.1). We have created home districts by adding households, sorted on ID, until a number of 2000 people was reached. We assumed that household IDs are based on geographic proximity and the threshold was adopted from Chao *et al*. [14]. The day districts have been created analogously. The Nassau population consists of 1.31 million people in 448 519 households and 140 861 day clusters. Brooklyn has 2.46 million people and the cluster sizes range from 1 up to 62 962 people. More details on the study populations are listed in Table 5.1.

Table 5.1: Population statistics. Legend: [min - max] and (median).

Name	Nassau, New York	Brooklyn, New York
Age (years)	[0 - 94]	[0 - 94]
Day districts	386	630
Home districts	656	1231
Day clusters	140 861	183 451
Households	448 519	916 831
Population size	1 313 103	2 463 651
Household size	[1 - 18] (3)	[1 - 16] (2)
Day cluster size	[1 - 25 339] (1)	[1 - 62 962] (2)
Home district size	[1565 - 2009] (2002)	[660 - 2 013] (2002)
Day district size	[1071 - 26 458] (2021)	[1370 - 62 962] (2002)

The population data file determined the initial ordering of the person data in the *Population* object. We used seven different orderings for the same population details: the original sequence from the RTI database, a fully randomized order

and population data sorted according to household, day cluster, and both household (first) and day cluster (second), and vice versa. To minimize the effect of random draws, we created 5 different files for each ordering with a random component.

Algorithmic Extension: Sorting

The open source models [14, 15] process disease transmission by looping over all members of a cluster and if a member is infectious, to match them with all susceptibles. To reduce the total number of operations, we introduced a modified algorithm in which the members of a cluster are first sorted according health status before the infectious members are matched with the susceptible members. A newly infected member is moved ahead of the first susceptible. The member list obtains the following structure: First, recovered and infected (exposed and infectious) members and second, susceptible members. The following pseudo-code shows the sort algorithm for FRED and SID (the algorithm is structured analogously for FLUTE):

Transmission algorithm with sorting in FRED and SID:
1. **Loop** over all members $\{x\}$
2. **If** x is not susceptible
3. **If** index of $x >$ number infected + recovered
4. **Swap** x with first susceptible in the list
5. **Loop** over all non-susceptible members $\{x\}$
6. **If** x is infectious
7. **Loop** over all susceptible members $\{y\}$
8. **If** random number $< P_{tr}(age_x, age_y)$
9. **Start** disease in y

Parallelization: Scheduling

The OpenMP API is often used for shared memory parallel programming in C++ [155]. In this programming model, subsets of a process are managed independently (=threads) and share a global address space of a single or multiple processors which they read and write asynchronously. For each cluster type (household, day district, ...) in an area, a person is a member of only one cluster. Therefore, clusters are stored per type so that these containers can be processed in parallel without synchronization. Parallel processing within one cluster would lead to synchronization overhead. The workload distribution over the threads can be static or dynamic [155]. With static scheduling, a fixed number of tasks are assigned to each thread. In dynamic scheduling, the

workload is distributed over the idle threads until all tasks are done. We have used workloads in chunks of 1 and 10 clusters.

Inputs And Work Environment

We used a 2x2 transmission matrix and assumed that the transmission probability (P_{tr}) is doubled for contacts between children (<18y) [67,71,122]. Similar to the literature [14,15], we estimated the relationship between P_{tr} and the basic reproduction number R_0 by counting the number of secondary cases of one infected in a complete susceptible population. Based on 4000 realizations with seven P_{tr}, we approximated R_0 by $exp(5507 * P_{tr} - 0.1911)$. The total run time depends on the clinical attack rate (AR, total fraction of the population initially at risk that got infected) and for this reason, we performed benchmarks for a range of R_0 values (1.1, 1.25, 1.4, 1.8 and 3). Each simulation was performed for 100 days. To start the epidemic, we infected a random fraction of the population. After testing seeding rates of $1e^{-2}$, $1e^{-3}$, $1e^{-4}$ and $1e^{-5}$, we observed limited impact on the number of cases for these ranges and selected $1e^{-4}$ as baseline setting.

We included the pseudo-random number generator (PRNG) from an open source software package called TRNG [152,156]: a portable and highly optimized library of parallelizable generators. To prevent synchronization and latency, independent streams of random numbers are required for each thread. We used the robust and versatile "leapfrog" method where the PRNG sequence is distributed over p processes by calculating for draw i the $(i * (p-1))^{\text{th}}$ number in the sequence. There are no recommendations to select PRNG seeds to obtain different stochastic results, except that those seeds have to be different. Therefore, the run index has been used to seed the PRNG.

An extended class diagram and a user manual to make use of the project software can be found in Appendix B. During development, we used the Google C++ Testing Framework [157] to perform detailed tests. These tests were applied in automated fashion with every change in the code base via a continuous integration server [158]. The Templatized C++ Command Line Parser library [159] was used to transfer configurations to the executable. The project-code is standard C++11 throughout, independent of external libraries and portable over all platforms that have a GNU compiler (version 4.8 or later) available.

Timings presented in this chapter were obtained from benchmarks on a cluster with Intel® Xeon® E5-2680 v2 2.80GHz CPU's (release Q3'13) from the HPC core facility CalcUA at the University of Antwerp. We confirmed our results

with benchmarks on quad-core Intel® Xeon® W5580 3.2 GHz (release Q1'09) CPU's and AMD Opteron® 6274 CPU's. The GNU compiler (4.8) was used in release mode with compiler optimization "-O3". Appendix B contains more info on the hardware and extra results. The open source tool PerfExpert [160] was used for profiling, as installed on the CalcUA cluster.

We performed additional benchmarks to explore the effect of cluster size, dynamic clusters and increased model complexity on model performance. Methods and results can be found in Appendix B.

5.3 Results And Discussion

The number of infected people is the dominant factor in determining the computational workload and the required simulation time. Therefore, we needed to incorporate distinct epidemic curves in our benchmarks by using different R_0 values. Small deviations in the AR were observed for each R_0 as a result of different stochastic paths with and without the sort algorithm and given the different processing in FLUTE. To prevent stochastic fade-out, which is not appropriate for benchmarks, we used relatively high epidemic seeding rates to introduce new infected people in the population [17]. The benchmarks all report elapsed wall clock times as is appropriate for parallel programs. All results in this chapter are based on mean timings from 10 runs with a different random number generator seed. With intervention strategies, we expect more stochastic fade-out and would require more realizations. Benchmarks are performed on idle computing nodes and results on other hardware can be found in Appendix B.

Simulations with the basic models without concerns of the population order clearly required the longest run times. Figure 5.3 illustrates the total run time for FRED simulations with the Nassau population. Similar results were obtained with the other models (Appendix B). We observed a large decrease in run time when the population is structured according to day cluster and household. The workload for a cluster of size N with I infectious and S susceptible members can be approximated by N health checks to select the infectious members + $I*(N)$ health checks to select susceptible potential contacts + $I*S$ random draws to match the infectious member with the susceptible members. The number of susceptibles decreases with each new case, which explains the decreasing curve in Figure 5.3 for epidemics with high AR. Next, sorting cluster members on health status before processing disease transmission, had a large impact on the performance. The run times for Nassau were reduced with 26% to 79% compared to the basic models, depending on AR and population ordering. The

algorithm with sorting has overhead because of swapping infected and healthy members but overcomes *I*(N)* health checks on susceptibility, which explains the reduced run times.

Given the impact of the AR on the simulation time, we needed to monitor the benchmarks closely. The stochastic transmission process is altered by the sorting algorithm, which has limited impact on the AR. Also, the population ordering determines the initial sequence of the cluster members and thus the random path of the simulator. Figure 5.4 presents the AR from 10 Nassau simulations using different models and population structures. The AR distributions were overlapping, which suggested similar transmission dynamics and approved run time comparisons. To validate the transmission model presented here, we performed simulations with the open source FRED software from Greffenstette *et al*. [15] using population data distributed with the source code for Allegheny, Pennsylvania. We observed ARs of $\pm 68\%$ if $R_0 = 1.4$ and ill people could not stay home, which was close to our results.

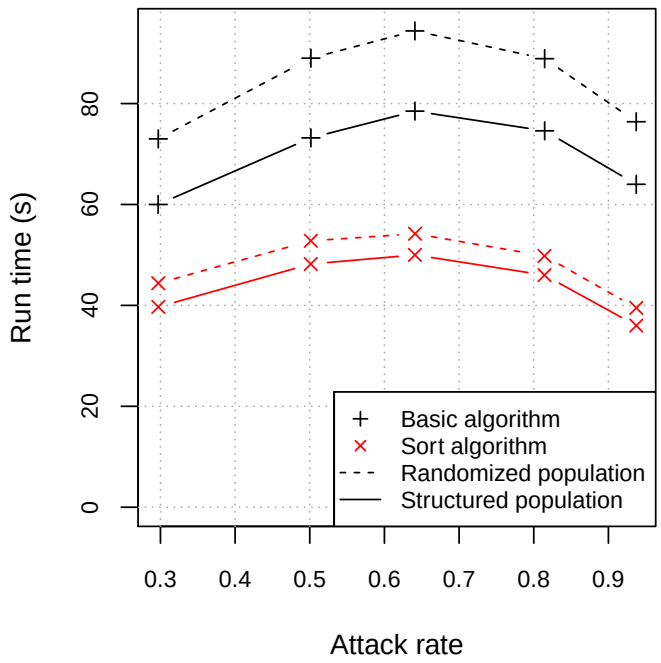


Figure 5.3: Run time according to attack rate and population structure for Nassau simulations using the FRED implementation with/without the sorting algorithm. Structured population: sorted according to day cluster (first) and household (second).

The population ordering appeared to have a large impact on model performance. To examine the effect on an epidemic with R_0 = 1.4 (AR $\pm$64%), we used different versions of the population data with and without sorting according to household and/or day cluster. We repeated our benchmarks multiple times and did not observe large differences in ARs (Figure 5.4). Table 5.2 presents the mean timings from multiple runs with each population ordering using the three basic models. The randomized populations gave the highest run times for all basic models. Using the original structure of the RTI population files slightly decreased the run time. Sorting the population on household ID improved the performance though most optimal was to sort the population on day cluster (first) and household (second). With this sorted population structure, we observed a reduction up to 59% for FLUTE and FRED compared with the randomized population. The effect of the population structure was

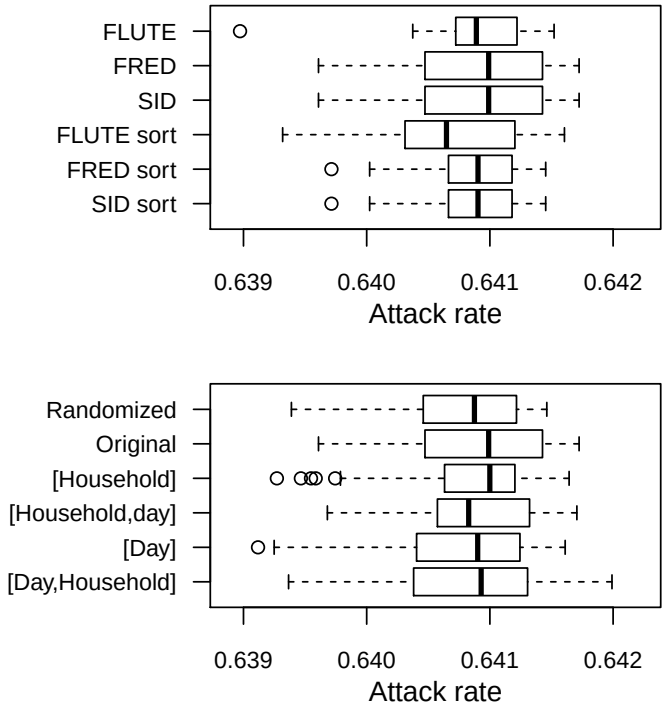


Figure 5.4: Attack rates for Nassau simulation using R_0 = 1.4 and seeding rate = 1e^{-4} according to implementation (top) and population structure (bottom). Results from 10 simulations. The original population structure is used to compare implementations (top) and FRED to compare population structures (bottom). Box = upper and lower quartile, wisker = minimum and maximum excluding outliers, circle: outliers (>1.5 interquartile range), "[ ]" = structured population. E.g., [Day, Household] = sorted population according to day cluster (first) and household (second).

less for SID. The original open source FluTE model [14] uses a population sorted according to household. With our FLUTE implementation, we observed a decrease in run time of 20% by using a population for Nassau sorted by day cluster and household. The population of the original FRED model [15] follows the structure of the RTI population files. A decrease of 6% in total run time can be achieved with our FRED implementation by sorting the population file once. The impact of the population ordering was limited for the models with the sort algorithms.

Table 5.2: Timings for Nassau simulations with different population structures. Results in seconds with R_0 = 1.4 and seeding rate = $1e^{-4}$ (AR = $\pm$64%). E.g., [Day, Household] represents populations sorted according to day cluster (first) and household (second).

Population Structure	**FLUTE**	**FRED**	**SID**
Randomized	108	91	95
Original RTI sequence	101	94	92
[Household]	94	86	87
[Household, Day]	93	80	87
[Day]	89	80	82
[Day, Household]	81	69	78

The general trends from the Nassau simulations were also valid for Brooklyn. The improvement of the sorting algorithm ranged from 34% to 63%. For Brooklyn, we reduced the simulation time by sorting the population once with respectively 15% and 19% compared to the original FLUTE and FRED population. The highest improvement with the population structuring was 39%. Table 5.3 presents the mean run times from 10 Brooklyn simulations with R_0 = 1.4. The ranking of the basic models based on total run time differed between Nassau and Brooklyn simulations due to the different population size and cluster size distribution. For the models with the sorting algorithm in the cluster class, the ranking was independent of the population structure. The extra effort to manage separated household and day clusters in FRED and SID improved the simulators' performance compared to the district-approach from FLUTE. The SID model with the sort algorithm performed best for all benchmarks, especially with the structured population.

On today's multi-core chips, memory access is a critical performance-limiting factor [161]. Therefore, we have analyzed software behavior and memory access patterns with a profiling tool for high-performance computing applications, PerfExpert [160]. We found that the function in *Cluster* to process disease transmission takes on average 98% of the run time. Therefore, optimizations

in this part of the code can have large impact. Since a member cannot be infectious and susceptible at the same time, it is not necessary to check whether a member tries to infect himself/herself. We observed that adding a simple comparison of two C++ pointers or two integer indices in FRED and SID respectively, increased the simulation time with 25%. Table 5.3 presents a selection of the PerfExpert output. FLUTE had the highest penalty for branch instructions (if-then-else structures), which limits the CPU to pipeline instructions and to execute different stages (fetching, decoding, processing and store data) at the same time. A mispredicted branch instruction disturbs this optimization. FRED and SID required less cycles for branch instructions, especially with the sort algorithm. Sorting the cluster members before processing transmission also reduced the data access. Regarding the cache-coherency, we have observed that structuring the population according to the social contact clusters decreased the number of last level cache misses. The sorting algorithm disrupts the memory consistency by relocating references to cluster members. By comparing FRED and SID profiling results, we can confirm the targeted data management strategy from struct-of-array vs array-of-structs: the SID models have fewer last level cache misses.

Table 5.3: Profiling results for Brooklyn simulations. Results with R_0 = 1.4 and seeding rate = $1e^{-4}$ (AR = $\pm$64%). All metrics, except run time, are given in local cycles per instruction (LCPI). LLC: last level cache. [Day, Household] populations are sorted according to day cluster (first) and household (second).

	FLUTE		**FRED**		**SID**	
	Basic	**Sort**	**Basic**	**Sort**	**Basic**	**Sort**
Randomized population						
- Branch instructions	0.29	0.14	0.23	0.07	0.18	0.07
- Data access	1.85	0.99	2.12	0.97	1.69	0.78
- LLC misses	0.14	0.14	0.27	0.48	0.13	0.29
- Run time (s)	229	114	237	103	222	102
[**Day, Household**] **population**						
- Branch instructions	0.26	0.14	0.22	0.07	0.17	0.07
- Data access	1.23	0.85	1.42	0.67	0.66	0.35
- LLC misses	0.05	0.07	0.12	0.25	0.04	0.12
- Run time (s)	168	103	188	94	147	88

Processing disease transmission requires many iterations over independent clusters and therefore seems suited for distributed programming. We observed that the effect of parallelization was dependent of the epidemic curve. Figure 5.5 presents differences in the speedup using FLUTE with 4 threads according to the AR and the epidemic seeding rate (= initial fraction of infected people). The different rates we used did not have impact on the total number of cases but only on the length of the initial phase with a small amount of infected clusters. Simulations with a high epidemic seeding rate and a large AR gave the best speedups using multiple threads. To illustrate the possibilities of parallelization, we compared simulation times using 1 to 20 threads for epidemics with R_0 = 1.4 and seeding rate = $1e^{-2}$ (AR $\pm$64%). Figure 5.6 presents the speedup for SID with basic and sort algorithm using a structured population according to day cluster and household. Similar results were obtained for the other implementations and using the randomized population, which can be found in Appendix B. We observed good speedup for all models and scheduling options with 2 threads. With 3 or more threads, the added

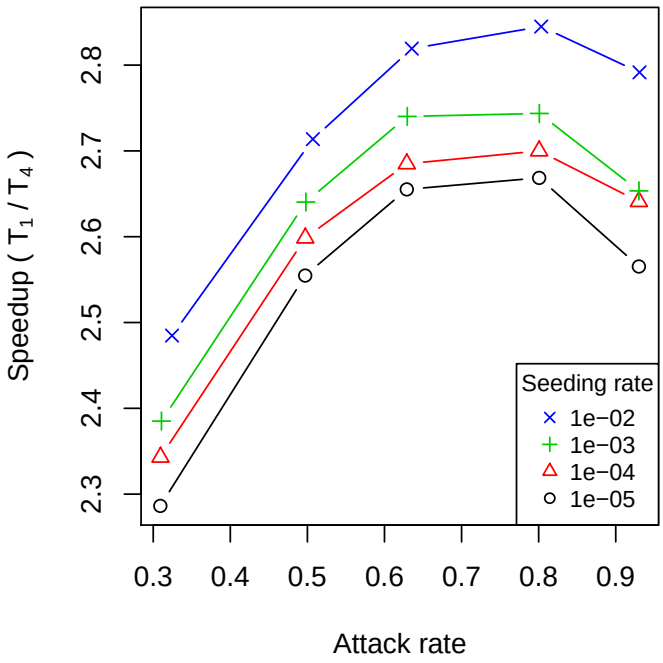


Figure 5.5: Speedup according to attack rate and epidemic seeding rate using the (basic) FLUTE implementation with 4 threads. All simulations were performed with a structured Brooklyn population sorted according to day cluster and household using dynamic scheduling with workload chunks of 1 cluster.

value of extra threads decreased due to memory bandwidth saturation. Making the clusters more self-contained by replacing the member references by actual person data would reduce this limitation although it requires much synchronization between the clusters and extra memory. All basic models had most benefit of dynamic scheduling with workload chunks of 1 cluster. With sorting, FRED and SID seemed to operate more optimally with static scheduling or dynamic scheduling with workload chunks larger than 1 cluster. For FLUTE, the dynamic scheduling with chunks of 1 cluster gave the best speedup. We tested the models on other hardware and observed similar results.

By increasing model complexity, more different cluster types can be used and sorting the population might be less effective. If more person attributes are required for the disease transmission, co-locating these in a person object will be beneficial. On the other hand, the increased amount of person data will reduce the number of persons that fit in the high-speed cache, so more data

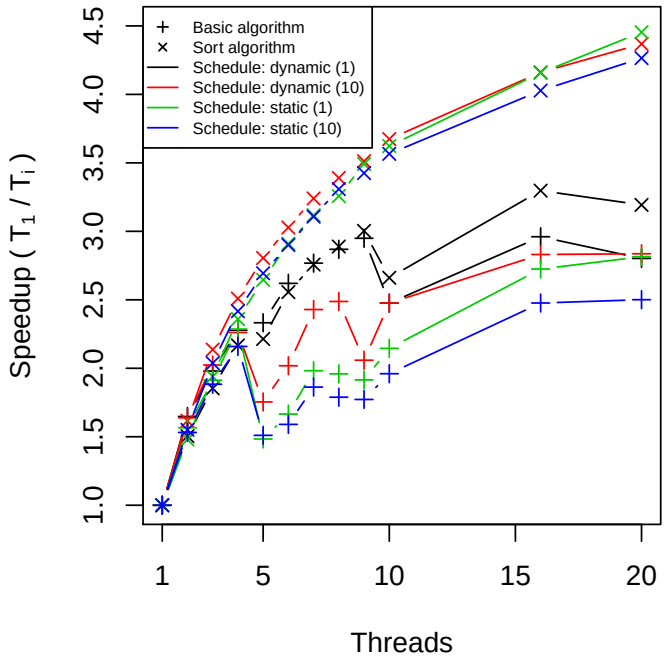


Figure 5.6: Speedup according to thread number and scheduling for Brooklyn simulations using the SID implementation. Results are shown for the basic and sort algorithm with dynamic and static parallel scheduling using workload chunk size of 1 and 10 clusters. All simulations were performed with a structured population sorted according to day cluster and household with R_0 = 1.4 and seeding rate of $1e^{-2}$ (AR = $\pm$64%).

needs to be fetched with higher latency. We explored these model aspects (Appendix B) and observed that cluster size had a large impact on run time. Though, the differences regarding population sorting, model design (FLUTE, FRED and SID) and the sorting algorithm scaled with cluster size. To estimate the effect of dynamic clusters on model performance, we implemented a model with changing cluster membership over time. This way, the run time increased but the overall conclusions remained valid. Increasing model complexity by adding extra person attributes in FLUTE and FRED reduced the impact of the population sorting. The run times for SID remained constant if these attributes were not used, which confirmed the targeted data strategy of struct-of-array vs array-of-structs. The SID design became a disadvantage regarding model performance and workload for the programmer if these extra person attributes were involved in the transmission process.

5.4 Conclusions

ABMs offer a very powerful and flexible framework to analyze infectious disease transmission. Unfortunately they come with a large computational cost. Investing time in code optimization and adaptation to hardware innovations reduces time available for adding new features although it can save much time during testing and in production.

We compared different ABM implementations for close-contact disease transmission models for two U.S. counties. Our ABM consisted of household, school, workplace and district clusters and people in a cluster can have social contacts and transmit an influenza-like disease. The transmission probability was assumed to be age dependent. We observed reductions up to 59% by structuring the model population once according to the largest social contact cluster. Next, sorting the cluster members based on health status before processing disease transmission appeared also very beneficial for the model performance (reduction up to 79% compared to the basic model).

Data movement and access require much more cycles than floating point operations and therefore data layout has impact on run time. We compared models that handle the population in large districts with models that also process the household and day clusters separately. The latter seemed beneficial for the performance especially in combination with the sorting on health status in the clusters. The storage of person data in separate containers instead of per person improved the data locality and cache-coherency and reduced modeling time. Models that sort cluster members on health-status before processing disease transmission are scalable with multiple threads if the epidemics have

a limited initial phase. The parallel scheduling and workload chunk size had significant impact on the simulation time.

Increasing model complexity may reduce the impact of the population ordering. We describe the core of the simulator but more research is needed to assess the role of data layout and sorting algorithms together with mitigation strategies. Although improving data layout by using a separate container for each person feature might increase the model performance, it is counter intuitive for an ABM and requires extra effort from the modeler. The current software does not predict the workload before scheduling the chunks over multiple threads. We believe this scheduling would be a valuable extension to the parallel implementation because the cluster sizes and the amount of infected individuals per cluster can be very heterogeneous. In conclusion, large performance gains can be achieved with limited effort by structuring the population once, adding an algorithm that sorts by health status and selecting appropriate parallel settings.

5.5 Acknowledgments

We thank Geert Bostlap and Franky Backeljauw from the HPC core facility Calcua at the University of Antwerp for their inspiring lectures and support.

CHAPTER 6

Synthetic Population

This chapter is based on unpublished research.

Summary

The predictive value of an ABM depends on the credibility of the synthetic population. Widely used is the "synthetic reconstruction" method, based on a limited seed set derived from survey data together with aggregated census data covering the complete target population. We explored opportunities and identified pitfalls in the construction of a representative seed set based on survey data from Belgium. We extrapolated the survey data over time, corrected ages in single-person households and calculated household-weights based on their size. A weighted least square scoring system based on census data for age and household size demonstrated that our methods did improve the seed set. We believe this chapter presents a valuable extension to the synthetic reconstruction methods and provides avenues for further research.

6.1 Introduction

Agent-based modeling has recently emerged as promising tool to assess real-world problems regarding transport planning and the evaluation of mitigation strategies for infectious diseases. Essential for these models are realistic synthetic populations that match the distribution of individuals and households located in the study area [162, 163]. The synthetic population should be able to represent the agent behavior that is of interest to the modeling purpose. The structure of social contact clusters, especially households, is fundamental to evaluate the impact of epidemic outbreaks. Households are crucial units in the dynamics of infectious diseases since relations between household members are typically characterized by frequent and intimate contacts [125, 164]. Also, households with children have a bridging function allowing an infection to spread from schools to workplaces and visa versa. Transmission among children plays a large role in influenza dynamics since both school opening and closing events have been associated with changes in influenza transmission [14, 121, 122, 126].

Many techniques to generate synthetic populations have been described and most of them belong to the "synthetic reconstruction" method [165]. The main idea is to merge disaggregated data from a survey with aggregated census data covering the whole population of interest. The survey data, referred as the "seed set" should provide individual information about the attributes of interest, though for a subsample of the population. This method has several limitations: First of all, most analysts do not have access to survey data detailing the joint distribution of individuals and households simultaneously due to budget or privacy reasons. Second, it is highly unlikely that the seed set is able to represent the complete interdependency of all the desired individual characteristics [163]. If a demographic or socio-economic class is not represented in the seed set, it will remain absent in the final synthetic population. These limitations have motivated researchers to develop methods entirely based on census data but issues remain [162, 163, 166–169]. For example, reconstructing households based on aggregated data is hard since the members' ages are correlated with each other and with the household size.

Different approaches are used to generate synthetic households in published ABMs. The population of FluTE [14] is based on census tracts and households are sampled from a hardcoded list of age combinations. Tract sizes and commuting data has to be provided as model input to reconstruct the population. The FRED model [15] uses fully characterized synthetic populations from the U.S. Synthetic Population Database from RTI International [153, 154]. This

database provides individual characteristics regarding age, household, school, locality, etc. and can be downloaded per state or city.

In this chapter, we explore opportunities to obtain a representative seed set for the synthetic reconstruction of the Flemish population. First, we describe the survey to obtain household data. Next, methods to adapt the survey data into a more realistic seed set are explained. Finally, we present the results and discuss avenues for further research.

6.2 Methods

We conducted a survey in Flanders between September 2010 and February 2011 with 1772 participants to collect household compositions by age and gender. We removed 14 households with missing data from the database and excluded 2 households of size 1 with ages below 10 years. The cleaned database contained 15 853 household members and was used as core for the seed set.

As reference data, we used aggregated 2001 census data for Belgium containing the number of people by age with respect to household size [136]. This data was provided in 5 years cohorts and converted into 1 year cohorts using the general age-distribution from Belgium in 2001, provided per year. The total number of people per household per size was calculated as the sum over all ages.

To increase the possible household combinations with respect to age, we extrapolated the households 1 and 2 years before and after the survey. We inferred households 1 year before the survey by decreasing all ages by 1 and removing members with age <0. Households 2 years before the survey were deduced similarly. Households 1 year after the survey were extrapolated by increasing all ages by 1 and sampling newborns for each female between 16 and 50 year based on age-dependent fertility rates from Belgium in 2001 [136]. We also accounted for mortality based on age-dependent mortality rates from Belgium in 2001 [136]. To extrapolate households 2 years after of the survey, we need to use the +1y households to obtain realistic numbers of 1 year old children.

The age of a household member is correlated with the household size and with the ages of the other members. Only for single-person households, we could improve the age-distribution by replacing the ages by sampled ages from the reference distribution for household size 1.

To account for household size, we calculated household-weights based on reference data from the Belgian census in 2001. The weight for households of

size i was calculated as:

$$w_i = \left(\frac{frequency_{i,observed}}{frequency_{i,census}} \right)^{-1}. \quad (6.1)$$

Weights were constrained to a maximum of 3 to limit influence of single households and normalized by dividing the weights through the sum over all weights. Normalized weights were used to sample households in the synthetic reconstruction.

Weighted least squares (WLS) scores were used to compare synthetic populations to reference data on age and household size. First, we sampled 300 000 households with a (weighted) seed set. Second, we calculated the observed frequency for each age and household size. These observed and expected frequencies were used to calculate WLS scores as follows, with n the maximum household size or age:

$$WLS = \sum_{i}^{n} frequency_{i,census} * (frequency_{i,observed} - frequency_{i,census})^2. \quad (6.2)$$

6.3 Results and Discussion

We obtained the age and gender of 15 853 household members distributed over 1756 households by a survey conducted in Flanders between September 2010 and February 2011. The survey data provided synthetic populations with a WLS score of 27.3x10^{-5} for the household sizes and 0.88x10^{-5} for the age-distribution of the total population. The low magnitude of the WLS scores can be explained by the multiplication of frequencies, rates between 0 and 1, with a second order of even smaller differences between the expected and observed frequency. The WLS scores for age per household size ranged between 1.71x10^{-5} and 24.1x10^{-5} and are summarized in Table 6.1. The increased WLS scores for the large household sizes were attributed to data sparseness. In households of size $\geq$6, which are uncommon in the total population, some ages were under-represented or completely absent (Figure 6.1). The survey contained an over-representation of household sizes 2 and 4.

To reduce data sparseness, we extrapolated the household combinations in a 5 years interval. The procedure, described in the methods section, is stochastic and subject to limited optimization. We performed 1000 different runs and

selected one stochastic outcome (with seed 589) that balances the WLS scores by age and household size (Figure 6.2). Both WLS scores decreased with this procedure, to 23.1x10^{-5} and 0.68^{-5} for household size and age respectively.

We could improve the age-distribution in single-person households by replacing the ages by sampled ages from the reference distribution. This correction decreased the WLS score for age for this type of household from 2.00x10^{-5} to 1.59x10^{-5}. This was not possible for other household types due to interdependencies between the members' ages.

Normalized weights based on household size decreased the WLS score for household size from 27.3x10^{-5} to 0.24x10^{-5} but had a slightly negative effect on the WLS score for the age-distribution of the total population. The combination of the household extrapolation over time, the age-correction for household size 1 and the size-dependent weights decreased the WLS score for household size and the age-distribution per household type, but did not affect the WLS score for age in the total population. The conclusions based on WLS scores can be confirmed by the age and size distributions of the survey data and seed set in Figure 6.1.

A limitation of these results is the inconsistency between census data of 2001 and survey data of 2010. The age distribution of the total population from 2001 presents a peak around age 40 years, whereas this peak appears around age 50 years in the survey data. There has been a new census in Belgium in 2011,

Table 6.1: Weighted least square (WLS) score of the survey data and final seed set compared to census data with respect to household size and age. The WLS score regarding age is given for the total population and per household size. Median values and 95% confidence intervals are based on 30 repetitions and multiplied with factor 10^5.

Characteristic	Survey		Final seed set	
	Median	95% CI	Median	95% CI
SIZE	27.3	[26.7 ; 27.9]	0.17	[0.16 ; 0.19]
AGE: total	0.88	[0.87 ; 0.88]	0.89	[0.89 ; 0.90]
AGE: size 1	6.54	[6.52 ; 6.55]	1.50	[1.49 ; 1.52]
AGE: size 2	2.82	[2.81 ; 2.83]	1.37	[1.35 ; 1.39]
AGE: size 3	1.71	[1.70 ; 1.72]	0.56	[0.55 ; 0.56]
AGE: size 4	1.94	[1.93 ; 1.96]	1.16	[1.14 ; 1.17]
AGE: size 5	2.54	[2.53 ; 2.54]	1.11	[1.10 ; 1.13]
AGE: size 6	10.2	[10.2 ; 10.2]	3.28	[3.23 ; 3.33]
AGE: size $\geq$7	24.1	[24.0 ; 24.2]	4.03	[4.00 ; 4.06]

but the required data containing the number of people per age with respect to household size are not publicly available (yet). We tried to correct for age in households of size $\geq$2 but it was difficult to estimate over-fitting.

A valuable extension to this research would be a comparison with the available methods that use only aggregated census data [162,163,166–169] and to explore opportunities and pitfalls. The latter can be used in the usual situations where no seed set is available but it is unclear what the accuracy loss is. Next to the initialization of synthetic populations, aging and household transitions over time are challenging tasks for future ABMs that want to simulate over many years.

6.4 Conclusions

Synthetic populations are key elements in ABM to obtain trustworthy results. To evaluate the impact of epidemic outbreaks, the structure of social contact clusters, especially households, are crucial. Many techniques to generate synthetic populations have been described but issues remain. We elaborated on the seed set containing household combinations used in synthetic reconstruction methods to sample new populations.

We conducted a household survey and obtained an adjusted seed set for household size and age-distribution per household type. Therefore, we extrapolated the survey data, corrected ages in single-person households and calculated household-weights based on their size. We were not able to improve the age-distribution for the total population.

We believe we present a valuable extension to the synthetic reconstruction methods. More research is needed to assess the role of the seed set and compare our results with methods based on aggregated data.

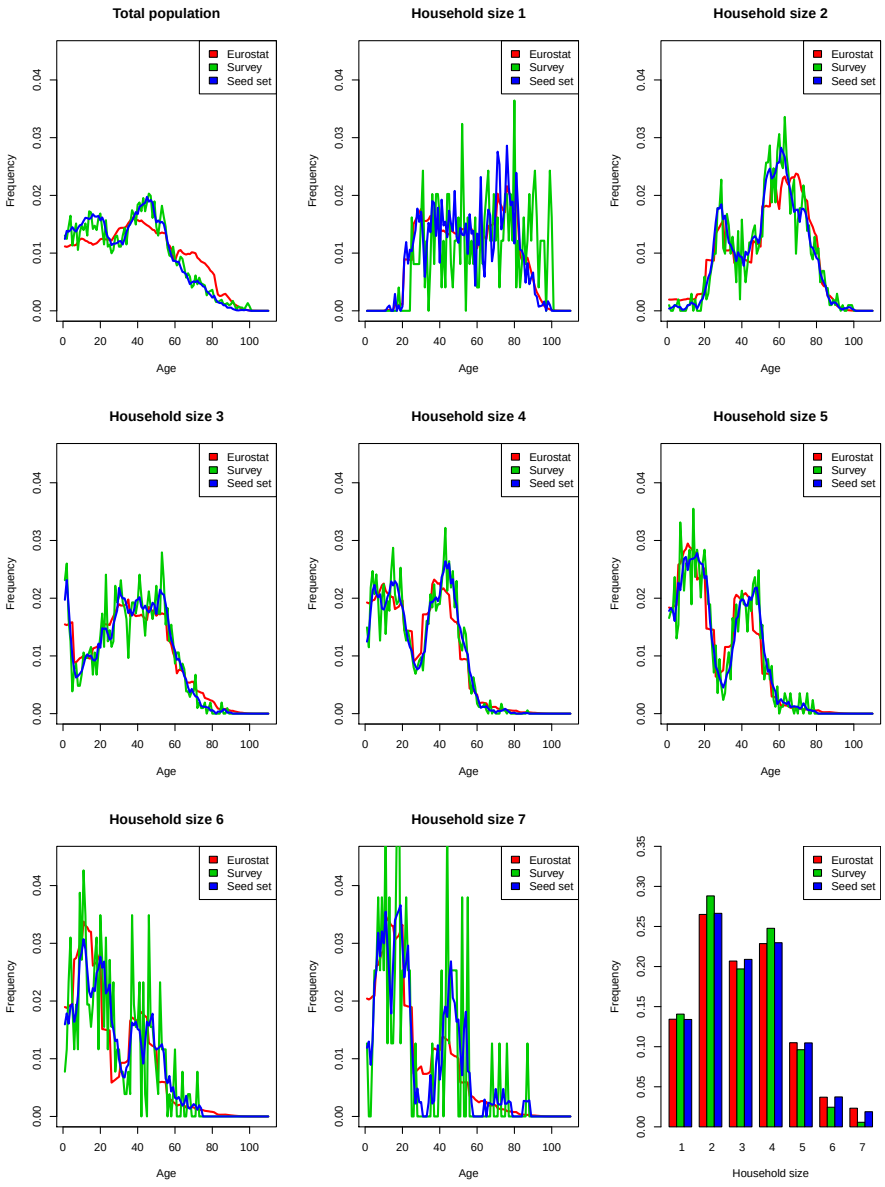


Figure 6.1: Comparing age and household size distributions from Eurostat (=reference) with synthetic populations reconstructed with the original survey data and the final seed set.

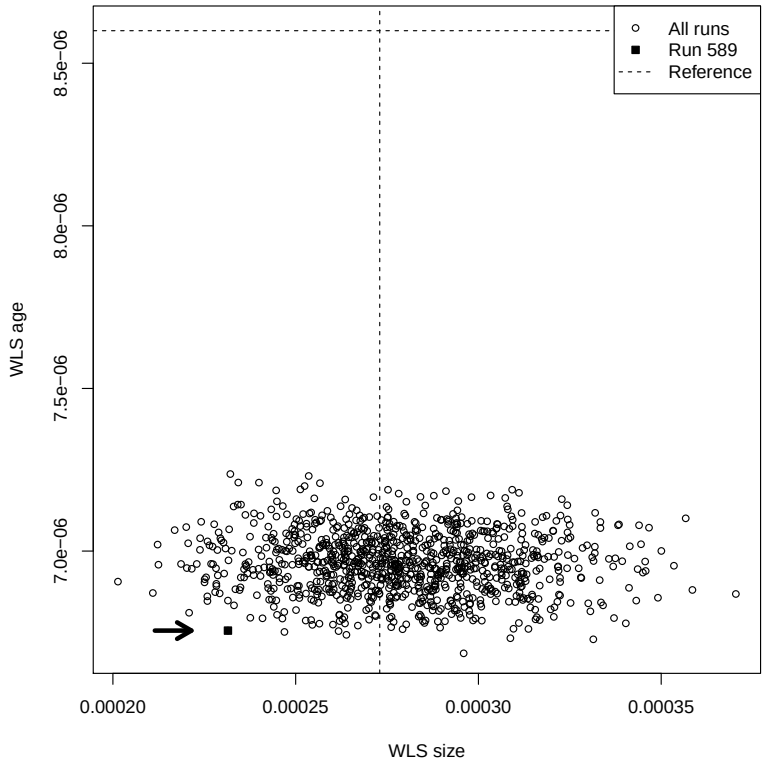


Figure 6.2: Weighted least square (WLS) scores regarding age and household size from 1000 survey data extrapolations over a 5 years interval. The dashed lines represented the WLS scores from the initial survey data. The square, indicated by the arrow, presents the run we selected to construct the seed set.

CHAPTER 7

Discussion

7.1 Introduction

The potential impact of emerging diseases on public health and macroeconomics can be overwhelming and demands monitoring and evaluation of (future) mitigation strategies. For example, influenza causes mostly mild disease in healthy adults but is responsible for significant morbidity and mortality among elderly people and children each year. This age-specific severity may be substantially different and higher during occasional pandemics. Large-scale trials documenting the indirect effects of vaccination are often unfeasible due to budget, time, logistical and/or ethical reasons. Mathematical models are therefore inevitable as pragmatic tools to inform policy [10, 170]. Current epidemic models can be divided into three categories: compartmental models, meta-population models and agent-based models (ABMs). In this first section, we discuss opportunities and drawbacks of the different modeling approaches (illustrated in Figure 7.1) supported by an historical overview.

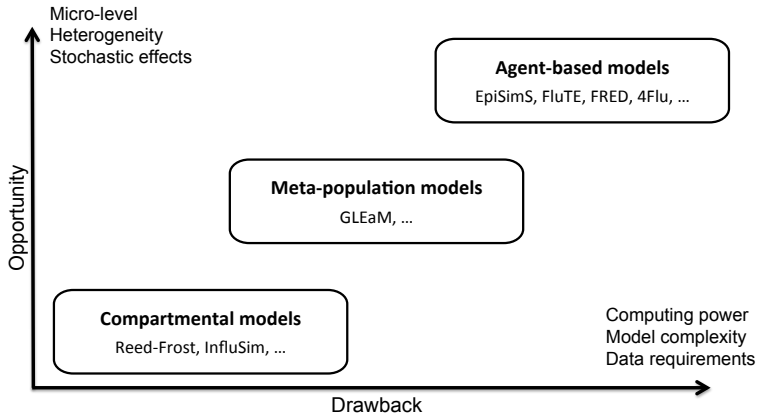


Figure 7.1: Modeling approaches.

Compartmental models. In 1766, Daniel Bernoulli formulated one of the first documented mathematical expressions to evaluate the effect of vaccination against smallpox by expressing the gain in life expectancy based on case fatality and endemic prevalence of susceptible individuals [171]. In 1911, the first model consisting of differential equations was published by Ross to investigate strategies for malaria prevention [172]. In addition, a formal description of disease spread across a population was published in 1920, known as the Reed-Frost model [173]. In this model, each individual in the population had a fixed probability of contacting any other individual and these direct contacts were considered as the source of disease transmission. The population was aggregated into compartments according to health state following a non-immune, infected and immune transition. The progression of the epidemic was fully characterized by the initial population size, number of cases and transmission probability. Many extensions of the original Reed-Frost model have been created and have substantially contributed to the understanding of transmission dynamics [10, 174]. One example is the online available InfluSim model; a deterministic compartmental model based on a system of over 1000 differential equations for several clinical and demographic parameters relevant for pandemic preparedness planning [175]. Although compartmental models are currently the most widely used epidemic models, they mainly focus on categorizing the population in health states and typically ignore heterogeneity and random effects, which are generally most influential in the initial or final stage of an epidemic [10].

Meta-population models. Over the last decade, remarkable progress has been made in the creation of network models to study epidemics [19, 176]. Epidemic systems are described as complex networks where nodes represent entities and interactions are represented by links. Individuals are still divided into different groups according to health states though the infectivity of an infectious node is determined by the network topology instead of assuming complete homogeneous mixing. Meta-population models consider compartmental subpopulations and stochastic inter-population mobility. For example, the Global Epidemic and Mobility (GLEaM) model, initiated in 2006, became publicly available in 2010 and has been used to address many public health issues regarding global epidemic spreading [22, 177–182].

Agent-based models. More detailed than the previous models, ABMs track each individual separately and allow for heterogeneous interactions with respect to social contact behavior and other person characteristics. The approach works bottom-up by representing micro-behavior and social contact patterns to obtain epidemic spread on the macro-level. The first attempt to develop an

ABM for the 1918 pandemic influenza was performed by Elveback *et al*. in 1976 [183]. This study defined the transmission risk as a function of contact time between individuals and modeled the interaction of 1000 people in a community, mixed in different clusters such as households, neighborhoods and schools. Next, Longini *et al*. [18] estimated in 2004 the effectiveness of interventions such as vaccination to reduce the attack rate of an epidemic in a virtual population of 2000 agents. In parallel, EpiSimS was developed based on an urban transport simulation model (TranSimS [184]) by applying graph theory on human mobility patterns to analyze infectious disease transmission [19,185]. Driven by the potential threat of avian influenza in 2005, Ferguson *et al*. [20] used an ABM for influenza transmission in Southeast Asia to evaluate the potential of targeted mass prophylactic use of antiviral drugs. Afterwards, they adapted their ABM to perform large-scale simulations for Great Britain and the U.S. [13]. Germann *et al*. [21] also modeled mitigation strategies for pandemic influenza in the U.S. at that time. They fitted the age-dependent attack rates to the historical 1957-1958 "Asian" influenza and fine-tuned contact probabilities. The model is also used to project the potential economic impact of pandemic influenza strategies in the U.S. from a societal perspective and assess strategies for vaccinating children [186, 187]. Furthermore, Das *et al*. [142] developed another epidemic simulation in 2008 with numerous demographic and community features of 1.1 million agents to help health care executives developing mitigation strategies related to vaccination, prophylaxis, social distancing and hospital admission. Next, an ABM with 57 million inhabitants was constructed by degli Atti *et al*. [141] to consider reactive strategies for influenza epidemics in Italy. In 2010, Chao *et al*. [14] extended the model of German *et al*. [21] and were one of the first to make their ABM, called FluTE, open source. FluTE is based on demographic data to recreate populations organized in communities consisting of census-households. Grefenstette *et al*. [15] released another open source ABM for pandemic influenza in 2013, with more detailed social contact locations. Infectious disease modeling and computational power have evolved such that large-scale ABMs have been used to address several influenza related issues in different countries. Merler *et al*. [188] used an ABM to evaluate strategies to contain an accidental pathogen escape from a biosafety laboratory in Rotterdam. To assess the effectiveness of pre-pandemic influenza vaccination, Halder *et al*. [189] constructed a detailed ABM of a community in Western Australia (Albany, 30 000 people). More recently, Eichner *et al*. [190] published an ABM study on the effect of quadrivalent vaccination on seasonal influenza in Germany with their 4Flu model. Laskowski *et al*. [191] used an ABM in 2014 to evaluate the effect of antiviral treatment and prophylaxis of close contacts in a Canadian remote northern community.

Next to influenza, also other public health issues have been studied with ABMs: infectious diseases like HIV [192], measles [193], ebola [194], tuberculosis [195, 196], malaria [197] and dengue [198] but also life style topics like obesity [199] and alcohol abuse [200]. ABMs have also been used in cancer research to investigate for example drug tolerance [201], cell interactions [202], tumor growth [203] and the link with inflammation [204]. The diversity of ABMs demonstrates the usefulness of this approach for studying complex systems and the evaluation of reactive strategies. The level of information that can be extracted from ABMs is more detailed compared to deterministic or meta-population models, however the computational cost is much higher. The increased granularity of ABMs also requires more advanced input data. Progression in research on social contact behavior and stochastic events contribute to the application and usefulness of current ABMs. In the remainder of this chapter, we discuss key elements of agent-based modeling together with the scientific contribution of the research presented in this PhD thesis.

7.2 Model Exploration (Chapter 2)

Considerable efforts have been made to build realistic simulation models, but few have undergone sufficient testing across a range of settings and situations to be fully validated [10]. What effect do the model parameters have on model outcomes and to which extent are model parameters correlated? How do model assumptions interact and affect the systems behavior? To answer these questions, many scenarios and parameter values need to be explored to account for structural, methodological and parameter uncertainty. Parameter uncertainty is routinely explored with static health models but is often lacking in dynamic models due to the computational complexity [27–29]. Both common and edge manifestations need to be systematically explored to improve confidence in complex models and to obtain system understanding. Therefore, an iterative and active learning approach is needed based on parameter sampling, executing the simulation model, surrogate modeling of the input-response data and interpreting the surrogate models. Model insights need to be used to enhance the experimental design of subsequent simulations by adapting the sampling strategy or reducing dimensionality.

In Chapter 2, we explored vaccination strategies for influenza with an open source ABM [14]. Five iterations were needed to optimize the sampling strategy in the multidimensional parameter space. Our initial experimental setup for the pandemic influenza model showed only a subset of the systems behavior but provided insights leading to an improved design. Finally, we observed a clear inverse relationship between vaccination coverage and clinical attack

rate reinforced by herd immunity. Recommendations for policy making can be explored instantaneously for various settings with an emulator based on the surrogate models.

Active learning based on surrogate modeling with symbolic regression is capable to handle high dimensionality and correlated inputs. As a proof, we identified the most influential variables with symbolic regression from a deterministic compartmental model for varicella vaccination with 185 inputs, including 100 correlated transmission rates.

We demonstrated the active learning approach on two vaccination programs based on two distinctly different dynamic models (stochastic agent-based and deterministic compartmental), though we are convinced that these methods are relevant to address a wide range of public health problems that are informed by modeling. Surrogate modeling with symbolic regression identified the most important variables. Decreasing the uncertainty of these driving parameters will improve the robustness of the simulation results (applied in Chapter 3). Feature selection can also be useful during the development of the simulator.

7.3 Parameter Estimation (Chapter 3)

Herd immunity, also called community immunity, is central to dynamic transmission models [58]. Social contacts are essential for disease transmission and recovered (or immune) individuals disrupt this process. Vaccinating subpopulations that dominate disease transmission can be more (cost-)effective compared to protecting vulnerable groups like elderly people or patients with underlying health conditions. Previous models for seasonal influenza vaccination typically used parameter values chosen *ad hoc* without formal model validation or sensitivity analyses and ignoring between-season variability [23, 63]. However, disease parameters such as transmission or waning rates can also be estimated by fitting model outcomes to observational data, such as age-specific influenza-like illness (ILI) incidence data over multiple influenza seasons. Substantial between-season transmission variability has been shown for influenza, partly explained by season dominance of influenza A and B strains and climate variables, such as temperature and humidity [98, 99, 102]. In many countries, ILI incidence is monitored via surveillance systems such as sentinel networks or online surveys like the "Great Influenza Survey" [75]. Web-based surveys are fast and collect data at the individual level, which are limitations of current surveillance networks [103].

In Chapter 3, we developed a dynamic transmission model for seasonal influenza starting from [23] to assess the impact of vaccination strategies on the incidence of infection, disease and mortality. We incorporated Belgian data regarding age-dependent baseline transmission rates derived from a social contact survey (cf. Chapter 4) assuming that contacts taking longer than 15 minutes and involving physical skin-to-skin touching are a good proxy for potential influenza transmission events. Previous modeling work demonstrated that this type of contact fits well the observed seroprevalence profiles for airborne infections such as varicella zoster virus and parvovirus B19 [72,74,85,86]. The use of empirical observations to inform the 'who acquires infection from whom' matrix has been successfully applied to model the transmission of different airborne infections [16,73,74].

Estimating the parameters in the dynamic transmission model was a non-trivial computationally intensive task, even without season-specific parameters. Using the model-guided experimentation process with symbolic regression from Chapter 2, a set of candidate models of differing complexity according to the number of season-specific parameters was selected. This analysis greatly improved upon the existing practice of imputing pre-specified values to the parameters, without adequately validating the dynamic model outcome. Several studies have already indicated the importance of parameter estimation and accounting for uncertainty when using dynamic transmission model outcomes as input for health economic evaluation [26,97]. The dynamic transmission model from the study in Chapter 3 was further used to evaluate the impact of various age-stratified vaccination scenarios in Belgium [62].

7.4 Social Contact Patterns (Chapter 4)

Infectious disease transmission is driven by social contacts, which create opportunities for new infections. Transmission models need to be well informed on social contact behavior and infectious disease dynamics to obtain trustworthy results [16,71,120–124]. First of all, age is a dominant factor determining social contact behavior [125]. Also day-type is important since people tend to have fewer contacts but with more intergenerational mixing during weekends and holiday periods compared to working days [122,125]. Social activity had been linked to health status as well. Adults with a symptomatic influenza-like illness have fewer community contacts compared to healthy adults [128,205]. Analysis of human activity data showed that also temperature and precipitation are linked with daily time-usage [131]. Seasonal fluctuations in social contact patterns could interact with other seasonal effects and give diseases like influenza a greater opportunity to spread during winters [108].

Over the last decade, mixing patterns have been studied with time-use surveys, network analyses and social contact surveys [16, 129, 130]. Important advances were made in the collection of social contact data, such as the large population-based survey conducted in eight European countries as part of the POLYMOD project [71]. All participants were asked to complete a paper diary recording their contacts during one randomly assigned day. More recently, additional social contact surveys have been performed in e.g., Belgium, France, Peru and Vietnam to improve our knowledge on social contact patterns in different seasons and community settings [72–74, 206]. To elaborate on contextual circumstances that reduce or enhance disease transmission, we analyzed social contact patterns with respect to weather conditions.

The effects of weather conditions on disease transmission have been explained mostly by biological and physical mechanisms [115]. Breathing cold, dry air reduces mucociliary clearance and cellular phagocytic activity of the nasal passage, which normally filters most pathogens from the upper respiratory tract [117]. Respiratory particles shed by infected hosts partially evaporate when the air is dry, become smaller and are more likely to stay airborne [109]. Sunlight exposure interacts on vitamin D, which has been connected to the body's immune response regulation [118]. Viral envelope lipids encasing the virus remain longer intact in cold and dry air, which favors airborne transmission [109,119]. Furthermore, weather conditions also seem to influence social contact patterns since people tend to congregate indoors during periods of bad weather [108, 120]. Time-use data showed that people spend 1–2 hours longer indoors during cold weather and about 0.5 hour longer during rainy days [131, 132].

We were the first to combine social contact survey data with weather conditions and identified opportunities and pitfalls. The survey was conducted in Belgium (Flanders) between September 2010 and February 2011 covering two school holidays. All participants were asked to fill in a paper diary recording their contacts during one randomly assigned day without changing their usual behavior. We collected and analyzed social contact data from 1752 participants. Apparently, high temperatures coincided predominantly with a holiday period in early November. Changes between week/weekend and regular/holiday dominated social contact behavior, so we needed to stratify the analysis according to day-type. We observed an increase in long duration contacts ($>$1 hour) for regular workdays with low temperatures, almost no precipitation and low absolute humidity of the air. Interestingly, these conditions are often assumed to be beneficial for virus survival and transmission [114].

Our results regarding contact duration do not guarantee a clear-cut causal relation between weather and influenza seasonality. More studies are needed assessing the role of climatic conditions and contact patterns on observed epidemiological transmission patterns to provide adequate information to simulate and evaluate mitigating strategies. Chan *et al*. [207] applied the methodology presented here using data from Taiwan and confirmed the link between weather conditions and social contact patterns. Other studies on changing social contact patterns and influenza incidence with respect to weather conditions in China, Finland, The Philippines, Peru, Swiss and The United Kingdom confirmed our findings [208–213]. Insights on mixing behavior and seasonality effects provide opportunities for future transmission models. Modeling should be an integral part in the ongoing cycle of hypothesis generation, data collection and hypothesis refinement [170]. Additional data on the location of the contacts can lead to new insights in this challenging research field.

7.5 Household Networks

The overwhelming majority of transmission models are based on random mixing within the population [214]. However, the number of contacts per individual is considerably smaller compared to the total population size and random mixing does not occur in real life. Models that incorporate network structures assign a fixed set of contacts to each individual instead of constantly changing the contacts with random mixing. Although the total number of contacts per time unit is the same, transmission dynamics are different. Recently published ABMs are based on network structures with social contact clusters such as households, schools, workplaces and communities [13–15, 21]. Still, almost all of them assume random mixing within these social contact clusters.

Households are crucial units in the dynamics of infectious diseases since relations between household members are typically characterized by frequent and intimate contacts [122, 164]. Also, households with children have a bridging function allowing an infection to spread from schools to workplaces and vice versa. Transmission among children plays a key role in influenza dynamics since both school opening and closing events have been associated with changes in influenza transmission [14, 121, 126]. Until now, there is no published evidence to support the assumption of homogenous mixing within households. Egocentric contact surveys demonstrated that within-household networks are important to understand disease dynamics [215, 216].

We performed the first social contact survey designed to study contact networks within households. The survey was conducted in 2010-2011 to study

social contact behavior in households with young children in Flanders and Brussels (Belgium) and was similar to the survey discussed in Chapter 4. In total, we collected data of 1266 participants from 318 households of size 2 to 7. Within-household contacts were identified and matched with other household members. As expected, contacts with household members were typically of long duration and high frequency [125]. We used exponential-family random graph models to fit the within-household networks to infer the processes driving social contact mixing. By performing a simulation study based on the "household model" from Ball *et al*. [164], we did not observe significant differences in terms of total incidence between random mixing and network structures in the households. However, we observed interesting trends regarding household size with respect to the incidence inside households. Further research is required to explore this dataset and network models more thoroughly, especially with respect to contact duration and mitigation strategies.

7.6 Computational Efficiency (Chapter 5)

ABMs are suited to explore heterogeneous problems since each individual has a unique set of characteristics and health status. Increasing hardware performance makes large-scale ABMs more frequently used but little attention is given in the literature to computational efficiency [142–145]. Especially model exploration (cf. Chapter 2) and sensitivity analysis [56] require bulk calculations and benefit from efficient algorithms. This is why, in Chapter 5, we compared different implementations of close-contact disease simulators.

Advances in central processing unit (CPU) technology have enabled performance/speed improvements with about a factor of two for every eighteen months during a few decades [147]. For cost reasons, a performance gap between processing and accessing data occurred, though countered by a hierarchy of small high-speed cache memories close to the CPU. Computing performance is most often memory-bounded, which is important for the design and development of new software [148]. Recently, sequential improvements have encountered physical limits and processor manufacturers have turned to multi-core and hyper-threading architectures to increase the accumulated peak performance [149]. This novel architecture requires adaptations of existing software and new programming approaches to fully exploit the performance potential. Extra attention is needed for shared resources like population data or random numbers [150].

Random numbers are a key resource of ABMs and the more randomness they exhibit, the better [151]. The only true randomness in a sequence of

pseudo-random numbers is the "seed" value that gets the series started. Most models use "random seeding" by using a different seed for each parallel thread, hoping that they will generate non-overlapping series [14,15,56]. More robust and versatile is the "leapfrog" method where one pseudo-random number generator sequence is distributed over the different threads [152].

Free open source software is valuable as application but also in terms of implementation. We had the opportunity to analyze the C++ implementation of two ABMs for pandemic influenza: FluTE from Chao *et al*. [14] and FRED (Framework for Reconstructing Epidemic Dynamics) from Grefenstette *et al*. [15]. These simulation models run in discrete time steps of 12-hours representing daytime with work, school, workplace and community contacts and nighttime with household and home community contacts. Contact between infectious and susceptible people may lead to disease transmission, which is a stochastic process based on social contact rates, infectiousness and susceptibility. We compared different implementations of close-contact disease simulators and observed that data locality had a large impact on model performance. Significant run time reductions can be achieved with limited adjustments of the ABM code. Increasing model complexity reduced the impact of data locality but had no impact on the benefit of the transmission algorithm with prior sorting. Speedup using parallel programming techniques was significant but dropped quickly with the number of cores. We recommend that future studies evaluate the impact of data management, algorithmic procedures and parallelization on model performance. The increasing hardware complexity requires adapted software to achieve the full potential of current high-performance workstations.

7.7 The Need For Efficient Models

Model performance is relevant for all simulation models. In addition to the ABM research described above, we illustrate in this section the need for efficient models by using examples of recently published research on varicella zoster virus (VZV) and rotavirus [59,101,217].

Varicella (chickenpox) is a typical childhood disease caused by VZV, which can reactivate later in life to cause herpes zoster (shingles). Immunological protection against zoster decreases with time but improves with natural re-exposure to VZV (= boosting) [43]. Infant VZV vaccination reduces chickenpox morbidity and mortality but the impact on herpes zoster remains unclear, especially in combination with immunization programs for adults. We tackled part of this problem for Belgium with a deterministic compartmental model (described in [59]). Unlike previous studies, this analysis accounted for major method-

ological and model uncertainties simultaneously and presented outcomes for 26 different target ages at which vaccination can be considered (ages 60-85). The multivariate scenario analysis was only possible due to optimization of the model code. The model was implemented in R [218] and performance was drastically improved in three steps: replacing for-loops by matrix multiplications, using temporary variables for duplicated operations and replacing uniform vectors and matrices by one value. These adjustments are straightforward though easily forgotten during the construction of complex models. Therefore, the development of simulation tools benefits from multidisciplinary teams to unite know-how on system dynamics and model efficiency.

A second example where computational efficiency did matter is a multiple logistic regression model [217]. The objective was to analyze herpes zoster incidence with herpes simplex, other diagnoses (malignancy, depression, diabetes mellitus, auto-immune diseases, asthma, multiple sclerosis, HIV, fractures), several medical interventions (systemic corticosteroids, biological, vaccination) and more infections. Given the high dimensionality, efficient computations were required. Optimizations in terms of data layout and using embedded multi-thread features of MATLAB [219] really improved the modeling performance and made the analysis possible within a reasonable timing.

Next, we used a dynamic transmission model for rotavirus vaccination in England and Wales, which was one of the first to account for uncertainty with Markov Chain Monte Carlo (MCMC) methods [101]. MCMC is a sampling technique that uses a probability distribution based on previous steps in the Markov chain. Initial conditions play an important role so many chains need to be conducted to prevent local maxima. Each chain requires many sampling steps to reach equilibrium and therefore computational cost is one of the main reasons MCMC is not often used for estimating parameters of complex disease transmission models. We estimated rotavirus transmission and vaccination parameters with MCMC algorithms implemented in R [218, 220]. Programming languages for scientific and mathematical computing like R run slower compared to compiled languages like C or C++ but provide many embedded features. The compiled languages are translated into optimized native code adapted to the hardware, which improves computational efficiency. We observed a 16x reduction in run time by using a compiled C-function to evaluate parameter estimations in the MCMC procedure instead of using only R-code. We tried to further optimize the model performance with GPU programming techniques. The transition of one compartmental in a given time step only relied on the previous time step and can be processed in parallel with the other compartments. Unfortunately, the workload per parallel thread was too small

to overcome the overhead of copying the data from the CPU to the GPU and back. Running one MCMC chain to estimate the epidemiological parameters still required about 2 weeks on current state-of-the-art hardware. Although the procedure remained computationally expensive, the model gave more insight in several aspects of the rotavirus vaccines [101].

7.8 Synthetic Population (Chapter 6)

Social contact clusters in an ABM are determined by the synthetic population and need to represent the social behavior in the study area. To evaluate the impact of epidemic outbreaks, the structure of households is of main interest since they strongly affect the transmission potential of a disease and final epidemic size [122, 164]. The population of FluTE [14] is based on census tracts and households are sampled from a hardcoded list of age combinations. No information is provided regarding the probabilities for the household combinations. The FRED model [15] uses fully characterized synthetic populations from the U.S. Synthetic Population Database from RTI International [153, 154]. This database provides individual characteristics regarding age, household, school, locality, etc. and can be downloaded per state or city.

Many techniques for synthetic populations have been described in the literature and most of them belong to the "synthetic reconstruction" method [165, 167–169]. The main idea is to merge survey data with aggregated census data covering the whole population of interest. The survey data, referred as the "seed set" should provide full information about the attributes of interest, though only for a sample of the population. This method has two major limitations. First, most analysts do not have access to a single data set detailing the joint distribution of individuals and households simultaneously due to budget or privacy constraints. Second, it is highly unlikely that this seed set provides a complete description of the interdependency of all the desired synthetic population attributes [163]. If certain demographic or socio-economic classes are not represented in the seed set, they will remain absent in the final synthetic population. These limitations have motivated researchers to develop other methods only based on aggregated census data but issues remain [162, 163, 166–169]. For example, reconstructing households based on aggregated data is hard since the members' ages are correlated with each other and with household size.

In Chapter 6, we explored the methodology to reconstruct synthetic households with survey and aggregated census data to target disease transmission in Belgium (Flanders). We conducted a household survey, along the social

contact survey from Chapter 4, which could be used as "seed set" in the synthetic reconstruction. First, we needed to improve the representativeness of the dataset since we observed data sparseness regarding age and an unbalanced distribution regarding household size. This was done by extrapolating the households over a 5 years period, correcting the ages in single-person households and calculating sampling weights based on household size. We believe we obtained an improved seed set to generate synthetic populations for Flanders and provide a general contribution to synthetic reconstructions. More research is needed to validate our results.

7.9 Future Perspectives

First, future research should focus on model exploration using iterative surrogate modeling by determining a standardized framework in order to speed up the process and make it more accessible. The total number of iterations and the required number of model realizations per parameter dimension should be analyzed in more detail. Especially when discrete and continuous input parameters are used. It would be interesting to further explore ABMs for influenza considering social interventions (e.g., isolation and social distancing) separately or in combination with vaccination.

Second, our findings on parameter estimations with the deterministic model need to be transferred to the ABM context. Likelihood-based methods are already a large improvement with respect to *ad hoc* parameter values but more research is required regarding Bayesian approaches using MCMC techniques. Recently, advances have been made in using Bayesian approaches to estimate dynamic transmission model parameters, mainly in the context of rotavirus infection [80, 100, 101].

Third, further research is needed to establish the impact of weather conditions on social contacts in combination with biological and physical mechanisms associated with disease transmission. Future social contact surveys need to sample over a broad spectrum of weather conditions and day types and include a sufficiently large proportion of holiday periods and weekends. Especially, household networks are of interest and need to be explored with respect to contact duration and mitigation strategies.

Fourth, more research is needed to assess the role of data layout and sorting algorithms on model performance together with increased model complexity and mitigation strategies. Improving data layout by using a separate container for each person attribute increased the model performance of our simulator, but

it is counter intuitive for an ABM and requires extra effort from the modeler. Workload prediction would be a valuable extension in the parallel context because cluster sizes and the number of infectious members, which regulate the number of operations per cluster, are very heterogeneous.

Fifth, given the importance of synthetic populations on social contact behavior and transmission dynamics, more research is required to compare the synthetic reconstruction methods with or without survey data and to evaluate the effect on disease transmission. Next to initialization, aging and household transitions over time are challenging tasks for future ABMs that simulate herd effects over many years.

CHAPTER 8

Conclusions

The use of agent-based models (ABMs) to address public health problems is expanding. In this dissertation, we identified a number of areas in which additional research was necessary to facilitate and support this process.

ABMs are computationally expensive, which makes them hard to fully explore and understand. To gain thorough system understanding, systematic model exploration is required to analyze both common and edge cases. An active learning approach, based on iterative surrogate modeling, provides model insight to address research questions and to improve the modeling process (Chapter 2).

The quality of the ABM outcome depends heavily on the quality of the simulation input. Parameter estimations based on model validation and sensitivity analysis are crucial to obtain trustworthy results. We observed that variability in seasonal influenza transmission is important to consider when future influenza vaccination programs are evaluated (Chapter 3).

Characterization of social contact patterns is needed to accurately simulate transmission dynamics. We confirmed general dominance of day-type on social contact patterns, though also observed a link between long duration contacts and weather conditions widely regarded as being beneficial for influenza transmission. Fluctuations in social contact patterns could interact with other biological and physical processes to regulate seasonality of infectious diseases (Chapter 4).

Improving model performance facilitates ABM development and exploration. Data locality has a large impact on model performance and adapted software is required to fully exploit current high-performance workstations. Large performance gains can be achieved with limited effort by (re)structuring data and selecting appropriate parallel settings (Chapter 5).

To simulate epidemic outbreaks in synthetic populations, households are of main interest since they strongly affect the transmission potential of a disease and the final epidemic size. We observed that extrapolated survey data and weighted household sampling might contribute to current synthetic reconstruction methods (Chapter 6).

Summary

Infectious diseases like influenza cause significant morbidity and mortality during annual epidemics and occasional pandemics. Mathematical models are widely used as pragmatic tools to inform policy on health care interventions when pre-introduction clinical trials are unfeasible for budget or ethical reasons. The spread of infectious diseases in human populations is usually simulated with deterministic compartmental models that partition the population according to health state. Household compositions, locality and social mobility are typically ignored. Agent-based models (ABMs) track each individual in the population separately and allow for random events and heterogeneous behavior, which are large improvements compared to deterministic compartment models. The aim of this PhD was to explore and improve agent-based transmission models for infectious diseases and elaborates on five key aspects: model exploration, parameter estimation, social contact patterns, computational efficiency and synthetic populations.

Model exploration needs to be systematic to gain a thorough system understanding. We investigated the usefulness of model-guided experimentation, called active learning, based on machine learning techniques such as iterative surrogate modeling to systematically analyze both common and edge manifestations of complex model runs. Model exploration is required to focus research by reducing dimensionality and decision uncertainty. We illustrate the methodology for agent-based and compartmental modeling and demonstrate that active learning is needed to fully understand complex system behavior. Surrogate models can be readily explored at no computational expense and be used as emulator to improve rapid policy making in various settings.

Parameter estimations profoundly affect model results. In previous seasonal influenza models, these parameter values were commonly chosen *ad hoc*, ignoring between-season variability and without formal model validation or sensitivity analyses. We propose to directly estimate the parameters by fitting the model to age-specific influenza-like illness incidence data over multiple influenza seasons. The results demonstrate the importance of between-season variability in influenza transmission and the transmission estimates are in line with the classification of influenza seasons according to intensity and vaccine matching.

Social contacts patterns are crucial for infectious disease transmission. Annual epidemics are associated with significant morbidity and mortality, though contextual circumstances that affect seasonal transmission remain unclear. Seasonal fluctuations in social contact patterns could interact with biological and physical adjustments to catalyze transmission of diseases like influenza during winters. We analyzed social contact data and observed general dominance of day-type (weekend, holiday, working day) over weather conditions, but also an increase in long duration contacts (>1 hour) on regular workdays with low temperatures, almost no precipitation and low absolute humidity of the air. Interestingly, these conditions are often assumed to be beneficial for virus survival and transmission. The combination of weather conditions and social contact patterns provides opportunities to improve future transmission models.

Computational performance is relevant to address a wide range of public health problems that are informed by modeling. Advances in high-performance computing enable the use of complex and computational expensive ABMs. However, the increasing hardware complexity requires adapted software to achieve the full potential of current high-performance workstations. We observed large performance gains by re-implementing discrete-time ABM for close-contact disease transmission. Key steps are straightforward: improve data locality and reduce the number of operation with prior sorting. We recommend that future studies evaluate the impact of data management, algorithmic procedures and parallelization on model performance.

Synthetic populations inside ABMs should be able to adequately represent the individual characteristics of interest. The structure of social contact clusters, especially households, and their interactions are essential to simulate epidemic outbreaks. Many techniques to generate synthetic populations have been described but some issues remain. Most methods make use of disaggregated survey data (=seed set) in combination with census data to sample new populations. We elaborated on methods to adjust household survey data into a representative seed set: extrapolation over time, age-correction in single-person households and household-weights based on size. In-dept comparison with existing methods has to be performed, though we believe this approach is a valuable contribution to the synthetic reconstruction methods.

In conclusion, agent-based models offer endless possibilities but systematic model exploration is needed to focus research and optimize parameter estimations through fitting to data. The incorporation of realistic social contact behavior based on synthetic population structures is essential to obtain trust-

worthy results. Improving model performance facilitates model opportunities to evaluate control strategies for emerging infectious diseases and protect public health.

Samenvatting

Infectieziektes zoals de griep veroorzaken jaarlijks een aanzienlijke ziektelast en sterfte onder de oudere bevolking en bij kinderen. Daarnaast vormen deze ook een bedreiging op wereldvlak door occasionele pandemieën zoals de Mexicaanse griep in 2009. Om het beleid inzake gezondheidszorg te informeren worden wiskundige modellen ingezet als pragmatisch hulpmiddel wanneer grootschalige studies onmogelijk zijn omwille van ethische of budgettaire redenen. Tot nu toe werden voornamelijk deterministische compartiment-modellen gebruikt om de verspreiding van infectieziektes te simuleren. Deze modellen focussen op de globale gezondheidstoestand van de populatie en negeren doorgaans individuele heterogeniteit en stochastische effecten die nochtans van groot belang zijn tijdens de initiële of finale fase van een epidemie. Individugebaseerde of agent-gebaseerde modellen (ABM) bieden wel de mogelijkheid om unieke gedragingen van een individu en toeval in rekening te brengen. Het doel van dit doctoraatsonderzoek is het gebruik van een ABM voor de verspreiding van infectieziektes te verbeteren door in te gaan op vijf belangrijke aspecten: model exploratie, parameterschatting, sociale contactpatronen, computationele efficiëntie en synthetische populaties.

Model exploratie moet systematisch gebeuren om grondig inzicht te verwerven in de normale en uitzonderlijke gedragingen van een simulatiemodel. We onderzochten het nut van een model-gestuurde onderzoeksmethode, genaamd "actief leren": een iteratief proces gebaseerd op bestaande machine learning technieken zoals surrogaat-modelleren. Model exploratie is vereist om onderzoek bij te sturen en de dimensionaliteit en structurele onzekerheid te reduceren. Toepassingen met een ABM en een compartiment-model hebben aangetoond dat actief leren nodig is om complexe systemen beter te begrijpen. Surrogaat modellen, die de resultaten van complexe simulaties benaderen, kunnen eenvoudig gebruikt worden als emulator om snel en efficiënt beleidsvorming te informeren in verschillende contexten.

Parameterschatting speelt een grote rol in de betrouwbaarheid van een simulatiestudie. In voorgaand onderzoek betreffende seizoensgriep werden input parameters vaak *ad hoc* gekozen zonder onderscheid te maken tussen seizoenen en zonder formele validatie of sensitiviteitsanalyse. Wij hebben aangetoond dat het ook mogelijk is om parameters te schatten met het transmissie-model door de output te vergelijken met referentiedata. Voor een studie rond sei-

zoensgriep hebben we hiervoor gebruik gemaakt van de leeftijdsspecifieke incidentie van griepachtige symptomen over meerdere griepseizoenen. De resultaten van ons model hebben aangetoond dat seizoen-specifieke parameters belangrijk zijn om de transmissie van griep te simuleren en te voorspellen in de toekomst. De parameterschattingen met ons model inzake transmissie worden bevestigd door de algemene indeling van griepseizoenen volgens intensiteit en vaccinmatching.

Sociale contactpatronen vormen de motor voor de verspreiding van infectieziektes. Ondanks de jaarlijks terugkerende griepepidemie, blijven contextuele omstandigheden die deze seizoensgebonden transmissie beïnvloeden onduidelijk. De hypothese is dat biologische en fysische effecten interageren met schommelingen in sociale contactpatronen en de overdracht van ziektes bevorderen tijdens de winter. In dit kader hebben wij sociale contactgegevens geanalyseerd en observeerden we naast het effect van week/weekend en reguliere/vakantie periode een toename van langdurige contacten (>1uur) tijdens reguliere werkdagen met lage temperaturen, bijna geen neerslag en een lage absolute luchtvochtigheid. Deze weercondities werden in voorgaand onderzoek gelinkt aan een verhoogde kans op virus overleving en overdracht. De combinatie van weersomstandigheden en sociale contactpatronen biedt mogelijkheden voor toekomstige transmissiemodellen.

Computationele efficiëntie is relevant voor een breed gamma van wiskundige modellen. Vooruitgang in computertechnologie stimuleert het gebruik van complexe en veeleisende ABM toepassingen, maar de toenemende hardware complexiteit vereist aangepaste software om het volledige potentieel van de huidige rekenkracht te benutten. We hebben grote efficiëntiewinsten geboekt met de optimalisatie van ABM code in C++ voor een besmettelijke ziekte zoals griep. Kort samengevat: gegevens die vaak gecombineerd worden in het model moeten ook samen opslagen worden in het geheugen en bijkomende sortering kan het aantal intensieve bewerkingen reduceren. Wij raden aan dat toekomstige studies de impact van data management, algoritmische procedures en parallellisatie evalueren om optimaal gebruik te maken van de simulatie modellen.

Synthetische populaties moeten in staat zijn om de individuele kenmerken, die van belang zijn voor het onderzoek, na te bootsen. In de context van infectieziektes zijn de clusters waarin sociale contacten plaatsvinden, met name de huishoudens, van groot belang samen met hun onderlinge uitwisseling. De meeste methoden voor synthetische reconstructie maken gebruik van een beperkte set van individuele huishoudgegevens in combinatie met geaggre-

geerde gegevens over de volledige populatie. Wij beschrijven een methode om de representativiteit van huishouddata te verhogen zodat deze meer geschikt is als basis voor een synthetische populatie. Hiervoor is nood aan extrapolatie van de data over een periode van 5 jaar, correctie van leeftijden in 1-persoonshuishoudens en sample-gewichten op basis van huishoudgrootte. Een grondige vergelijking van onze resultaten met bestaande methoden dient nog te gebeuren maar we zijn ervan overtuigd een waardevolle bijdrage te hebben geleverd aan het onderzoeksdomein rond synthetische reconstructie.

Tot slot, individu-gebaseerde modellen bieden eindeloze mogelijkheden maar systematische model-exploratie is vereist om analyses te focussen en parameterschattingen met behulp van referentiedata te optimaliseren. De integratie van realistische sociale contactpatronen, geënt op de synthetische populatie, is essentieel om betrouwbare resultaten te verkrijgen. Het verbeteren van de computationele efficiëntie van een model vergroot de mogelijkheden om strategieën tegen opkomende infectieziekten te evalueren en het beleid voor de toekomst te informeren.

Scientific Resume

Publications In Peer-Reviewed Scientific Journals

- **Willem L**, Stijven S, Tijskens E, Beutels P, Hens N, Broeckhove J. (2015) Optimizing agent-based transmission models for infectious diseases. BMC Bioinformatics, 16:183.
- Goeyvaerts N*, **Willem L***, Van Kerckhove K, Vandendijck Y, Hanquet G, Beutels P, Hens N. (2015) Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-specific influenza-like illness incidence, Epidemics, 13, p. 1-9.
- Bilcke J, Chapman R, Atchison C, Cromer D, Johnson H, **Willem L**, Cox M, Edmunds WJ, Jit M. (2015) Quantifying parameter and structural uncertainty of dynamic disease transmission models using MCMC: an application to rotavirus vaccination in England and Wales. Medical Decision Making, 0272989X14566013.
- **Willem L***, Stijven S*,Vladislavleva E, Broeckhove J, Beutels P, Hens N. (2014) Active learning to understand infectious disease models and improve policy making. PLoS Computational Biology 10, e1003563.
- Ogunjimi B, Buntinx F, Bartholomeeusen S, Terpstra I, De Haes I, **Willem L**, Elli S, Bilcke J, Van Damme P, Coenen S, Beutels P. (2014) Herpes zoster is associated with herpes simplex and other infections in under 60 year-olds. The Journal of Infection 70:2, p. 171-177.
- **Willem L**, Van Kerckhove K, Chao DL, Hens N, Beutels P. (2012) A nice day for an infection? Weather conditions and social contact patterns relevant to influenza transmission. PLoS ONE 7:11, e48695.
- Bilcke J, Marais C, Ogunjimi B, **Willem L**, Hens N, Beutels P. (2012) Cost-effectiveness of vaccination against herpes zoster in adults aged over 60 years in Belgium. Vaccine 30:3, p. 675-684.

*Both authors contributed equally.

Submitted Manuscripts

- Kifle YW, Goeyvaerts N, Van Kerckhove K, **Willem L**, Faes C, Leirs H, Hens H, Beutels P. Animal ownership and touching enrich the context of social contacts relevant to the spread of human infectious diseases. *Under revision by PLoS ONE*.
- Ogunjimi B, **Willem L**, Beutels P, Hens N. Integrating between-host transmission and within-host immunological response dynamics to estimate the impact of chickenpox vaccination on herpes zoster. *Under revision by eLIFE*.

Reports with Peer-Review

- Beutels P, Vandendijck Y, **Willem L**, Goeyvaerts N, Blommaert A, Van Kerckhove K, Bilcke J, Hanquet G, Neels P, Thiry N, Liesenborgs J, Hens N. (2013) Seasonal influenza vaccination: prioritizing children or other target groups? Part 2: cost-effectiveness analysis. Belgian Health Care Knowledge Centre Reports 204, 247 p.

Publications In Scientific Journals Without Peer-Review

- **Willem L**, Stijven S, Niel H, Vladislavleva E, Broeckhove J, Beutels P. (2015) Individual-based models: powerful or power struggle? Communications in Agricultural and Applied Biological Sciences 80, 97-102.

Oral Presentations

- **Willem L**, Stijven S, Niel H, Vladislavleva E, Broeckhove J, Beutels P. Agent-based models: powerful or power struggle? SIMID Workshop on infectious disease modeling and economic evaluation of vaccines, Antwerp, Belgium, 28-29 April, 2015.
- **Willem L**. Individu-gebaseerde simulaties van griepverspreiding in Vlaanderen. Masterclass Science Communication 2015. University of Antwerp, Antwerp, Belgium, 12 February, 2015.
- **Willem L**, Stijven S, Niel H, Vladislavleva E, Broeckhove J, Beutels P. Individual-based models: powerful or power struggle? 20th National

Symposium for Applied Biological Sciences, Louvain-la-Neuve, Belgium, 30 January, 2015.

- **Willem L**, Goeyvaerts N, Potter G, Van Kerckhove K, Beutels P, Hens N. Within-household contact networks and their implications for epidemic modeling. Contagion '14, Satellite Meeting of the European Conference of Complex systems '15, Lucca, Italy, 25 September, 2014.
- **Willem L**. Active learning: exploring infectious disease models and improve policy making. Research Day GGW University of Antwerp, Antwerp, Belgium, 25 April, 2014.
- **Willem L**, Stijven S, Vladislavleva E. Symbolic regression for modeling epidemiological systems: Applications. SIMID Workshop on infectious disease modeling and economic evaluation of vaccines, Hasselt, Belgium, 25-27 April, 2012.

Poster Presentations

- **Willem L**, Stijven S, Vladislavleva E, Broeckhove J, Beutels P, Hens N. Exploring infectious disease models through active learning to improve policy making. Epidemics, Amsterdam, The Netherlands, 19-22 November, 2013.
- **Willem L***, Van Kerckhove K*, Chao DL, Hens N, Beutels P. The influence of weather conditions on social contact patterns relevant for the spread of influenza. Epidemics, Amsterdam, The Netherlands, 19-22 November, 2013.

*Both authors contributed equally.

Awards

- Runner-up at the "Masterclass Science Communication 2015" organized by the University of Antwerp, Antwerp, Belgium, 12 February, 2015.
- "Best Presentation Award" for "Individual-based models: powerful or power struggle?" at the 20th National Symposium for Applied Biological Sciences, Louvain-la-Neuve, Belgium, 30 January, 2015.
- "Best Oral Presentation" for "Active learning: exploring infectious disease models and improve policy making" at the Research Day GGW University of Antwerp, Antwerp, Belgium, 25 April, 2014.
- "Outstanding Poster Presentation" for "Exploring infectious disease models through active learning to improve policy making" at Epidemics, Amsterdam, The Netherlands, 19-22 November, 2013.

Press Coverage

- "Kinderen jaarlijks vaccineren?" De Standaard, 12 July, 2013.
- "Geef alle kinderen een griepvaccin. Griepvaccin bij kinderen spaart levens en kosten" Nieuwsblad, 12 July, 2013.
- "Nog een draagvlak voor griepvaccin vinden." De Redactie, 12 July, 2013.
- "Winter bugs: A yearly battle for dominance." BBC news, 06 January, 2013.
- "Koud en droog weer? Sneller ziek." EOS, 26 December, 2012.
- "Op een koude, droge dag worden we sneller ziek." De Standaard, 26 December, 2012.

Curriculum Vitae

Lander Willem was born in Antwerp, Belgium, on the 21^{th} of July 1987 and studied Bioscience-Engineering at the KU Leuven (2005-2010). He graduated cum laude in 2010 with a growing interest in the interdisciplinary field of biology and computer science. His master thesis elaborated on protein identification by combining extensive gene expression and sequencing datasets with wet lab validation. This research contributed to several PhD-projects and a paper with already 17 citations. In the philosophy of engaging in interdisciplinary research, he started a PhD on modeling infectious diseases in 2010 at the University of Antwerp at the Centre of Health Economics Research & Modeling Infectious Diseases (CHERMID) at the Faculty of Medicine in collaboration with the Computational Modeling and Programming group at the Faculty of Science. His work focused on agent-based modeling in the field of public health and infectious disease transmission: model exploration, parameter estimation, social contact patterns, computational efficiency and synthetic populations. To improve the statistical know-how in his research, he initiated a joint-PhD in 2013 with the Centre for Statistics at Hasselt University. He studied mainly influenza transmission models and was also successfully involved in other projects at CHERMID to improve agent-based modeling and computational efficiency. His research has been published in peer-reviewed journals (7 publications of which 4 as main author, and 2 publications submitted). The PLoS ONE paper on the association between social contact patterns and weather conditions was referenced in a BBC News article and had impact on succeeding research worldwide. Additionally, he co-authored a peer-reviewed Government Report and gave 6 oral and 2 poster presentations at national and international meetings and conferences. He received 4 awards for his contributions with as highlight the "Outstanding Poster Award" out of 278 posters at the international Epidemics^3 conference in Amsterdam in 2013.

Dankwoord

Dit doctoraat is tot stand gekomen in, met voorsprong, de 5 meest intense jaren van mijn leven. Uit verschillende hoeken heb ik steun gekregen om deze beproeving tot een goed einde te brengen. Ik zal mijn eerste werkdag nooit vergeten. Als een gek raasde er door mijn hoofd "gedaan met de vrijheid voor de komende 40 jaar!". Al snel ondervond ik dat het werkleven helemaal niet zo slecht was...

Dit doctoraat zou nooit gelukt zijn zonder de hulp van mijn promotoren! Philippe, hartelijk dank voor de academische en persoonlijke ondersteuning. Je was een ankerpunt tijdens mijn doctoraat. Je breed kennisveld en blijk van vertrouwen hebben ongetwijfeld bijgedragen tot het welslagen van dit werk. Niel, je enthousiasme en inzet blijven me verbazen. Het is een genoegen om met je samen te werken. Je nuchtere kijk en doelgericht oplossend vermogen apprecieer ik ten zeerste. Jan, je passie voor informatica en C++ heb ik gretig overgenomen. Het was pittig om aanvankelijk mee te kunnen maar je bent een prima leermeester. Onze regelmatige meetings en je drive om code clean te krijgen hebben zeker hun vruchten afgeworpen.

Bedankt Katya om me te introduceren en te begeleiden in de wereld van symbolische regressie en me in contact te brengen met dit doctoraatsproject. Samen met Guido, Tine en Janick heb je tijdens mijn studies de interesse in onderzoek aangescherpt zodat ik met de nodige nieuwsgierigheid aan dit doctoraat gestart ben.

Ik heb het genoegen om in een zeer diverse groep van collega's te mogen werken. Velen van hen wil ik bij deze bedanken: Joke, voor het brengen van rust en inspiratie tijdens koffiepauzes en de goede samenwerking. Ik zal de zomer met "dat verdient taart!" niet snel vergeten. Sean, voor de grote hulp bij de computationele kant van mijn doctoraat. Menig uur hebben we samen achter de computer doorgebracht om allerhande programma's en clusters aan de praat te krijgen. Adriaan, voor de boeiende discussies en aangename tijd op de 3 bureau's die we in 5 jaar gedeeld hebben. Nele, voor de nauwe samenwerking voor het KCE rapport en de geslaagde match tussen statistiek en informatica. Benson, voor de interessante gesprekken rond varicella, immuniteit en Apple. Jeroen, voor de filosofische insteek in het (meestal droge) verhaal van vaccinatie en infectieziekten. Kim, voor de vele uren die we samen gewerkt hebben aan de

cleaning van de dagboekjes. Yannick, voor het feilloos overnemen van het KCE project en de Grote GriepMeting insteek. Steven en Eva voor de frisse wind die jullie vanuit Hasselt meebrachten. Thank you Yimer, Stefano, Matthieu en Tony for the interesting discussions and nice afterwork moments. Bert, voor de overtuiging dat code-optimalisatie eenvoudig kan zijn. Jessica, voor het aanreiken van maatschappelijke topics. Mijn dank gaat ook uit naar vele CEV collega's voor de fijne tijd samen. Thank you Dennis Chao for making the FluTE model open source. It has been of great value for this PhD-project!

Vele vrienden en vriendinnen hebben me de afgelopen jaren bijgestaan. Zeer veel mensen ben ik dankbaar voor hun ondersteuning en sportieve uitlaatklep waaronder Katleen, Tom, Laura W, Koen, Geert J, Geert B, Brecht, Sarah, Charlotte, Aline, Ruth, Griet, Marie, Sara, Hannes V, Ellen, Anke, Laura S, Inte, Veronique, Ann-Sophie, Intse, Bart, Nick, Jo, Madeleen, Iwan, Saar, Jos, Tim, Tanja, Hannes VH, Pieter-Jan, Inge, Julie, Sofie, Jan,... Met veel trots en genoegen heb ik dit doctoraat gecombineerd met een rol als vriend/speler/trainer/coach/kapitein/secretaris én caféganger. Speciale dank gaat uit naar Senne om me intens op te vangen wanneer dat nodig was en de vele impulsieve uitstappen die er op volgden. Ook Joris voor de nodige ondersteuning om positief te blijven. Mijn studiegenoten van Leuven liggen me nog steeds nauw aan het hart. Ik ben blij dat we elkaar nog steeds zien en (te) zotte weekends inplannen. Bedankt Nico, Tim, Yleni, Stijn, Jan, Joaquin, Steven, Philip, Eline, Kristien, Anouk, Bernd,... !

Mijn ouders en broers hebben een speciale rol gespeeld tijdens het verloop van dit doctoraat. Naast de theoretische aspecten van "de medische wereld" op het werk is de praktische kant nooit veraf geweest. Door dik en dun hebben we elkaar gesteund. Het heeft ons dichter naar elkaar gebracht en jammer genoeg ook uit elkaar gehaald. Dankuwel mama en papa voor de enorme steun de afgelopen jaren, het ga jullie goed! Dankuwel Floris en Korneel om mee "de Jackies" te vormen en samen sterk te staan! Dankuwel Opa en Maria, tantes, nonkels, neven, nichten en Justine voor de steun aan elk van ons! Bedankt Marc, Brigitte, Kristof, Carolien, Kris, Sara, Jeroen, Wouter en Sophie om me zo warm te ontvangen in jullie midden!

Tot slot wil ik graag Laura bedanken voor haar aangenaam gezelschap en liefdevolle toets in mijn leven als "ondernemende doctoraatsstudent". Je steun en daadkracht zijn me enorm waardevol. Dat er nog vele jaren mogen volgen!

Lander
juni 2015

Bibliography

1. Taubenberger JK, Morens DM (2010). Influenza: The once and future pandemic. Pub Health Rep. 125(3):16–26.

2. Beutels P, Edmunds W, Smith R (2008). Partially wrong? Partial equilibrium and the economic analysis of public health emergencies of international concern. Health Econ. 17(11):1317–1322.

3. Beutels P, Jia N, Zhou QY, Smith R, Cao WC, De Vlas SJ (2009). The economic impact of SARS in Beijing, China. Trop Med Int Health. 14(s1):85–91.

4. Keogh-Brown MR, Smith RD, Edmunds JW, Beutels P (2010). The macroeconomic impact of pandemic influenza: estimates from models of the United Kingdom, France, Belgium and The Netherlands. Eur J Health Econ. 11(6):543–554.

5. Keogh-Brown MR, Wren-Lewis S, Edmunds WJ, Beutels P, Smith RD (2010). The possible macroeconomic impact on the UK of an influenza pandemic. Health Econ. 19(11):1345–1360.

6. Neuzil KM, Zhu Y, Griffin MR, Edwards KM, Thompson JM, Tollefson SJ, Wright PF (2002). Burden of interpandemic influenza in children younger than 5 years: A 25-year prospective study. J Infect Dis. 185(2):147–152.

7. Poehling KA, Edwards KM, Weinberg GA, Szilagyi P, Staat MA, Iwane MK, Bridges CB, Grijalva CG, Zhu Y, Bernstein DI, et al. (2006). The underrecognized burden of influenza in young children. N Eng J Med. 355(1):31–40.

8. Nair H, Brooks WA, Katz M, Roca A, Berkley JA, Madhi SA, Simmerman JM, Gordon A, Sato M, Howie S, Krishnan A, Ope M, Lindblade KA, Carosone-Link P, Lucero M, et al. (2011). Global burden of respiratory infections due to seasonal influenza in young children: A systematic review and meta-analysis. Lancet Infect Dis. 378(9807):1917–1930.

9. Chen R, Holmes EC (2006). Avian influenza virus exhibits rapid evolutionary dynamics. Mol Biol Evol. 23(12):2336–2341.

10. Garnett G, Cousens S, Hallet T, Steketee R, Walker N (2011). Mathematical models in the evaluation of health programmes. Lancet. 378(9790):515–525.

11. Longini Jr IM, Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, Cummings DAT, Halloran ME (2005). Containing pandemic influenza at the source. Science. 309(5737):1083–1087.

12. Eubank S, Kumar VA, Marathe MV, Srinivasan A, Wang N (2006). Structure of social contact networks and their impact on epidemics. DIMACS Ser Discrete Math Theory Comput Science. 70:181.

13. Ferguson NM, Cummings DAT, Fraser C, Cajka JC, Cooley PC, Burke DS (2006). Strategies for mitigating an influenza pandemic. Nature. 442(7101):448–452.

14. Chao DL, Halloran ME, Obenchain VJ, Longini IM (2010). FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Comput Biol. 6(1):e1000656.

15. Grefenstette JJ, Brown ST, Rosenfeld R, DePasse J, Stone NT, Cooley PC, Wheaton WD, Fyshe A, Galloway DD, Sriram A, Guclu H, Abraham T, Burke DS (2013). FRED (A Framework for Reconstructing Epidemic Dynamics): An open-source software system for modeling infectious diseases and control strategies using census-based populations. BMC Public Health. 13(1):940.

16. Wallinga J, Teunis P, Kretzschmar M (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. Am J Epidemiol. 164(10):936–944.

17. Willem L, Stijven S, Vladislavleva E, Broeckhove J, Beutels P, Hens N (2014). Active learning to understand infectious disease models and improve policy making. PLoS Comput Biol. 10(4):e1003563.

18. Longini Jr IM, Halloran ME, Nizam A, Yang Y (2004). Containing pandemic influenza with antiviral agents. Am J Epidemiol. 159(7):623–633.

19. Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, Toroczkai Z, Wang N (2004). Modelling disease outbreaks in realistic urban social networks. Nature. 429(6988):180–184.

20. Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature. 437(7056):209–214.

21. Germann TC, Kadau K, Longini Jr IM, Macken CA (2006). Mitigation strategies for pandemic influenza in the United States. Proc Natl Acad Sci. 103(15):5935–5940.

22. Colizza V, Barrat A, Barthelemy M, Valleron AJ, Vespignani A (2007). Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. PLoS Med. 4(1):e13.

23. Vynnycky E, Pitman R, Siddiqui R, Gay N, Edmunds WJ (2008). Estimating the impact of childhood influenza vaccination programmes in England and Wales. Vaccine. 26(41):5321–5330.

24. Samsuzzoha M, Singh M, Lucy D (2012). Uncertainty and sensitivity analysis of the basic reproduction number of a vaccinated epidemic model of influenza. Appl Math Model. 37(3):903–915.

25. Milne GJ, Halder N, Kelso JK (2013). The cost effectiveness of pandemic influenza interventions: A pandemic severity based analysis. PLoS ONE. 8(4):e61504.

26. Bilcke J, Beutels P, Brisson M, Jit M (2011). Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: A practical guide. Med Decis Making. 31(4):675–692.

27. Seaholm SK, Ackerman E, Wu SC (1988). Latin hypercube sampling and the sensitivity analysis of a Monte Carlo epidemic model. Int J Biomed Comput. 23(1):97–112.

28. Ma J, Ackerman E, Yang J (1993). Parameter sensitivity of a model or viral epidemics simulated with Monte Carlo techniques. I. ILLNESS ATTACK RATES. Int J Biomed Comput. 32:237—253.

29. Okaïs C, Roche S, Kürzinger ML, Riche B, Bricout H, Derrough T, Simondon F, Ecochard R (2010). Methodology of the sensitivity analysis used for modeling an infectious disease. Vaccine. 28(51):8132–8140.

30. Smits G, Kotanchek M (2005). Pareto-front exploitation in symbolic regression. Genetic programming theory and practice II. 8:283–299.

31. Vladislavleva E (2008). Model-based problem solving through symbolic regression via Pareto genetic programming. Tilburg, The Netherlands: Tilburg University.

32. Keijzer M (2003). Improving symbolic regression with interval arithmetic and linear scaling. Lecture Notes Comput Sci. 2610:70–82.

33. Poli R, Langdon WWB, McPhee NF (2008). A field guide to genetic programming. Lulu Enterprises UK Limited.

34. McRee RK (2010). Symbolic regression using nearest neighbor indexing. In: Proc Genet Evol Comput Conf. GECCO '10. Portland, Oregon, USA p. 1983–1990.

35. Stijven S, Minnebo W, Vladislavleva K (2011). Separating the wheat from the chaff: On feature selection and feature importance in regression random forests and symbolic regression. In: Proc Genet Evol Comput Conf. GECCO '11. Dublin, Ireland p. 623–630.

36. Crombecq K, Gorissen D, Deschrijver D, Dhaene T (2011). A novel hybrid sequential design strategy for global surrogate modeling of computer experiments. SIAM J Sci Comput. 33(4):1948–1974.

37. Gorissen D, Couckuyt I, Demeester P, Dhaene T (2010). A surrogate modeling and adaptive sampling toolbox for computer based design. J Mach Learn Res. 11:2051–2055.

38. Hendrickx W, Dhaene T (2005). Sequential design and rational metamodelling. In: Proc Winter Simul Conf. WSC '05. Orlando, Florida, USA p. 290–298.

39. Yu K, Bi J, Tresp V (2006). Active learning via transductive experimental design. In: Proc Int Conf Mach Learn. ICML '06. Pittsburgh, Pennsylvania, USA p. 1081–1088.

40. Zuluaga M, Krause A, Sergent G, Püschel M (2013). Active learning for multi-objective optimization. In: Proc Int Conf Mach Learn. ICML '13. Atlanta, Georgia, USA p. 462–470.

41. Tong S, Koller D (2001). Support vector machine active learning with applications to text classification. J Mach Learn Res. 2:45–66.

42. van Hoek AJ, Melegaro A, Gay N, Bilcke J, Edmunds WJ (2012). The cost-effectiveness of varicella and combined varicella and herpes zoster vaccination programmes in the United Kingdom. Vaccine. 30:1225–1234.

43. Ogunjimi B, Van Damme P, Beutels P (2013). Herpes zoster risk reduction through exposure to chickenpox patients: A systematic multidisciplinary review. PLoS ONE. 8(6):e66485.

44. Thiry N, Beutels P, Van Damme P, Van Doorslaer E (2003). Economic evaluations of varicella vaccination programmes: A review of the literature. Pharmacoecon. 21(1):13–38.

45. Rozenbaum MH, van Hoek AJ, Vegter S, Postma MJ (2008). Costeffectiveness of varicella vaccination programs: An update of the literature. Expert Rev Vaccines. 7(6):753–782.

46. Schmid DS, Jumaan AO (2010). Impact of varicella vaccine on varicella-zoster virus dynamics. Clin Microbiol Rev. 23(1):202–217.

47. van Hoek A, Melegaro A, Zagheni E, Edmunds W, Gay N (2011). Modelling the impact of a combined varicella and zoster vaccination programme on the epidemiology of varicella zoster virus in England. Vaccine. 29:2411–2420.

48. Bilcke J, Van Hoek AJ, Hens N, Aerts M, Edmunds W, Beutels P (2010). Cost-effectiveness of varicella and zoster vaccination in England&Wales: Importance measures for correlated parameters. Procedia Soc Behav Sci. 2(6):7611–7612.

49. Santner TJ, Williams BJ, Notz WI (2003). The design and analysis of computer experiments. Springer.

50. Husslage B, Rennen G, Van Dam ER, Den Hertog D (2006). Space-filling Latin hypercube designs for computer experiments. Tilburg University.

51. Crombecq K, De Tommasi L, Gorissen D, Dhaene T (2009). A novel sequential design strategy for global surrogate modeling. In: Proc Winter Simul Conf. WSC '09. Austin, Texas, USA p. 731–742.

52. Crombecq K (2011). Surrogate modelling of computer experiments with sequential experimental design. Antwerp, Belgium: University of Antwerp.

53. Koza JR (1992). Genetic programming: On the programming of computers by means of natural selection. vol. 1. MIT press.

54. Evolved Analytics LLC (2011). DataModeler Release 8.0 [Software]. Evolved Analytics LLC. Available from: `www.evolved-analytics.com`.

55. Staelens N, Deschrijver D, Vladislavleva E, Vermeulen B, Dhaene T, Demeester P (2012). Constructing a no-reference H. 264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression. IEEE Trans Circuits Syst Video Technol. 23(8):1322–1333.

56. Halloran ME, Ferguson NM, Eubank S, Longini IM, Cummings DAT, Lewis B, Xu S, Fraser C, Vullikanti A, Germann TC, Wagener D, Beckman R, Kadau K, Barrett C, Macken CA, et al. (2008). Modeling targeted layered containment of an influenza pandemic in the United States. Proc Natl Acad Sci. 105(12):4639–4644.

57. Andradóttir S, Chiu W, Goldsman D, Lee ML, Tsui KL, Sander B, Fisman DN, Nizam A (2011). Reactive strategies for containing developing outbreaks of pandemic influenza. BMC Public Health. 11(1):S1.

58. Piedra PA, Gaglani MJ, Kozinetz CA, Herschler G, Riggs M, Griffith M, Fewlass C, Watts M, Hessel C, Cordova GW J (2005). Herd immunity in adults against influenza-related illnesses with use of the trivalent-live attenuated influenza vaccine (CAIV-T) in children. Vaccine. 23(13):1540–1548.

59. Bilcke J, van Hoek AJ, Beutels P (2013). Childhood varicella-zoster virus vaccination in Belgium: Cost-effective only in the long run or without exogenous boosting? Hum Vaccin Immunother. 9(4):812–822.

60. Epstein JM, Goedecke DM, Yu F, Morris RJ, Wagener DK, Bobashev GV (2007). Controlling pandemic flu: The value of international air travel restrictions. PLoS ONE. 2(5):e401.

61. Goeyvaerts N, Willem L, Van Kerckhove K, Vandendijck Y, Hanquet G, Beutels P, Hens N (2015). Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-specific influenza-like illness incidence. Epidemics. 13:1–9.

62. Beutels P, Vandendijck Y, Willem L, Goeyvaerts N, Blommaert A, Van Kerckhove K, Bilcke J, Hanquet G, Neels P, Thiry N, Liesenborgs J, Hens N. Seasonal influenza vaccination: Prioritizing children or other target groups? Part II: Cost-effectiveness analysis. Health Technology Assessment (HTA) Brussels: Belgian Health Care Knowledge Centre (KCE) (2013).

63. Glasser J, Taneri D, Feng Z, Chuang JH, Tull P, Thompson W, Mason McCauley M, Alexander J (2010). Evaluation of targeted influenza vaccination strategies via population modeling. PLoS ONE. 5(9):e12777.

64. Finkenstädt BF, Morton A, Rand DA (2005). Modelling antigenic drift in weekly flu incidence. Stat Med. 24(22):3447–3461.

65. Fuhrmann C (2010). The effects of weather and climate on the seasonality of influenza: What we know and what we need to know. Geography Compass. 4(7):718–730.

66. Shaman J, Kohn M (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. Proc of the Natl Acad Sci USA. 106(9):3243–3248.

67. Willem L, Van Kerckhove K, Chao DL, Hens N, Beutels P (2012). A nice day for an infection? Weather conditions and social contact patterns relevant to influenza transmission. PLoS ONE. 7(11):e48695.

68. Ballesteros S, Vergu E, Cazelles B (2009). Influenza A gradual and epochal evolution: Insights from simple models. PLoS ONE. 4(10):e7426.

69. Andreasen V (2003). Dynamics of annual influenza A epidemics with immuno-selection. J Math Biol. 46(6):504–536.

70. Prosper O, Saucedo O, Thompson D, Torres-Garcia G, Wang X, Castillo-Chavez C (2011). Modeling control strategies for concurrent epidemics of seasonal and pandemic H1N1 influenza. Math Biosci Eng. 8(1):141–170.

71. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, Massari M, Salmaso S, Tomba GS, Wallinga J, Heijne J, Sadkowska-Todys M, Rosinska M, Edmunds WJ (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 5(3):e74.

72. Goeyvaerts N, Hens N, Ogunjimi B, Aerts M, Shkedy Z, Van Damme P, Beutels P (2010). Estimating infectious disease parameters from data on social contacts and serological status. J R Stat Soc Ser C Appl Stat. 59(2):255–277.

73. Kretzschmar M, Teunis PF, Pebody RG (2010). Incidence and reproduction numbers of pertussis: estimates from serological and social contact data in five European countries. PLoS Med. 7(6):e1000291.

74. Ogunjimi B, Hens N, Goeyvaerts N, Aerts M, Van Damme P, Beutels P (2009). Using empirical social contact data to model person to person infectious disease transmission: an illustration for varicella. Math Biosci. 218(2):80–87.

75. Vandendijck Y, Faes C, Hens N (2013). Eight Years of the Great Influenza Survey to Monitor Influenza-Like Illness in Flanders. PLoS ONE. 8(5):e64156.

76. Hsieh YH (2010). Age groups and spread of influenza: Implications for vaccination strategy. BMC Infect Dis. 10:106.

77. Poletti P, Ajelli M, Merler S (2011). The effect of risk perception on the 2009 H1N1 pandemic influenza dynamics. PLoS ONE. 6(2):e16460.

78. Pitman RJ, White LJ, Sculpher M (2012). Estimating the clinical impact of introducing paediatric influenza vaccination in England and Wales. Vaccine. 30(6):1208–1224.

79. Pitman RJ, Nagy LD, Sculpher MJ (2013). Cost-effectiveness of childhood influenza vaccination in England and Wales: Results from a dynamic transmission model. Vaccine. 31(6):927–942.

80. Baguelin M, Flasche S, Camacho A, Demiris N, Miller E, Edmunds WJ (2013). Assessing optimal target populations for influenza vaccination programmes: An evidence synthesis and modelling study. PLoS Med. 10(10):e1001527.

81. Hanquet G, Jonckheer P, Vlayen J, Vrijens F, Thiry N, Beutels P. Seasonal influenza vaccination: Priority target groups - Part I. Good Clinical Practice (GCP). Brussels: Belgian Health Care Knowledge Centre (KCE) (2011).

82. Bollaerts K, Antoine J, Van Casteren V, Ducoffre G, Hens N, Quoilin S (2013). Contribution of respiratory pathogens to influenza-like illness consultations. Epidemiol Infect. 141(10):2196–2204.

83. Eurostat (2009). Population table for Belgium [Online Database]. Eurostat. Available from: `http://epp.eurostat.ec.europa.eu/`.

84. Vynnycky E, White RG (2010). An Introduction to Infectious Disease Modelling. Oxford University Press, USA.

85. Goeyvaerts N, Hens N, Aerts M, Beutels P (2011). Model structure analysis to estimate basic immunological processes and maternal risk for parvovirus B19. Biostatistics. 12(2):283–302.

86. Melegaro A, Jit M, Gay N, Zagheni E, Edmunds WJ (2011). What types of contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns. Epidemics. 3(3-4):143–151.

87. Earn D J D LSA Dushoff J (2002). Ecology and evolution of the flu. Trends Ecol Evol. 17(7):334–340.

88. Kissling E, Valenciano M, Larrauri A, Oroszi B, Cohen JM, Nunes B, Pitigoi D, Rizzo C, Rebolledo J, Paradowska-Stankiewicz I, Jimenez-Jorge S, Horvath JK, Daviaud I, Guiomar R, Necula G, et al. (2013). Low and decreasing vaccine effectiveness against influenza A(H3) in 2011/12 among vaccination target groups in Europe: Results from the I-MOVE multicentre case-control study. Euro Surveill. 18(5).

89. Skowronski DM, Tweed SA, De Serres G (2008). Rapid decline of influenza vaccine-induced antibody in the elderly: Is it real, or is it relevant? J Infect Dis. 197(4):490–502.

90. European Commission (2010). European public assessment reports. Summary of product characteristics: FLUENZ nasal spray suspension. [Report]. European Commission. Available from: `http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/001101/WC500103711.pdf`.

91. Basta NE, Halloran ME, Matrajt L, Longini Jr IM (2008). Estimating influenza vaccine efficacy from challenge and community-based study data. Am J Epidemiol. 168(12):1343–1352.

92. Halloran ME, Longini IM, Struchiner CJ (2010). Design and analysis of Vaccine Studies. New York: Springer.

93. Vynnycky E, Edmunds WJ (2008). Analyses of the 1957 (Asian) influenza pandemic in the United Kingdom and the impact of school closures. Epidemiol Infect. 136(2):166–179.

94. Carrat F, Vergu E, Ferguson NM, Lemaitre M, Cauchemez S, Leach S, Valleron AJ (2008). Time lines of infection and disease in human influenza: A review of volunteer challenge studies. Am J Epidemiol. 167(7):775–785.

95. Cauchemez S, Carrat F, Viboud C, Valleron A, Boelle P (2004). A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Stat Med. 23(22):3469–3487.

96. Paget WJ, Meerhoff TJ, Meijer A (2005). Epidemiological and virological assessment of influenza activity in Europe during the 2003-2004 season. Euro Surveill. 10(4):107–111.

97. Jit M, Brisson M (2011). Modelling the epidemiology of infectious diseases for decision analysis: A primer. Pharmacoeconomics. 29(5):371–386.

98. Chowell G MMA, Viboud (2008). Seasonal influenza in the United States, France, and Australia: Transmission and prospects for control. Epidemiol Infect. 136:852–864.

99. Edlund S, Kaufman J, Lessler J, Douglas J, Bromberg M, Kaufman Z, Bassal R, Chodick G, Marom R, Shalev V, Mesika Y, Ram R, Leventhal A (2011). Comparing three basic models for seasonal influenza. Epidemics. 3(3-4):135–142.

100. Weidemann F, Dehnert M, Koch J, Wichmann O, Hhle M (2014). Bayesian parameter inference for dynamic infectious disease modelling: Rotavirus in Germany. Stat Med. 33(9):1580–1599.

101. Bilcke J, Chapman R, Atchison C, Cromer D, Johnson H, Willem L, Edmunds WJ, Jit M (2015). Quantifying parameter and structural uncertainty of dynamic disease transmission models using MCMC: An application to rotavirus vaccination in England & Wales. Med Decis Making. p. pii0272989X14566013.

102. Axelsen JB, Yaari R, Grenfell BT, Stone L (2014). Multiannual forecasting of seasonal influenza dynamics reveals climatic and evolutionary drivers. Proc Natl Acad Sci. 111(26):9538–9542.

103. Paolotti D, Carnahan A, Colizza V, Eames K, Edmunds J, Gomes G, Koppeschaar C, Rehn M, Smallenburg R, Turbelin C, et al. (2014). Web-based participatory surveillance of infectious diseases: The Influenzanet participatory surveillance experience. Clin Microbiol Infect. 20(1):17–21.

104. Beutels P, Edmunds WJ, Antonanzas F, De Wit GA, Evans D, Feilden R, Fendrick AM, Ginsberg GM, Glick HA, Mast E, Pechevis M, Van Doorslaer EKA, van Hout BA (2002). Economic evaluation of vaccination programmes - A consensus statement focusing on viral hepatitis. Pharmacoeconomics. 20(1):1–7.

105. Brotherton J, McIntyre P, Puech M, Wang H, Gidding H, Hull B, Lawrence G, MacIntyre R, Wood N, Armstrong D (2004). Vaccine preventable diseases and vaccination coverage in Australia, 2001 to 2002. Commun Dis Intell. 28(Suppl 2):S41–S45.

106. Thompson WW, Shay DK, Weintraub E, Brammer L, Bridges CB, Cox NJ, Fukuda K (2004). Influenza-associated hospitalizations in the United States. J Am Med Assoc. 292(11):1333–1340.

107. Iskander M, Booy R, Lambert S (2007). The burden of influenza in children. Curr Opin Infect Dis. 20(3):259–263.

108. Lofgren E, Fefferman NH, Naumov YN, Gorski J, Naumova EN (2007). Influenza seasonality: Underlying causes and modeling theories. J Virol. 81(11):5429–5436.

109. Lowen AC, Mubareka S, Steel J, Palese P (2007). Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathog. 3(10):e151.

110. Flasche S, Hens N, Boëlle PY, Mossong J, van Ballegooijen WM, Nunes B, Rizzo C, Popovici F, Santa-Olalla P, Hrubá F, Parmakova K, Baguelin M, van Hoek AJ, Desenclos J, Bernillon P, et al. (2011). Different transmission patterns in the early stages of the influenza A(H1N1)v pandemic: A comparative analysis of 12 European countries. Epidemics. 3(2):125–133.

111. Tamerius J, Nelson MI, Zhou SZ, Viboud C, Miller MA, Alonso WJ (2011). Global influenza seasonality: Reconciling patterns across temperate and tropical regions. Environ Health Perspect. 119:439–445.

112. Steel J, Palese P, Lowen AC (2011). Transmission of a 2009 pandemic influenza virus shows a sensitivity to temperature and humidity similar to that of an H3N2 seasonal strain. J Virol. 85(3):1400–1402.

113. Davis RE, Rossier CE, Enfield KB (2012). The impact of weather on influenza and pneumonia mortality in New York City, 1975–2002: A Retrospective Study. PLoS ONE. 7(3):e34091.

114. Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M (2010). Absolute humidity and the seasonal onset of influenza in the continental United States. PLoS Biol. 2(8):e1000316.

115. van Noort SP, Águas R, Ballesteros S, Gomes MGM (2012). The role of weather on the relation between influenza and influenza-like illness. J Theor Biol. 298:131–137.

116. Alonso WJ, Viboud C, Simonsen L, Hirano EW, Daufenbach LZ, Miller MA (2007). Seasonality of influenza in Brazil: A traveling wave from the Amazon to the subtropics. Am J Epidemiol. 165(12):1434–1442.

117. Eccles R (2002). An explanation for the seasonality of acute upper respiratory tract viral infections. Acta Otolaryngol. 122(2):183–191.

118. Cannell JJ, Zasloff M, Garland CF, Scragg R, Giovannucci E (2008). On the epidemiology of influenza. Virol J. 5(1):29.

119. Polozov IV, Bezrukov L, Gawrisch K, Zimmerberg J (2008). Progressive ordering with decreasing temperature of the phospholipids of influenza virus. Nat Chem Biol. 4(4):248–255.

120. Mikolajczyk RT, Akmatov MK, Rastin S, Kretzschmar M (2008). Social contacts of school children and the transmission of respiratory-spread pathogens. Epidemiol Infect. 136:813–822.

121. Cauchemez S, Valleron AJ, Boëlle PY, Flahault A, Ferguson NM (2008). Estimating the impact of school closure on influenza transmission from Sentinel data. Nature. 452(7188):750–754.

122. Hens N, Ayele GM, Goeyvaerts N, Aerts M, Mossong J, Edmunds JW, Beutels P (2009a). Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries. BMC Infect Dis. 9(1):187.

123. Cauchemez S, Donnelly CA, Reed C, Ghani AC, Fraser C, Kent CK, Finelli L, Ferguson NM (2009). Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. N Eng J Med. 361(27):2619–2627.

124. Eames KTD, Tilston NL, Brooks-Pollock E, Edmunds WJ (2012). Measured dynamic social contact patterns explain the spread of H1N1v influenza. PLoS Comput Biol. 8(3):e1002425.

125. Hens N, Goeyvaerts N, Aerts M, Shkedy Z, Damme PV, Beutels P (2009). Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. BMC Infect Dis. 9(1):5.

126. Monto AS, Koopman JS, Bryan ER (1986). The Tecumseh study of illness. XIV: Occurrence of respiratory viruses, 1976–1981. Am J Epidemiol. 124(3):359–367.

127. Chao DL, Halloran ME, Longini IM (2010). School opening dates predict pandemic influenza A(H1N1) outbreaks in the United States. J Infect Dis. 202(6):877–880.

128. DeStefano F, Haber M, Currivan D, Farris T, Burrus B, Stone-Wiggins B, McCalla A, Guled H, Shih H, Edelson P, Wetterhall S (2011). Factors associated with social contacts in four communities during the 2007–2008 influenza season. Epidemiol Infect. 139:1181–1190.

129. Zagheni E, Billari FC, Manfredi P, Melegaro A, Mossong J, Edmunds WJ (2008). Using time-use data to parameterize models for the spread of close-contact infectious diseases. Am J Epidemiol. 168(9):1082–1090.

130. Del Valle SY, Hyman JM, Hethcote HW, Eubank SG (2007). Mixing patterns between age groups in social networks. Soc Networks. 29(4):539–554.

131. Mccurdy T, Graham SE (2003). Using human activity data in exposure models: Analysis of discriminating factors. J Expo Anal Environ Epidemiol. 13(4):294–317.

132. Graham S, Mccurdy T (2004). Developing meaningful cohorts for human exposure models. J Expo Sci Environ Epidemiol. 14:23–43.

133. National Oceanic and Atmospheric Administration (2010). Global summary of the day [Online Database]. NOAA. Available from: `http://www.ncdc.noaa.gov/`.

134. Wallace JM, Hobbs PV (2006). Atmospheric science: An introductory survey. 2nd ed. New York: Academic press.

135. Skinner CJ, Holt D, Smith TMF (1989). Analysis of complex surveys. Chichester: Wiley.

136. Eurostat (2001). Census data for Belgium [Online Database]. Eurostat. Available from: `http://epp.eurostat.ec.europa.eu/`.

137. Eurostat (2010). Population table for Belgium [Online Database]. Eurostat. Available from: `http://epp.eurostat.ec.europa.eu/`.

138. Diekmann O, Heesterbeek JAP, Metz JAJ (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. J Math Biol. 28(4):365–382.

139. Smieszek T, Fiebig L, Scholz RW (2009). Models of epidemics: When contact repetition and clustering should be included. Theor Biol Med Model. 6(1):11.

140. Willem L, Stijven S, Tijskens E, Beutels P, Hens N, Broeckhove J (2015). Optimizing agent-based transmission models for infectious diseases. BMC Bioinformatics. 16:183.

141. Degli Atti MLC, Merler S, Rizzo C, Ajelli M, Massari M, Manfredi P, Furlanello C, Tomba GS, Iannelli M (2008). Mitigation measures for pandemic influenza in Italy: An individual based model considering different scenarios. PLoS ONE. 3(3):e1790.

142. Das TK, Savachkin AA, Zhu Y (2008). A large-scale simulation model of pandemic influenza outbreaks for development of dynamic mitigation strategies. IIE Transactions. 40(9):893–905.

143. Roche B, Drake JM, Rohani P (2011). An agent-based model to study the epidemiological and evolutionary dynamics of influenza viruses. BMC Bioinformatics. 12(1):87.

144. Laskowski M, Demianyk BC, Witt J, Mukhi SN, Friesen MR, McLeod RD (2011). Agent-based modeling of the spread of influenza-like illness in an emergency department: A simulation study. IEEE Trans Inf Technol Biomed. 15(6):877–889.

145. Aleman DM, Wibisono TG, Schwartz B (2011). A nonhomogeneous agent-based simulation approach to modeling the spread of disease in a pandemic outbreak. Interfaces. 41(3):301–315.

146. Drepper U (2007). What every programmer should know about memory. http://wwwakkadiaorg/drepper/cpumemorypdf. 11.

147. Giles M, Reguly I (2014). Trends in high-performance computing for engineering calculations. Phil Trans R Soc A. 372(2022):20130319.

148. Giles MB, Mudalige GR, Sharif Z, Markall G, Kelly PH (2012). Performance analysis and optimization of the OP2 framework on many-core architectures. Comput J. 55(2):168–180.

149. Sutter H (2005). The free lunch is over: A fundamental turn toward concurrency in software. Dr Dobb s journal. 30(3):202–210.

150. Sutter H, Larus J (2005). Software and the concurrency revolution. Queue. 3(7):54–62.

151. Hayes B (2001). Randomness as a resource. Am Sci. 89(4):300–304.

152. Bauke H, Mertens S (2007). Random numbers for large-scale distributed Monte Carlo simulations. Phys Rev E. 75(6):066701.

153. RTI International (2014). 2010 RTI U.S. Synthetic Population Ver. 1.0 [Online Database]. RTI International. Available from: `http://www.epimodels.org/midas/pubsyntdata1.do`.

154. Wheaton W (2014). 2010 U.S. Synthetic Population Quick Start Guide. RTI International. Available from: `http://www.epimodels.org/midasdocs/SynthPop/2010_synth_pop_ver1_quickstart.pdf`.

155. Chapman B, Jost G, Van Der Pas R (2008). Using OpenMP: Portable shared memory parallel programming. vol. 10. Massachusetts, USA: MIT Press.

156. Bauke H, Brown WE, Fischler M, Kowalkowski J, Paterno M, Knuth DE, Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2011). Tina's random number generator library [Software]. Number Crunch. Available from: `http://numbercrunch.de/trng/`.

157. Google (2014). Google C++ Testing Framework [Software]. Google. Available from: `http://code.google.com/p/googletest`.

158. Smart J (2011). Jenkins The definitive guide. O'Reilly Media, Sebastopol, California.

159. TCLAP (2014). The Templetized C++ Command Line Parser Library [Software]. SourceForge. Available from: `http://www.tclap.sourceforge.net`.

160. Fialho L, Browne J (2014). Framework and modular infrastructure for automation of architectural adaptation and performance optimization for HPC systems. In: Supercomputing. Springer, Cham, Switzerland p. 261–277.

161. Rane A, Browne J (2014). Enhancing performance optimization of multicore/multichip nodes with data structure metrics. ACM Trans Par Comput. 1(1):3.

162. Barthelemy J, Toint PL (2013). Synthetic population generation without a sample. Transp Science. 47(2):266–279.

163. Namazi-Rad MR, Huynh N, Barthelemy J, Perez P (2014). Synthetic population initialization and evolution-agent-based modelling of population aging and household transitions. PRIMA 2014: Prin Pract Multi Agent Syst. p. 182–189.

164. Ball F, Mollison D, Scalia-Tomba G (1997). Epidemics with two levels of mixing. Ann Appl Prob. p. 46–89.

165. Beckman RJ, Baggerly KA, McKay MD (1996). Creating synthetic baseline populations. Transp Res Part A Policy Pract. 30(6):415–429.

166. Ye X, Konduri K, Pendyala RM, Sana B, Waddell P (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In: Ann Meeting Transp Res Board p. 1–24.

167. Gargiulo F, Ternes S, Huet S, Deffuant G (2010). An iterative approach for generating statistically realistic populations of households. PLoS ONE. 5(1):e8828.

168. Fumanelli L, Ajelli M, Manfredi P, Vespignani A, Merler S (2012). Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. PLoS Comput Biol. 8(9):e1002673.

169. Yu Z, Liu J, Zhu X (2015). Inferring a district-based hierarchical structure of social contacts from census data. PLoS ONE. 10(2):e0118085.

170. Lessler J, Edmunds WJ, Halloran ME, Hollingsworth TD, Lloyd AL (2014). Seven challenges for model-driven data collection in experimental and observational studies. Epidemics. p. 3–7.

171. Dietz K, Heesterbeek J (2002). Daniel Bernoulli's epidemiological model revisited. Math Biosci. 180(1):1–21.

172. Ross R (1911). The prevention of malaria. Murray, London.

173. Abbey H (1952). An examination of the Reed-Frost theory of epidemics. Hum Biol. 24(3):201–233.

174. Keeling MJ, Danon L (2009). Mathematical modelling of infectious diseases. Br Med Bull. 92(1):33–42.

175. Eichner M, Schwehm M, Duerr HP, Brockmann SO (2007). The influenza pandemic preparedness planning tool InfluSim. BMC Infect Dis. 7(1):17.

176. Duan W, Fan Z, Zhang P, Guo G, Qiu X (2014). Mathematical and computational approaches to epidemic modeling: a comprehensive review. Front Comput Sci. p. 1–21.

177. Colizza V, Barrat A, Barthélemy M, Vespignani A (2006). The modeling of global epidemics: Stochastic dynamics and predictability. Bull Math Biol. 68(8):1893–1921.

178. Balcan D, Gonçalves B, Hu H, Ramasco JJ, Colizza V, Vespignani A (2010). Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. J Comput Sci. 1(3):132–145.

179. Ajelli M, Gonçalves B, Balcan D, Colizza V, Hu H, Ramasco JJ, Merler S, Vespignani A (2010). Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. BMC Infect Dis. 10(1):190.

180. Broeck WV, Gioannini C, Gonçalves B, Quaggiotto M, Colizza V, Vespignani A (2011). The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. BMC Infect Dis. 11(1):37.

181. Bajardi P, Poletto C, Ramasco JJ, Tizzoni M, Colizza V, Vespignani A (2011). Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. PLoS ONE. 6(1):e16591.

182. Tizzoni M, Bajardi P, Poletto C, Ramasco JJ, Balcan D, Gonçalves B, Perra N, Colizza V, Vespignani A (2012). Real-time numerical forecast of global epidemic spreading: Case study of 2009 A/H1N1pdm. BMC Med. 10(1):165.

183. Elveback L, Fox J, Ackerman E, Langworthy A, Boyd M, Gatewood L (1976). An influenza simulation model for immunization studies. Am J Epidemiol. 103(2):152–162.

184. Barrett C, Beckman R, Berkbigler K, Bisset K, Bush B, Campbell K, Eubank S, Henson K, Hurford J, Kubicek D, et al. (2001). TRANSIMS: Transportation analysis simulation system. Los Alamos National Laboratory Unclassified Report.

185. Mniszewski SM, Del Valle SY, Stroud PD, Riese JM, Sydoriak SJ (2008). EpiSimS simulation of a multi-component strategy for pandemic influenza. In: Proc Spring Sim Multiconf. Society for Computer Simulation International p. 556–563.

186. Basta NE, Chao DL, Halloran ME, Matrajt L, Longini Jr IM (2009). Strategies for pandemic and seasonal influenza vaccination of schoolchildren in the United States. Am J Epidemiol. 170(6):679–686.

187. Sander B, Nizam A, Garrison LP, Postma MJ, Halloran ME, Longini IM (2009). Economic evaluation of influenza pandemic mitigation strategies in the United States using a stochastic microsimulation transmission model. Value Health. 12(2):226–233.

188. Merler S, Ajelli M, Fumanelli L, Vespignani A (2013). Containing the accidental laboratory escape of potential pandemic influenza viruses. BMC Med. 11(1):252.

189. Halder N, Kelso JK, Milne GJ (2014). A model-based economic analysis of pre-pandemic influenza vaccination cost-effectiveness. BMC Infect Dis. 14(1):266.

190. Eichner M, Schwehm M, Hain J, Uphoff H, Salzberger B, Knuf M, Schmidt-Ott R (2014). 4Flu-an individual based simulation tool to study the effects of quadrivalent vaccination on seasonal influenza in Germany. BMC Infect Dis. 14(1):365.

191. Laskowski M, Greer A, Moghadas S, Vespignani A (2014). Antiviral strategies for emerging influenza viruses in remote communities. PLoS ONE. 9(2):e89651.

192. Tolentino SL, Meng F, Delva W (2013). A simulation-based method for efficient resource allocation of combination HIV prevention. In: Proc Int ICST Conf Sim Tools Techn p. 31–40.

193. Marguta R, Parisi A (2015). Impact of human mobility on the periodicities and mechanisms underlying measles dynamics. J R Soc Interface. 12(104):20141317.

194. Merler S, Ajelli M, Fumanelli L, Gomes MF, y Piontti AP, Rossi L, Chao DL, Longini IM, Halloran ME, Vespignani A (2015). Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. Lancet Infect Dis.

195. Guzzetta G, Ajelli M, Yang Z, Merler S, Furlanello C, Kirschner D (2011). Modeling socio-demography to capture tuberculosis transmission dynamics in a low burden setting. J Theor Biol. 289:197–205.

196. Kasaie P, Andrews JR, Kelton WD, Dowdy DW (2014). Timing of tuberculosis transmission and the impact of household contact tracing. An agent-based simulation model. Am J Respir Crit Care Med. 189(7):845–852.

197. Arifin SN, Zhou Y, Davis GJ, Gentile JE, Madey GR, Collins FH (2014). An agent-based model of the population dynamics of Anopheles gambiae. Malar J. 13(1):424.

198. Thavara U, Tawatsin A, Nagao Y (2014). Simulations to compare efficacies of tetravalent dengue vaccines and mosquito vector control. Epidemiol Infect. 142(06):1245–1258.

199. Wang Y, Xue H, Chen Hj, Igusa T (2014). Examining social norm impacts on obesity and eating behaviors among US school children based on agent-based model. BMC Public Health. 14(1):923.

200. Speybroeck N, Van Malderen C, Harper S, Müller B, Devleesschauwer B (2013). Simulation models for socioeconomic inequalities in health: A systematic review. Int J Environ Res Public Health. 10(11):5750–5780.

201. Chisholm RH, Lorenzi T, Lorz A, Larsen AK, de Almeida LN, Escargueil A, Clairambault J (2015). Emergence of drug tolerance in cancer cell populations: An evolutionary outcome of selection, nongenetic instability, and stress-induced adaptation. Cancer Res. 75(6):930–939.

202. Uppal A, Wightman SC, Ganai S, Weichselbaum RR, An G (2014). Investigation of the essential role of platelet-tumor cell interactions in metastasis progression using an agent-based model. Theor Biol Med Model. 11:17.

203. Norton KA, Popel AS (2014). An agent-based model of cancer stem cell initiated avascular tumour growth and metastasis: The effect of seeding frequency and location. J R Soc Interface. 11(100):20140640.

204. An G, Kulkarni S (2015). An agent-based modeling framework linking inflammation and cancer using evolutionary principles: Description of a generative hierarchy for the hallmarks of cancer and developing a bridge between mechanism and epidemiological data. Math Biosci. 260:16–24.

205. Van Kerckhove K, Hens N, Edmunds WJ, Eames KT (2013). The impact of illness on social networks: Implications for transmission and control of influenza. Am J Epidemiol. 178(11):1655–1662.

206. Horby P, Thai PQ, Hens N, Yen NTT, Thoang DD, Linh NM, Huong NT, Alexander N, Edmunds WJ, Duong TN, et al. (2011). Social contact patterns in Vietnam and implications for the control of infectious diseases. PLoS ONE. 6(2):e16965.

207. Chan T, Fu Y, Hwang J (2015). Changing social contact patterns under tropical weather conditions relevant for the spread of infectious diseases. Epidemiol Infect. 143(2):440–451.

208. Grijalva CG, Goeyvaerts N, Verastegui H, Edwards KM, Gil AI, Lanata CF, Hens N, RESPIRA P (2015). A household-based study of contact networks relevant for the spread of infectious diseases in the highlands of Peru. PLoS ONE. 10(3):e0118457.

209. Juanico DEO (2014). Recurrent epidemic cycles driven by intervention in a population of two susceptibility types. J Phys Conf Ser. 490(1):12188.

210. Meyer S, Held L, et al. (2014). Power-law models for infectious disease spread. Ann Appl Stat. 8(3):1612–1639.

211. Goodwin R, Sun S (2014). Early responses to H7N9 in southern Mainland China. BMC Infect Dis. 14(1):8.

212. Jaakkola K, Saukkoriipi A, Jokelainen J, Juvonen R, Kauppila J, Vainio O, Ziegler T, Rönkkö E, Jaakkola JJ, Ikäheimo TM, et al. (2014). Decline in temperature and humidity increases the occurrence of influenza in cold climate. Environ Health. 13:22.

213. Zhang Y, Yan C, Kan H, Cao J, Peng L, Xu J, Wang W (2014). Effect of ambient temperature on emergency department visits in Shanghai, China: a time series study. Environ Health. 13(1):100.

214. Keeling MJ, Eames KT (2005). Networks and epidemic models. J R Soc Interface. 2(4):295–307.

215. Potter GE, Handcock MS, Longini Jr IM, Halloran ME (2011). Estimating within-household contact networks from egocentric data. Ann Appl Stat. 5(3):1816–1838.

216. Potter GE, Hens N (2013). A penalized likelihood approach to estimate within-household contact networks from egocentric data. J R Stat Soc Ser C Appl Stat. 62(4):629–648.

217. Ogunjimi B, Buntinx F, Bartholomeeusen S, Terpstra I, Willem L, Elli S, Bilcke J, Van Damme P, Coenen S, Beutels P, et al. (2015). Herpes zoster is associated with herpes simplex and other infections in under 60 year-olds. J Infect. 70(2):171–177.

218. R Core Team (2014). R: A language and environment for statistical computing [Software]. R Core Team. Available from: `http://www.R-project.org/`.

219. MathWorks Inc (2014). MATLAB version 8.3.0 (R2014a) [Software]. MathWorks Inc. Available from: `www.mathworks.com`.

220. Gilks WR, Richardson S, Spiegelhalter DJ (1996). Introducing Markov Chain Monte Carlo. Markov Chain Monte Carlo Pract. 1:1–19.

APPENDIX A

Supplementary Information Chapter 2

Summary

This appendix contains additional results and a step-by-example of active learning to extend Chapter 2: "Model Exploration". We performed five iterations with the FluTE model and analyzed the input-response data with symbolic regression. Here, we present statistics for all obtained surrogate models and for the optimized high-quality ensemble. Next, all surrogate models from the optimized high-quality model ensemble from the first iteration are listed. Section A.3 contains similar results for the surrogate modeling with the varicella model. Surrogate models can be readily explored, which is illustrated with a visualization tool in Section A.4. Finally, we present a step-by-step example of active learning, implemented in Mathematica.

A.1 Symbolic Regression Analyses FluTE

We performed five iterations with the FluTE model (parameters are listed in Table 2.2 on page 22) and analyzed the input-response data with Pareto-aware symbolic regression (SR). This section contains a summary of the SR results. The SR parameters are presented in Table 2.1 on page 20. We present for each iteration:

1. Descriptive statistics for all obtained surrogate models and for the optimized high-quality ensemble.
2. The obtained surrogate models according complexity and model error and the optimized high-quality ensemble.
3. Surrogate model examples from the optimized high-quality ensemble for the attack rate (AR) and the peak day.

RUN 1

The first run with FluTE contained four transmission parameters and was based on demographic data from Seattle. All surrogate models obtained with SR are presented according to complexity and model error in Figure A.1. We selected models in the knee of the Pareto front to obtain a high-quality model ensemble with limited complexity. This ensemble is optimized using nonlinear techniques to end up with a predictive set, presented in Figure A.1. Descriptive statistics for all surrogate models and for the optimized high-quality model ensemble are given in Table A.1. Table A.6 illustrates surrogate models for the AR and the epidemic peak day, arbitrary chosen from the Pareto front. The complete optimized high-quality model ensemble for the AR is given in section A.2.

Table A.1: Specifications of all surrogate models and the optimized high-quality ensemble from RUN 1. The generation count represents the total number of generations in each independent evolution (average).

Response	Generation Count	Max Model Age	Mean Model Age	Min Model Error	Mean Model Complexity
AR	232	139	32	0.000860588	186
AR (ensemble)	-	122	45	0.000638384	171
Peak Day	227	130	29	0.022570700	140
Peak Day (ensemble)	-	130	38	0.026274200	191

RUN 2

The second FluTE run was again based on demographic data from Seattle and four transmission parameters but the seeding parameter was different. We sampled low seeding values more into detail to include more edge cases from the simulation model and increase stochasticity of the results. All surrogate models obtained with SR are presented according to complexity and model error in Figure A.2. We selected models in the knee of the Pareto front to obtain a high-quality model ensemble with limited complexity. This ensemble is optimized using nonlinear techniques and is presented in Figure A.2. Descriptive statistics for all surrogate models and for the optimized high-quality model ensemble are given in Table A.2. We present two model examples for the AR in Table A.7 to illustrate the trade-off between model error and complexity.

Table A.2: Specifications of all surrogate models and the optimized high-quality ensemble from RUN 2. The generation count represents the total number of generations in each independent evolution (average).

Response	Generation Count	Max Model Age	Mean Model Age	Min Model Error	Mean Model Complexity
AR	231	173	36	0.159324	184
AR (ensemble)	-	165	50	0.158718	104
Peak Day	221	181	41	0.319475	156
Peak Day (ensemble)	-	130	38	0.0262742	191

RUN 3

The third FluTE run was based on demographic data from LA County and four transmission parameters with the low seeding numbers. All surrogate models obtained with SR are presented according to complexity and model error in Figure A.3. We selected models in the knee of the Pareto front to obtain a high-quality model ensemble with limited complexity. This ensemble is optimized using nonlinear techniques to end up with a predictive set, presented in Figure A.3. We performed SR analysis with two time budgets, 1000s and 2400s and obtained similar results. We observe a very small decrease in model error at a high computational cost to find more complex functions. Descriptive statistics for all surrogate models and for the optimized high-quality model from the 1000s and 2400s analysis are given in Table A.3. Next, Table A.8 illustrates surrogate model examples for the AR and epidemic peak day.

Table A.3: Specifications of all surrogate models and the optimized high-quality ensemble from RUN 3. The generation count represents the total number of generations in each independent evolution (average).

Response	Generation Count	Max Model Age	Mean Model Age	Min Model Error	Mean Model Complexity
AR (1000s)	773	542	89	0.177380	245
AR (2400s)	-	1286	171	0.175891	327
AR (1000s, ensemble)	-	401	102	0.175269	198
AR (2400s, ensemble)	-	828	259	0.174229	320
Peak Day	693	355	80	0.412795	247
Peak Day (ensemble)	-	340	113	0.425443	136

RUN 4

The fourth FluTE run was performed with demographic data from Seattle, four transmission parameters with the low seeding numbers and eight vaccination parameters. All surrogate models obtained with SR are presented according to complexity and model error in Figure A.4. We selected models in the knee of the Pareto front to obtain a high-quality model ensemble with limited complexity. This ensemble is optimized using nonlinear techniques to end up with a predictive set, presented in Figure A.4. Descriptive statistics for all surrogate models and for the optimized high-quality model ensemble are given in Table A.4. Table A.9 illustrates surrogate models for the AR and epidemic peak day.

Table A.4: Specifications of all surrogate models and the optimized high-quality ensemble from RUN 4. The generation count represents the total number of generations in each independent evolution (average).

Response	Generation Count	Max Model Age	Mean Model Age	Min Model Error	Mean Model Complexity
AR	416	240	51	0.184299	152
AR (ensemble)	-	203	57	0.184299	164
Peak Day	441	254	57	0.391609	188
Peak Day (ensemble)	-	203	77	0.391609	251

RUN 5

The fifth FluTE run was performed with demographic data from LA, four transmission parameters with the low seeding numbers and six vaccination parameters to model instant reaction strategies. All surrogate models obtained with SR are presented according to complexity and model error in Figure A.5. We selected models in the knee of the Pareto front to obtain a high-quality model ensemble with limited complexity. This ensemble is optimized using nonlinear techniques to end up with a predictive set, presented in Figure A.5. Descriptive statistics for all surrogate models and for the optimized high-quality model ensemble are given in Table A.5. Table A.10 illustrates surrogate models for the AR and the epidemic peak day.

Table A.5: Specifications of all surrogate models and the optimized high-quality ensemble from RUN 5. The generation count represents the total number of generations in each independent evolution (average).

Response	Generation Count	Max Model Age	Mean Model Age	Min Model Error	Mean Model Complexity
AR	763	573	97	0.115576	230
AR (ensemble)	-	570	138	0.105285	229
Peak Day	751	500	92	0.349411	238
Peak Day (ensemble)	-	91	31	0.412062	291

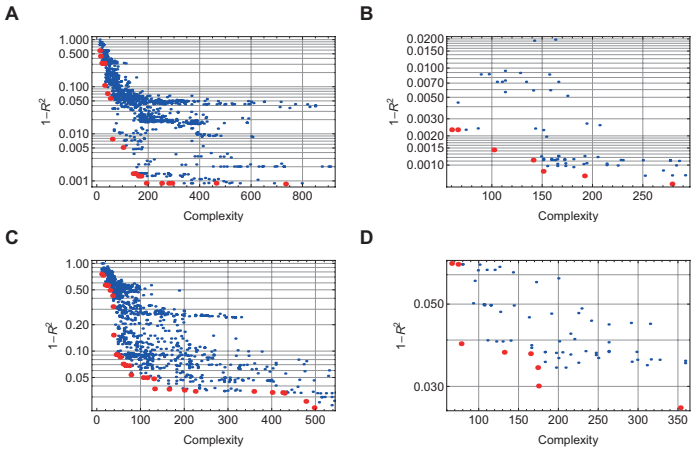


Figure A.1: Model error (1-R^2) and complexity of the surrogate models for the AR and epidemic peak from RUN 1. The model error is presented on a log scale. A: AR, all models - B: AR, optimized high-quality ensemble - C: peak day, all models - D: peak day, optimized high-quality ensemble.

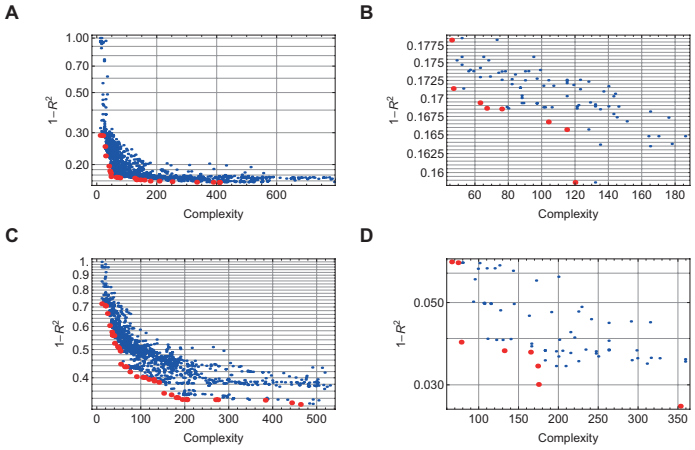


Figure A.2: Model error (1-R^2) and complexity of the surrogate models for the AR and epidemic peak from RUN 2. The model error is presented on a log scale. A: AR, all models - B: AR, optimized high-quality ensemble - C: peak day, all models - D: peak day, optimized high-quality ensemble.

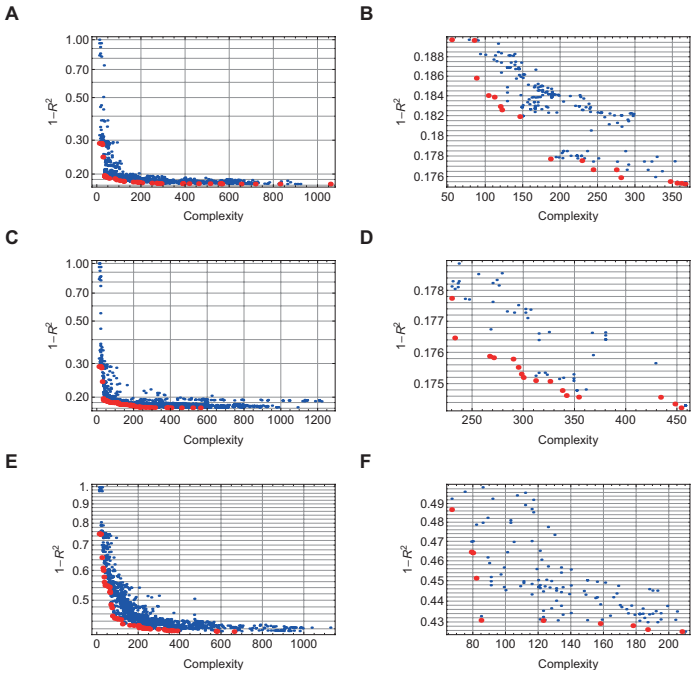


Figure A.3: Model error (1-R^2) and complexity of the surrogate models for the AR and epidemic peak from RUN 3. The model error is presented on a log scale. A: AR, all models after 1000s - B: AR, optimized high-quality ensemble after 1000s - C: AR, all models after 2400s - D: AR, optimized high-quality ensemble after 2400s - E: peak day, all models - F: peak day, optimized high-quality ensemble.

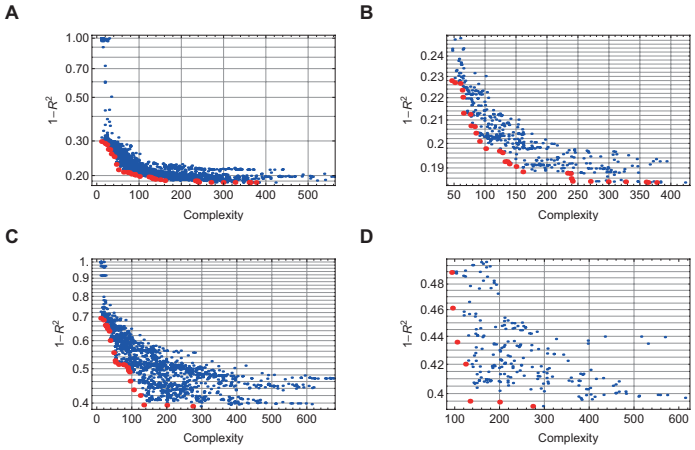


Figure A.4: Model error (1-R^2) and complexity of the surrogate models for the AR and epidemic peak from RUN 4. The model error is presented on a log scale. A: AR, all models - B: AR, optimized high-quality ensemble - C: peak day, all models - D: peak day, optimized high-quality ensemble.

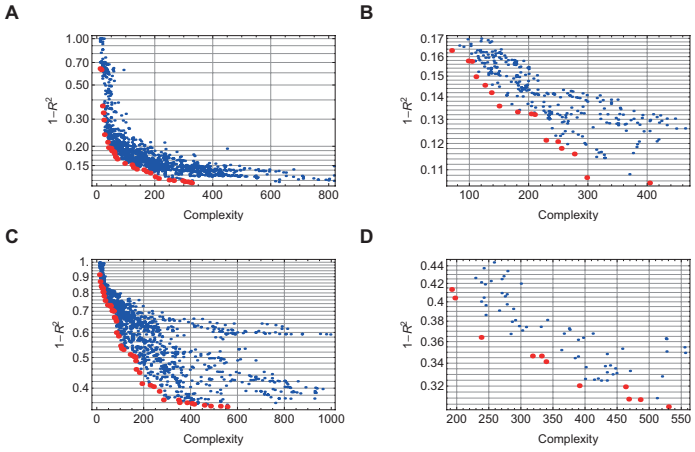


Figure A.5: Model error (1-R^2) and complexity of the surrogate models for the AR and epidemic peak from RUN 5. The model error is presented on a log scale. A: AR, all models - B: AR, optimized high-quality ensemble - C: peak day, all models - D: peak day, optimized high-quality ensemble.

Table A.6: Surrogate model examples from RUN 1. The models are arbitrary chosen from the Pareto front of the optimized high-quality ensemble from RUN 1. Parameters are listed in Table 2.2 on page 22.

Response	Complexity	1-R^2	Function
AR	60	0.002	$\frac{-0.737}{\frac{-2.899}{3.450+\text{SeedInfected}}+\text{SeedInfected*SeedInfectedDaily}\left(6.015\times10^{-4}\right)+\text{R}_0}+0.799$
Peak Day	78	0.039	$\frac{54.305}{\frac{1}{-0.025+\text{SeedInfected}\left(-4.202\times10^{12}\right)}+0.029\sqrt{\text{SeedInfected}}-1.007+\text{R}_0+\text{SeedInfectedDaily}}+1.429$

Table A.7: Surrogate model examples from RUN 2. The models are arbitrary chosen from the Pareto front of the optimized high-quality ensemble from RUN 2. Parameters are listed in Table 2.2 on page 22.

Response	Complexity	1-R^2	Function
AR	63	0.169	$0.931+\frac{-6.549}{\frac{-1.067}{\text{SeedInfected}}+5.591*\text{R}_0+\sqrt[8]{\text{SeedInfected}}+\text{SeedInfectedDaily}}$
AR	120	0.159	$\frac{-2.571*\text{R0}^{-0.339}3.612^{-0.818*\text{R}_0}\left(3.794+\frac{\text{R0}}{\text{SeedInfected}}\right)}{0.004*\text{SeedInfected}*(0.217+\text{SeedInfectedDaily})+5.561+\text{SeedInfectedDaily}}+0.593$
Peak Day	78	0.039	$\frac{54.305}{\frac{1}{-0.025+\text{SeedInfected}\left(-4.202\times10^{12}\right)}+0.029\sqrt{\text{SeedInfected}}-1.007+\text{R}_0+\text{SeedInfectedDaily}}+1.429$

Table A.8: Surrogate model examples from RUN 3. The models are arbitrary chosen from the Pareto front of the optimized high-quality ensemble from RUN 3. Parameters are listed in Table 2.2 on page 22. Legend: SeedInfected (a) and R_0 (b).

Response	Complexity	1-R^2	Function
AR	187	0.178	$0.233\left(-0.162*\mathrm{b}+\frac{-0.714}{\mathrm{a}}+\frac{-0.921}{15.001-2.554*\mathrm{a}+\mathrm{a}^{\mathrm{b}}+\mathrm{b}}+\frac{1}{\frac{1}{-0.983\mathrm{a}+\mathrm{b}}+9.653+\mathrm{b}}+\frac{-4.092}{\mathrm{b}}\right)+0.963$
Peak Day	85	0.431	$\frac{100.767}{1.819*\mathrm{b}-3.367+\log(\log(4.064*\mathrm{a}))+\frac{1}{\mathrm{a}*\log(\mathrm{b})}}+13.869$

Table A.9: Surrogate model examples from RUN 4. The models are arbitrary chosen from the Pareto front of the optimized high-quality ensemble from RUN 4. Parameters are listed in Table 2.2 on page 22. Legend: SeedInfected (a), R_0 (b), SeedInfectedDaily (c), Travel (d), AscertainmentFraction (e), AscertainmentDelay (f), ResponseThreshold (g), VaccinationCoverage (h),$VE_{susceptibility}$ (i).

Response	Complexity	1-R^2	Function
AR	161	0.188	$0.823+\frac{-16.853\left(-1.048*\mathrm{g}+\frac{1}{\mathrm{b}}\right)}{\frac{-12.507}{-1.674*\mathrm{i}+1.964+6.888\mathrm{c}+\mathrm{a}+\log(\mathrm{a})}-1.623*\mathrm{e}+18.609+\sqrt{\mathrm{f}}+\mathrm{c}+\mathrm{d}}$
Peak Day	134	0.395	$\frac{464.427}{\mathrm{b}*\left(1.434+\sqrt{\mathrm{a}}+\mathrm{h}\right)*\left(3.243^{\frac{2.002}{\mathrm{c}*\left(1.522\times10^{6}\right)+\mathrm{a}}}-1.754+\mathrm{b}+\mathrm{c}\right)}+26.035$

Table A.10: Surrogate model examples from RUN 5. The models are arbitrary chosen from the Pareto front of the optimized high-quality ensemble for RUN 5. Parameters are listed in Table 2.2 on page 22. Legend: SeedInfected (a), R_0 (b), VaccinationCoverage (c), $VE_{susceptibility}$ (d) and $VE_{infectiousness}$ (e).

Response	Complexity	1-R^2	Function
AR	150	0.136	$-0.003 + 0.579 * (0.959 + -1.085\text{cd} + -1.255\text{ce})2.918^{\frac{-0.814}{\text{b}^2 \log(\text{b})(-1.120*\text{c}+1.122)}}$
PeakDay	197	0.404	$\frac{251.047}{\left(-1.041*\text{b}+6.153^{-1.238*\text{b}+2.641^{\text{c}}+\text{b}*\text{c}*\text{d}+\text{e}}+\frac{4.672}{\text{a}}\right)^2+\text{b}+\frac{1}{\text{a}}+\text{d}^2} + 17.969$

A.2 Optimized High-Quality Model Ensemble RUN 1

Table A.11: Optimized high-quality surrogate model ensemble for the attack rate from RUN 1 with SeedInfected (a), R_0 (b) and SeedInfected-Daily (c).

Complexity	1-R^2 ($\times 10^3$)	Function
60	2	$\frac{-0.737}{\frac{-2.899}{3.450+\mathrm{a}}+\mathrm{a}\,\mathrm{c}\left(6.015\times10^{-4}\right)+\mathrm{b}}+0.799$
66	2	$\frac{-0.740}{\frac{2.283\times10^{-10}}{-1.763+\mathrm{b}+\mathrm{a}}+\mathrm{a}\,\mathrm{c}\left(6.012\times10^{-4}\right)+\mathrm{b}}+0.800$
66	4	$\frac{-0.748}{\frac{8.837}{-11.274+\mathrm{a}+\mathrm{c}}+\mathrm{a}\,\mathrm{c}\left(6.021\times10^{-4}\right)+\mathrm{b}}+0.801$
74	2	$\frac{-0.737}{\frac{32204.600}{-11108.500\mathrm{a}-38325.800}+\mathrm{a}\,\mathrm{c}\left(6.015\times10^{-4}\right)+\mathrm{b}}+0.799$
86	2	$-0.505\left(\frac{1.464}{\mathrm{a}\,\mathrm{c}\left(6.005\times10^{-4}\right)+\mathrm{b}}+\frac{1}{\sqrt{\mathrm{b}}+\mathrm{a}}\right)+0.800$
89	9	$0.697+\frac{-2.389}{\left(\frac{\mathrm{a}}{10.596+\mathrm{a}}+\mathrm{a}\,\mathrm{c}\left(6.620\times10^{-4}\right)+\mathrm{b}\right)^2}$
97	9	$0.697+\frac{-2.389}{\left(\mathrm{a}\,\mathrm{c}\left(6.614\times10^{-4}\right)+\frac{\mathrm{a}}{9.058+\mathrm{b}+\mathrm{a}}+\mathrm{b}\right)^2}$
102	1	$0.638+\frac{-4.405}{-0.644+2.721^{\frac{\mathrm{a}}{0.719+\mathrm{a}}+\mathrm{a}\,\mathrm{c}\left(6.184\times10^{-4}\right)+\mathrm{b}}+\mathrm{b}}$
105	7	$0.619-3.777\exp\left(\mathrm{a}\,\mathrm{c}\left(-7.849\times10^{-4}\right)-\mathrm{b}-e^{-\sqrt{\frac{1}{\mathrm{a}}}}\right)$
110	7	$0.619-11258.800\exp\left(\mathrm{a}\,\mathrm{c}\left(-7.849\times10^{-4}\right)-\mathrm{b}-e^{-\sqrt{\frac{1}{\mathrm{a}}}}-8\right)$
113	6	$0.618+\frac{-3.758}{-1.494+\exp\left(\mathrm{a}\,\mathrm{c}\left(7.849\times10^{-4}\right)+\mathrm{b}+e^{-\sqrt{\frac{1}{\mathrm{a}}}}\right)+\mathrm{b}}$
113	7	$0.628+\frac{-4.454}{\exp\left(\mathrm{a}\,\mathrm{c}\left(7.849\times10^{-4}\right)+\mathrm{b}+e^{-\sqrt{\frac{1}{\mathrm{a}}}}\right)+7.849\times10^{-4}+\mathrm{b}}$
113	9	$0.620+\frac{-3.851}{\exp\left(\mathrm{a}\,\mathrm{c}\left(7.849\times10^{-4}\right)+\mathrm{b}+e^{-\sqrt{\frac{1}{\mathrm{a}}}}\right)+7.849\times10^{-4}+\mathrm{c}}$
128	9	$-3.538\exp\left(\exp\left(\mathrm{a}\,\mathrm{c}\left(-7.849\times10^{-4}\right)-\mathrm{b}-e^{-\sqrt{\frac{1}{\mathrm{a}}}}\right)\right)+4.153$
134	7	$-11258.600\exp\left(\exp\left(\mathrm{a}\,\mathrm{c}\left(-7.849\times10^{-4}\right)-\mathrm{b}-e^{-\sqrt{\frac{1}{\mathrm{a}}}}-8\right)\right)+11259.200$
141	1	$\frac{-88.811}{\left(-0.028+5.393\left(\frac{\mathrm{a}}{3.037+\mathrm{a}}+\mathrm{a}\,\mathrm{c}\left(7.285\times10^{-4}\right)+\mathrm{b}\right)+\mathrm{b}\right)^2}+0.702$
141	6	$-0.203\left(\frac{1.853}{0.011\mathrm{c}(0.020\mathrm{a}+\mathrm{b})+\sqrt{\mathrm{b}}}+\frac{1}{\mathrm{b}+\mathrm{a}}\right)^2+0.772$
142	19	$\frac{-109295\exp\left(e^{-0.253\mathrm{a}}-3.833-\mathrm{b}\right)}{1.081\sqrt{\mathrm{a}^2\mathrm{c}^2+1484.140+\mathrm{b}}}+0.665$
143	2	$-0.367\left(\frac{2.015}{\mathrm{a}\,\mathrm{c}\left(6.007\times10^{-4}\right)+\mathrm{b}}+\frac{1}{\sqrt{\frac{1}{\sqrt[8]{\mathrm{b}}+\mathrm{a}}}+\mathrm{a}}\right)+0.800$
149	1	$\frac{2.844\left(\frac{3.075}{2.887+\sqrt{\mathrm{b}}}+\frac{1}{\sqrt{\mathrm{b}}+\mathrm{a}}\right)}{-0.002\mathrm{ac}^2-3.176\mathrm{b}+\mathrm{a}\,\mathrm{c}\left(-7.586\times10^{-4}\right)}+0.753$
151	0.9	$\frac{0.867\left(\frac{1.509}{0.501+\sqrt{\mathrm{b}}}+\frac{1}{\sqrt[4]{\mathrm{b}}+\mathrm{a}}\right)}{-0.340-0.929\mathrm{b}+\mathrm{a}\,\mathrm{c}^2\left(-9.338\times10^{-4}\right)}+0.742$

continued on next page

Table A.11 – continuation from previous page (a: SeedInfected, b: R_0 and c: SeedInfectedDaily)

Complexity	1-R^2 ($\times 10^3$)	Function
151	1	$\frac{-88.223}{\left(-0.035+5.376\left(\frac{a}{1.489+b+a}+a\,c\left(7.283\times10^{-4}\right)+b\right)+b\right)^2}+0.702$
151	1	$\frac{-88.196}{\left(-0.037+5.377\left(\frac{a}{2.596+a+c}+ac\left(7.298\times10^{-4}\right)+b\right)+b\right)^2}+0.702$
151	1	$\frac{-87.310}{\left(-0.028+5.339\left(a\,c\left(7.297\times10^{-4}\right)+\frac{a}{b+a+c}+b\right)+b\right)^2}+0.702$
151	2	$0.710+\frac{-322.488}{\left(10.680\left(a\,c\left(6.140\times10^{-4}\right)+\frac{a}{b+a+c}+b\right)+b+c\right)^2}$
154	2	$0.696+\frac{-75.710}{\left(5\left(a\,c\left(7.849\times10^{-4}\right)+b+e^{-\sqrt{\frac{1}{a}}}\right)+7.849\times10^{-4}+b\right)^2}$
155	1	$\frac{-110.383}{\left(0.182c+6.043\left(a\,c\left(6.927\times10^{-4}\right)+\frac{a}{4.702+a}+b\right)+b\right)^2}+0.705$
157	9	$0.697+\frac{-2.388}{\left(\frac{a}{10.531+\frac{a}{a\left(-8.600\times10^{6}\right)+6.140\times10^{9}}+a}+a\,c\left(6.619\times10^{-4}\right)+b\right)^2}$
158	6	$-4.163\exp\left(\frac{1}{-0.140+\exp\left(a\,c\left(7.849\times10^{-4}\right)+b+e^{-\sqrt{\frac{1}{a}}}\right)+b}\right)+4.787$
159	1	$0.703+\frac{-71.711}{\left(0.895b+4.834\left(\frac{a}{1.443+b+a}+ac\left(7.271\times10^{-4}\right)+b\right)\right)^2}$
159	1	$0.703+\frac{-107.726}{\left(1.104b+5.918\left(a\,c\left(7.287\times10^{-4}\right)+\frac{a}{b+a+c}+b\right)\right)^2}$
162	1	$\frac{-89.867}{\left(5.414\left(0.011a\,c+\frac{a}{3+a}+b\right)-0.057a\,c+b\right)^2}+0.703$
163	1	$\frac{-87.680}{\left(-0.037+5.357\left(\frac{a}{1.829b+a}+a\,c\left(7.281\times10^{-4}\right)+b\right)+b\right)^2}+0.702$
163	20	$0.013\log\left(0.306a\left(0.306+b^{4.625}+\sqrt{ba\,c}+b-c\right)^{9.249}+40.635\right)-0.058$
165	1	$\frac{-109.989}{\left(0.182c+6.030\left(\frac{a}{3.149+b+a}+ac\left(6.921\times10^{-4}\right)+b\right)+b\right)^2}+0.706$
165	1	$\frac{-104.291}{\left(0.158c+5.851\left(a\,c\left(6.984\times10^{-4}\right)+\frac{a}{b+a+c}+b\right)+b\right)^2}+0.705$
165	1	$0.703+\frac{-84.771}{\left(5.230\left(a\,c\left(7.332\times10^{-4}\right)+\frac{a}{b+a+c}+b\right)+a\left(8.899\times10^{-6}\right)+b\right)^2}$
166	1	$\frac{14.449\left(\frac{3.558}{2.901+\sqrt{b}}+\frac{1}{\sqrt[4]{b}+a}\right)}{0.022ac-0.036a\,c^2-18.599b}+0.753$
166	1	$\frac{3.182\left(\frac{3.826}{2.907+\sqrt{b}}+\frac{1}{\sqrt[8]{b}+a}\right)}{0.113ac-0.117a\,c^2-4.397b}+0.753$
166	7	$-11259.300\exp\left(\frac{1}{-0.140+\exp\left(a\,c\left(7.849\times10^{-4}\right)+b+e^{-\sqrt{\frac{1}{a}}}+8\right)+b}\right)+11259.900$
172	1	$\frac{-89.544}{\left(5.402\left(0.002a\,c+\frac{a}{1.443+b+a}+b\right)-0.009a\,c+b\right)^2}+0.703$
172	1	$\frac{-89.595}{\left(5.405\left(0.123a\,c+\frac{a}{2.546+a+c}+b\right)-0.663a\,c+b\right)^2}+0.703$
173	1	$\frac{-109.989}{\left(0.182c^2+6.030\left(\frac{a}{3.149+b+a}+ac\left(6.921\times10^{-4}\right)+b\right)+b\right)^2}+0.706$

continued on next page

Table A.11 – continuation from previous page (a: SeedInfected, b: R_0 and c: SeedInfectedDaily)

Complexity	1-R^2 ($\times 10^3$)	Function
173	1	$\frac{-313.318}{\left(0.340b^2+10.426\left(\frac{a}{2.848+a+c}+ac\left(7.479\times10^{-4}\right)+b\right)+b\right)^2}+0.695$
175	5	$0.680+\frac{-73.045}{\left(5\left(a\,c\left(7.849\times10^{-4}\right)+b+e^{-\sqrt{\frac{1}{a}}}\right)+ac\left(7.849\times10^{-4}\right)+b\right)^2}$
181	1	$\frac{-233.224}{\left(0.110b^3+9.046\left(\frac{a}{2.760+a+c}+ac\left(7.559\times10^{-4}\right)+b\right)+b\right)^2}+0.693$
181	1	$0.703+\frac{-84.765}{\left(0.970b+5.261\left(\frac{a}{\frac{3.892}{b}+a+c}+ac\left(7.293\times10^{-4}\right)+b\right)\right)^2}$
191	1	$0.699+\frac{-91.273}{\left(5.484\left(\frac{a}{1.722b+a}+a\,c\left(6.459\times10^{-4}\right)+b\right)+ba\,c\left(3.472\times10^{-4}\right)+b\right)^2}$
192	0.8	$0.707+\frac{-87.716}{\left(a^2c\left(-1.433\times10^{-7}\right)+5.306\left(\frac{a}{2.468b+a}+a\,c\left(7.823\times10^{-4}\right)+b\right)+b\right)^2}$
192	1	$\frac{0.723\left(\frac{3.828}{2.909+\sqrt{b}}+\frac{1}{\sqrt[8]{b}+a}\right)}{-1b+c\left(-1.450\times10^{-4}\right)(2.501a+\text{Travel})+a\,c^2\left(-3.951\times10^{-4}\right)}+0.753$
194	1	$\frac{-91.636}{\left(5.474\left(0.025a\,c+\frac{a}{6.255-6.752c+a}+b\right)-0.132a\,c+b\right)^2}+0.703$
194	3	$-0.010\left(\frac{11.236+\frac{81}{\left(a\,c\left(3.931\times10^{-4}\right)+\sqrt{b}\right)^2}}{0.291+\sqrt{b}}+\frac{29.120}{\sqrt{0.516+a}}\right)+0.745$
205	1	$\frac{0.725\left(\frac{4.202}{3.305+\sqrt{b}}+\frac{1}{\sqrt[8]{b}+a}\right)}{-0.993b+c\left(7.231\times10^{-6}\right)\left(-2378+3.599a+\sqrt{b}\right)+a\,c^2\left(-7.523\times10^{-4}\right)}+0.758$
207	3	$-0.010\left(\frac{\frac{81}{\left(a\,c\left(3.931\times10^{-4}\right)+\sqrt{b}\right)^2}+7.389+\sqrt{b}}{0.291+\sqrt{b}}+\frac{29.120}{\sqrt{0.516+a}}\right)+0.732$
214	1	$0.705+\frac{-111.800}{\left(6.092\left(c(22.882b+a)\left(7.740\times10^{-4}\right)+\frac{a}{4.078+a+c}+b\right)+a\,c\left(-4.615\times10^{-4}\right)+b\right)^2}$
214	1	$0.704+\frac{-105.475}{\left(5.894\left(c(20.345b+a)\left(7.622\times10^{-4}\right)+\frac{a}{b+a+c}+b\right)+ac\left(-3.491\times10^{-4}\right)+b\right)^2}$
225	1	$\frac{-111.882}{\left(6.090\left(0.013c\left(b+\frac{1}{\sqrt{b}}+a\right)+\frac{a}{4.227+a+c}+b\right)-0.073a\,c+b\right)^2}+0.705$
228	1	$\frac{-82.814}{\left(\frac{-0.673}{b\left(\frac{20.425}{b}+a+c\right)}+5.176\left(\frac{a}{1.427+b+a}+a\,c\left(7.327\times10^{-4}\right)+b\right)+b\right)^2}+0.702$
230	1	$0.702+\frac{-82.505}{\left(\frac{-0.758}{b(12.657+a+c)}+5.165\left(\frac{a}{\frac{-2.862}{b}+4.828+a}+ac\left(7.329\times10^{-4}\right)+b\right)+b\right)^2}$
250	1	$0.702+\frac{-82.566}{\left(\frac{-0.711}{b\left(\frac{21.092}{b}+a+c\right)}+5.167\left(\frac{a}{\frac{-2.831}{b}+4.813+a}+a\,c\left(7.328\times10^{-4}\right)+b\right)+b\right)^2}$
252	0.8	$0.707+\frac{-87.316}{\left(a^2c\left(-1.405\times10^{-7}\right)+5.292\left(\frac{a}{\frac{a}{-33.302b+a}+b+a}+ac\left(7.823\times10^{-4}\right)+b\right)+b\right)^2}$
270	0.9	$0.692+\frac{-189.836}{\left(b^3\left(4.522\times10^{-4}\right)\left(\frac{a}{-106.593+a+c}+91.310b\right)+8.152\left(a\,c\left(7.618\times10^{-4}\right)+\frac{a}{b+a+c}+b\right)+b\right)^2}$
277	1	$\frac{-83.997}{\left(5.202\left(\frac{a}{\frac{1}{-0.003a+a\,c\left(3.777\times10^{9}\right)+b}+b+a}+a\,c\left(7.335\times10^{-4}\right)+b\right)+a\left(1.002\times10^{-5}\right)+b\right)^2}+0.703$

continued on next page

Table A.11 – continuation from previous page (a: SeedInfected, b: R_0 and c: SeedInfectedDaily)

Complexity	1-R^2 ($\times 10^3$)	Function
277	1	$0.703 + \frac{-86.242}{\left(5.283\left(\mathrm{a\,c}\left(7.310\times10^{-4}\right)+\frac{\mathrm{a}}{\frac{\mathrm{a}}{4.781\mathrm{b}^3+\mathrm{b}+\mathrm{a}}+\mathrm{b}+\mathrm{a}}+\mathrm{b}\right)+\mathrm{a}\left(6.976\times10^{-6}\right)+\mathrm{b}\right)^2}$
279	0.6	$\frac{-205.480}{\left(0.110\mathrm{b}^2\left(0.004\mathrm{a\,c}+\frac{\mathrm{a}}{\mathrm{b}+\mathrm{a}+\mathrm{c}}+\mathrm{b}\right)+8.509\left(\frac{\mathrm{a}}{3.319+\mathrm{a}+\mathrm{c}}+\mathrm{a\,c}\left(7.138\times10^{-4}\right)+\mathrm{b}\right)+\mathrm{b}\right)^2}+0.680$
279	0.8	$0.707 + \frac{-87.117}{\left(5.285\left(\frac{\mathrm{a}}{\frac{\mathrm{a}}{\frac{1}{3.256\times10^{10}}+\mathrm{b}+\mathrm{a}}+\mathrm{b}+\mathrm{a}}+\mathrm{a\,c}\left(7.825\times10^{-4}\right)+\mathrm{b}\right)+\mathrm{a}^2\mathrm{c}\left(-1.398\times10^{-7}\right)+\mathrm{b}\right)^2}$
283	1	$\frac{-82.028}{\left(5.120\left(\frac{\mathrm{a}}{-0.053\mathrm{a\,c}+\frac{\mathrm{a}}{-377.952+\mathrm{a}+\mathrm{c}}+\mathrm{b}+\mathrm{a}+\mathrm{c}}+\mathrm{ac}\left(7.498\times10^{-4}\right)+\mathrm{b}\right)+\mathrm{a\,c}\left(-2.850\times10^{-4}\right)+\mathrm{b}\right)^2}+0.705$
286	1	$\frac{-86.776}{\left(\mathrm{b}^2\mathrm{a}\left(1.417\times10^{-6}\right)\left(-0.008\mathrm{a\,c}+\frac{\mathrm{a}}{-397.533+\mathrm{b}+\mathrm{a}}+\mathrm{b}\right)+5.297\left(\frac{\mathrm{a}}{1.976+\mathrm{a}+\mathrm{c}}+\mathrm{a\,c}\left(7.443\times10^{-4}\right)+\mathrm{b}\right)+\mathrm{b}\right)^2}+0.703$
292	0.8	$0.706 + \frac{-87.975}{\left(\mathrm{a}^2\mathrm{c}\left(-1.383\times10^{-7}\right)+5.319\left(\frac{\mathrm{a}}{\frac{\mathrm{a}}{-20.512\mathrm{b}^3+\mathrm{b}+\mathrm{a}}+\mathrm{ac}\left(7.826\times10^{-4}\right)+\mathrm{b}+\mathrm{a}}+\mathrm{b}\right)+\mathrm{b}\right)^2}$

A.3 Symbolic Regression Analysis QALY

In this section, we focus on the surrogate modeling step of the active learning approach with a dynamic transmission model to explore the cost-effectiveness of infant and adult VZV vaccination options. We analyzed the results from an economic evaluation with 185 inputs, 100 of which are correlated transmission rates using 1000 different configurations. First we performed a SR analysis with all 185 variables to model vaccination benefits, expressed in incremental QALYs. The SR parameters are listed in Table 2.1 on page 20. We observed that most surrogate models in the optimized high-quality ensemble contained twelve parameters. In some cases, high dimensionality cause problems with the parameter selection in the sense that some less important variables remain captured in the surrogate models. Therefore, we performed a second SR analysis with these twelve variables to select the most important variables. All surrogate models obtained with SR are presented according to complexity and model error in Figure A.6. We selected models in the knee of the Pareto front to obtain a high-quality model ensemble with limited complexity. This ensemble is optimized using nonlinear techniques to end up with a predictive set, presented in Figure A.6. Descriptive statistics for all surrogate models and for the optimized high-quality model ensemble are given in Table A.12.

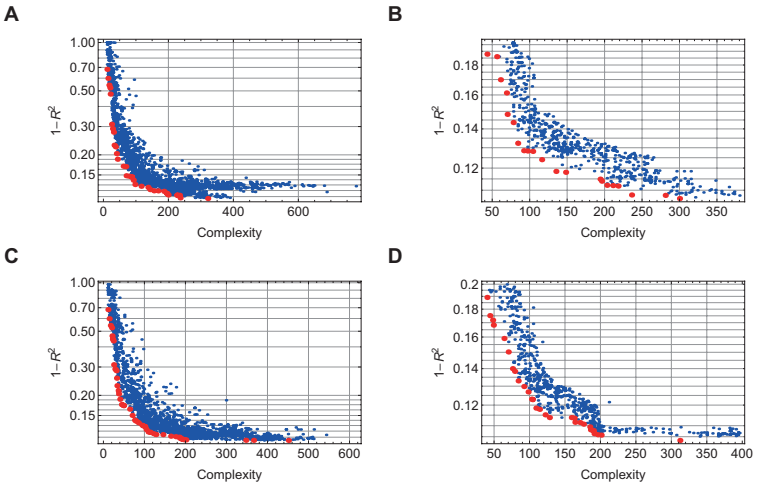


Figure A.6: Model error (1-R^2) and complexity of the surrogate models for the QALY. The model error is presented on a log scale. A: SR analysis with all variables - B: SR analysis with all variables, optimized high-quality ensemble - C: SR analysis with twelve selected variables - D: SR analysis with twelve selected variables, optimized high-quality ensemble.

Table A.12: Specifications of all surrogate models and the optimized high-quality ensemble for the QALY response.

Response	Max Model Age	Mean Model Age	Min Model Error	Mean Model Complexity
All variables	417	72	0.106965	166
All variables (ensemble)	417	113	0.106454	169
Variable selection	709	95	0.104985	161
Variable selection (ensemble)	709	138	0.103342	153

Table A.13: Variable presence in the optimized high-quality ensemble for the QALY response. With variables chi_allages: change in zoster reactivation in varicella vaccinees versus non vaccinees, durboost_allages: duration of immunity to zoster after varicella infection, takezos_7579: zoster vaccine uptake for ages 75-79y, wanzos_allages: zoster immunity waning and to*j.i*: VZV transmission rate between a person from age class *i* to a person of age class *j* with age classes {0-1y, 1-4y, 5-14y, 25-44y, 45-64y, 65-74y, 75-95y, +95y}.

Variable Name	% Models (Variable Selection)	% Models (All Variables)
durboost_allages	100.0	100.0
chi_allages	100.0	100.0
to3.10	99.2	98.7
to2.10	97.8	99.5
takzos_7579	97.0	85.9
to3.9	93.1	74.4
to2.9	64.3	43.6
wanzos_allages	45.7	34.8
to4.8	41.3	22.6
to8.10	20.7	13.0
to8.8	14.3	18.4
to5.10	8.9	8.9

A.4 Response Plot Explorer

In order to improve confidence in model-based conclusions, it is necessary to gain a thorough understanding of the system and assess how model assumptions and parameters alter the results. Surrogate models are very useful for this purpose since they can be readily explored. We used Pareto-aware symbolic regression to analyze input-response data to obtain algebraic expressions. These can easily be used in a visualization tool to explore the response surfaces from multiple parameters. Parameter must be chosen and the predicted response plots are shown for every parameter with all others fixed. Figure A.7 and A.8 presents screen-shots of a response exploration tool we installed at *www.idm.uantwerpen.be* with surrogate model for the cumulative clinical attack rate from FluTE. Surrogate modeling is relevant for many public health problems. Figure A.9 shows an analogous tool to estimate the quality-adjusted life year gain of varicella-zoster virus vaccination.

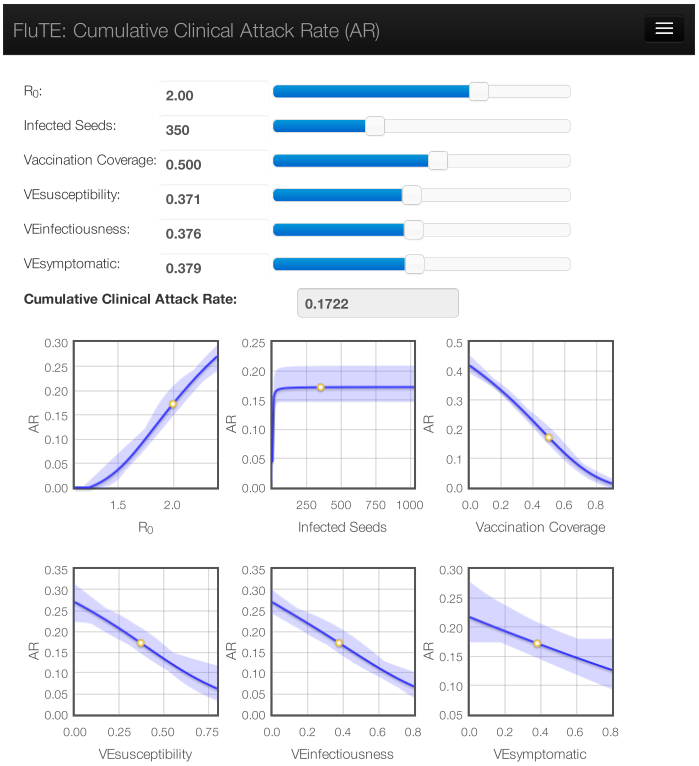


Figure A.7: Response plot explorer for the cumulative clinical attack rate. The shaded area presents the model ensemble divergence, a representation of the prediction uncertainty. An interactive version is available at *www.idm.uantwerpen.be*.

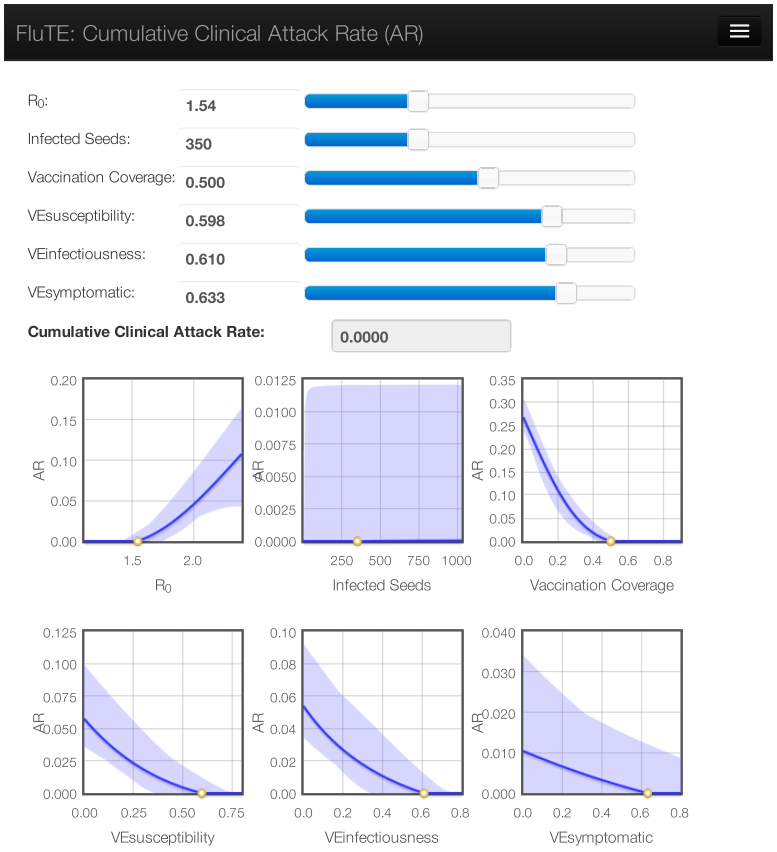


Figure A.8: Observing herd immunity with the response plot explorer for the cumulative clinical attack rate. Limited R_0 and high vaccine efficacies result in total protection of the population (AR=0) with only 50% vaccination coverage. This community effect is known as community immunity.

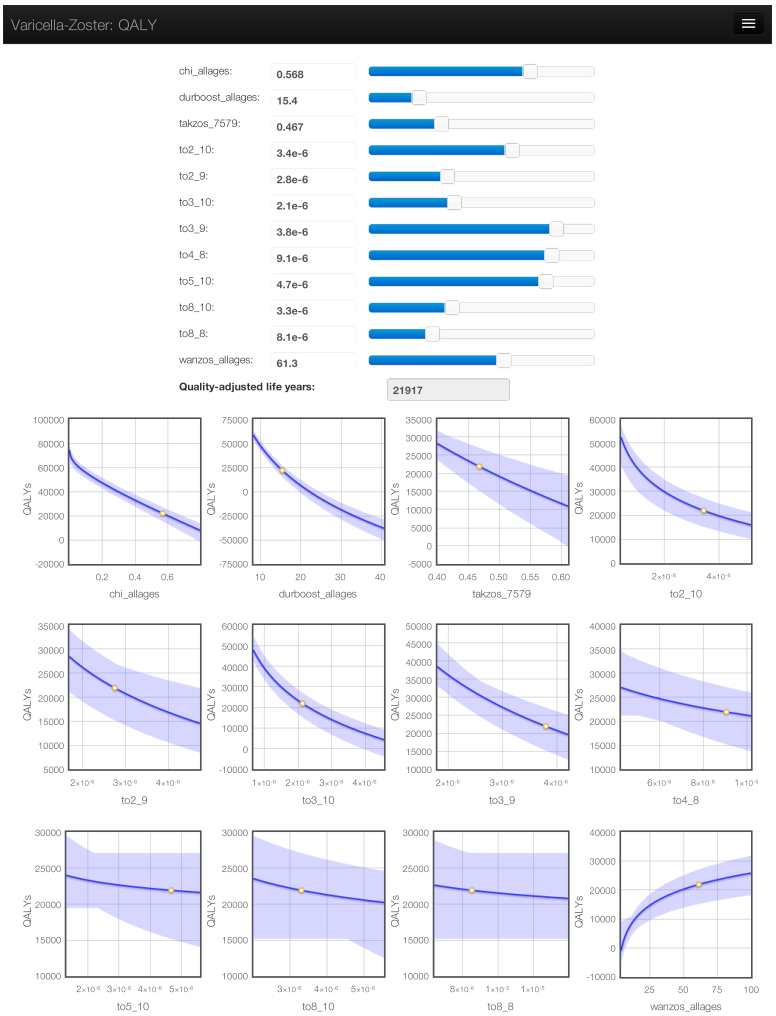


Figure A.9: Response plot explorer for the quality-adjusted life year gain of varicella-zoster vaccination. An interactive version is available at *www.idm.uantwerpen.be*.

A.5 Step-By-Step Example of Active Learning

This section presents a step-by-step example of active learning through symbolic regression in Mathematica. The grey boxes present Mathematica code. The original Mathematica notebook and required data sources can be retrieved from the PLoS Computational Biology website: `http://doi.org/10.1371/journal.pcbi.1003563`.

Problem Domain

In this example, we use the ACKLEY function as core for our simulation model [Ackley D.H. "A connectionist machine for genetic hillclimbing". Boston: Kluwer Academic Publishers, 1987.]:

$$f(x,a,b,c,d) = -a * e^{-b\sqrt{\frac{1}{d}*\sum_{i=1}^{d}(x_i\text{\^{}}2)}} - e^{\frac{1}{d}\sum_{i=1}^{d}(\mathrm{Cos}[cx_i])} + a * e \qquad \text{(A.1)}$$

This can be implemented in Mathematica as follows:

```
f[x_, a_, b_, c_, d_]:= - a * Exp [-b * Sqrt [(1/d) * Sum_{i=1}^{d} (x[[i]]^2)]]
                        - Exp [(1/d) * Sum_{i=1}^{d} (Cos[c * x[[i]]])] + a + Exp[1];
```

We defined a simulator with parameters x_1 to x_{15} where only x_1 and x_2 are really used.

```
Simulator[x1_, x2_, x3_, x4_, x5_,
          x6_, x7_, x8_, x9_, x10_,
          x11_, x12_, x13_, x14_,
          x15_]:=f[{x1, x2}, 20, 0.2, 2 * Pi, 2];
```

Plot of the ACKLEY function (Figure A.17):

```
Plot3D[
  f[{x1, x2}, 20, 0.2, 2 * Pi, 2],
  {x1, -5, 5}, {x2, -5, 5},
  PlotRange → Full, PlotPoints → 100,
  AxesLabel → {"x1", "x2", "f(x1, x2)"},
  BaseStyle → {FontWeight → Bold, FontSize → 12},
  ColorFunction → DarkRainbow, AspectRatio → 1]
```

Step 1: Design Of Experiment

It is important to use a high-quality space filling design generator or a pre-computed design. Here, we selected an initial design with 50 points for 15 parameters:

```
SetDirectory[NotebookDirectory[]] ; (** Set the working directory**)
spaceFillingDesign = Import[
    "Design15D_50points.csv",
    "FieldSeparators" → {","}
  ];
```

Scale the initial design [0,1] to fit the parameter ranges of the simulator in]-5, 5[and plot x_1 versus x_2 (Figure A.10):

```
initialDesign = spaceFillingDesign/5 − 5;
ListPlot [initialDesign[[All, {1, 2}]], FrameLabel → {x1, x2} , Frame → True]
```

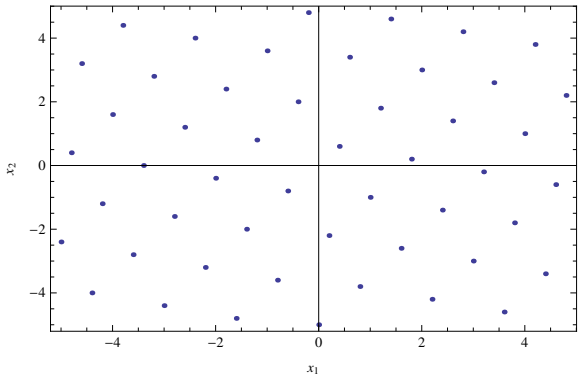


Figure A.10: Initial design of x_1 and x_2.

Step 2: Simulation Model

Run the simulator with the input combinations from the design:

```
simulatorOutput = Table[Apply[Simulator, x], {x, initialDesign}];
```

Join the input parameters with the simulator output:

```
initialDataSet = MapThread[Join[#1, {#2}]&, {initialDesign, simulatorOutput}];
```

To plot the simulator output (Figure A.18):

```
ListPlot3D[
  initialDataSet[[All, {1, 2, 16}]],
  PlotRange → Full, AxesLabel → {"x1", "x2", "f(x1, x2)"},
  BaseStyle → {FontWeight → Bold, FontSize → 12},
  ColorFunction → DarkRainbow, AspectRatio → 1]
```

Step 3: Surrogate Modeling

The DataModeler package is required for the symbolic regression function that we use. See `http://www.evolved-analytics.com/?q=datamodeler` for more information.

Load the DataModeler package:

```
<< DataModeler`
modelingRunName1 = RUN1;
```

Create surrogate models with symbolic regression for the response variable in column 16, allow power and sinusoids functions and run each evolution for 500 seconds:

```
modelSet1 =
  SymbolicRegression[initialDataSet, 16, TimeConstraint → 500,
      FunctionPatterns → {1, BuildFunctionPatterns[PowerMath, Sinusoids]}
    ];
```

Step 4: System Understanding

Plot the surrogate models with the model accuracy on the vertical axis and the complexity on the horizontal axis (Figure A.11):

```
ParetoFrontLogPlot[modelSet1, PlotLabel → LabelForm@modelingRunName1]
```

Only models with high accuracy and low complexity are used for predictions. Select the high-quality models:

```
highQualityModels1 = SelectModels[modelSet1, QualityBox → {650, 0.04}];
```

Plot the high-quality surrogate models (Figure A.12):

```
ParetoFrontLogPlot[
```

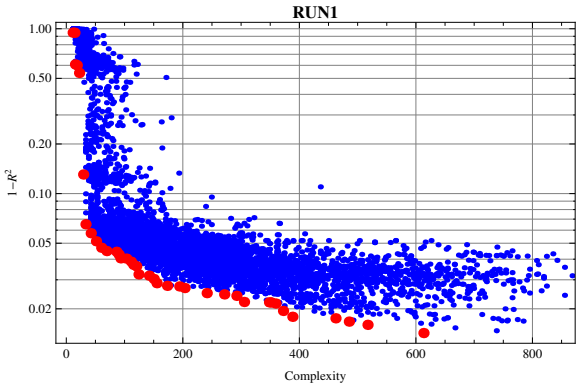


Figure A.11: Pareto front log plot of "modelSet1".

```
highQualityModels1,
PlotLabel → LabelForm@(modelingRunName1 <> " High Accuracy Models")]
```

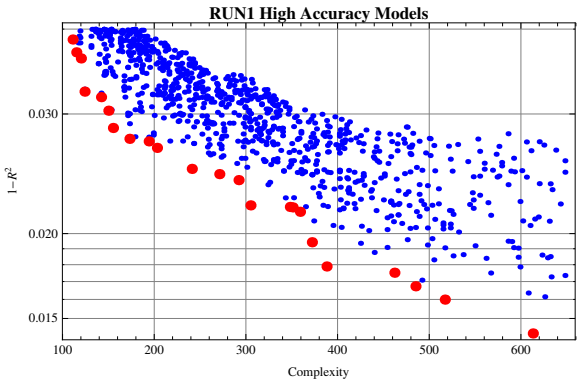


Figure A.12: Pareto front log plot of the high accuracy models of "modelSet1".

Variable importance is related to variable presence in the high-quality model ensemble. In this example, only variables $x_1$and x_2 seem important. This is in accordance with the setup of our toy simulator.

Display the variable combination table (Figure A.13):

```
VariableCombinationTable[modelSet1, SignificanceLevel → 0.1]
```

To make predictions, we need an "ensemble" data structure with only the robust models, without asymptotic behavior or discontinuities.

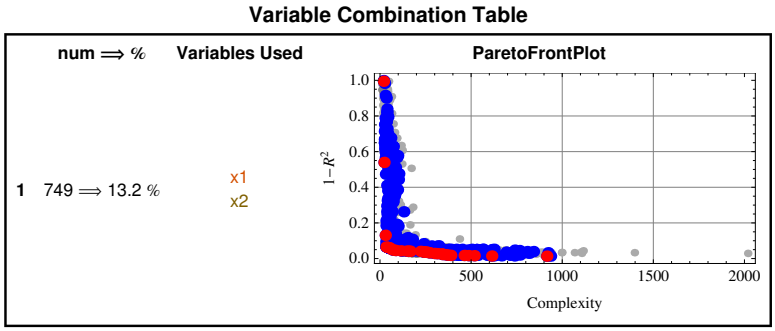


Figure A.13: Variable combination table of "modelSet1".

```
ensemble1 = CreateModelEnsemble[RobustModels@highQualityModels1];
```

Plot the model response surface (Figure A.14):

```
ResponsePlot[ensemble1, VariablesToPlot → {1,2}]
```

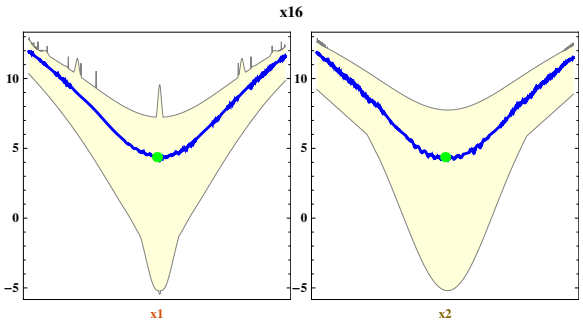


Figure A.14: Response surface behavior of the high-quality model ensemble.

Next, plot the predicted 3D surface based on the high-quality model ensemble (Figure A.19):

```
ResponseSurfacePlot[
  ensemble1,
  VariablesToPlot → {1,2},
  PlotRange → Full, AxesLabel → {"x1", "x2", "f̃ (x1, x2)"},
  BaseStyle → {FontWeight → Bold, FontSize → 12},
  ColorFunction → DarkRainbow, AspectRatio → 1, PlotLabel → None
]
```

We observe large model disagreement in the global minimum of the function. This area requires more sampling in the next modeling iteration. The 3D response prediction captures the general behavior of the simulator though refinements are needed in the next iteration.

RUN 2, Step 1: Initial Design

To improve the quality of our surrogate models we need to focus our design on x_1 and x_2. We added 250 points to the original space filling design for x_1 and x_2 homogeneously. It is also possible to sample especially the area of the predicted global minimum more into detail. The new design is presented in Figure A.15.

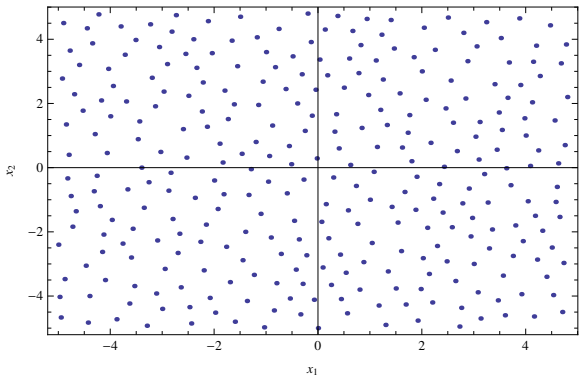


Figure A.15: Second design of x_1 and x_2.

We repeated the previously described approach with the new input design...

RUN 2, Step 2: Simulation Model

...

RUN 2, Step 3: Surrogate Modeling

...

RUN 2, Step 4: System Understanding

... Plot the model response surface (Figure A.16):

```
ResponsePlot[ensemble1, VariablesToPlot → {1, 2}]
```

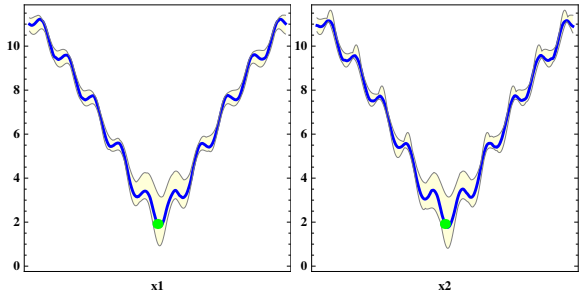


Figure A.16: Response surface behavior of the high-quality model ensemble from RUN 2.

Next, plot the predicted 3D surface based on the surrogate models (Figure A.20):

```
ResponseSurfacePlot[
  ensemble1,
  VariablesToPlot → {1, 2},
  PlotRange → Full, AxesLabel → {"x1", "x2", "f̃ (x1, x2)"},
  BaseStyle → {FontWeight → Bold, FontSize → 12},
  ColorFunction → DarkRainbow, AspectRatio → 1, PlotLabel → None
]
```

Conclusion

Active learning with feature selection and adaptive designs did improve the predictions with symbolic regression for the ACKLEY function.

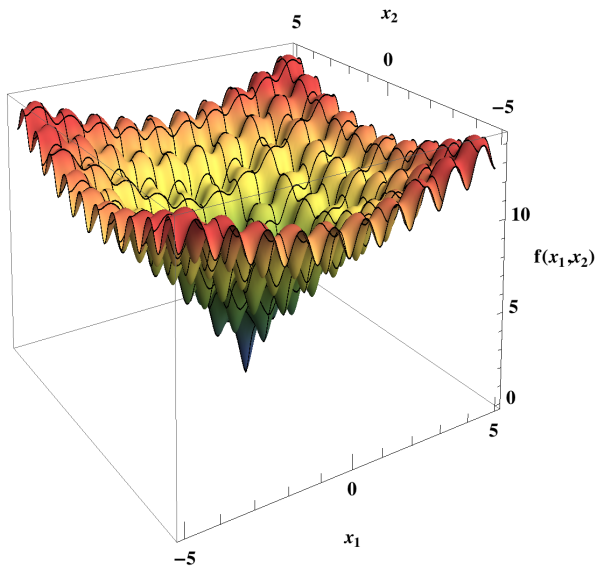


Figure A.17: ACKLEY function.

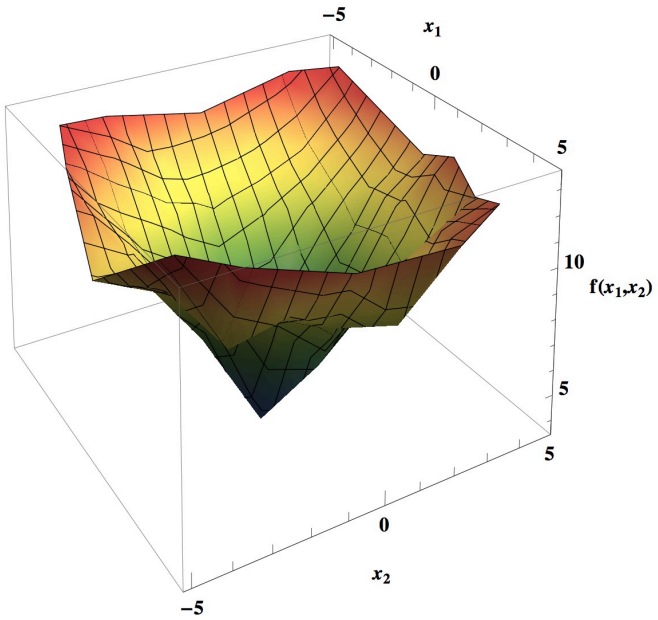


Figure A.18: Simulator output from RUN 1.

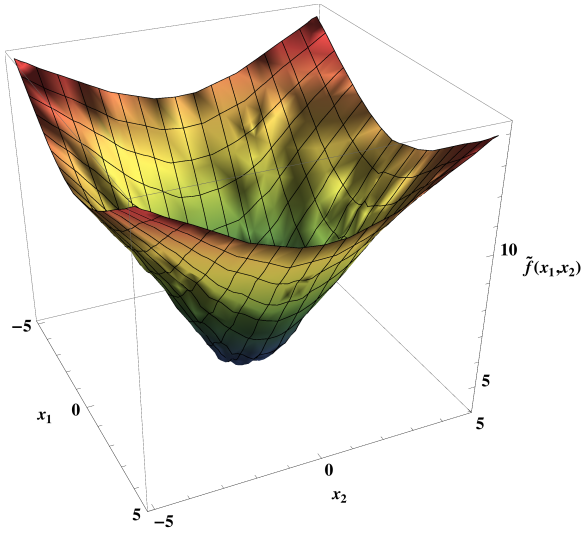


Figure A.19: Predicted 3D surface with the surrogate models from RUN 1.

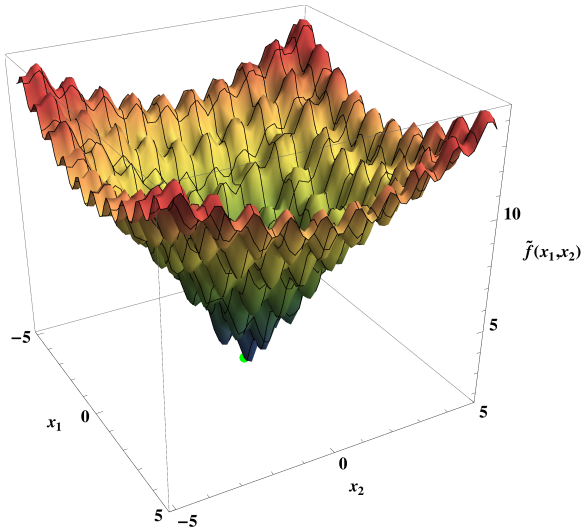


Figure A.20: Predicted 3D response of the high-quality model ensemble from RUN 2.

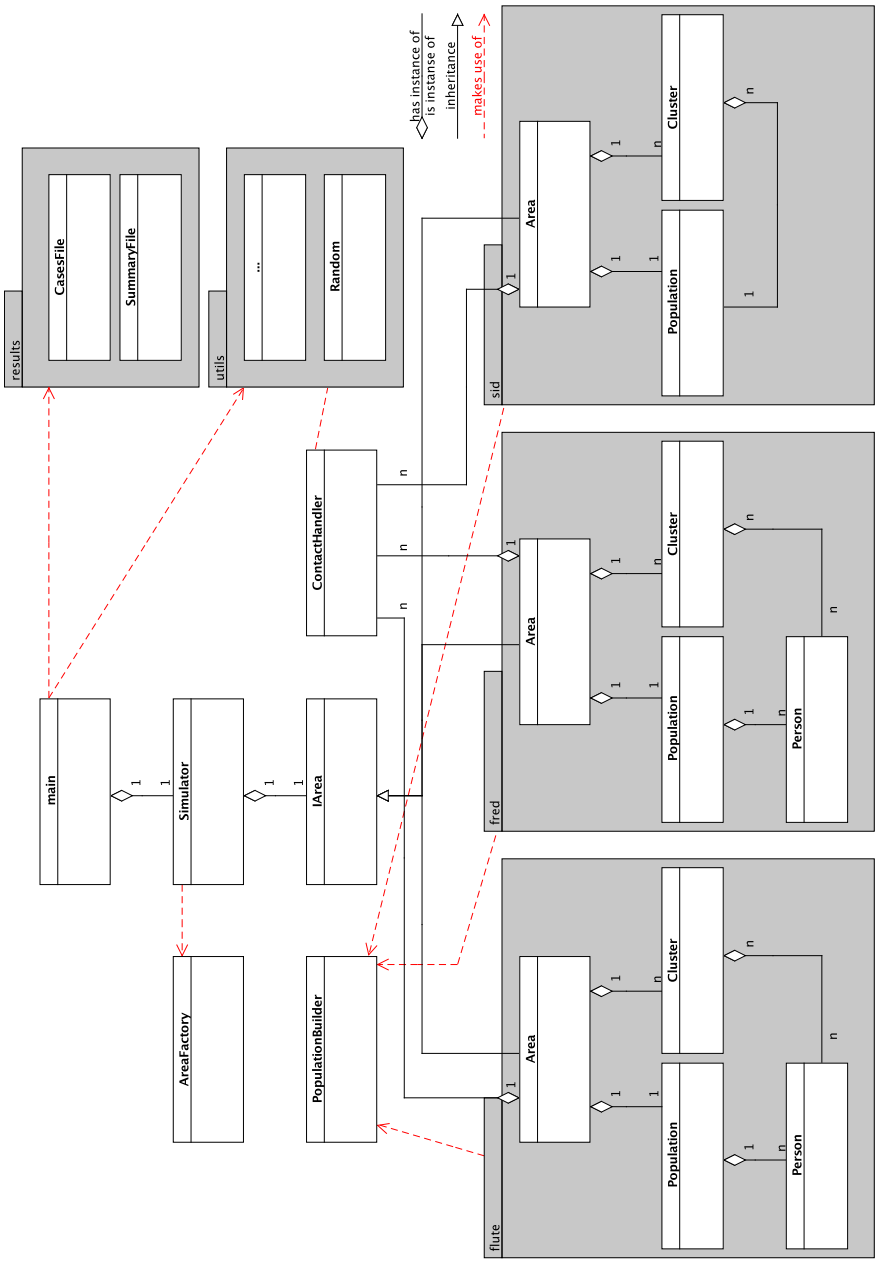
APPENDIX B

Supplementary Information Chapter 5

Summary

To generate the results discussed in Chapter 5 ("Computational Efficiency"), we created an open source agent-based modeling system for close-contact disease transmission, called indismo. Here, we present the class diagram of the software together with the user manual. Next, many additional simulations have been conducted to approve the results in Chapter 5, which are listed in Section B.3 along the computer hardware details. Finally, we present additional model exploration and validation to assess the effect of increasing model complexity on the modeling system.

B.1 Class Diagram



B.2 User Manual

Introduction

This manual provides a brief description of the indismo software and its features. Indismo is an open source, agent-based modeling system for close-contact disease transmission developed by researchers at the University of Antwerp and Hasselt University, Belgium. The simulator uses census-based synthetic populations that capture demographic and geographic distributions, as well as detailed social networks. The software provides data structures and algorithms to compare and discuss model performance. Indismo is open source in the hope of making large-scale agent-based epidemic models more useful to the community.

The model population consists of households, schools, workplaces and districts, which represent a group of people we define as a "cluster". Social contacts can only happen within a cluster. At night, people are present in their household and home district and can make social contacts with the other members. During daytime, people can stay at home or be assigned to a workplace or school in a specific district.

We use a *Simulator* class to organize the activities from the people in an *Area*. The *Area* class has a *Population*, different *Cluster* objects and a *Contact Handler*. The *Contact Handler* performs Bernoulli trials to decide whether a contact between an infectious and susceptible person leads to disease transmission. People transit through Susceptible-Exposed-Infected-Recovered states, similar to an influenza-like disease. Each *Cluster* contains a link to its members and the *Population* stores all person data.

The indismo software contains three core implementations for an agent-based simulator:

- FLUTE: Implementation based on the open source model from Chao *et al*. [14]. The *Area* contains only clusters at the district level. The household and workplace or school ID needs to be checked whether two district members make extra contact. The *Population* is a collection of *Person* objects.
- FRED: Implementation based on the open source model from Grefenstette *et al*. [15]. The household, workplace and school clusters are handled separately from the district clusters, which are only used to model general community contacts. The *Population* is equally structured as in FLUTE.

- SID: Similar cluster structure as FRED but with an innovative data-layout. The *Population* has a separate container for each person feature instead of collection of person objects.

Two algorithmic approaches are incorporated into the software to handle social contacts: with and without sorting the cluster members on health state. In total, the indismo software enables the comparison of six different implementations.

Software

System requirements

Indismo is written in C++ and portable over all platforms that have the GNU C++ compiler. De software has no dependencies on external libraries. The following tools needs to be installed:

- g++
- make
- CMake
- Python (optional, for automatization)
- Doxygen (optional, for documentation)
- LaTeX (optional, for documentation)

Installation

To install the project, first obtain the source code by cloning the repository to a directory (e.g. "git clone `https://bitbucket.org/indismo/indismo`") or download a zip file with all project material from the Bitbucket website and decompress the archive. The build system for indismo uses the CMake tool. This is used to build and install the software at a high level of abstraction and almost platform independent (see `http://www.cmake.org/`). The project includes the conventional make targets to "build", "install", "test" and "clean" the project. There is one additional target "configure" to set up the CMake/make structure that will actually do all the work. For those users that do not have a working knowledge of CMake, a front end Makefile has been provided that invokes the appropriate CMake commands. More details on building the software can be found in "INSTALL.txt" in the source folder.

Documentation

The Application Programmer Interface (API) documentation is generated automatically using Doxygen from documentation instructions embedded in the code (see `www.doxygen.org`). The developer documentation is written in Doxygen syntax and can be generated in HTML format. Figure B.1 presents the home page of the API documentation. The user manual distributed with the source code is written in LaTeX (see `www.latex-project.org`).

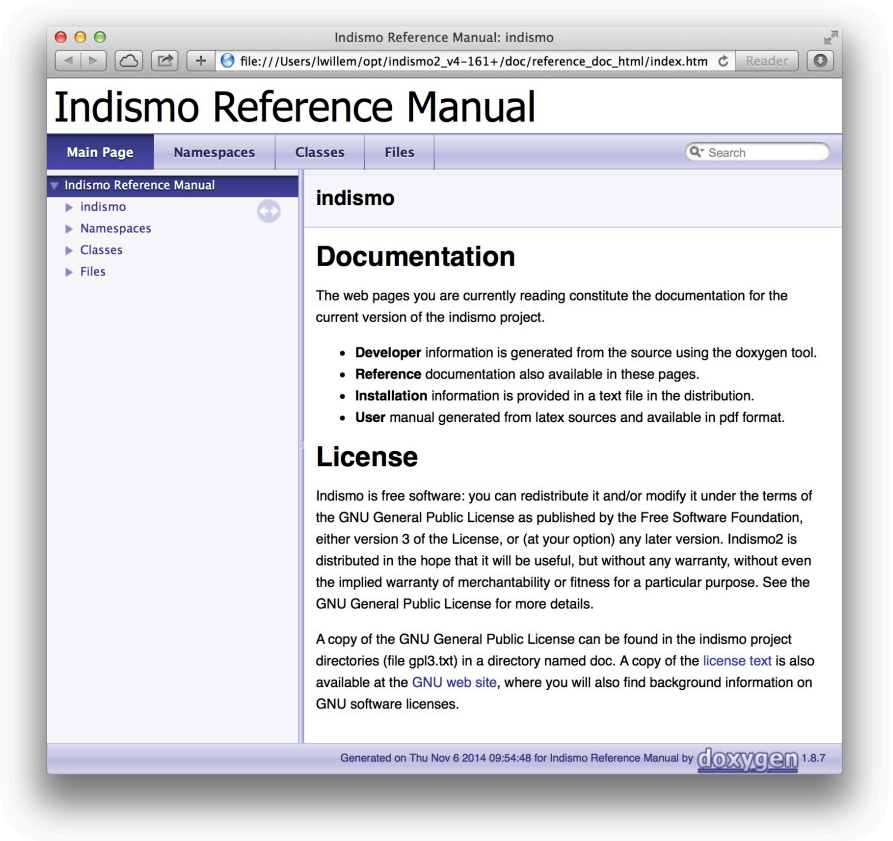


Figure B.1: Screenshot of API documentation generated with Doxygen.

Directory layout

The project directory structure is designed following maven conventions and presented in Figure B.2. Everything used to build the software is stored in directory `./src`:

- `src/main`: Code related files (sources, third party libraries and headers)
  - `src/main/"language"`: source code, per coding language
  - `src/main/resources`: third party resources
- `src/doc`: documentation files (API, manual, ...)
  - `src/doc/"tool"`: files per document processing tools
- `src/test`: test related files (scripts, regression files, ...)

Every artefact during the build procedure is generated in directory `./target` and is completely removed when the project is cleaned.

File Formats

The indismo software supports two file formats:

- CSV: comma separated values, used for population input data and the simulator output.
- JSON: JavaScript Object Notation, an open standard format that uses human-readable text to transmit objects consisting of attribute-value pairs.
  (see `www.json.org`)

Testing

Unit tests and install checks are added to indismo based on Google's "gtest" framework and CMake's "ctest" tool. In addition, the code base contains assertions to verify the simulator logic. They are activated when the application is built in debug mode and can be used to catch errors at run time.

Results

The software generates two output files:

- Log: cumulative number of cases per day.
- Output: aggregated results on the number of cases, configuration details and timings.

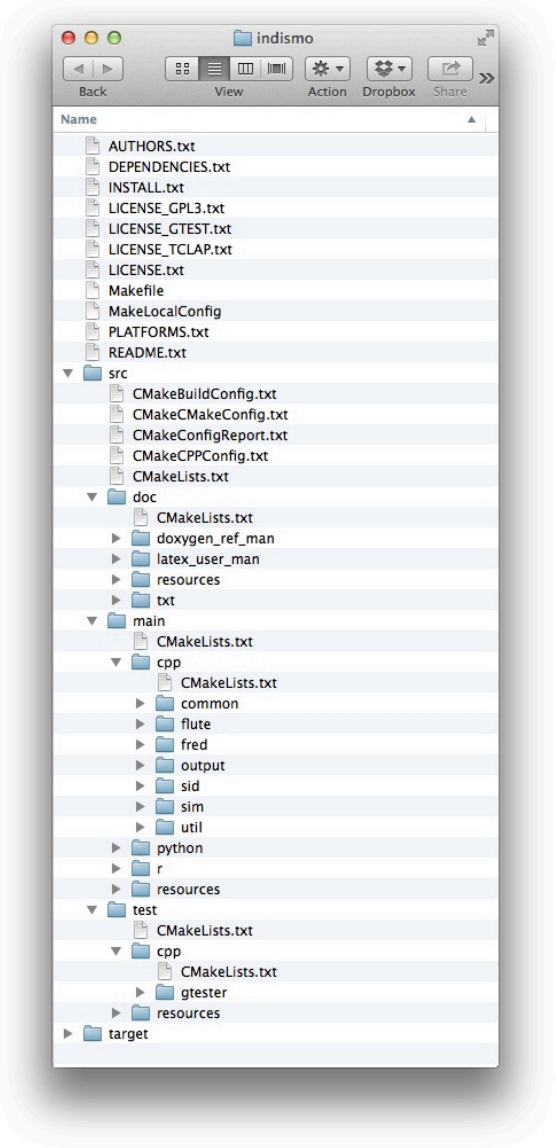


Figure B.2: Screen shot of the indismo software main directory.

Simulator

Workspace

By default, indismo is installed in `./target/installed/` inside de project directory though this can be modified using the CMakeLocalConfig.txt file (example is given in `./src/main/resources/make`). Compilation and installation of the software will create the following files and directories: (illustrated in Figure B.3):

- Binaries in directory `./target/installed/bin`
  - *indismo*: sequential executable.
  - *indismo_omp*: OpenMP executable.
  - *indismo_r*0: executable to measure the basic reproduction number (R_0).
  - *gtester*: regression tests for the sequential code.
  - *gtester_omp*: regression tests for the OpenMP code.
  - *sim_wrapper.py*: Python simulation wrapper
- Configuration files (json) in directory `./target/installed/config`
  - *config_ar_brooklyn.json*: configuration file for the sim_wrapper to perform Brooklyn simulations with different attack rates.
  - *config_ar_nassau.json*: configuration file for the sim_wrapper to perform Nassau simulations with different attack rates.
  - *config_pop_brooklyn.json*: configuration file for the sim_wrapper to perform Brooklyn simulations using population files with a randomized, sorted and original RTI person sequence.
  - *config_pop_nassau.json*: configuration file for the sim_wrapper to perform Nassau simulations using population files with a randomized, sorted and original RTI person sequence.
- Input data files (csv) in directory `./target/installed/data`
  - *brooklyn_synt_pop_original*: Synthetic population data extracted from the 2010 U.S. Synthetic Population Database (Version 1) from RTI International for Brooklyn, New York [153, 154]. The original person ordering is retained.
  - *brooklyn_synt_pop_randomized*: Synthetic population data for Brooklyn in a randomized order.
  - *brooklyn_synt_pop_sorted*: Synthetic population data for Brooklyn sorted according to day cluster (first) and household (second).

  - *nassau_synt_pop_original*: Synthetic population data extracted from the 2010 U.S. Synthetic Population Database (Version 1) from RTI International for Nassau, New York [153, 154]. The original person ordering is retained.
  - *nassau_synt_pop_randomized* : Synthetic population data for Nassau in a randomized order.
  - *nassau_synt_pop_sorted*: Synthetic population data for Nassau sorted according to day cluster (first) and household (second).
- Documentation files in directory `./target/installed/doc`
  - Reference manual
  - User manual

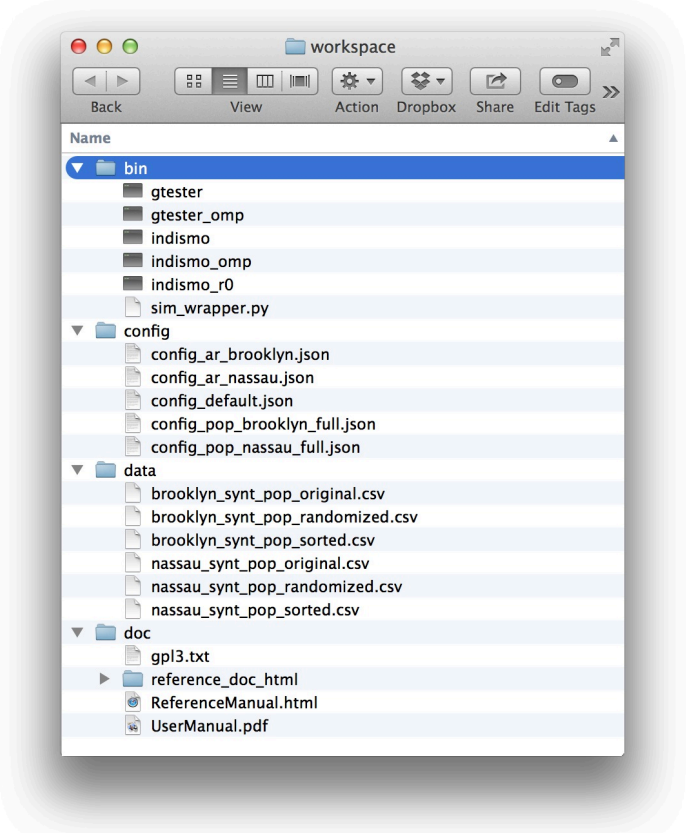


Figure B.3: Screen shot of the workspace directory.

Run the simulator

From the workspace directory, the simulator can be started with default configuration using the command "`./bin/indismo`". Settings can be passed to the simulator using one or more command line arguments:

- `-m` or `--model`: Model you want to run: flute, flute_sort, fred, fred_sort, sid, sid_sort.
- `-o` or `--output_prefix`: Prefix for the output files, by default a time stamp.
- `-p` or `--population_file`: Population file.
- `-r` or `--r0`: Basic reproduction number: the number of secondary cases by a typical primary case in a complete susceptible population.
- `-n` or `--rng_seed`: Random number generator seed.
- `-s` or `--seeding_rate`: Epidemic seeding rate: fraction of initially infected people to start the epidemic.
- `-t` or `--transmission_rate`: Transmission rate: the probability that an infection is transmitted during a contact between two adults (+18 years) in the same social contact cluster.
- `-d` or `--days`: Number of days to simulate.

Sim Wrapper

A Python wrapper is provided to perform multiple runs with the C++ executable. The wrapper forwards the model configurations with command line arguments and merges the output. The wrapper is designed to be used with .json configuration files and examples are provided with the source code. For example: `./bin/sim_wrapper --config ./config/config_ar_nassau.json` will start the simulator with each configuration in the file illustrated in Figure B.4. It is important to note the input notation: values given inside brackets can be extended (e.g. "rng_seeds"=[1,2,3]) but single values can only be replaced by one other value (e.g. "days": 100).

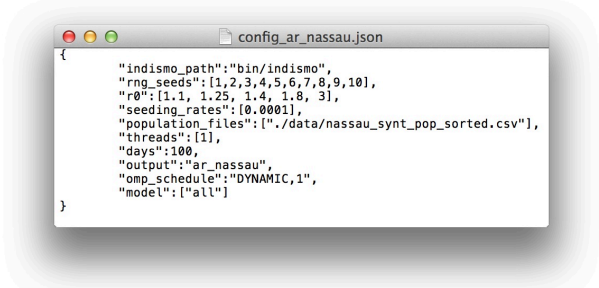


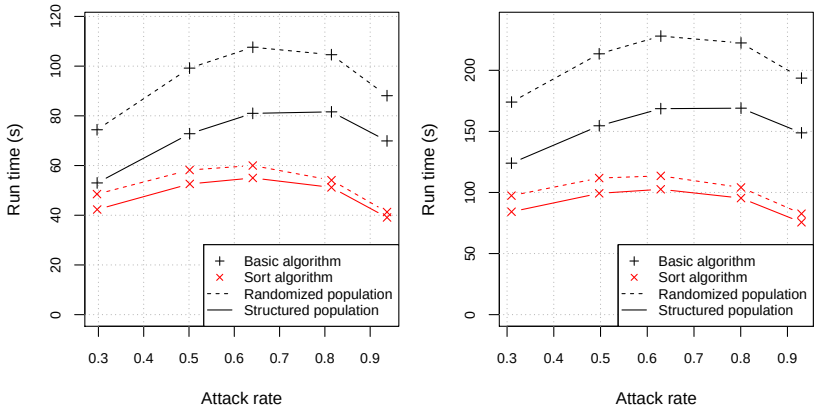
Figure B.4: Screen shot of a *sim_wrapper* configuration file.

B.3 Hardware Specifications And Extra Results.

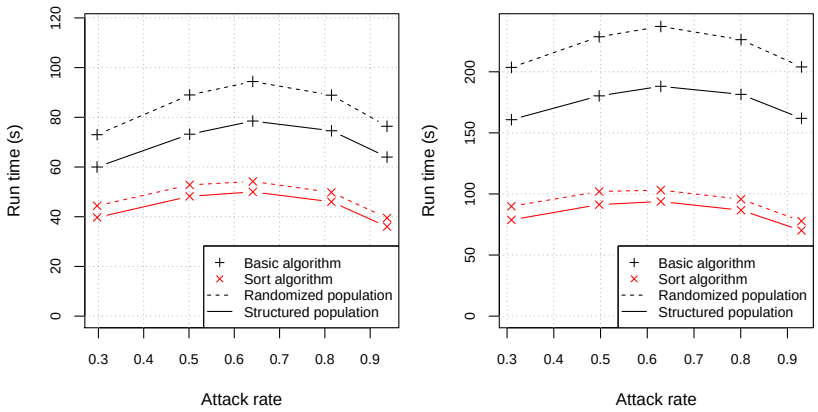
This section presents the hardware specifications of the platforms we used and extra results in addition to Chapter 5. Figure B.5 presents all benchmarks with the sequential models for the Nassau and Brooklyn population on CPU type 1. Figure B.6 and B.7 present similar results for CPU type 2 and 3 respectively. Figure B.8 shows the run time according to the number of threads on CPU type 1 for all models with the Brooklyn population. Figure B.9 presents the OpenMP results for CPU type 2.

Table B.1: Hardware specifications

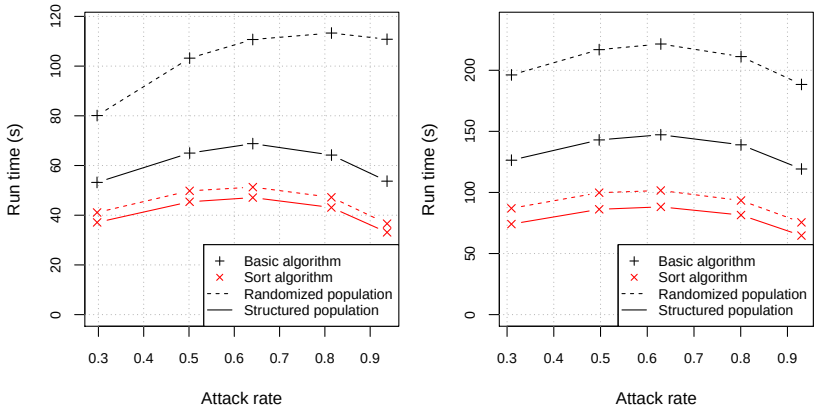
Name	**CPU type 1**	**CPU type 2**	**CPU type 3**
Configuration	Cluster	Desktop	Cluster
CPU type	Intel® Xeon® CPU E5-2680V2	Intel® Xeon® W5580	AMD Opteron® 6274
CPU speed	2.8 GHz	3.20 GHz	2.2 GHz
CPU cores	20 (2x10)	4	64 (4x16)
RAM	64 GB	24 GB	192 GB
L1 cache	10x64 KB	4x64 KB	16x48 KB
L2 cache	10x256 KB	4x256 KB	16x1024 KB
L3 cache	25 MB	8 MB	16 MB
Compiler	GNU gcc 4.8.2	GNU gcc 4.8.1	GNU gcc 4.8.1
Platform	Red Hat 4.4.7-3	Ubuntu 12.04.4 LTS	Ubuntu 12.04.1 LTS



(a) FLUTE, Nassau. **(b)** FLUTE, Brooklyn.

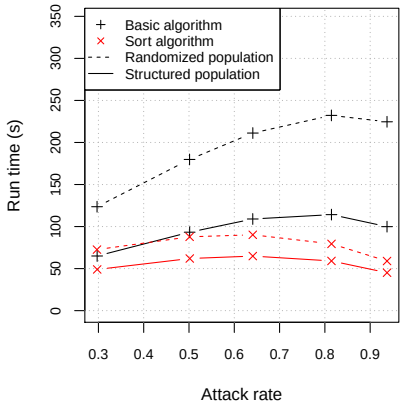


(c) FRED, Nassau. **(d)** FRED, Brooklyn.

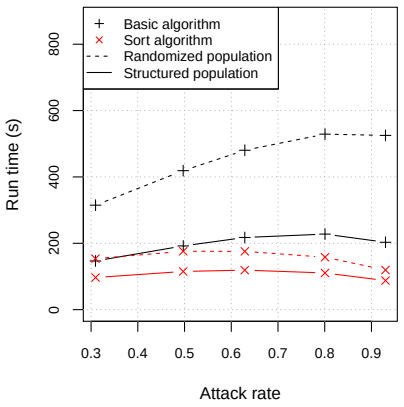


(e) SID, Nassau. **(f)** SID, Brooklyn.

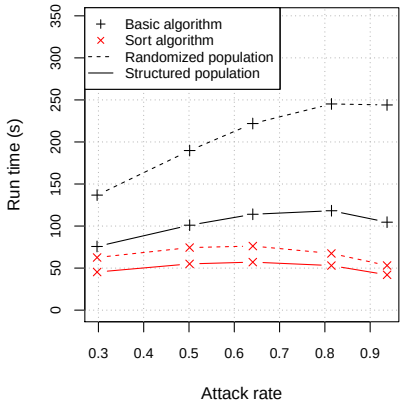
Figure B.5: Run time according to attack rate for Nassau and Brooklyn simulations on CPU type 1 from Table B.1.



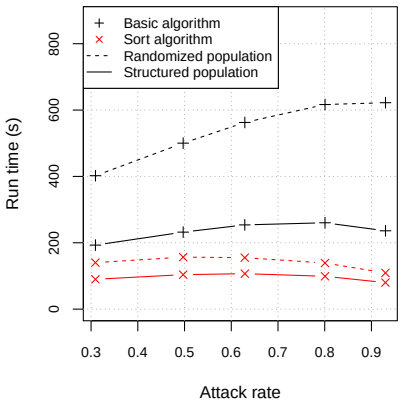
(a) FLUTE, Nassau.



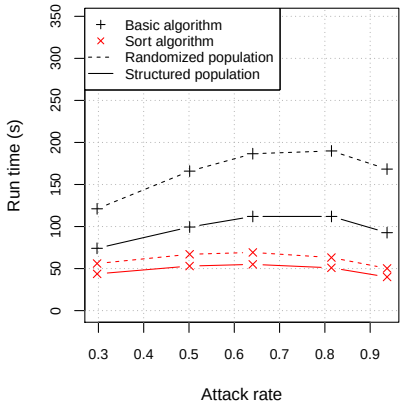
(b) FLUTE, Brooklyn.



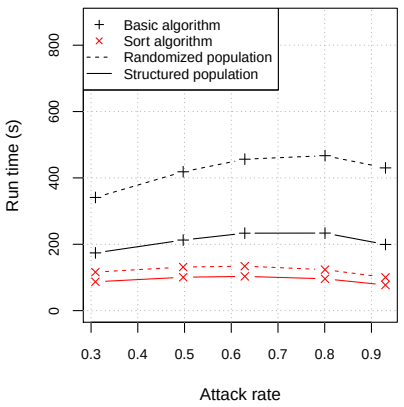
(c) FRED, Nassau.



(d) FRED, Brooklyn.

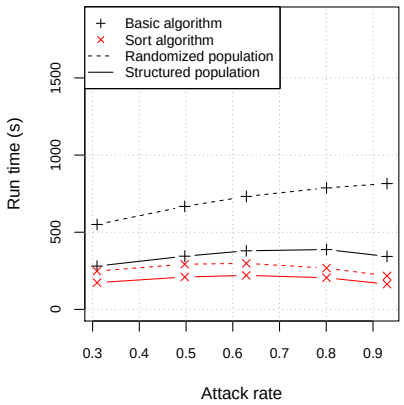


(e) SID, Nassau.

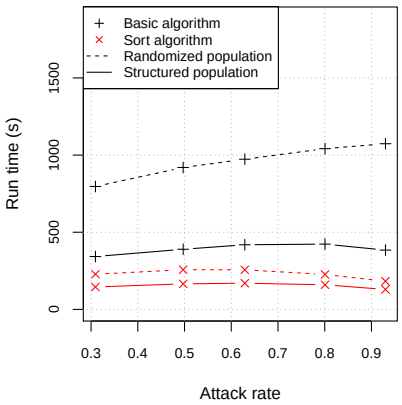


(f) SID, Brooklyn.

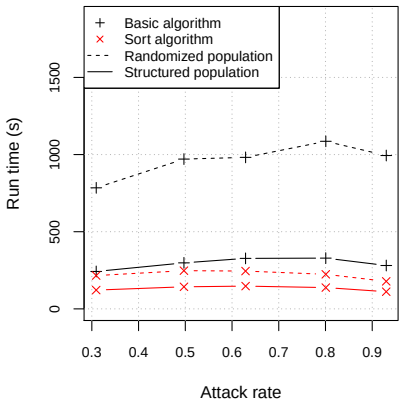
Figure B.6: Run time according to attack rate for Nassau and Brooklyn simulations on CPU type 2 from Table B.1.



(a) FLUTE, Brooklyn.

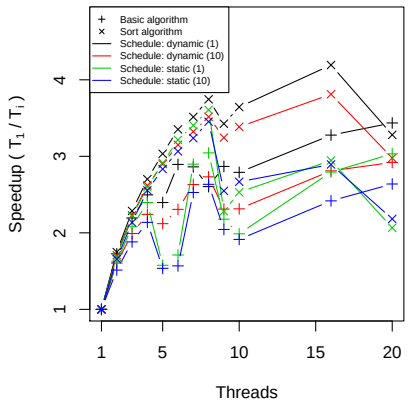


(b) FRED, Brooklyn.

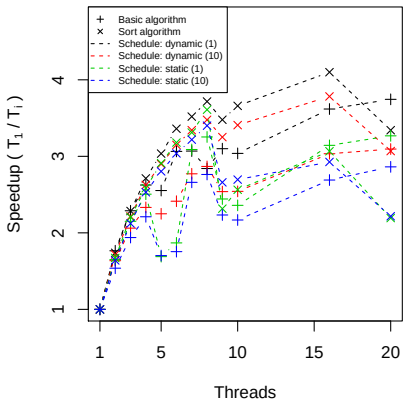


(c) SID, Brooklyn.

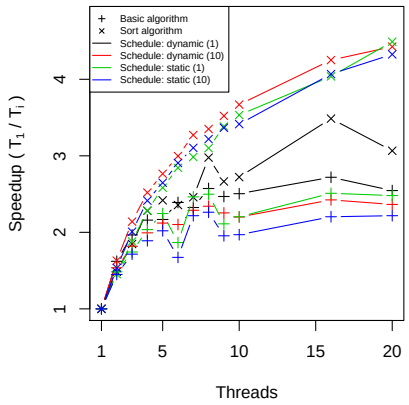
Figure B.7: Run time according to attack rate for Brooklyn simulations on CPU type 3 from Table B.1.



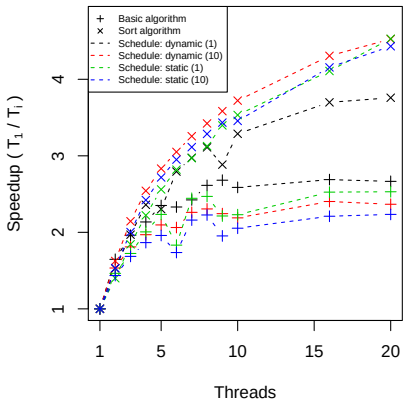
(a) FLUTE, sorted population.



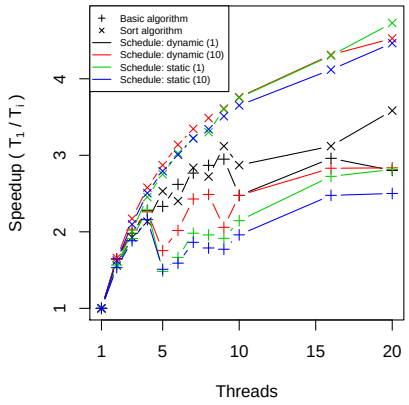
(b) FLUTE, randomized population.



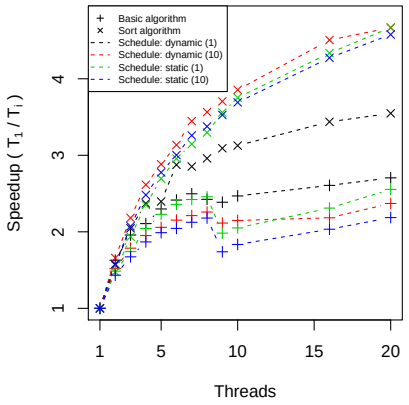
(c) FRED, sorted population.



(d) FRED, randomized population.

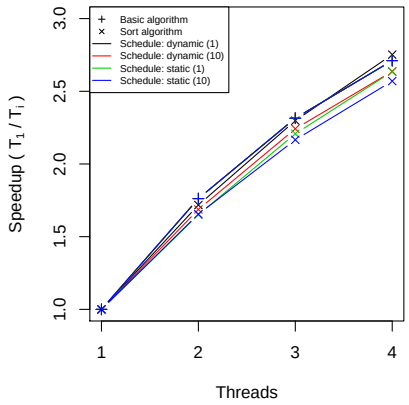


(e) SID, sorted population.

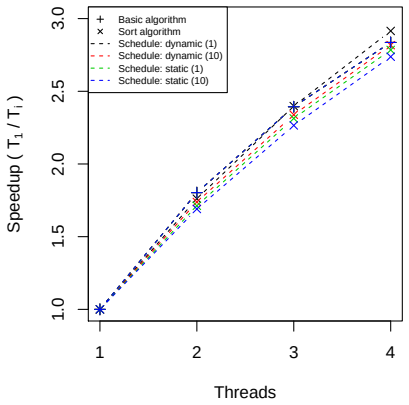


(f) SID, randomized population.

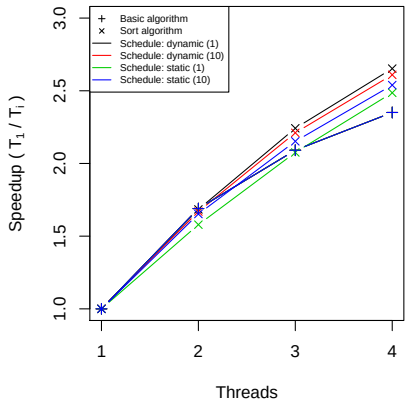
Figure B.8: Speedup according to thread number and scheduling for Brooklyn simulations on CPU type 1 from Table B.1. Timings are shown for the basic and sort algorithm with dynamic and static parallel scheduling, using workload chunks of size 1 and 10 clusters.



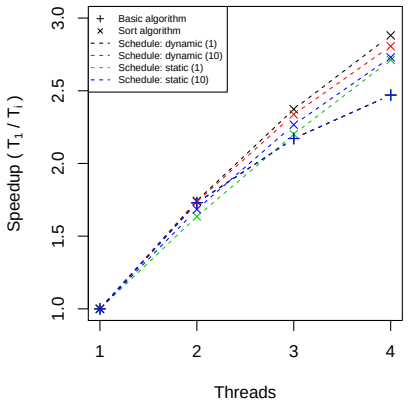
(a) FLUTE, sorted population.



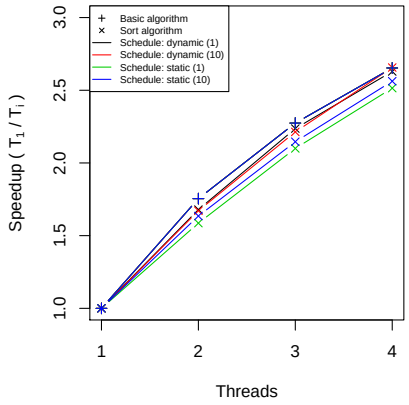
(b) FLUTE, randomized population.



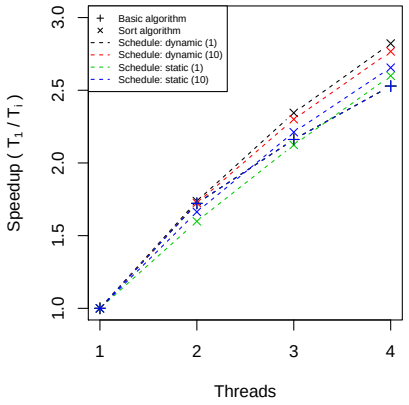
(c) FRED, sorted population.



(d) FRED, randomized population.



(e) SID, sorted population.



(f) SID, randomized population.

Figure B.9: Run times according to thread number and scheduling for Brooklyn simulations on CPU type 2 from Table B.1. Timings are shown for the basic and sort algorithm with dynamic and static parallel scheduling, using workload chunks of size 1 and 10 clusters.

B.4 Model Exploration And Validation.

The complexity of the transmission models presented in Chapter 5 is limited and person autonomy is constraint. To demonstrate that these findings are also useful for more complex models, we explored four aspects of transmission models: cluster size, dynamic clusters, number of person attributes, person autonomy.

Cluster size

Instead of using population data from RTI International, we created a virtual population of 1 million people with ages between 0-94 years and assigned them randomly to predefined home and day district clusters. Households, schools and workplaces were not included in this population structure. Each district cluster had the same size and we repeated the procedure for size 20, 500 and 10000. These populations were used in a randomized and sorted order to simulate epidemics with seeding rate $1e^{-4}$ and attack rate ± 0.65. To obtain similar attack rates with different cluster sizes, we needed to adjust the transmission rates: E.g., to obtain on average 2 secondary cases per index case, the transmission rate for cluster size 20 should be around 0.1 whereas this rate should be around 0.0002 for cluster size 10000. Also the network dynamics differ with the altered cluster sizes. Figure B.10 illustrates the attack rate distributions from 10 simulations with different cluster sizes and population structures using adapted transmission rates.

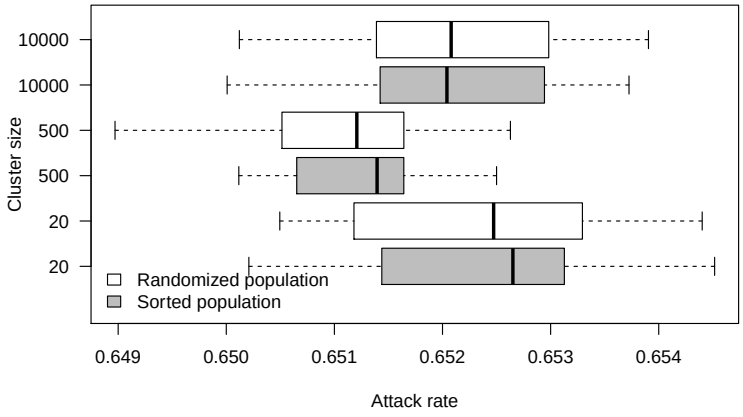


Figure B.10: Attack rate according to cluster size and population structure using adapted transmission rates. Results from 10 simulations. Box: upper and lower quartile, wiskers: minimum and maximum excluding outliers.

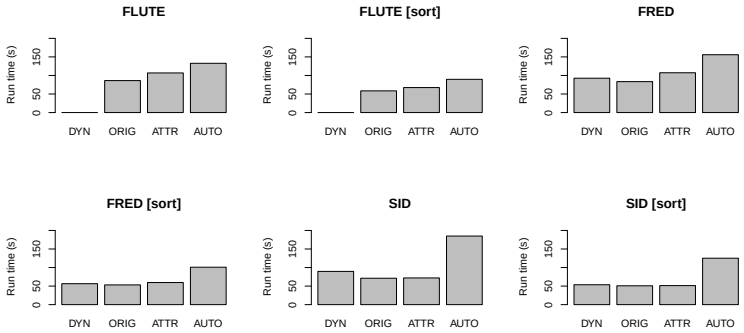
Table B.2 presents the mean run time for simulations with attack rate = ± 0.65 given the cluster size and population structure. We observed that cluster size had a large impact on run time because it affects the size of the 2 nested loops in the disease transmission processing (for each infected $=>$ for each susceptible). The run times for simulations with clusters size 20 are dominated by model initialization ($\pm$5s). Sorting the population was beneficial for model performance irrespective of cluster size or model design and the benefit increased with incremented cluster size. The differences in run time regarding model design (FLUTE, FRED and SID) did scale with cluster size and the sort algorithm performed better than the basic algorithm.

Dynamic clusters

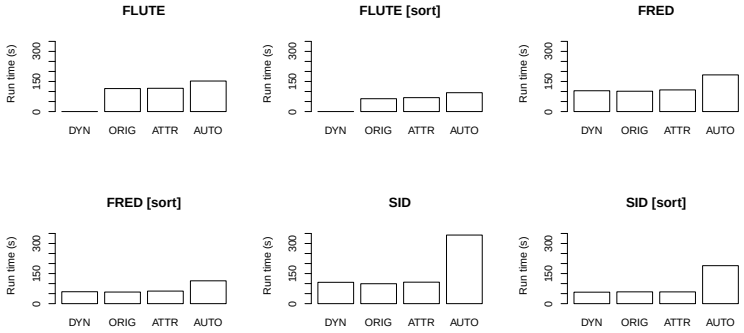
To estimate the effect of dynamic clusters on model performance, we adapted the original implementation so that membership to a social-contact cluster can change over time. To focus this analysis, cluster sizes remained constant and only day districts could change. All simulations started with the original RTI International population for Nassau and after each time step 200 000 people (15% of the population) were assigned to a new day district. Therefore, we randomly selected two people from the population and switched their day district. Within FLUTE, it is not possible to separate district clusters from subclusters like schools or workplaces and therefore we implemented this extension only for FRED and SID. Dynamic clusters enhance disease transmission so we needed to adjust transmission rates to obtain similar attack rates. Figure B.13, at the end of this document, illustrates attack rates from 10 simulations with different implementations. Figure B.11 presents the run time for simulations with the original and dynamic clusters using a sorted or randomized population. The run time increased with the introduction of dynamic clusters although the ranking of the models on run time did not change (from high to low: FRED, SID, FRED [sort], SID [sort]). The impact of the dynamic clusters was similar for sorted and randomized populations.

Table B.2: Mean run time from 10 simulations with attack rate = ± 0.65 using different cluster sizes and population structures.

Cluster size	Population	FLUTE	FLUTE [sort]	FRED	FRED [sort]	SID	SID [sort]
20	Randomized	8	8	7	7	7	6
20	Sorted	6	6	6	6	6	5
500	Randomized	23	19	22	15	21	15
500	Sorted	20	16	19	14	18	13
10000	Randomized	519	287	414	200	432	205
10000	Sorted	373	229	307	176	308	166



(a) Sorted population.



(b) Randomized population.

Figure B.11: Run time according to model implementation. Mean run times from 10 simulations with attack rate = ± 0.64. DYN: dynamic clusters, ORIG: original implementation, ATTR: extended person attributes, AUTO: extended person autonomy.

Person attributes

People in the original implementation had 11 attributes: id, age, household, home district, day cluster, day district, disease counter and four health state booleans (susceptible? / infected? / infectious? / recovered?). To estimate model performance with more elaborated person implementations, we added 30 data members to each person. For FLUTE and FRED, these values were added to the *Person* class as an array and initialized with random doubles in the constructor. For SID , we added 30 vectors to the *Population* class and each vector was extended with a random double when a person was added to the population. The new person attributes were not used in the program. The transmission dynamics or attack rate did not change compared with the original implementation (Figure B.13).

Figure B.11 presents the mean run time for simulations with and without the extra person attributes. The run time for FLUTE and FRED increased using the extended person features but the run time for SID remained constant. Interestingly, the timings for FLUTE and FRED did not change with the randomized population. We analyzed the effect of the extra person attributes for FLUTE and FRED with the profiling tool PerfExpert and observed that the amount of low level cache misses using the sorted population became similar to the amount of cache misses when the randomized population was used. We did not observe an increase in cache misses due to the extra person attributes with the randomized population. These results confirmed the targeted data management strategy using struct-of-array vs array-of-structs.

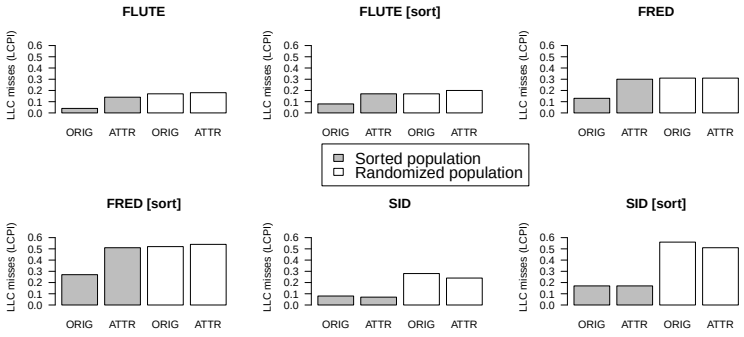


Figure B.12: Last level cache (LLC) misses with the original and extended person attribute implementations with respect to population structure. ORIG: original implementation, ATTR: extended person attributes, LCPI: local cycles per instruction.

Person Autonomy

To explore increased person autonomy, we used the random values that were stored in the extra person attributes (see previous section). We did not include an extra random number engine to focus in this extension on person autonomy rather than on calculating random numbers. We extended the decision process whether an infectious person is able to transmit the disease. This might be equivalent for choices like: to stay home, taking precautions,... To be able to compare model performance, we needed to increase the workload and branch instructions without changing the disease dynamics and simulation outcome. Therefore we implemented the following procedure with $disease\_counter \in \{0-6\}$ and attributes $A[i] \in \{0-10\}$:

Function to check whether person x is able to transmit a disease:

1. **Set** $p = 0$
2. **For** all attributes $A[i]$ in *Person*
3. **Set** $p = p + 1/(disease\_counter * A[i] + 1)$
4. **If** $p > 1$
5. **Set** p to 0
6. **If** $p < 0$
7. **Return** false
8. **Else**
9. **Return** $health\_state\_is\_infectious$?

The original function contained only the last statement. The extended function did not alter the transmission dynamics since p could never be < 0 and the random number generator is not additionally used. The extra person attribute values and math function (line 3) were chosen so that program errors could not occur. For FLUTE and FRED , we introduced this extension in *Person* with a loop over the array containing the extra person attributes. Modifying SID required a lot more work since each attribute vector needed to be consulted separately.

All run times increased with the extended person autonomy, independent of the population structure, though we observed most difference in the timings with SID (Figure B.11). The latter seemed not suited to handle many person attributes. Together with the increased workload, these timings demonstrate the disadvantage of the SID implementation.

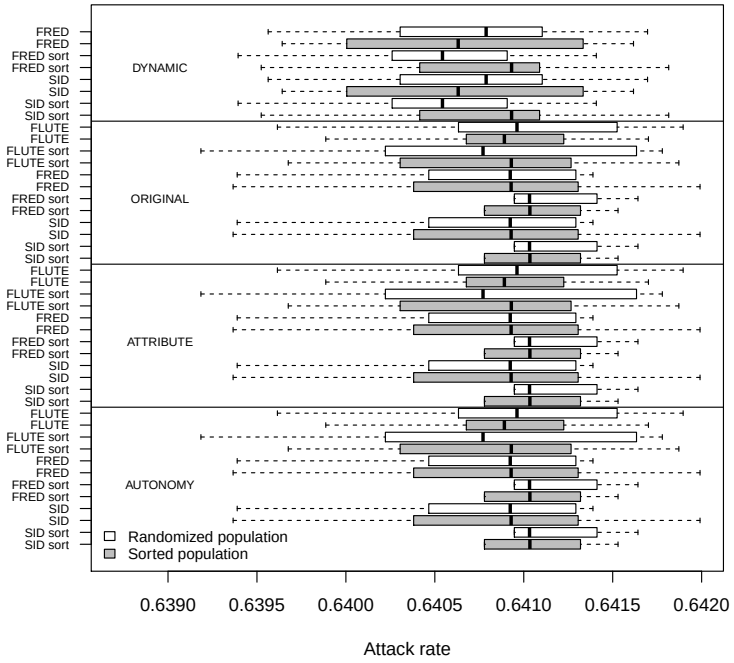


Figure B.13: Attack rate distribution according to model implementation and population structure. Results from 10 simulations. Adapted transmission rates were used for the simulations with dynamic clusters. Box: upper and lower quartile, wiskers: minimum and maximum excluding outliers.