

DATA INFORMED HEALTH SIMULATION MODELING

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Yang Qin

©Yang Qin, November/2019. All rights reserved.

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building
110 Science Place
Saskatoon, Saskatchewan
Canada
S7N 5C9

ABSTRACT

Combining reliable data with dynamic models can enhance the understanding of health-related phenomena. Smartphone sensor data characterizing discrete states is often suitable for analysis with machine learning classifiers. For dynamic models with continuous states, high-velocity data also serves an important role in model parameterization and calibration. Particle filtering (PF), combined with dynamic models, can support accurate recurrent estimation of continuous system state. This thesis explored these and related ideas with several case studies. The first employed multivariate Hidden Markov models (HMMs) to identify smoking intervals, using time-series of smartphone-based sensor data. Findings demonstrated that multivariate HMMs can achieve notable accuracy in classifying smoking state, with performance being strongly elevated by appropriate data conditioning. Reflecting the advantages of dynamic simulation models, this thesis has contributed two applications of articulated dynamic models: An agent-based model (ABM) of smoking and E-Cigarette use and a hybrid multi-scale model of diabetes in pregnancy (DIP). The ABM of smoking and E-Cigarette use, informed by cross-sectional data, supports investigations of smoking behavior change in light of the influence of social networks and E-Cigarette use. The DIP model was evidenced by both longitudinal and cross-sectional data, and is notable for its use of interwoven ABM, system dynamics (SD), and discrete event simulation elements to explore the interaction of risk factors, coupled dynamics of glycemia regulation, and intervention tradeoffs to address the growing incidence of DIP in the Australia Capital Territory. The final study applied PF with an SD model of mosquito development to estimate the underlying *Culex* mosquito population using various direct observations, including time series of weather-related factors and mosquito trap counts. The results demonstrate the effectiveness of PF in regrounding the states and evolving model parameters based on incoming observations. Using PF in the context of automated model calibration allows optimization of the values of parameters to markedly reduce model discrepancy. Collectively, the thesis demonstrates how characteristics and availability of data can influence model structure and scope, how dynamic model structure directly affects the ways that data can be used, and how advanced analysis methods for calibration and filtering can enhance model accuracy and versatility.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Prof. Nathaniel Osgood, who has inspired and guided me during my graduate study at University of Saskatchewan. His knowledge, expertise, patience, and understanding taught me immensely in becoming a computer scientist and simulation modeler. He always encouraged and supported me to work on projects I was interested in. Without his guidance and assistance, finishing research projects and writing thesis would never have been possible.

I would like to thank my colleague for their help in many projects and sharing knowledge in computer science and simulation modeling field. I am also grateful to Mosquito Control Program of the City of Saskatoon and Australia Capital Territory and Saskatchewan Ministry of Health for providing data.

Dedicated to Guiying Xu, Xigui Qin, and Weicheng Qian, for always encouraging me and making me smile.

CONTENTS

Permission to Use	i
Abstract	ii
Acknowledgements	iii
Contents	v
List of Tables	viii
List of Figures	ix
List of Abbreviations	x
1 Introduction	1
1.1 Motivation	4
1.2 Problem	6
1.3 Contributions	7
1.4 Thesis Outline	8
1.5 Publications	10
2 Background	12
2.1 Basics of Hidden Markov Models	12
2.1.1 Forward and Backward Probabilities	13
2.1.2 Baum-Welch Algorithm	15
2.1.3 Viterbi Algorithm	16
2.2 Particle Filtering	17
2.2.1 Basics of Particle Filtering	17
2.2.2 Application of Particle Filtering in the Proposed Model	19
2.3 Introduction to Simulation Modeling	20
2.3.1 System Dynamics Models	20
2.3.2 Agent Based Modeling	21
2.3.3 Discrete Event Simulation	21
2.4 Literature Review	22
2.4.1 Models of Gestational Diabetes Mellitus and Type 2 Diabetes Mellitus	22
2.4.2 Smoking Detection	23
2.4.3 Agent Based Model of Smoking or E-Cigarette Use	23
3 Identifying Smoking from Smartphone Sensor Data and Multivariate Hidden Markov Models	24
3.1 Introduction	24
3.2 Data Processing and Algorithm	25
3.3 Results	28
3.3.1 Results with Single Feature	29
3.3.2 Results with Three Features	29
3.3.3 Results with Five Features	29
3.4 Discussion	31
3.4.1 Thoughts on Labeled Data	31
3.4.2 Emission Probability for Multi-dimensional Observations	31
3.5 Related Work	32

3.6	Limitations and Future Work	32
3.7	Conclusions	32
4	Effect of E-cigarette Use and Social Network on Smoking Behavior Change: An agent-based model of E-cigarette and Cigarette Interaction	34
4.1	Introduction	34
4.2	Methods	35
4.2.1	Model Overview	35
4.2.2	Model Formulation	36
4.2.3	Model Calibration	39
4.2.4	Model Scenarios	39
4.2.5	Sensitivity Analysis	40
4.3	Results	41
4.3.1	Comparison between Scn1 and Scn2	41
4.3.2	Comparison between Scn2 and Scn3	41
4.3.3	Sensitivity Analysis	42
4.4	Discussion	42
5	Multi-Scale Simulation Modeling for Prevention and Public Health Management of Diabetes in Pregnancy and Sequelae	44
5.1	Introduction	44
5.2	Methods	46
5.2.1	Model Overview	46
5.2.2	Model Formulation	47
5.3	Model Development Process	52
5.4	Model Calibration	54
5.5	Performance Optimization	55
5.6	Logic-Presentation Separation	56
5.7	Results	57
5.7.1	Individual Trajectory	57
5.7.2	Population-Level Outcomes	58
5.8	Discussion	59
6	Particle Filter Applied to System Dynamics Model for Mosquito Population Surveillance	64
6.1	Introduction	64
6.2	Empirical Data	66
6.3	Methods	67
6.3.1	Overview of the System Dynamics Model of Mosquito Population	67
6.3.2	Model of Mosquito Control (VectoBac)	68
6.3.3	Probability of Capturing a Mosquito by CDC Light Trap	68
6.3.4	State Equation Model	70
6.3.5	Particle Filtering Model	71
6.4	Model Calibration	71
6.5	Model Scenarios	72
6.6	Results	72
6.7	Discussion	75
7	Conclusion and Future Work	85
7.1	Solutions	87
7.2	Summary of Findings	88
7.2.1	Identifying Smoking from Smartphone Sensor Data and Multivariate Hidden Markov Models	88
7.2.2	Effect of E-cigarette Use and Social Network on Smoking Behavior Change: An Agent-Based Model of E-cigarette and Cigarette Interaction	88

7.2.3	Multi-Scale Simulation Modeling for Prevention and Public Health Management of Diabetes in Pregnancy and Sequelae	89
7.2.4	Particle Filter Applied to System Dynamics Modeling for Mosquito Population Surveillance	89
7.3	Contributions	89
7.4	Limitation and Future Work	90
7.5	Conclusion	92
References		93
Appendix A Calibrated parameters of the hybrid multi-scale model of diabetes in pregnancy		101
Appendix B Linear regression on weather-related variables for the probability of a <i>Culex</i> mosquito being captured by CDC light trap		102
Appendix C Calibrated parameters of the particle filtering model applied to the SDM of <i>Culex</i> mosquito development		103
Appendix D Parameters of the SDM of <i>Culex</i> mosquito development		104

LIST OF TABLES

6.1	Discrepancy between the model outputs and the empirical data in Scn1	73
6.2	Discrepancy between the model outputs and the empirical data in Scn2	74
6.3	Discrepancy between the model outputs and the empirical data in Scn3 and Scn4	75
A.1	Calibrated parameters of the hybrid multi-scale model of diabetes in pregnancy	101
B.1	Coefficients of weather-related variables for the probability of a <i>Culex</i> mosquito being captured by CDC light trap	102
B.2	Summary of the linear regression on weather-related variables for the probability of a <i>Culex</i> mosquito being captured by CDC light trap	102
C.1	Calibrated parameters of the particle filtering model applied to SDM of <i>Culex</i> mosquito population	103
D.1	Parameters of the particle filtering model applied to SDM of <i>Culex</i> mosquito population . . .	104

LIST OF FIGURES

3.1	Plot of time series sensor data	26
3.2	ECDF of raw and transformed accel, wifi, and GPS data	28
3.3	AUC and error rate of univariate HMMs	30
3.4	AUC and error rate of multivariate HMMs	31
4.1	Smoking statechart	37
4.2	ECig use statechart	37
4.4	PFCS, PMCS, PFCEU and PMCEU of Scn1 and Scn2	39
4.5	SA of rate of ECigUC on PCS	40
4.6	SA of rate of ECigUR on PCS	40
4.7	Population breakdown by smoking category and fraction of CEU	41
4.8	Fraction of CS and others	42
5.1	Illustration of hybrid model structure	46
5.2	Pregnancy statechart	48
5.3	Dysglycemia classification statechart	48
5.4	Primary care and ACT health service statechart	48
5.5	Population statechart	48
5.6	DES of the ACT health clinical service pathway	48
5.7	Individual trajectory of BMI and corresponding K_{xgI} change over age without PLIs and Services	50
5.8	Individual trajectory of BMI and corresponding K_{xgI} change over age with DRI	51
5.9	Model simulated incidence of DIP and empirical data	55
5.10	Agent Person after model refactoring	57
5.11	Individual trajectory of BMI and corresponding K_{xgI} change over age under PHMSI	59
5.12	Prevalence of DIP in Scn1, Scn4, Scn3, and Scn2	60
5.13	Prevalence of DIP from the scenario assuming all children are either overweight or obese	60
5.14	Average of beta-cell mass in Scn1, Scn4, Scn3, and Scn2	60
5.15	Average of glycemia generated by the model in Scn1, Scn4, Scn3, and Scn2	61
5.16	Average of KxgI generated by the model in Scn1, Scn4, Scn3, and Scn2	61
5.17	Prevalence of DIP from the model in Scn1 showing uncertainty of model stochastics	62
6.1	SDM of <i>Culex</i> mosquito development	65
6.2	Posterior distribution of the number of <i>Culex</i> mosquitoes being trapped in Scn1	76
6.3	Prior distribution of the number of <i>Culex</i> mosquitoes being trapped in Scn1	77
6.4	Posterior distribution of the number of <i>Culex</i> mosquitoes being trapped in Scn2	78
6.5	Prior distribution of the number of <i>Culex</i> mosquitoes being trapped in Scn2	79
6.6	Distribution of <i>Culex</i> adult mosquito population in Scn1	80
6.7	Count of <i>Culex</i> mosquitoes being trapped in Scn3	81
6.8	Count of <i>Culex</i> mosquitoes being trapped in Scn4	82

LIST OF ABBREVIATIONS

ABM	Agent Based Model
ACF	Autocorrelation Function
ACT	Australian Capital Territory
ADIPS	Australasian Diabetes in Pregnancy Society
AG	Age Group
AUC	Area Under the ROC Curve
ATSI	Aboriginal and Torres Strait Islander
BMI	Body Mass Index
BMID	BMI Distribution
BSSID	Basic Service Set Identifier
CEU	Current ECig user
CS	Current Smokers
DES	Discrete Event Simulation
DGR	Dynamics of Glycemic Regulation
DIP	Daibetes in Pregnancy
DRI	Diet Review Intervention
ECDF	Empirical Cumulative Distribution Function
ECig	E-Cigarette
ECigUC	ECig Use Cessation
ECigUR	ECig Use Relapse
ECigUI	ECig Use Initiation
EM	Expectation-Maximization
FEU	Former ECig User
FIFO	First-In-First-Out
FS	Former Smoker
G	Glycemia
G1GC	Garbage First Garbage Collector
GDM	Gestational Diabetes Mellitus
HMMs	Hidden Markov Models
HPSI	Health Professional Support Intervention
IT	Insulin Treatment
KDE	Kernel Density Estimation
LC	Lifestyle Change

MAC address	Media Access Control Address
MBD	Mosquito-Borne Disease
MCMC	Markov chain Monte Carlo
MT	Metformin Treatment
MVC	Model-View-Control
NaN	Not a Number
NS	Never Smoker
NEU	Never ECig User
ODE	Ordinary Differential Equation
PACT	Personal Automatic Cigarette Tracker
PCEU	Prevalence of Current E-Cigarette User
PCS	Prevalence of Current Smokers
PF	Particle Filtering
PFS	Prevalence of Former Smoker
PHMMASI	Public Health Messaging and Mobile App Support Intervention
PHMSI	Public Health Messaging and Support Intervention
PLI	Population-Level Intervention
PMCMC	Particle Markov chain Monte Carlo
ROC Curve	Receiver Operating Characteristic Curve
RSSI	Received Signal Strength Indicator
SA	Sensitivity Analysis
SC	Smoking Cessation
Scn1	Scenario 1
Scn2	Scenario 2
Scn3	Scenario 3
Scn4	Scenario 4
SDM	System Dynamics Model
SI	Smoking Initiation
SN	Social Network
SR	Smoking Relapse
SSID	Service Set Identifier
SVM	Support Vector Machine
T2DM	Type 2 Diabetes Mellitus
T1DM	Type 2 Diabetes Mellitus
WNv	West Nile virus

CHAPTER 1

INTRODUCTION

Reflecting the heavy associated health burdens, a high amount of effort is currently expended in collecting and analyzing empirical evidence in health areas such as communicable illness, obesity, tobacco-related diseases, mental health, and health service delivery.

While traditionally conducted surveys have for a long time provided key support for collecting reliable evidence on health behaviors and exposures in the community, such surveys suffer from practical limits in the volume and type of information that can be gathered without overly burdening the participant; in addition, large-scale efforts dependent on such surveys are both encountering increasing difficulty in reaching potential respondents, and becoming increasingly expensive. At the same time, there is an increasing need for higher-resolution understanding of health behaviours and exposures due to the growing sophistication and diversity of analytic methods, including machine learning and other computational modeling methods such as simulation. To address both the growing challenge in data collection with traditional instruments and the increasing need for research evidence, a growing number of researchers seek to support health insights by supplementing traditional tools with combinations of “Big Data” sources. The availability of rich and reliable data provides supports for understanding public and population health and health care insights via analytic methods drawing on data science, system science, and computational science.

As an important contemporary example of Big Data, sensor data collected from smartphones allows less intrusive and – for many measurands – automatic data collection on participant’s behavior. Examples of such automatically-collected smartphone data include location data via GPS and Wi-Fi signal strength, physical activity data via processed pedometer data or lower-level motion sensors such as accelerometry and gyroscope readings, contact network information gathered using Bluetooth beacons, digital footprint data – such as information on whether the smartphone’s screen is on – and environmental data involving temperature or humidity measurements. In addition to the smartphone sensor data, the high prevalence and continuity of smartphone usage further makes attractive and lower-burden self-reporting from participants via ecological momentary assessments. Cross-linking multiple sensor types – and sensor data to self-reported data – allows researchers to assess behaviors, health outcomes, and elevated understanding of the dynamics of implemented interventions.

Fine-grained data generated from underlying processes generally sheds light on the causal structure of the dynamic and unobserved system by providing evidence regarding the dynamics of generative pathways

underlying observed impacts. Models, by comparing model-generated with empirical data, test the consistency of the behavior resulting from the hypothesized causal structure and evidence for the world. This can, in turn, provide support for studying the dynamics and evolution of the data generating process and the system-wide implication of the data sources. The dynamic model further provides a way to examine and estimate the system-wide outcomes for counter-factual situations and interventions in a complex system.

While the characteristics of data inform the selection of appropriate methods for analyzing it, the dynamic model structure and level of sophistication directly shape the ability to effectively utilize data, and constrain the analytic methods that can be used to exploit data with the dynamic model. Machine learning classifiers such as Hidden Markov models (HMMs) and recurrent neural network models are suitable for inferring categories, classifying the (generally latent) underlying state over time, and predicting outcomes. Such classifiers require large volumes of data concerning the character of discrete (e.g., categorical) state. Furthermore, the noisy and sparse observations collected by smartphones are frequently individually insufficient to conclusively identify the unobserved state at a given time, particularly in light of the continuing evolution between states over time. Therefore, uniformly-spaced time series data from smartphone sensors in contexts well-characterized by discrete (e.g., categorical) underlying states is often suitable for analysis with machine learning classifiers such as HMMs. However, HMMs incorporate important limitations, including linearity, and the assumption of memoryless transitions according to constant probabilities between discrete states in discrete time. Models seeking to characterize memoryful states, time-varying transition probabilities, and continuous data require recourse to other analytic approaches.

High velocity data plays an important role in grounding dynamic models offering continuous state, including nonlinear aggregate System Dynamics (ODE) models, and agent-based models (ABM). Beyond supporting a continuous state space, nonlinear models in general characterize a notably richer behavioural repertoire. ABMs further allow for capturing memoryful states and richer context, such as those associated with social networks, individual characteristics, and geography, and exploiting cross-sectional data points and statistics from individual-level longitudinal data for model parameterization and calibration of model results. For a dynamic model, the marked variability in model dynamics and the sophistication of the data interface further shape the appropriate methods to be used to exploit the data. Beyond parameterization and calibration, filtering can be used with dynamic models to estimate the distribution of system state given time-series data for one or more measurands, and to estimate evolving parameter values.

The objective of this thesis is to use case studies to investigate different levels of sophistication by which dynamic models can be combined with data, and further discuss the characteristics and availability of data in shaping model structure and scope, as well as the ways in which dynamic model structure affects the data that can be used. The case studies of this thesis that demonstrate combinations of dynamic modeling with data, include machine learning classifiers informed by transformed time series of smartphone sensor data, and dynamic models informed by longitudinal individual-level data via calibration and particle filtering and calibration used to integrate an aggregate ordinary differential equation/System Dynamics model (SDM) with

daily data representing samples from the natural environment. The paragraphs below briefly characterize these applications.

To lend meaning to collected time series of individual-level longitudinal data, the first case study employed multivariate HMMs to dichotomously classify smoking and non-smoking intervals, making use of binned time-series of transformed smartphone-based Wi-Fi, GPS and accelerometer sensor data. To support supervised learning, these data were labeled with self-reported smoking status collected with the same smartphone data collection app. With the character of data in this study, HMMs support identifying a set of discrete (categorical) states assumed to be memoryless according to the transitions between states, with the particular focus here being the estimation of the most likely sequence of latent states. To evaluate the use of data in machine learning classifiers, this work further assessed the sensitivity and specificity of univariate HMMs when compared with that of multivariate HMMs, and investigated the influence of tailoring training and test datasets to preserve complete smoking periods on model performance.

Individual-based dynamic models further allow for capturing memoryful states and richer context, such as those associated with network and geography, and for utilizing longitudinal data for calibration and parameter value. In the second and third case studies, an ABM and (separately) multi-scale hybrid simulation model were constructed to explore the potential for examining the causal structure of a system and interventions using data and dynamic models. Specifically, an ABM of smoking and E-Cigarette (ECig) use was constructed to examine the effects of ECig use and social networks on smoking behavior change. Separately, we further developed a multi-scale hybrid simulation model of diabetes in pregnancy (DIP). This model included as building blocks a SDM, an ABM and a discrete event simulation model (DES). The work extensively calibrated the model against empirical data. The SDM captured central physiological-level phenomena, by simulating the dynamics of glycemic regulation by β -cell and insulin resistance, and the evolution of prevailing insulin and glycemic levels, as well as beta cell mass and function. Continuous body mass index (BMI) evolution, dynamics of insulin sensitivity driven by pregnancy and weight change and birth, interventions, and diabetes classification were simulated by the ABM. Finally, DES was employed to characterize investigate the optimization of resource use in public health service delivery.

Due to the marked variability in Saskatchewan’s mosquito population dynamics year-to-year and within the course of a summer, the need to estimate continuous state in terms of the underlying mosquito population, and having detailed time-series data, the final case study applied the machine learning technique of particle filtering with a SDM of mosquito population to predict *Culex* mosquito adult population using various direct observations – including weather-related factors – and mosquito-related time series. In contrast to previous applications of particle filtering models to health-related dynamic models, this application of particle filtering required consideration of the marked temporal variation in exogenous (here, weather-related) factors affecting measurements of model state. As a result, the measurement model employed in the particle filtering included a complex likelihood function dependent on time-series data regarding such exogenous factors. This final work also made an innovative methodological contribution by extensive use of particle filtering in the

context of automated model calibration to deal with the uncertainties associated with mosquito dynamics and particularly the relationship between multiple weather-related factors and the per-day probability of trapping a mosquito – factors of key importance in reasoning about estimates of the abundance of mosquitoes captured in traps.

Levels of sophistication of model structure and by which dynamic models can be combined with data are considered in the presentation sequence used for the models. The multivariate HMMs support classifying a set of discrete states – whose emission probability densities are assumed to be memoryless, and with constant and memoryless transitions between states. A relatively rudimentary level of sophistication in interfacing between the model and the time-series data is demonstrated in this initial case study. Reflecting increasing model sophistication, ABM and hybrid models capture memoryful, continuous states and a richer behavioral repertoire. Reflecting these advantages, the thesis presents two case studies – an ABM of smoking and ECigs informed by cross-sectional data, and a hybrid multi-scale model of DIP using both cross-sectional data and longitudinal data. In both case studies, the model interfaces with the data in a relatively simple and traditional way – via model calibration. The richer data availability for the DIP model, by providing additional information for grounding model behavior, can considerably increase the level of sophistication of the model that can be evidenced. While having a simpler model structure than the hybrid model of DIP, the SDM of mosquito development employs a more sophisticated means of grounding the model in data – via Particle Filtering – and requires a more complex way of relating the empirical data to corresponding model outputs. The particle filtering recurrently regrounds the latent state of the dynamic model, based on a time series of empirical data treated as incoming observations, despite highly variable and environmentally-dependent data observation processes. The SDM notably further employs particle filtering in the context of model calibration to deal with the uncertainties associated with the evolution of the mosquito population, and particularly the complex and multifactorial dynamics involving use of the empirical data to estimate mosquito abundance.

1.1 Motivation

Smoking is one of the foremost public health threats, harming nearly every organ of the body, and the cause for many preventable diseases, e.g., lung cancer, coronary heart disease, chronic obstructive pulmonary disease, and other cardiovascular diseases [1, 2]. The detection of smoking behavior change is key to smoking surveillance and informing effective policy making. High prevalence and continuity of smartphone usage and mature smartphone sensor data collecting techniques make such sensor data a cheap, ubiquitous data source for automatic human activity recognition, and make that a strong candidate for allowing viable monitoring of smoking behavior [3]. However, an effective detection algorithm for smoking behaviors using fused sensor data from commodity smartphones remains elusive.

As a cigarette alternative, the ECig was introduced to the market in 2003, and its usage has subsequently

increased dramatically, due in part to the promotion and marketing by major tobacco companies in the last decade [4, 5]. An ECig vaporizes a liquid mixture as a substitute for tobacco leaves, and provides an aerosol for users to inhale [4, 6]. The use of ECigs amongst youth has exhibited a particularly dramatic and alarming rise [7]. The health behaviors associated with smoking and ECig use have been studied by many researchers, predominantly using self-reported surveys, in cohort studies and clinical trials [8, 9, 10]. Such studies are expensive and difficult to scale, and are typically far from real-time in character, being frequently associated with delays in obtaining results. Furthermore, such studies can be difficult to plan and execute, and their designs often exclude some factors or patterns of importance in the complex ecological context of smoking. Dynamic models play a key role in addressing such challenges by simulating complex social dynamics and behaviors with considerably high resolution [7]. ABMs can serve as an effective tool for investigating the impacts of counter-factual interventions by characterizing a high level of heterogeneity of individuals, and can further help prioritize data collection in a complex milieu of complex interactions of smoking and ECig use behaviors and choices regarding nicotine-containing products.

Gestational diabetes mellitus (GDM) poses challenge for public health in the Australian Capital Territory (ACT), particularly on account of its influence on future diabetes, elevating the risk of developing Type 2 Diabetes Mellitus (T2DM) across the population [11, 12] and in offspring [13, 14]. This problem is worsened by the increasing prevalence in the ACT – and in many developed countries – of established risk factors for GDM, including advanced maternal age [15], obesity [16], historical declines in physical activity, growing GDM risk factors in those with family history of diabetes, and a rising number of residents whose ethnicity group has traditionally been subject to elevated rates, which contribute to the increase in the prevalence of GDM [11, 17]. While earlier diagnostic screening improves the capacity for early case discovery, supporting early treatment and intervention, such benefits must be balanced with the increased resource demand this imposes on public health services. The health concerns associated with GDM increases and its risk factors have been studied in previous studies predominantly using cohort studies, administrative data or clinical trials [18, 19, 20]. The SDM constructed by Osgood et al. [14] offered a dynamic perspective on the inter and intragenerational interaction of GDM and T2DM, but was limited by the aggregate view of model, a difficulty in grounding the model by longitudinal data, and an inability to characterize a rich portfolio of interventions – including those depending on detailed patient risk factors, family context or history – or to address decision-making at multiple levels of the system (e.g., clinical, health services delivery, public health). While promising and filling some important research needs, the complex interactions among the risk factors and the underlying system including feedback, accumulations, delays, heterogeneity impose fundamental barriers for the use of aggregate studies in examining “what-if” questions and counter-factuals whose outcomes have not yet been observed, in reasoning about times-to-effect, scaling and other implementation science concerns, and in providing a timely evaluation of complex portfolios of interventions and treatments.

West Nile virus (WNV) infection is one of the leading causes of mosquito-borne disease (MBD) in Saskatchewan, and across Canada [21]. In many Northeast regions and within Saskatchewan, mosquitoes

in genus *Culex* [22] serve as a bridge vector of WNV between birds and humans, resulting in human cases of WNV infection [23, 24]. *Culex pipiens* is associated with the highest number of human cases in WNV infection in the northeast and north-central of the United States [23, 24]. *Culex pipiens* and *Culex restuans*, and *Culex tarsalis* are dominant vector in eastern Canada and in western Canada, respectively [25]. *Culex tarsalis* amplify the spread of infections as feeding preferences shifts from bird to human during the American robins (*Turdus migratorius*) migration period [23]. Control of *Culex* mosquitoes can greatly reduce the burden of WNV infection. Taken in isolation, current information or data about the abundance of mosquitoes offers insufficient evidence for public health entities to most efficiently control outbreaks of MBD [26]. Due to the marked variability in mosquito population dynamics year-to-year, complex weather-based influences on, and labour-intensive nature of trap-based estimation of mosquito abundance, and confounding effects of mosquito control measures, we were motivated to draw on a combination of a SDM of mosquito lifecycle with the machine learning technique of particle filtering to predict adult mosquito population. This approach used various direct observations, including mosquito-related time series and weather-related factors critically influencing both mosquito dynamics and the probability of mosquito detection in separate ways, to build a model capable of predicting mosquito populations and assessing potential interventions to effectively control the outbreaks of MBD.

1.2 Problem

To investigate use and challenges of different types of empirical data with appropriate advanced analysis methods – such as calibration, filtering – can aid in tasks such as classifying or estimating latent state, and predicting outcomes, the following problems need to be resolved.

- Various types of sensors from smartphones allow unobtrusive and automatic data collection on participant’s behavior; however, when collecting data with high granularity, the data contains noise and missing and sparse observations, e.g., the dominant invariant gravitational component in the readings from accelerometer. Without means of effectively handling it, such noisy data may adversely influence the accuracy of machine learning classifiers.
- Multivariate HMMs were informed by time-series observations from multiple types of smartphone sensor in this study. To improve the performance of our Multivariate HMMs, appropriate assumptions must be made about these multiple types of observations in terms of emission distributions associated with particular states.
- Model calibration processes using longitudinal and (alternatively) cross-sectional data involve “tuning” values of unknown or poorly-measured parameters to best match model outputs with observed data. Frequently during this process, missing assumptions were required to be identified, and incorporated new constraints or mechanisms into the model for better matching simulation results with empirical

data.

- Dynamic simulation models commonly include some parameters with no reliable information, but where other data is available for governing the emergent behavior of a system, e.g. cross-sectional data and time-specific data. While calibration can be employed for point estimation, the presence of pronounced stochastics makes it difficult to closely match observed patterns in a way that is desirable for projection of model behaviour, particularly because the calibration cannot reground the state of the model based on incoming observations.
- Dynamic modeling, particularly for extensive calibration, may require many realizations per parameter combination. Therefore, it is of high importance to employ an effective model configuration to reduce the computation cost and scale up the model for simulating large populations for a long time, e.g., in a ABM with networks and geography, or a multi-scale hybrid model.
- The particle filtering method recurrently regrounds estimates of dynamic model state by estimating latent state and dynamically evolving parameters based on incoming observations. The measurement process used to relate model outputs to empirical observations can be involved, and can depend on unfolding exogenous factors. While particle filtering can provide great value in estimating the latent state of the model, it does not directly address the need to estimate static parameters, including pronounced uncertainties regarding model outputs associated with bias in the measurement process; to address such cases, there is a need to combine particle filtering with other parameter estimation methods, such as calibration.

1.3 Contributions

The main contributions of this thesis are listed as follows:

- **Classifying smoking intervals using smartphone sensor data and multivariate HMMs.**
We fused various types of smartphone sensor data – including Wi-Fi, GPS, and accelerometer time series – after transformation of the raw data, and applied multivariate HMMs to identify smoking and non-smoking intervals. The fidelity of predictions from univariate and multivariate HMMs were compared using an evaluation metric of Area Under the Curve (AUC) and error rate. In terms of novel contributions, multivariate HMMs used problem-specific features to support more effective distinguish the underlying patterns of smoking behavior. The improvements in data conditioning achieved by the transformations and the identification of the need to tailor the analysis to entire smoking cycles serves an important role in elevating classifier performance.
- **Agent based modeling to investigate the effect of ECig use and social networks on smoking behavior change.**

An ABM was constructed to simulate complex interactions between smoking and ECig use, and further implemented a dynamic physical proximity network and examined its effect on smoking and ECig use initiation. Aggregated individual outcomes were collected to demonstrate the influences of network and ECig use. The model was calibrated against empirical data from 2013 to 2017 from the Canadian Tobacco and Alcohol and Drugs Survey [27].

- **Multi-scale simulation modeling for prevention and public health management of DIP and sequelae.**

Extending and elaborating a previously built model, an open population ABM enclosing a previously contributed SDM module was built to represent the population dynamics, evolution of continuous BMI as age cohorts shift, and diabetes classification. Offspring outcomes were refined based on the maternal hyperglycemia status and modifications were made to the equations to make insulin sensitivity over time dependent on pregnancy and weight status from other areas of the model. This work further identified a missing assumption and elaborated the equations giving the spontaneous recovery rate of the pancreatic beta cells for diabetes high-risk ethnicity groups. We built a novel portfolio of population-level interventions and clinical health service pathways in the ACT. The model was extensively calibrated against empirical data, and incorporated missing assumptions during the calibration process. The model was scaled up to simulate 200,000 agents over 93 years for 23.25min per realization, using 2.2 GHz Intel Core i7 and 16GB RAM, by modifying the model configuration. To enhance modularity and flexibility, the `Person` agent was refactored by separating the SDM module, weight change dynamics, and interventions as three types of agents, and encapsulating the three agents in the `Person`.

- **Particle Filter Applied to SDM for Mosquito Population Surveillance.**

We applied particle filtering to a previously built SDM of a *Culex tarsalis* mosquito population using various direct observations, including weather-related factors and mosquito-related time series. Addressing key elements of model parameter uncertainty, model calibration was combined with particle filtering for optimizing values of parameters to minimize the discrepancy following particle filtering. For the novel contribution, the resulting SDM model used both particle filtering and automated model calibration to deal with the uncertainties associated with the dynamics of Mosquito population and complex and multifactorial dynamics involving the number of mosquitoes being captured by mosquito traps used to estimate mosquito abundance, with the calibration reducing that discrepancy by more than a factor of 3 (from 336.84 to 110.287). Furthermore, linear regression was employed as an alternative approach to estimate the relationship of each type of weather factor and influences on the dynamics of number of mosquitoes being captured.

1.4 Thesis Outline

The structure of the remainder of the thesis is as follows:

- Chapter 2 provides background information on the methodologies employed by the models in this thesis, including a brief introduction of simulation modeling, HMMs, and particle filtering. The final section in this chapter consists of a literature review regarding previous studies on modeling of GDM and T2DM, smoking identification, and ABMs of smoking or ECig use.
- Chapter 3 presents a study investigating classification of smoking and non-smoking intervals using smartphone sensor data and multivariate HMMs. As a key enabler for this contribution, data pre-processing methods to yield better performance of the HMM are discussed in this chapter. Furthermore, taking advantage of the labeled data available in the supervised learning approach applied, emission probability distributions were estimated using kernel density estimation based on empirical distribution of three types of independent observations. Comparisons of the accuracy of multivariate HMMs and univariate HMMs are further provided.
- Chapter 4 introduces an ABM to examine the effect of ECig use and social networks on smoking initiation, relapse and cessation. This chapter further characterizes the use of model calibration to adjust select model assumptions so as to best match simulation outcomes with empirical data. Results from different scenarios and sensitivity analysis on model parameters are also presented in this chapter.
- Chapter 5 describes a multi-scale hybrid model that includes an SDM simulating an underlying physiological regulation of glycemic status based on beta-cell dynamics and insulin resistance, nested in an ABM depicting a dynamics of continuous BMI evolution, glycemic status change during pregnancy, and diabetes classification driven by the individual-level SDM. Furthermore, public health service pathways using DES are presented to explore the optimization of resource use during service delivery. The impact of interventions of varying levels of the system – clinical, health service delivery and population level – on health outcomes at the physiological, health service and population levels are further discussed. Calibration, model scalability, and model refactoring are also discussed and explored in this chapter.
- Chapter 6 describes an application of particle filtering for an SDM of a species-specific mosquito population using weather-related time-series on a daily basis. The chapter further includes modeling the effect of mosquito control on the abundance of mosquito larvae, and linear regression to estimate the probability of capturing *Culex* mosquito adult per trap. The distinctive work using both model calibration and particle filtering for optimizing values of parameters is also presented in this chapter. Finally, the results of particle filtering are presented and discussed.
- Chapter 7 includes following elements: A summary of motivation and conclusion of each chapter, the possible solutions for the problems encountered in this work, a discussion of major insights and findings, a brief statement of the contribution of this work, and a discussion of limitations and future work.

1.5 Publications

- Chapter 3 includes a manuscript of "Identifying Smoking from Smartphone Sensor Data and Multivariate Hidden Markov Models" by Yang Qin, Weicheng Qian, Narjes Shojaati, and Nathaniel Osgood, published in: Lee D., Lin YR., Osgood N., Thomson R. (eds) Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2017. Lecture Notes in Computer Science, vol 10354. Springer, Cham [3].

Yang Qin wrote the draft of this paper. Yang Qin obtained and processed the raw data. Yang Qin and Weicheng Qian contributed to implementing the HMM and validating the results. Narjes Shojaati contributed in collecting the raw data. Weicheng Qian and Nathaniel Osgood helped in editing the manuscript. Nathaniel Osgood supervised the study and advised on feature selection and building multivariate HMMs.

- Chapter 4 includes a manuscript of "Effect of E-cigarette Use and Social Network on Smoking Behavior Change: An agent-based model of E-cigarette and Cigarette Interaction" by Yang Qin, Rojiemiahd Edjoc, and Nathaniel Osgood, published in: Thomson R., Bisgin H., Dancy C., Hyder A. (eds) Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2019. Lecture Notes in Computer Science, vol 11549. Springer, Cham [7].

Yang Qin wrote the manuscript of this paper and contributed to building the model, obtained data from a website for model calibration. Rojiemiahd Edjoc provided empirical data and contributed the domain knowledge in the health aspect of this model. Nathaniel Osgood supervised the study and advised on the model building process and edited the manuscript.

- Chapter 5 includes a manuscript of "Multi-Scale Simulation Modeling for Prevention and Public Health Management of Diabetes in Pregnancy and Sequelae" by Yang Qin, Louise Freebairn, Jo-An Atkinson, Weicheng Qian, Anahita Safarishahrbijari, and Nathaniel Osgood, published in: Thomson R., Bisgin H., Dancy C., Hyder A. (eds) Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2019. Lecture Notes in Computer Science, vol 11549. Springer, Cham [17].

Yang Qin drafted the manuscript. Louise Freebairn contributed by project leadership, provision of empirical data, and advised on modeling and helped with drafting the manuscript. Jo-An Atkinson advised on the modeling process. Anahita Safarishahrbijari engaged in physiological model implementation, contributed to implementing a fast system dynamic solver applying Newton-Raphson numeric method and the mechanism of insulin resistance changing with pregnancy and treatment. Yang Qin contributed to an extensive model debugging session, model performance optimization, model calibration, model refactoring (logic-presentation separation), and modeling continuous BMI dynamics, refining the effect of hyperglycemia of mother on child insulin sensitivity, building DES of ACT health service pathways, and implementation of five population-level intervention mechanisms. Weicheng Qian advised on modeling and led the implementation of mechanisms towards achieving model scalability. Yang Qin

obtained and visualized the results. Nathaniel Osgood helped with the manuscript, supervised and advised on the entire study.

CHAPTER 2

BACKGROUND

This chapter provides an introduction to the methodologies employed by the models in this thesis and corresponding background knowledge. Section 2.1 and Section 2.2 describes the basics of HMMs and particle filtering, respectively. Section 2.3 introduces the basics of simulation models, including ABM, SDM, and DES. Section 2.4.1, Section 2.4.2, and Section 2.4.3 provide a brief introduction to GDM, smoking detection and ABM of smoking and ECig use, respectively.

2.1 Basics of Hidden Markov Models

HMMs, initially introduced in the late 1980s, are models in which the distribution that generates some observations depends on the state of a process satisfying the Markov property. The underlying process is posited to transition in discrete time between categorical states that are not directly observable, but at any given time, the probability distribution of possible observations depends on only on the state at that time. Each observation in the univariate and multivariate time series of observations is generally individually ambiguous, but collectively – taken in context – such observations can provide insight into the unobserved state. This capacity to make inferences in HMMs based on a broader context of observations at nearby points in time is limited by the fact that the system is shifting over time between the underlying states. The features of systems addressed by HMMs make it particular suitable for signal-processing application, especially in speech recognition, and other fields such as bioinformatics, environment, finance and biophysics [28, 29].

Given an HMM, the fundamental problems for model design are the evaluation of the probability of a sequence of observations, decoding the most likely sequence of hidden states underlying such a sequence, and training the HMM model parameters based on those observations [28].

An HMM has a functional form with the representation of Equation 2.3 [30], and consists of following parameters: (1) count of hidden states, m , (2) the count of types of observations, n (3) the count of successive observation time points, T , (4) the initial distribution $P(C_0)$, where $P(C_0) = P(S_i), 1 < i < m$, (5) the parameter related with a sequence of hidden states, transition probability distribution $P(C_t|C_{t-1})$, and (6) the emission probability distribution $P(X_t|C_t)$ [30]. The hidden states and observations of a HMM are denoted as sets $S = \{S_1, S_2, \dots, S_m\}$ with a size of m and $X = \{x_1, x_2, \dots, x_n\}$ with a size of n , respectively. The sequence of hidden states, $C = \{C_1, C_2, \dots, C_T\}$ where $C_t = S_i, 1 \leq i \leq m$, is assumed to satisfy the

Markov property, namely, the conditional probability distribution of state at time t of a process depends only on the most recent state at time $t - 1$ and is thus (conditionally) independent of the value of that state at times before $t - 1$, as specified by Equation 2.1 [28, 29].

$$P(C_t|C_{t-1}, \dots, C_1) = P(C_t|C_{t-1}), t = 2, 3, \dots, T \quad (2.1)$$

Each of the observations at time t , $\{X_t : t \in T\}$, is given as a vector $X_t = x_i, 1 \leq i \leq n$. The HMM formalism further assumes that the distribution of possible observations at time t ($\{X_t : t \in T\}$) depends purely the current hidden state C_t , given by the Equation 2.2, and is thus conditionally independent of, for example, both previous states of the process, and previous observations.

$$P(X_t|X_{(t-1)}, \dots, X_1, C_{(t-1)}, \dots, C_1) = P(X_t|C_t), 2 \leq t \leq T. \quad (2.2)$$

The joint distribution of both a posited state sequence C and the observations X is thus given by the following:

$$P(C, X) = P(C_0) \prod_{t=1}^T P(C_t|C_{(t-1)}) \prod_{t=1}^T P(X_t|C_t) \quad (2.3)$$

In practice, the transition probability $P(C_t|C_{(t-1)})$ is an $m \times m$ square matrix of probabilities, as shown in Equation 2.4 [28]:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix} \quad (2.4)$$

where $a_{ij} = P(C_t = S_j|C_{t-1} = S_i), 1 \leq i, j \leq m$, and $\sum_{j=1}^m a_{ij} = 1$. The probability distribution of an observation X_t being generated from a state C_t , $P(X_t|C_t)$, called the emission probability, can be a custom density function according to a specific HMM; when the possible observations for a particular observation type are discrete, this can be expressed by a emission matrix, but for continuous such observations – such as those employed for some measurands within this thesis – continuous distributions are employed.

2.1.1 Forward and Backward Probabilities

An given application of an HMM can be described as a sequence of hidden states $C = \{C_1, C_2, \dots, C_T\}$, a sequence of observations $X = \{X_1, X_2, \dots, X_T\}$, and a set of parameters $\theta = (A, B, \pi)$, where $A = a_{ij}$ is transition probability matrix as shown in Equation 2.4, $B = b_{S_i}(X_t), b_{S_i}(X_t) = P(X_t = x_j|C_t = S_i)$ where $1 \leq i \leq m, 1 \leq j \leq n$, is the likelihood of observing X_t considering the current state S_i , and π represents $P(C_0)$, the initial distribution of states.

Following Rabiner's notation convention in [28, 29, 31], the forward propagation computes $P(X_{1..t}, C_t | \theta)$, the probability of observation sequence to time t by summing the probabilities of all possible sequences of hidden states that generate the observation sequence available up to and including time t . Based on the definitions of hidden states C , a sequence of observations X and HMM parameters $\theta = (A, B, \pi)$, in the forward propagation, the probability of being in state C_t at time t having observed the observation sequence $\{X_1, X_2, \dots, X_t\}$, and given the HMM parameter θ , can be represented as Equation 2.5, and calculated as Equation 2.6.

$$\alpha_t(S_i) = P(X_1, X_2, \dots, X_t, C_t = S_i | \theta), \quad t = 1, 2, \dots, T \quad (2.5)$$

$$\alpha_t(S_i) = \sum_{j=1}^m \alpha_{t-1}(S_j) a_{ji} b_{S_i}(X_t), \quad t = 2, 3, \dots, T \quad (2.6)$$

where $\alpha_{t-1}(S_i)$ is the previous forward probability from previous time step $t-1$, a_{ji} is transition probability from state S_j to state S_i . Note that with this notation system, initialization happens at $t = 1$, rather than $t = 0$. At $t = 1$ hidden states are distributed according to π and immediately observation correction S_1 kicks in. The algorithm for computing the forward probabilities is as follows:

- Initialization:

$$\alpha_1(S_i) = \pi_{S_i} b_{S_i}(X_1), \quad 1 \leq i \leq m \quad (2.7)$$

- Recursion:

$$\alpha_t(S_i) = \sum_{j=1}^m \alpha_{t-1}(S_j) a_{ji} b_{S_i}(X_t), \quad 1 \leq i \leq m, 1 < t \leq T \quad (2.8)$$

- Termination:

$$P(X | \theta) = \sum_{i=1}^m \alpha_T(S_i) \quad (2.9)$$

Similar to the forward procedure, the backward probability for time t is defined as the probability of ending partial observation sequence from time $t+1$ to the end time T , given its presence in starting state S_i at that time t and the HMM parameter θ , as shown in Equation 2.10.

$$\beta_t(S_i) = P(X_{t+1}, X_{t+2}, \dots, X_T | C_t = S_i, \theta) \quad (2.10)$$

The backward probability is computed by following the steps.

- Initialization:

$$\beta_T(S_i) = 1 \quad (2.11)$$

- Recursion:

$$\beta_t(S_i) = \sum_{j=1}^m \beta_{t+1}(S_j) a_{ij} b_{S_j}(X_{t+1}), \quad 1 \leq i \leq m, 1 < t \leq T \quad (2.12)$$

- Termination:

$$P(X | \theta) = \sum_{j=1}^m \pi_{S_j} b_{S_j}(X_1) \beta_1(S_j) \quad (2.13)$$

2.1.2 Baum-Welch Algorithm

The Baum-Welch algorithm, a special case of expectation-maximization (EM) algorithm, is a standard algorithm employed to answer the third fundamental problem in HMM, namely, training the HMM to estimate parameters that maximize the likelihood of observation from a training set. There is no proper solution to maximize the probability of the entire observation sequence [28], but the EM algorithm can choose the HMM parameters which locally maximize the probability of an observation sequence [28] using forward and backward procedures [29, 32].

The Baum-Welch algorithm employs the forward and backward probabilities to compute the estimated transition probability \hat{a}_{ij} defined in Equation 2.14 and estimated emission probability $\hat{b}_{S_i}(x_k)$ defined in Equation 2.15

$$\hat{a}_{ij} = \frac{\text{expected count of transition from } S_i \text{ to } S_j}{\text{count of possible transitions from } S_i} \quad (2.14)$$

$$\hat{b}_{S_i}(x_k) = \frac{\text{expected count of times observation } x_k \text{ is observed in state } S_i}{\text{expected count of states } S_i \text{ of hidden state sequence}} \quad (2.15)$$

To compute the numerator in Equation 2.14, the probability of being in state X_i and X_j at time t and $t+1$, respectively, is defined and calculated by Equation 2.16, given the observation sequence X and θ .

$$\xi_t(S_i, S_j) = P(C_t = S_i, C_{t+1} = S_j | X, \theta) = \frac{\alpha_t(S_i) a_{ij} \beta_{t+1}(S_j) b_{S_j}(X_{t+1})}{\sum_{j=1}^m \alpha_t(S_j) \beta_t(S_j)} \quad (2.16)$$

Thus, the \hat{a}_{ij} is computed using Equation 2.17.

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(S_i, S_j)}{\sum_{t=1}^{T-1} \sum_{k=1}^m \xi_t(S_i, S_k)} \quad (2.17)$$

For the estimation of emission probability, the Equation 2.18 defines the probability of being in state S_i at time t , given the observation sequence X and parameter θ . And $\gamma_t(S_i)$ is calculated by Equation 2.19.

$$\gamma_t(S_i) = P(C_t = S_i | X, \theta) \quad (2.18)$$

$$\gamma_t(S_i) = \frac{P(C_t = S_i, X | \theta)}{P(X | \theta)} = \frac{\alpha_t(S_i) \beta_t(S_i)}{P(X | \theta)} \quad (2.19)$$

Therefore, the estimated emission probability $\hat{b}_{S_i}(x_k)$ is computed by Equation 2.20. The numerator in Equation 2.20 represents summing the times of seeing observation x_k in state S_i for all time steps, and the denominator represents the expected number of times in state S_i .

$$\hat{b}_{S_i}(x_k) = \frac{\sum_{t=1, X_t=x_k}^T \gamma_t(S_i)}{\sum_{t=1}^T \gamma_t(S_i)} \quad (2.20)$$

Starting with some estimated HMM parameters θ , the Baum-Welch algorithm iteratively runs the expectation step and maximization step to re-estimate the transition probability and emission probability until some convergence criterion has been met. In expectation step, all the quantities $\xi_t(S_i, S_j)$ and $\gamma_t(S_i)$ are

replaced by their conditional expectations, given observation sequence X and the current estimated parameters. In the maximization step, the ξ and γ from the expectation step are employed to recompute the HMM parameters using Equation 2.17 and Equation 2.20.

2.1.3 Viterbi Algorithm

With a trained model, the Viterbi dynamic programming algorithm is used to “decode” observations emitted from hidden states to get an estimated sequence of the hidden states [28, 33]. The equations in this section were derived by L.R. Rabiner [28].

To implement this method, the best estimated sequence of hidden states is denoted as $C = \{C_1, C_2, \dots, C_T\}$. The observation sequence which needs to be decoded is denoted as $X = \{X_1, X_2, \dots, X_T\}$. The idea of the Viterbi algorithm is to compute the highest probability single path along the observation sequence left to right at time t . Therefore, given HMM parameter θ , the maximum probability of being in state S_i given the observation sequence until time t in light of the highest probability state sequence $\{C_1, C_2, \dots, C_{t-1}\}$ is defined as Equation 2.21.

$$\delta_t(S_i) = \max_{\{C_1, C_2, \dots, C_{t-1}\}} P(\{C_1, C_2, \dots, C_{t-1}\}, \{X_1, X_2, \dots, X_t\}, C_t = S_i | \theta) \quad (2.21)$$

Notably, the most probable sequence is identified by taking the maximum over all possible previous state sequences, $\max_{\{C_1, C_2, \dots, C_{t-1}\}}$. To decode the state sequence with the highest probability, as a dynamic programming algorithm, the argument $\delta_{t+1}(S_j)$ in Equation 2.22, needs to be maximized at each time t over state S_j at time t . Vector $\psi_t(S_j)$ is used to keep track of the index associated with $\delta_{t+1}(S_j)$ for each successive time t .

$$\delta_{t+1}(S_j) = \max_{i=1}^N \delta_t(S_i) a_{ij} \cdot b_{S_j}(X_{t+1}) \quad (2.22)$$

Similarly, with the forward algorithm, the steps for a complete Viterbi algorithm are divided into three steps, initialization, recursion, and termination. In addition to these three steps, the Viterbi algorithm has state sequence backtracking steps, which makes it different from previous algorithms. The steps for the Viterbi algorithm are demonstrated as follows:

- Initialization:

$$\delta_1(S_i) = \pi_{S_i} b_{S_i}(X_1), 1 \leq i \leq m \quad (2.23)$$

$$\psi_1(S_i) = 0, 1 \leq i \leq m \quad (2.24)$$

- Recursion:

$$\delta_t(S_i) = \max_{j=1}^m \delta_{t-1}(S_j) a_{ij} b_{S_i}(X_t), 2 \leq t \leq T, 1 \leq i \leq N \quad (2.25)$$

$$\psi_t(S_i) = \arg \max_{j=1}^m \delta_{t-1}(S_j) a_{ij} b_{S_i}(X_t), 2 \leq t \leq T, 1 \leq i \leq N \quad (2.26)$$

- Termination:

$$\text{The best probability: } P^* = \max_{1 \leq i \leq m} \delta_T(S_i) \quad (2.27)$$

$$\text{The state sequence backtracking: } q_T^* = \arg \max_{1 \leq i \leq m} [\delta_T(S_i)] \quad (2.28)$$

The implementation of the Viterbi algorithm can be visually expressed by a trellis diagram, and the Viterbi path is the shortest path through the trellis diagram of the states of HMM.

In Chapter 3 of this work, each state (smoking or non-smoking state) has three types of observations corresponding to readings from Wi-Fi, accelerometer and GPS sensors. The probabilities of such observations follow an empirical distribution, and observations within a given state are assumed to be independent of each other, conditional on being in that state. So the customized density function of the emission probability for the multivariate HMM is:

$$b_{S_j}(X_t) = P^{\text{Accel}}(X_t|C_t = S_j) \times P^{\text{Wi-Fi}}(X_t|C_t = S_j) \times P^{\text{GPS}}(X_t|C_t = S_j) \quad (2.29)$$

2.2 Particle Filtering

2.2.1 Basics of Particle Filtering

Particle filtering is a sequential Monte Carlo methodology recursively calculating the approximation of the posterior probability distribution using the concepts of sequential importance sampling in the framework of a state-space model with discrete random measures defined by weighted particles [34,35]. Both particle filtering and HMMs assume an underlying Markov process that cannot be directly observed, and for both approaches, observations are generally individually ambiguous to give insights to the hidden state, but collectively the observations are able to pin down the hidden state. Particle filtering can be combined with state-space models having continuous states, while HMMs are applied to the state-space models having a set of discrete states. The state-space model can be non-linear in particle filtering applications, and tight distributional assumptions about the state-space model, measurement process – such as those seen in Kalman Filtering – are not required for generating samples represented by weighted particles from a distribution.

The basic equation and procedures for building a particle filtering model are summarized as follows [35]. In an important change of notation from the coverage of the HMMs above, for a given time t , let X_t and Y_t represent a vector of states and a vector of observations, respectively. The main task of particle filtering is to estimate $P(X_t|X_{0:t})$ given by Equation 2.30 from $P(X_{t-1}|X_{0:t-1})$, which can be calculated recursively by Equation 2.31. Notably, the denominator in Equation 2.31, $P(Y_t|Y_{0:(t-1)})$, is a constant.

$$P(X_{0:t}|Y_{0:t}) = P(X_0|Y_0) \prod_{k=1}^t P(Y_k|X_k)P(X_k|X_{k-1}). \quad (2.30)$$

$$P(X_{0:t}|Y_{0:t}) = \frac{P(Y_t|X_t)P(X_t|X_{t-1})}{P(Y_t|Y_{0:(t-1)})}P(X_{0:(t-1)}|Y_{0:(t-1)}) \quad (2.31)$$

In particle filtering, the particles and their normalized weights $\chi = \{x^{(m)}, w^{(m)}\}_{m=1}^M$ are employed to estimate – via sampling – the distribution of interest where $x^{(m)}$, $w^{(m)}$ and M are particle, weight assigned to the particle and count of particles used to approximate the probability distribution $P(x)$. $P(x)$ is approximated according to the principles of importance sampling by Equation 2.32, where δ is the Dirac delta function.

$$P(x) \approx \sum_{m=1}^M w^{(m)} \delta(x - x^{(m)}) \quad (2.32)$$

The principle of importance sampling plays an important role in obtaining χ_t . Given χ_{t-1} and the observation Y_t , χ_t can be obtained by generating particles $x_t^{(m)}$, and appending the $x_t^{(m)}$ to $x_{t-1}^{(m)}$, and updating the weight $w_t^{(m)}$ of the particle $x_t^{(m)}$. As directly sampling particles from $P(x)$ is unattainable, using importance sampling the particle $x^{(m)}$ can be sampled from a known (proposal) distribution via importance function $\pi(x)$, and the weight $w^{(m)}$ is calculated and normalized by Equation 2.33 and Equation 2.34, respectively.

$$w^{*(m)} = \frac{P(x)}{\pi(x)} \quad (2.33)$$

$$w^{(m)} = \frac{w^{*(m)}}{\sum_{i=1}^M w^{*(i)}} \quad (2.34)$$

Applying the $\pi(x)$ in the context of state and observation in particle filtering, the $\pi(x)$ can be re-factored according to Equation 2.37. It is worth noting that Equation 2.36 results from applying chain-rule on Equation 2.35, and Equation 2.36 can be further simplified to Equation 2.37 in sequential update.

$$\pi(X_{0:t}|Y_{0:t}) = \pi(X_{0:t-1}, X_t|Y_{0:t-1}, Y_t) \quad (2.35)$$

$$= \pi(X_t|X_{0:t-1}, Y_{0:t-1}, Y_t) \pi(X_{0:t-1}|Y_{0:t-1}, Y_t) \quad (2.36)$$

$$= \pi(X_t|X_{0:t-1}, Y_{0:t}) \pi(X_{0:t-1}|Y_{0:t-1}) \quad (2.37)$$

The trajectory of the particle $x^{(m)}$ from time 0 to time $t-1$, $x_{0:t-1}^{(m)}$, and its corresponding weight at time $t-1$, $w_{t-1}^{(m)}$, can be expressed as Equation 2.38 and Equation 2.39, respectively.

$$x_{0:t-1}^{(m)} \sim \pi(X_{0:t-1}|Y_{0:t-1}) \quad (2.38)$$

$$w_{t-1}^{(m)} = \frac{P(x_{0:t-1}^{(m)}|Y_{0:t-1})}{\pi(x_{0:t-1}^{(m)}|Y_{0:t-1})} \quad (2.39)$$

The particles and their corresponding weights can then be updated by performing the following two steps, given by Equation 2.40 and Equation 2.41, respectively.

$$x_t^{(m)} \sim \pi(X_t|x_{0:t-1}^{(m)}, Y_{0:t}) \quad (2.40)$$

$$w_t^{(m)} = \frac{P(Y_t|x_t^{(m)})P(x_t^{(m)}|x_{t-1}^{(m)})}{\pi(x_t^{(m)}|x_{0:t-1}^{(m)}, Y_{0:t})}w_{t-1}^{(m)} \quad (2.41)$$

Another important concept in particle filtering is resampling, which is used to avoid the degeneracy of the discrete random measures in particle filtering. The resampling method operates according to the “survival of the fittest” principle, removing the particles with small weight, replicating the particles with large weight, and assigning the weight to the particles so replicated. To be noted, before sampling particles for the next time step, in practice the resampling process is recommended to take place when the effective particle size falls below a predefined threshold [35,36]

2.2.2 Application of Particle Filtering in the Proposed Model

Within this work, the particle filtering method is employed to estimate the posterior distributions of the unobserved states in the state-space model, given that the observations depend on the current state in an evolving process. In our model, the unobservable state sequence satisfies the Markov property, namely, the conditional probability distribution of state at time t (X_t) of a process depends only on the most recent state at time $t - 1$, as shown in Equation 2.42.

$$P(X_t|X_{t-1}, X_{t-1}..., X_1) = P(X_t|X_{t-1}) \quad (2.42)$$

$P(Y_t|X_t)$, the probability of making observation Y_t at time t given the state variable X_t , provides insight into the state variable X_t .

A particle and its corresponding weight are employed to estimate the distribution of interest, $P(x)$. As it is unable to directly sample particles from $P(x)$, the particle filtering method utilizes the importance sampling approach to sample particles from the proposal distribution ($\pi(x)$) to make inference on sampling particles directly from $P(x)$ by associating particles sampled from $\pi(x)$ and weights. The particle filtering algorithm applied in this thesis is summarized as follows [37].

1. At time $t = 1$, $1 \leq m \leq M$, M is the total count of particles.

- (a) Assign each particles with equal weight $\{x_1^{(m)}, \frac{1}{M}\}$.
- (b) Sample particles $x_1^{(m)}$ from a uniform distribution. The process of sampling particles is achieved by running the simulation model forward on each particle.
- (c) Compute weights $w_1^{(m)}$ of the particles using $w_1^{(m)} = \frac{P(Y_1|X_1)P(X_1)}{\pi(x_1^{(m)}|Y_1)}$

2. At time $t \geq 2$, update particle weight as follows:

- (a) Sample particles $x_t^{(m)}$ from $\pi(x_t^{(m)}|X_{1:t-1}, Y_t)$. The process of sampling particles is achieved by running the simulation model forward on each particle.

(b) Compute weights of the particles using $w_t^{(m)} = \frac{P(Y_t|x_t^{(m)})P(x_t^{(m)}|x_{t-1}^{(m)})}{\pi(x_t^{(m)}|x_{t-1}^{(m)}, Y_t)}w_{t-1}^{(m)}$, and update particle weight $\{x_t^{(m)}, w_t^{(m)}\}$ to reflect the posterior distribution of the current state. In our model, the proposal distribution of $P(X_t|X_{t-1}) = \pi(X_t|X_{t-1}, Y_t)$ is employed and the particle weight update equation is simplified to $w_t^{(m)} = w_{t-1}^{(m)}P(Y_t|x_t^{(m)})$; that is, for a given particle of index m , the weight at time t is simply equal to the previous weight for that particle times the value of the likelihood function when applied to the observation in light of the state of particle m .

(c) Normalize particle weights: $w_t^{(m)} = \frac{w_t^{(m)}}{\sum_{m=1}^M w_t^{(m)}}$

3. Resampling step: During the update at time t , the resampling process takes place when the effective particle size falls below a predefined threshold k , which will occur when the model exhibits sufficiently high variance in particle weight ($\frac{1}{\sum_{i=1}^N (w_t^{(m)})^2} < k$) [38].

2.3 Introduction to Simulation Modeling

Complex systems usually behave in non-linear and thus often unexpected ways [39], consist of multiple behaviors and interactions, and other complex sub-systems [40], making the system expensive or impractical to experiment with interventions or study its unpredictable outcomes [41]. Simulation modeling methods employ mathematical characterization (optionally in the form of explicit equations) to represent the operation of a system or a process. Such models can be applied to study of the behavior of the actual system, evaluating and optimizing the performance of a system, and experimenting with the interventions and their corresponding impact over longer period in various scenarios [40,41]. There are three particularly prominent simulation paradigms within the sphere of Health and Health Care: SDM, ABM, and DES, which will be briefly sketched in the subsections below.

2.3.1 System Dynamics Models

The SDM technique is a modeling approach used to develop and design a set of differential and algebraic equations to represent and capture complex patterns of a system by feedback and accumulation of individual components [40, 42]. The core components of SDM consist of the elements representing state variables and accumulations (stocks), rates of change in such state variables (flows), causal loops involving stocks (feedbacks), and time delays (typically emerging from the accumulations) [40,42, 43]. While SDM can be applied at different scales [44], at any given level of analysis, the stocks – by virtue of being “well-mixed” and memoryless – make SDM an aggregate approach at that level. Thus, instead of describing specific characteristics of individual actors in underlying populations, SDM simulates over time the patterns in the evolution of the system composed of groups of such individuals. The rates represented by a set of interlocking algebraic equations are feed in and out of the stocks; such algebraic equations ultimately represent

a convenience, and the underlying dynamics can be captured via variable substitution in the underlying stocks and flows in order to form a set of ordinary differential equations. In some cases, the causal loops in SDM can exhibit non-linear complex relationships between stocks and flows [40]. In general, given the dynamically complex interaction between the stocks, flows, and feedbacks in an actual system, it is convenient to employ aggregate SDM to explore places to intervene in a system, to answer high-level “what-if” questions related with both clinical-level and population-level interventions whose outcomes have not yet been studied carefully, and examine the vulnerabilities or leverage points associated with system behaviour.

2.3.2 Agent Based Modeling

By contrast to SDM, ABM is a simulation method for modeling complex dynamic systems by characterizing the interactions between the autonomous and interacting objects – referred to as agents – within the environment of the system [40], and where behaviour of the system as a whole is emergent from diverse such interactions. Composing the core of an ABM, the agents determine their actions according to the defined decision rules based on the heterogeneous characteristics of agents and their context, including the current state of agents evolving along with multiple aspects of states and transitions, agent history, structured interactions between agents [44, 45] as it affects and is affected by networks, geography, and other components of the system.

ABM is tool well-suited for studying social epidemiology [46] and noncommunicable diseases and understanding public health risk factors [47]. The outputs of an ABM model can consist of individual-level trajectories (e.g., the simulated health outcomes) as well as aggregate level outcomes such as population health outcome indicators, prevalence, and incidence of diseases, costs, and resource utilization [40]. The key advantages of ABM includes its ability to take into account individual history and characteristics in decision making, probing the impacts of counter-factual interventions and health outcomes, and its capacity to interact via multiple networks or spatial context, to capture aspects of decision making given local situation or resources, and to scalably support heterogeneity of agent. ABM supports the nested model structure, extending an ABM to a multi-scale model to provide an effective way to explore complex dynamics at a fine-grained level.

2.3.3 Discrete Event Simulation

DES is a simulation paradigm used for characterizing agent resource-limited progression through structured workflows associated with service delivery, queuing processes, networks of queues, and waiting times with an emphasis on resource utilization [40]. Five levels of features, simple client or server, and First-in-first-out (FIFO) queues, decision making, service scheduling, queuing and entities encompass a coherent set of concepts about DES models [48]. Core concepts in DES modeling include of entities, attributes, queues, events, activities, delay and resources. An entity is a representation of an object having specific attributes and using the resources. Attributes are the specific properties of a given entity. In contrast to SDM – where

the emphasis is traditionally placed on patterns over time rather than occurrences of specific events – events are of central importance in DES; the occurrence of such an event can instantaneously change the state of a system (e.g., change resource use and entity characteristics). Queues occur at the interaction of several entities and help manage competition amongst entities for consuming limited resources.

2.4 Literature Review

2.4.1 Models of Gestational Diabetes Mellitus and Type 2 Diabetes Mellitus

Simulation models characterizing diabetes progression, the complications of diabetes at an aggregate level have been built by many researchers [14, 49, 50, 51, 52, 53], as have models of the underlying physiology between beta-cell, insulin, and glucose [54, 55]. By contrast, very few models have examined the dynamics of Gestational Type 2 Diabetes, with work of [14] being a notable exception.

Jones et al. developed an aggregated SDM of diabetes population [50]. The model characterized the dynamics of population with respect to stocks representing normoglycemia, prediabetes, uncomplicated diabetes, and complicated diabetes. Furthermore, the model simulated the influences (e.g., high-level interventions) over the rates of population flow. The simulation outputs addressed relationships between the dynamics in the prevalence of diabetes, mortality and diabetes control and primary prevention efforts. Osgood et al. [14] constructed a simulation model characterizing GDM and its contribution to T2DM. This SDM stratified by age-, sex- and in-utero exposure status characterized the inter- and intragenerational interaction of GDM and T2DM by specifying population flows between different possible health status. The population in the model flows through seven compartments including aspects of aging, pregnancy, onset of GDM, weight increase, and development of T2DM. The model results demonstrate the linkage between GDM and T2DM both within a given generation and between generations, suggesting that GDM may be an important factor for contributing to the growth of T2DM, and – by extension – that treatment of GDM may substantially reduce the burden of developing T2DM. T2DM can lead to a series of complications. Gao et al. [51, 52] constructed a hybrid simulation model to investigate tradeoffs between different intervention strategies for diabetes and diabetic end stage renal disease. The model utilized SDM for simulating diabetic progression and generated data for feeding into ABM to support some key functionality at the individual level. DES was used in the model for investigating health care processes. The simulated output matched closely with empirical data, and demonstrated the ability of the hybrid simulation model for forecasting outcomes under various scenarios, and evaluate the effectiveness of different interventions.

De Gaetano et al. [56] introduced a mathematical model of the physiological-level progression of T2DM over both rapid and longer periods of the disease, representing the physiological adaptation to developing insulin resistance due to the interaction between pancreatic islet compensation, beta-cell mass, and glycemia in healthy and diabetic individuals. The model formulated the pancreatic islet compensation mechanism and solved some fundamental qualitative characteristics based on refined physiological assumptions. DeGaetano

et al. [57] further extended the diabetes progression model, emphasizing formulation of physiological compensation for worsening insulin resistance in the glucose-insulin system. Hardy et al. [58] further used the model to simulate the trajectory of diabetes progression and characterized the mechanisms of anti-diabetic intervention, evaluating the effectiveness of the interventions in terms of improvements in insulin sensitivity or beta-cell protection. The model was well calibrated, and simulation results concerning antidiabetic interventions are consistent with the improvement of glycemia regulation observed in traditional treatment.

2.4.2 Smoking Detection

Approaches for smoking identification have been studied and analyzed by many researchers [59, 60, 61, 62]. Sazonov et al. [59] designed a wearable sensor system (Personal Automatic Cigarette Tracker PACT) and applied that system in investigating the unique pattern of smoking. The study further demonstrated the potential of using the signal received by the PACT in identifying smoking inhalation patterns and assessing smoking frequency. Lopez-Meyer et al. [62] demonstrated the feasibility of monitoring smoking using continuous breathing and hand-to-mouth gesture data collected by a wearable sensor system and the support vector machine classifier. The paper examined two classification models: a subject-independent model, and subject-dependent model, and assessed their performance. The subject-dependent classification model archived better average precision and recall (e.g., 90%) than the subject-independent model. Scholl and Laerhoven [60] presented a study of the feasibility of smoking detection using gesture data collected by an accelerometer device on the wrist.

2.4.3 Agent Based Model of Smoking or E-Cigarette Use

Cherng et al. [63] implemented an ABM to discuss the effect of ECig on the prevalence of smoking among adults in the United States. The ABM further simulated the effects of smoking initiation, cessation, and relapse on ECig use behavior change. The simulation results suggested that the effects of ECig use on smoking cessation changed smoking prevalence in a more pronounced manner than that on smoking initiation. Chao et al. [64] developed an ABM reproducing the historic smoking trends in Japan, to examine the effect of socioeconomic disparity of different gender groups on the prevalence of smoking. The study concluded that greater disparity in socioeconomic status among males increases the chance of reducing the prevalence of smoking, but females show exactly the opposite pattern.

CHAPTER 3

IDENTIFYING SMOKING FROM SMARTPHONE SENSOR DATA AND MULTIVARIATE HIDDEN MARKOV MODELS

The text presented in this chapter is from the manuscript "Identifying Smoking from Smartphone Sensor Data and Multivariate Hidden Markov Models" by Yang Qin, Weicheng Qian, Narjes Shojaati, and Nathaniel Osgood, published in Proceedings of the 2017 Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation. The contribution of each author is presented in Section 1.5 of Chapter 1 [3].

Hidden Markov models (HMMs) support identifying a set of discrete states – assumed to be memoryless – according to the transitions between states, using the observations emitted from the latent states, with the particular focus here being the estimation of the most likely sequence of latent states. A relatively rudimentary level of sophistication of interfacing with between a dynamic model and the empirical data are demonstrated in this chapter. This simplicity reflects several strong limitations of HMMs – the memoryless character of the assumed states, the linear character of the transition process and the assumption of time-invariant transition probabilities. This case study related the model to various types of smartphone sensor data – including Wi-Fi, GPS, and accelerometer time series using a supervised learning approach. The raw data was separated into smoking- and non-smoking time periods according to the “ground truth” data in the form of self-reported smoking status collected with the same smartphone data collection app. The resulting distributions serve as the probability distributions for observations emitted from the discrete underlying states of HMMs, and are fused and applied in multivariate HMMs to estimate model parameters and identify latent smoking and non-smoking intervals.

3.1 Introduction

Smoking is one of the biggest public health threats listed by the World Health Organization. Effective tools for smoking recognition can ensure public health surveillance for policy making [65], provide early detection before addiction [66], and aid former smokers to avoid relapse. Detection of smoking status has long relied on biomedical assays based around the detection of substances such as cotinine, nicotine [67], carbon monoxide [68], and respiration [69]. Application of such assays normally requires mildly to moderately invasive measurements, from breath tests to provision of saliva to clippings of hair, and many test results are

available only after delays measured in days or more.

Studies using wearable sensors for recognition of smoking [59, 62] have shown potentials to avoid the invasiveness of measurements and delays of test results to perform seamless online detection. These studies predominately based on hand-to-mouth gestures and breathing pattern and using specialized hardware to collect data, which can be costly and hard to comply continuously, informativeness from other aspects correlates of smoking have not yet been considered, such as presence outdoors (as required by regional regulations) or designated smoking areas, activity levels, and characteristic length of dwelling period correlated to the burning time of cigarettes.

In recent years, and paralleling their rapid penetration across diverse strata of society worldwide, the smartphone has become an attractive platform for the sensor-based data collection on human behavior. The use of such techniques has been enhanced by the growing maturity of data collection apps (such as the iEpi system [70], UPenn’s DREAM project) that make smartphone sensor data a highly available and easily accessible data source for many studies [71]. Feasibility studies on using the accelerometer sensor to detect smoking behaviors have been initiated [60], but published studies on fusing sensor data available on smartphones remain absent.

In this paper, we fused various types of sensor data commonly available on smartphone, after considering data completeness, accuracy, and informativeness, examined the effects of five transformations for GPS, Wi-Fi and accelerometer sensor data, and applied multivariate HMMs to classify periods to recognize whether smoking was taking place. Finally, we investigated the performance of univariate HMMs and multivariate HMMs and the impact of tailoring the training and test set to preserve entire smoking cycles.

3.2 Data Processing and Algorithm

Dataset Description

Data used in the project came from a previously conducted Behavioural Ethics Board-approved study that collected multiple types of sensor data together with self-reported ground truth on smoking behavior by four participants who were self-reported smokers. The dataset contains labeled data on segments of intervals of smoking and non-smoking periods. The sensor data was continuously collected from four participants for one month with a five-minute duty cycle [70, 72]. The position features (Wi-Fi and GPS) and a set of gesture and orientation related features (gyroscope, and accelerometer data with three directions) were collected by Ethica system [70, 72]. Given the potential interest in vehicular context, data from the magnetometer was also collected. There are 36 million records, 0.3 million records, 1.9 million records and 36 million records for gyroscope, GPS, Wi-Fi, and accelerometer, respectively. Due to limited time and resources, we have randomly picked three participants for this study, whose data were collected from April 04, 2015 to May 12, 2015. In this dataset, GPS data included data source (GPS and network), speed, accuracy meter, longitude, and latitude. Wi-Fi data included a basic service set identifier (BSSID), service set identifier (SSID) and

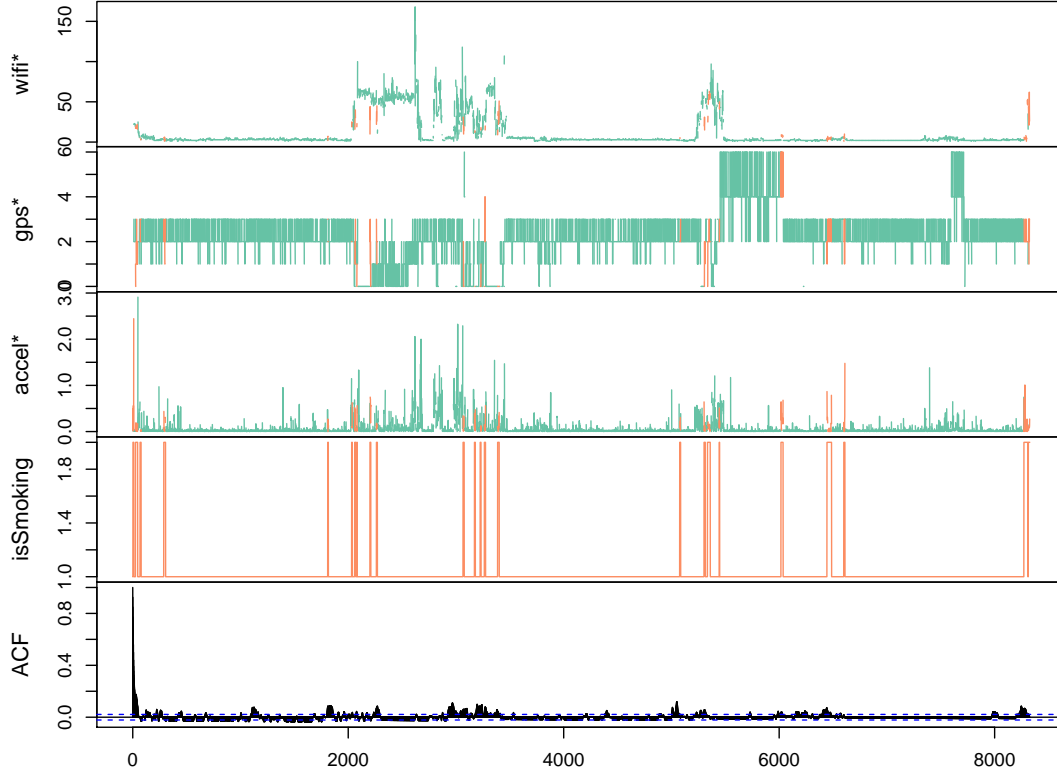


Figure 3.1: Plot of time series sensor data, labelled smoking and non-smoking periods and autocorrelation function (ACF)

received signal strength indicator (RSSI). Accelerometer and gyroscope data incorporated linear acceleration and angular velocity on X, Y and Z axes, respectively.

Data Processing

Time series of sensor data, especially when collected with high granularity (with bursts every 5 minutes in our case), is distinguished from normal features. Windowed aggregation (every 30 seconds in our case) can effectively reduce noise as well as extracting derived features with considering time dependency, rather than simply treat features as Markov process. Here for continuous measurements accelerometer data, GPS data and wifi readings, we successfully extracted derived features such as average and standard deviation for accelerometer data, count of GPS readings drawn from satellite sources, maximum RSSI and count of unique media access control address (MAC address) from Wi-Fi data.

Each participant has an extremely long smoking period at the end of their self-labeled smoking-nonsmoking periods, which are apparently outliers and need to be resolved. Smoking is a periodic behavior and there

exists a normal smoking period. Therefore we preserved only a period at the head of each of those extremely long periods, whose length equals to the average length of previous smoking periods of this very person, and trimmed the rest smoking periods in the end, as shown in 3.1.

For accelerometer data, the norm of readings across the X, Y and Z accelerometer axes was calculated to combine the three features of acceleration into one feature. The average norm of raw accelerometer data across 30s timeslots exhibits an empirical cumulative distribution function (ECDF) that is almost identical between smoking periods and nonsmoking periods. The overlap of the accelerometer norm is largely caused by the invariant dominant gravitational acceleration component, regardless of the position of the smartphone or smoking/non-smoking status. To separate out the dominant invariant gravitational component, we applied a high-pass filter, using a standard deviation of norms in the 30s timeslot. ECDFs for one participant are shown in 3.2.

For Wi-Fi data, RSSI in 30-second timeslots was considered. As shown in 3.2, the maximum RSSI in 30-second timeslots did not show a pronounced difference between the two states. Canadian laws require people to smoke outside buildings, and there is not much Wi-Fi coverage outside buildings. So the counts of unique MAC address during 30s timeslots were calculated for each state. In 3.2, ECDF of the counts of unique MAC address showed a better difference between two states than that of maximum RSSI. Maximum RSSI indicated the strongest Wi-Fi signal, and counts of unique MAC address represented the number of accessible networks for the smartphone. So both of the features were used to train the HMMs.

The source of location data, either GPS or network (using cell tower and Wi-Fi based location), can also help indicates whether the participant is indoor or outdoor. So the count of GPS readings across 30s timeslots specifically drawn from satellite (as opposed to network) sources were used. Its ECDF was shown in 3.2.

Multivariate HMMs

Using the transformed data described above, a multivariate HMMs was employed to classify smoking and non-smoking intervals based on real-world labeled observations. In this model, each state has multiple observations corresponding to readings from Wi-Fi, accelerometer and GPS sensors. The probabilities of observations follow empirical distributions, and observations are assumed to be independent of each other, conditional on being in a given state. So the likelihood of observing a given vector of observed quantities was approximated as the product of independent probability density functions as given by kernel density estimation (KDE).

Two-fold Cross Validation

Hidden Markov models expect sequential observations, therefore when choosing a training set and test set, we can not simply sample at random time intervals from data sequence, but rather need to divide the data sequence into disjoint contiguous sequences. Firstly, we made use of a two-fold cross-validation approach,

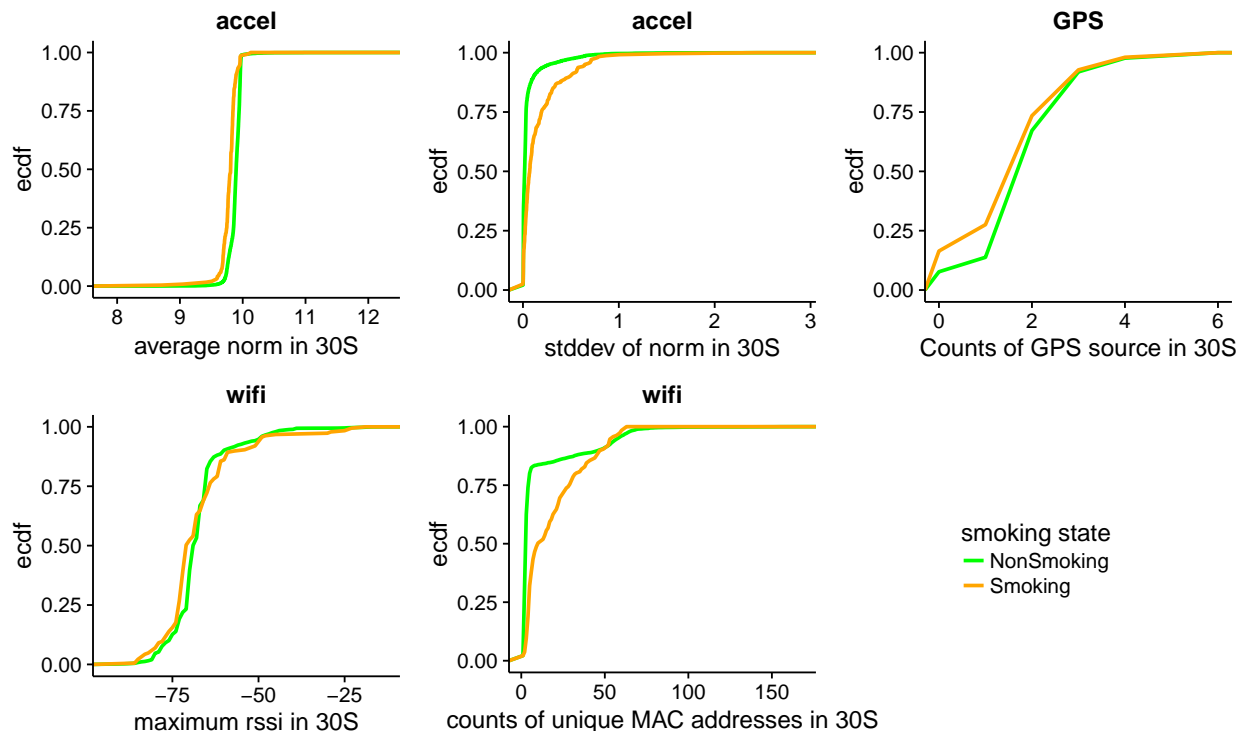


Figure 3.2: ECDF of raw and transformed accel, wifi, and GPS data

where we cut the sequence from the head to 50%, 55%, 60%, 65%, 70% and 75% of the sequence to ensure sequential observations (including NAs) for training, and used the balance of the observations for testing. Second, We swapped the training set and test set in the first step to feed the HMMs.

3.3 Results

Structured learning was used in this project. This work was conducted in several phases. In phase 1, maximum RSSI, counts of unique MAC address during 30s timeslots for Wi-Fi, average norms, a standard deviation of norms for accelerometer and counts of GPS reading from GPS source in 30s timeslots were considered as a single feature, respectively. Each of the five features was used to train univariate HMMs, which were then evaluated. In phase 2, Multivariate HMMs using three features and five features were trained and evaluated.

The HMMs were found to yield favorable results in the multivariate cases and in univariate cases considering accel sensor data as a feature. As shown below, multivariate HMMs exhibit accuracy over 0.9, and an area under the receiver operating characteristic curve (ROC curve) (AUC) above 0.8 when collected data is representative. The results of HMMs with a single feature are less favorable than those for multivariate HMMs.

3.3.1 Results with Single Feature

For using the average of the norm of the accelerometer as a feature, the AUC for the training set and test set with different size of training set range from 0.69 to 0.94 and from 0.60 to 0.79, respectively. And the error rates of the training set and test set range from 0.096 to 0.27 and from 0.057 to 0.357, respectively. For using the standard deviation of the norm of the accelerometer as a feature, the AUC for the training set and test set with different size of training set range from 0.76 to 0.92 and from 0.63 to 0.86, respectively. And the error rates for the training set and test set range from 0.05 to 0.12 and from 0.06 to 0.2, respectively, as shown in Figure 3.3.

For using readings from the Wi-Fi sensor as a single feature (maximum RSSI or counts of unique MAC address), for maximum RSSI, the range of AUC for the test set is from 0.43 to 0.69, and the range of error rate for the test set is from 0.23 to 0.65. For counts of unique MAC address, the range of AUC and error rate for test set ranges from 0.47 to 0.82 and from 0.17 to 0.86, respectively.

In this model, using only one feature derived from the GPS sensor, the range of AUC for the training set is from 0.52 to 0.64, for the test set is from 0.50 to 0.75. Error rates for the training set and test rate are from 0.04 to 0.92, and from 0.015 to 0.96, respectively.

3.3.2 Results with Three Features

Counts of unique MAC address for Wi-Fi, a standard deviation of accel norm and counts of GPS sourced signals were employed together to train the HMMs. The results offer AUC and error rates of the HMMs were shown in 3.4. The range of AUC for the test set is from 0.5 to 0.83 with an average of 0.65, and the error rate for test set ranged from 0.017 to 0.96 with an average of 0.23.

3.3.3 Results with Five Features

All five features derived from sensors were employed together to train HMMs. The results for AUC and error rates of the five feature HMMs were also shown in Figure 3.4. The range of AUC for the training set is from 0.5 to 0.98 with an average of 0.79, and for the test set, it is from 0.52 to 0.84 with an average of 0.66. The Wi-Fi and GPS data are location-based features, while the components of the accelerometer data are associated with the body gestures and orientation features of the participant. The combination of five features can capture a larger set of information on the current smoking activity of participants. This enlarged information can, in turn, enhance the performance of the HMMs.

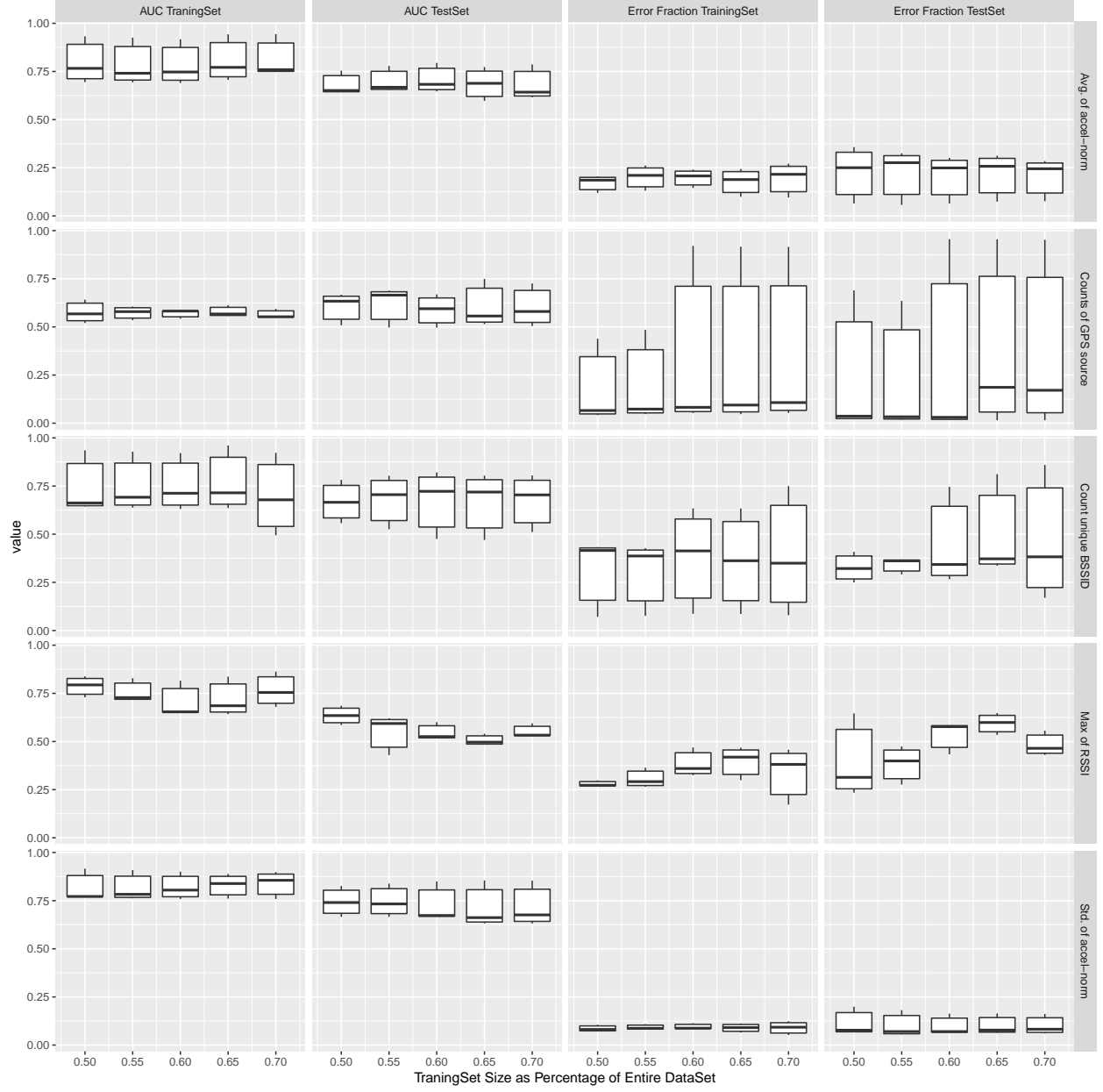


Figure 3.3: AUC and error rate of HMMs using avg. and std. of accel-norm, count of GPS source, count of unique BSSID and max of RSSI as single feature

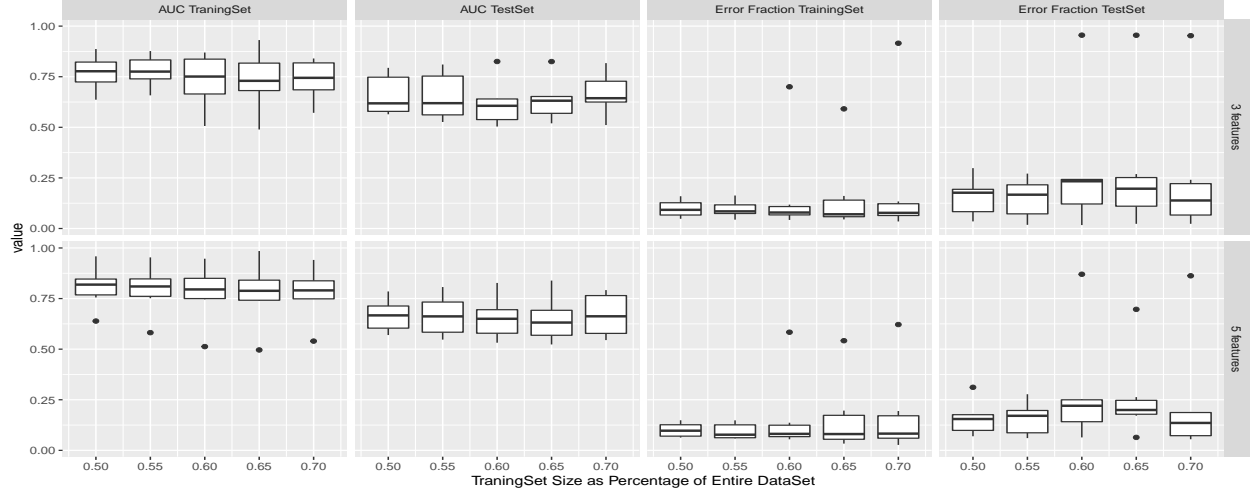


Figure 3.4: AUC and error rate of HMMs with three features and five features

3.4 Discussion

3.4.1 Thoughts on Labeled Data

The results of HMMs using information drawn from single features alone are worse than that of multivariate HMMs. As noted above, each different type of feature adds additional information, and the combination of three features can best describe the current smoking activity and gestures of participants. A more accurate description can result in better performance of HMMs. Additionally, at the data processing step, the ecdf of the standard deviation of the accelerometer norm in 30s timeslots demonstrated significant differences between the two states, compared to the average of the norm. So the results of HMMs with a standard deviation of the norm as a single feature are better than the results of average norm HMMs. Similarly, the results of HMMs using counts of unique MAC addresses, which has a distinct difference between the two states, are better than the results of HMMs with maximum RSSI. The difference between the ecdf of two states is crucial for the model to distinguish different states.

The AUC is varied with training set sizes used for the HMMs. Smoking is a periodic behavior, and subdividing the training set and test set so as to include entire smoking cycles can capture smoking patterns. For the smoking sequence, if the training set were cut in the middle of a smoking cycle, the long smoking state or non-smoking state sequence changes the predicted results by changing the transition probabilities.

3.4.2 Emission Probability for Multi-dimensional Observations

For emission probabilities, it was hard to find joint parametric distribution for 3 or higher dimensions. Investigation not shown here revealed that the results of using a product of univariate KDE are better than that of a single joint multivariate KDE. The AUC for a joint KDE is 0.51. The reason for the bad results of a

joint KDE likely reflects a high proportion of missing data. In a joint KDE, co-occurring wifi, accelerometer and GPS data are all required for the KDE. So at a time point, if one of three variables is missing, the joint KDE has to discard this time point. However, univariate KDEs are more robust in the context of incomplete data, as each such KDE is affected only by missing data with respect to a single variable.

3.5 Related Work

Sazonov et al. (2013) developed a wearable sensor system based on hand-to-mouth smoking gestures and breathing pattern [59], Lopez-Meyer et al (2013) further applied a support vector machine and achieved 87% and 80% of average user-independent precision and recall and 90% in user-independent precision and recall [62]. Scholl and Laerhoven (2012) used a wearable accelerometer device to collect data and applied a basic Gaussian classifier to detect smoking gestures with a precision of 51.2% and 70% of user-specific recall [60]. We have not yet found papers about smoking detection using multivariate HMMs based on various types of sensor data available from commodity smartphones.

3.6 Limitations and Future Work

Despite high granularity data for each participant, our study is limited by the number of participants, as future work, we will experiment with a larger participant size. We will also seek to depict the joint distribution of three observations using empirical 3D distribution estimated by using linear interpolated KDE. Perhaps covariance among these three observations can provide additional information to boost the performance of our Multivariate HMMs. Another interesting question for investigation is whether a personal empirical distribution can be reused on other persons if so, we only need to train the model once. While room for further improvement remains, the success of smartphone collected data in supporting the accurate classification of smoking behavior raises prospects for proactive prediction of smoking occurrences, and may also play an important role in smoking cessation program through behavior-triggered “nudges” that could be delivered at times proximate to smoking.

3.7 Conclusions

The results of multivariate HMMs demonstrated the classification and detection of smoking activity with high accuracy. Compared to single feature HMMs, the multivariate HMMs had higher accuracy, because additional types of sensor data can help us better describe smoking gesture and activity.

This work further suggests that tailoring training set and test set close to entire smoking cycles can improve the performance of HMMs. The large variation in results across participants further raises the possibility that significant components of remaining error rates may be due to limitations in the accuracy of self-reporting of ground truth data on smoking behavior.

The work presented in this chapter demonstrates how the characteristics of data affect the dynamic model structure. And that model structure further shapes the ability to use data. The HMM is a statistical model in which the underlying system being modeled is assumed to be a Markov process with a set of discrete unobservable states assumed to be memoryless according to the transitions between states. HMMs in general assume that the emission probabilities of observations are conditional on the hidden states, and that successive observations within a given such state are conditionally independent. The multivariate HMMs built in this work further assume that each type of observation themselves is independent of each other. There are reasons to use the data collected across the multiple types of sensors to infer the underlying smoking status. First, the signals collected with multiple types of sensors are changed with smoking activity and gestures of participants, as well as other behaviors common exhibited during smoking behaviour, such as low-vigour pacing or resting. While the smartphone-based sensor data is individually insufficient to infer the hidden states, collectively, it can inform the multivariate HMMs for identifying the unobserved state (smoking and non-smoking). Second, the time-series characteristic continuously collecting the sensor data with a five-minute duty cycle makes the data suitable for modeling discrete time process of the HMMs. In this regard, it bears noting that the availability of bursts of data for noisy sensors such as GPS and accelerometers can be helpful for inferring the underlying state with considerably greater reliability than would be possible using single observations from such sensors in isolation.

As discussed in previous sections, the multivariate HMMs examined here support the classification and detection of smoking activity with relatively high accuracy, despite the fact that observations across multiple sensors within a given observation are assumed to be independent of each other. Perhaps data collected with other types of sensor – e.g., heart rate and gyroscope data depicting hand movement, such as could be gathered by wrist-based wearable technologies – and more textured model assumption regarding the statistical dependence exhibited between these three observations can provide additional information on smoking activity.

CHAPTER 4

EFFECT OF E-CIGARETTE USE AND SOCIAL NETWORK ON SMOKING BEHAVIOR CHANGE: AN AGENT-BASED MODEL OF E-CIGARETTE AND CIGARETTE INTERACTION

The text presented in this chapter is from the manuscript of "Effect of E-cigarette Use and Social Network on Smoking Behavior Change: An agent-based model of E-cigarette and Cigarette Interaction" by Yang Qin, Rojiemiahd Edjoc, and Nathaniel Osgood, published in Proceedings of the 2019 Conference on Social Computing, Behavioral-Cultural Modeling Prediction and Behavior Representation in Modeling and Simulation. The contribution of each author is presented in Section 1.5 of Chapter 1 [7].

The work of this chapter demonstrates an ABM of conventional smoking and ECig use calibrated by matching cross-sectional data points to model outputs. The dynamic model examined here supports a description of rich context and captures continuous data. In contrast to the HMMs, the structure of ABM supports the model in capturing a high level of heterogeneity at an individual level. The cross-sectional data from 2013 to 2017 from the Canadian Tobacco and Alcohol and Drugs Survey [27] was employed in model calibration for estimating non-measured parameters and supporting the projection of future model outputs. The data considered here includes the prevalence of current smokers (PCS), the prevalence of former smoker (PFS), and prevalence of current ECig user (PCEU). With the characteristic of the data, and aggregating results across individuals, the model parameters were calibrated to match statistics from individual-level simulation outputs with the cross-sectional data.

4.1 Introduction

Smoking and secondhand smoke harm nearly every organ of the body and contribute to many preventable diseases, including lung cancer, coronary heart disease, chronic obstructive pulmonary disease, and other cardiovascular diseases [1, 2]. Nicotine products come in various forms, e.g., cigarettes, nicotine gum, patch, and ECig [73]. ECigs, vaporizing a liquid mixture which is used as a substitute for tobacco leaves and stored inside cartridges [4, 6], were introduced to the market in 2003, promoted and marketed by major tobacco companies in the last decade [4, 5]. The use of ECigs as a cigarette alternative has increased dramatically. The PCEU among US adults increased from 0.3% in 2010 to 6.8% in 2013 [63]. Within recent years, there has

been a particularly dramatic and alarming rise in the use of ECig amongst youth.

The health behaviors associated with smoking have been studied in detail. The majority of smokers attempt to quit smoking, but fewer than 5% of them remain quit for more than three months [8]. Effective tools for smoking cessation (SC) may help current smokers (CS) quit, and forestall an individual at risk of smoking, e.g., former smoker (FS), struggling with avoiding relapse. ECigs also allow never smoker (NS) seeking to experiment with nicotine as an alternative to cigarettes. The rise of ECig use is associated with a perception that ECig is safer than cigarettes and a useful SC device. However, there remains little solid scientific evidence confirming the effectiveness and safeness of ECig as a SC tool [9,63]. By surveying 2028 US smokers in 2012 and 2014 and two years of follow-up, Zhuang et al. [9] concluded that long-term ECig users had a higher rate of SC of 42.4% than short-term Ecig users and non-users (14.2% and 15.6%, respectively). Zhu et al. [10] concluded that ECig users have a higher rate of SC, and are more likely to remain quit than non-ECig users. Cherng et al. [63] proposed an ABM to examine the effect of ECig on the smoking prevalence of US adults and concluded that the simulated effects of ECig on SC largely changed smoking behavior. The ABM simulated the influences of smoking behavior on ECig use initiation and cessation, and how ECig reversely affected SC and smoking initiation (SI).

While promising, previous studies have predominantly relied upon self-reported surveys, cohort studies and clinical trials. Such larger studies are expensive, are associated with high delay until they show effect, and can be difficult to plan and execute given the wide variety of patterns of behavior possible (e.g., initiation of exclusive smoking following ECig use, initiation of exclusive ECig use following tobacco, dual-use, start of ECig use following quitting tobacco, etc.). Clinical trials often regulate or exclude factors that play a key role in shaping outcomes in society, such as switching of nicotine delivery modality, varying rates of compliance, and peer influence effects.

In this paper, extending the preliminary model structure introduced by Cherng et al. [63], we build an ABM of smoking and ECig use with modalities of initiation, cessation, and relapse to examine the effects of ECig use on individual-level smoking behavior change and population-level smoking patterns according to the aggregation of individual outcomes. Our model incorporates strong SN effects involving both the selection of networks and influence over networks, age, sex and history-dependent effects regarding the rate of initiation, cessation and relapse for both smoking and ECig use and individual decision-making effects based on characteristics of social contacts. In particular, we use the model to investigate whether the ECig is an effective SC device and the impact of ECigs on non-smokers with regards to SI.

4.2 Methods

4.2.1 Model Overview

ABM can simulate complex social dynamics and behaviors with considerably high resolution, and generate population-level results by aggregating individual outcomes in different scenarios [63]. Equally notable,

ABM is widely applied to probe the impacts of counter-factual interventions, as well as to help prioritize data collection in a complex milieu of complex interactions of behaviors and product types. In this study, a high level of heterogeneity characterizing both exogenous and endogenous components, specific traits at the individual level and modularity also strongly motivated the use of ABM.

Our model was built in AnyLogic (version 8.3.3), and used four interacting statecharts for each agent, featuring smoking states, ECig use states, birth, and mortality. The parameters, transition rates and statecharts in the `Person` class serve as influences from within an agent on smoking and ECig use behavior. The model further incorporates a distance-based network to simulate social contacts between agents.

The model simulates a population of 100,000 agents with age distribution based on the population pyramid of Canada [74]. The model time unit is 1yr, and the length of the time horizon is 70yrs. The initial states may misestimate the prevalence of each smoking and ECig use state, so a period of burn time (52yrs) is used for the model to achieve equilibrium. Over the continuous time of the simulation, agents either maintain their current state of smoking and ECig use or transit to other states based on the (hazard) rates discussed in the next section.

4.2.2 Model Formulation

Smoking statechart describes three smoking states: `NS`, `CS` and `FS`. An individual can switch its presence in each of the three states of statechart according to specified transition rates, namely the rate of `SI`, the rate of `SC`, and the rate of smoking relapse (`SR`).

ECig use statechart separates the states of ECig use as never ECig user (`NEU`), `CEU` and former ECig user (`FEU`). The transition of ECig use initiation (`ECigUI`) is fired with a hazard rate, transferring an agent from `NEU` to `CEU`. Other transitions are message triggered transitions, which will be activated only under scenarios when we consider: A `CS` who never used ECig may possibly initiate ECig use after quitting smoking, transiting from `CS^NEU` to `FS^CEU` by chance; An `FS` who is `CEU` may possibly quitting ECig after relapse to smoking, transferring from `FS^CEU` to `CS^FEU` by chance; And a `CS` who is `FEU` may possibly relapse to `CEU` after quitting smoking, transiting from `CS^FEU` to `FS^CEU` by chance. For the two statecharts, agents can occupy a specific, concrete state of one statechart at any one time, while being in any state of the other statechart.

Rate of `SI`, `SC` and `SR` denoted as r_{si} , r_{sc} and r_{sr} , respectively, are each the product of its corresponding hazard rate (α_{si} , α_{sc} and α_{sr} for the calculation of r_{si} , r_{sc} and r_{sr} , respectively), a multiplier (m_{si} , m_{sc} and m_{sr} for the calculation of r_{si} , r_{sc} and r_{sr} , respectively) and a coefficient (e_{si} , e_{sc} and e_{sr} for the calculation of r_{si} , r_{sc} and r_{sr} , respectively).

The hazard rates reflect the magnitude of the effect of age, gender and smoking history on r_{si} , r_{sc} and r_{sr} . We transformed the annual probabilities of `SI` and `SC` (p_{si} and p_{sc} , respectively) of male and female of 1970 birth cohort, reported by Holford et al. [75], into their corresponding α_{si} and α_{sc} as table functions in AnyLogic by using $p = 1 - e^{-\alpha}$. The model assumed that α_{sr} declines with growing time since quit; thus, individuals who only recently quit have far higher relapse risk than an agent who has remained as `FS` for a

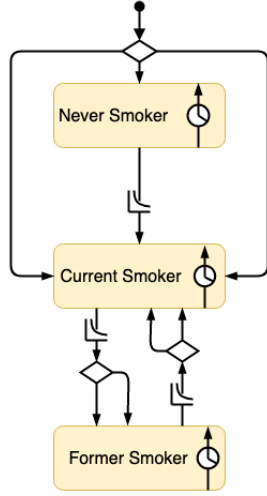


Figure 4.1: Smoking statechart

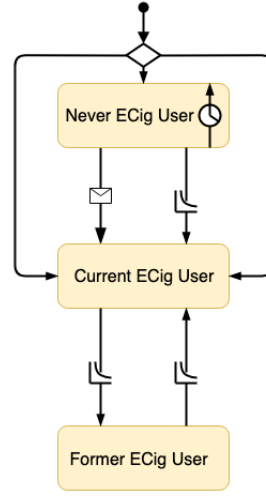


Figure 4.2: ECig use statechart

prolonged period. The value of multipliers is driven by the state of ECig use. Wills et al. [76] suggested that **NS** who tried ECig is three times more likely to start smoking. Leventhal et al. [77] reported that ECig users were four times likely to uptake cigarettes. McRobbie et al. [78] suggested that the rate for **SC** was significantly higher in the presence of ECig use (RR 2.3; 95%CI: 1.05 - 4.96). Based on the linkages between ECig use and r_{si} and r_{sc} mentioned above. As ECig use can help relieve the symptoms of nicotine withdrawal to some degree and might provide an additional avenue towards continued socialization with companions who remain, tobacco users, **CEUs** are less likely to relapse in smoking, compared to non-ECig users [79]. Therefore the model assumes m_{si} is 4.0 for agents who are **CS**^**CEU** [77], or is 2.87 for agents who are **FEU** [76], m_{sc} is 2.3 for agents who are **CS**^**CEU** [78], m_{sr} is 0.5 for agents who are **CS**^**CEU** [79]. If each rate is only the product of its hazard rate and a multiplier, the rate may misestimate the projection of smoking. Therefore, the coefficients e_{si} , e_{sc} and e_{sr} were calibrated to match simulation outcomes against historical data.

For the rate of **ECigUI**, the model adapted the time-based sigmoid function and divisors introduced by Cherng et al. [63], to characterize the increasing use of ECig after its introduction into the market and the influence of smoking status on **ECigUI**. Additionally, the rate of **ECigUI** is strongly related to the agent's smoking status and demographic factors [80], suggesting that ECig is popular in smokers and young people; thus, we assumed a hazard of **ECigUI** of male agents using a table function, which has an x-axis of age of the agents and y-axis of the hazard rate and follows the same pattern as for the hazard rate of **SI** for male. If an individual is female, the hazard of **ECigUI** of this agent is given by the corresponding point on the table function divided by the variable **divECigFemale** with a value of 1.5. The overall rate of **ECigUI** is the product of the hazard of **ECigUI** given by the time-based sigmoid function [63] and a coefficient (e_{ECig}), which was calibrated by matching model generated incidence of ECig use against corresponding historical data.

The model assumes that the transition of ECig use cessation (**ECigUC**) and ECig use relapse (**ECigUR**) are affected only by the smoking behavior, that is, the model assumed that individuals who are **CEUs** or **FEUs** would

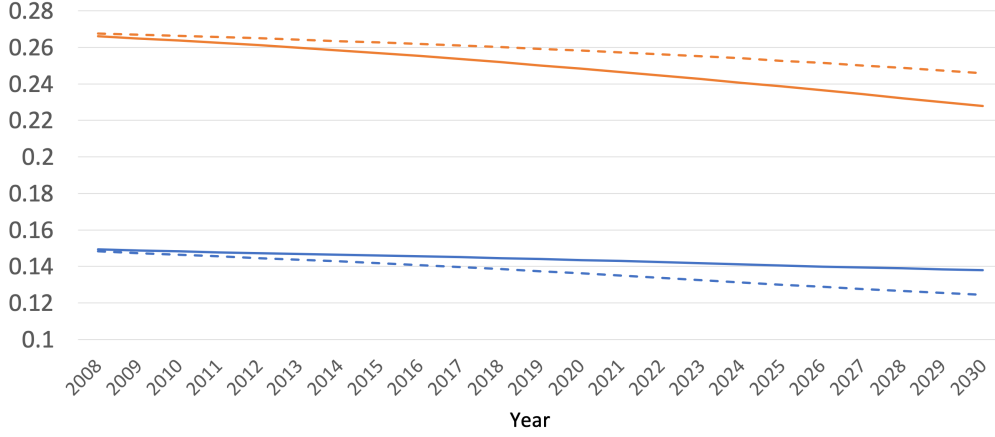


Figure 4.3: PCS and PFS of Scn1 and Scn2

Orange and blue solid line represent the PFS and PCS in Scn1, respectively. Orange and blue dashed line represents the PFS and PCS in Scn2, respectively.

remain so unless changes occurred in their smoking behavior. Specifically, in the absence of identified evidence with respect to the fraction of individuals whose state of ECig use will be affected by smoking behavior, the model posited that 85% of $CEUs \wedge FFS$ will quit ECig if they relapse to smoking, since their nicotine cravings were satisfied by smoking, and 80% of agents who are $FEU \wedge CS$ will transit to CEU if they quit smoking. As ECigs may be used as cessation tools, the model further assumed that 50% of smoking quitters would uptake ECig immediately after quitting smoking. Therefore, message dichotomous branching transitions were built for $ECigUI$ and $ECigUC$ under these assumptions in addition to the rate of $ECigUI$ discussed above.

Age-specific birth and mortality rates are drawn from Statistics Canada of 2016 [81, 82] are used in the model. The total fertility rate in Canada in 2016 is 1.54 per woman. To maintain population replacement (with a total fertility rate of 2.1) for successive years of the model run, we thus multiplied a coefficient (with a value of 1.357) by the fertility rate of each age group.

Smoking is well recognized as both an individual habit and a social phenomenon [8]. The baseline model was extended with a distance-based network to simulate the effect of social connection and peer pressure on the SI and $ECigUI$. To build a localized SN for each agent, connecting with its nearby agents, the model assumed that an agent establishes the network with the agents in proximity (50m). The SN was implemented as a dynamic network driven by agent mobility in continuous space with width and height both equal to 250,000m. Specifically, the agent moves to a new location within the space, and disconnects from the current network then re-establish a network based on agent layout by using a cyclic timeout event with an interval of 2 yrs. As dynamic network, the fraction of CS and $CEUs$ among its connected agents are modified with the change of the SN , therefore, influence the effect of SN on SI and $ECigUI$.

The effect of SN was modeled using multipliers (m_{net}), and applied them to the baseline r_{si} and r_{ECig} , respectively. The overall rate of SI and $ECigUI$ of a particular agent were increased by m_{net} , relative to the rates in the baseline scenario. Without a specific mathematical model to quantify the effects of connected neighbors of a particular agent, the model employed a sigmoid function to describe the progression of the

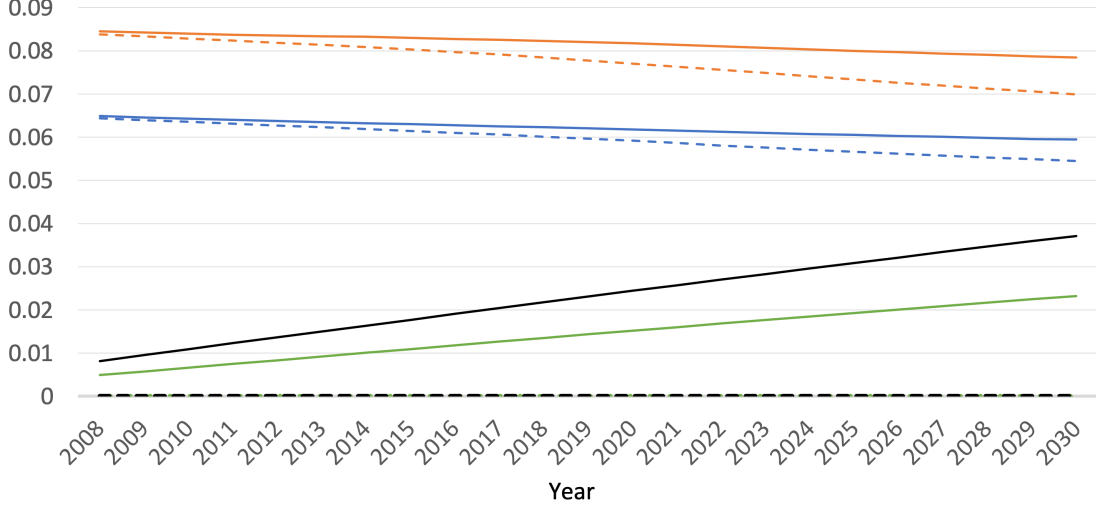


Figure 4.4: PFCs, PMCS, PFCEU and PMCEU of Scn1 and Scn2

Blue, orange, green, and black solid line represent the PFCs and PMCS of Scn1, PFCEU and PMCEU of Scn2, respectively. Blue, orange, green, and black dashed line represent the PFCs and PMCS of Scn2, PFCEU and PMCEU of Scn1.

influence from the connected neighbors, which increases small at the beginning then accelerates fast and reaches the plateau. We, therefore, assumed m_{net} follows a sigmoid function (Equation 4.1), where f is the fraction of CS (or, correspondingly, CEUs) among its connected agents if m_{net} is used to calculate the rate of SI or ECigUI, respectively, and f_0 , α and γ in Equation 4.1 are 0.25, 2.0 and 1.0, respectively. Similarly, in Equation 4.1, r_{average} represents the average rate of SI or rate of ECigUI of population for the calculation of rate of SI or ECigUI of this agent, respectively. The r_{average} s are re-calculated every year based on the smoking and ECig use status of the population at the beginning of each year.

$$m_{\text{net}} = \frac{\alpha + e^{-\gamma \times (f - f_0)}}{(1 + e^{-\gamma \times (f - f_0)}) \times r_{\text{average}}} \quad (4.1)$$

4.2.3 Model Calibration

e_{si} , e_{sc} , e_{sr} and e_{ECig} were calibrated to match the estimated PCS, PFS and PCEU generated by the rates (r_{si} , r_{sc} , r_{sr} and r_{ECig}) in the baseline model, against historical data of 2013-2017 from CTADS [27]. The calibrated result of e_{si} , e_{sc} , e_{sr} and e_{ECig} is 1.088, 2.435, 1.51 and 7.898, respectively.

4.2.4 Model Scenarios

We examine here simulated population-level smoking behaviour change and ECig use under following three scenarios: smoking behavior in scenario one (Scn1) which is in absence of ECig use and the SN, smoking behavior in scenario two (Scn2) which is under the use of ECig, and smoking behavior in scenario three (Scn3) which SN exists and supports the SI and ECigUI (Scn3). The outputs from these scenarios examined the difference in prevalence and incidence of smoking arising from considering ECigs as well as SN both separately

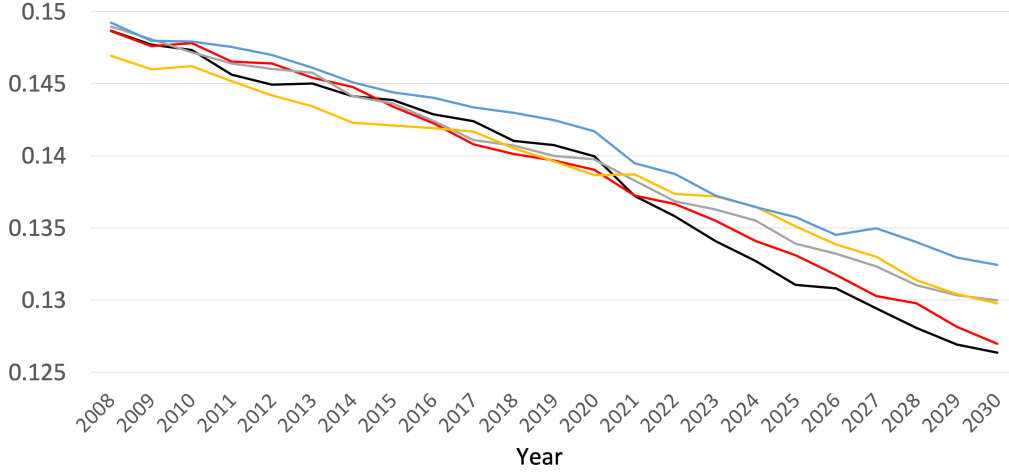


Figure 4.5: SA of rate of ECigUC on PCS

Black, red, grey, yellow, and blue line represent a successively larger rate of ECigUC of 0.2, 0.4, 0.6, 0.8, and 1.0, respectively.

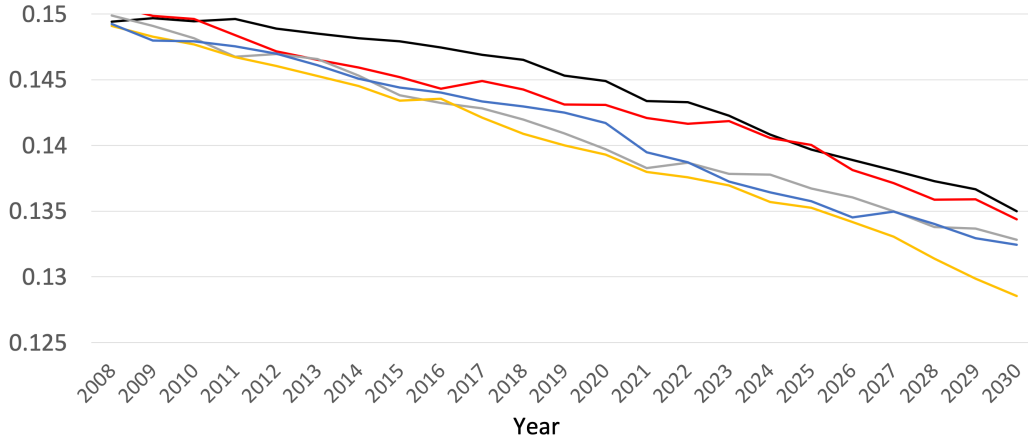


Figure 4.6: SA of rate of ECigUR on PCS

Black, red, grey, yellow, blue line represent a successively larger rate of ECigUR of 0.2, 0.4, 0.6, 0.8, 1.0, respectively.

and in combination. The simulation of Scn1 and Scn2 were run for 100 realizations, and simulation under Scn3 were run for 40 realizations with respect to the considerably large computation of SN in AnyLogic, with random seeds making each simulation run unique, then the means of the outputs of all runs were calculated for the comparison. Furthermore, to examine the statistical significance between the results from Scn1 and Scn2, we performed a Mann-Whitney-U test on the per-realization output (PCS), from the two scenarios.

4.2.5 Sensitivity Analysis

To assess the sensitivity of model parameters on model outputs, we performed sensitivity analysis (SA) on the parameters such as the rate of ECigUC and the rate of ECigUR. The message transitions for ECigUC and ECigUR were replaced by the rate transitions. In the SA of the rate of ECigUC, the model assumed the rate of

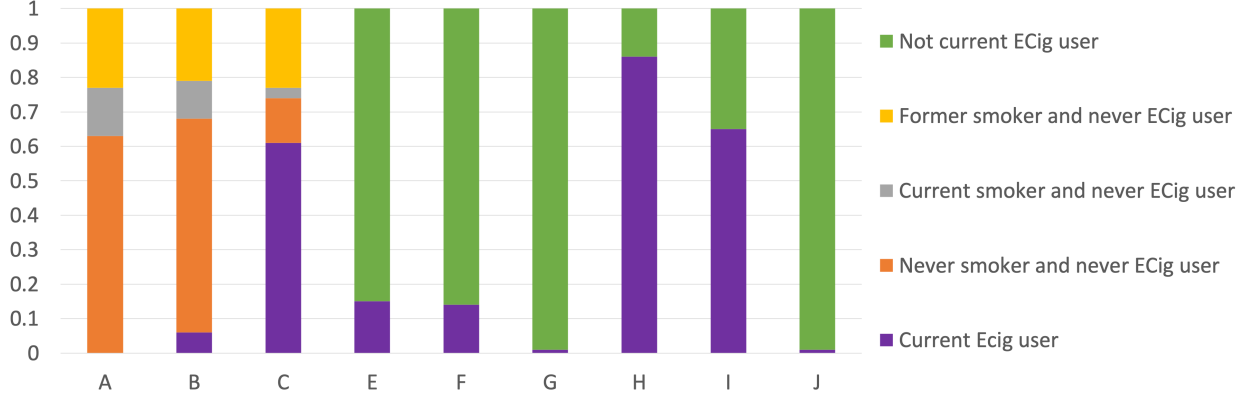


Figure 4.7: Population breakdown by smoking category and fraction of CEU

Panels A, B and C depict the population breakdown by smoking category in Scn1, Scn2, and Scn3, respectively. Panels E, F, G, H, I and J illustrate fraction of CEU among CS, among FS and among NS in Scn2 and those fractions in Scn3, respectively.

ECigUR is 1.0, and the range of the rate of ECigUC was 0.2 to 1.0 with a step of 0.2 for each iteration. Similarly, for the SA of the rate of ECigUR, the model assumed the rate of ECigUC is 1.0 and the rate of ECigUR had the same range and step with the rate of ECigUC in its SA experiment. The SA experiments examined the potential change of PCS resulting from changes in the value of the rate of ECigUC and the rate of ECigUR.

4.3 Results

4.3.1 Comparison between Scn1 and Scn2

Mean, median and standard deviation of the results for PCS, generated by the model realizations in Scn1, are 0.1438, 0.1440 and 0.0037, respectively, and those from the model realizations in Scn2 are 0.1369, 0.1374 and 0.0074, respectively. The results of a two-sided Mann-Whitney-U test for the results of two scenarios, $p < 2.2e^{-16}$, demonstrate that the distributions in the results of two scenarios differed significantly.

The message transitions in ECig use statechart were disabled in Scn1 and Scn2, therefore, in the stacked column chart showing the breakdown by smoking category (Figure 4.7A, B and C), the agents were divided into four categories: CEU regardless of their smoking status, NS^NEU, CS^NEU, and FS^NEU, respectively. The portion of FS and NEU (23%) in Scn1 is slightly higher than that (21%) in Scn2, due to a large portion (6%) of CEU, as shown in Figure 4.7 A and B. This reflects the fact that the FS in Scn1 is located within the FS^NEU category, whereas in Scn2 some of those individuals are located within the CEU category.

4.3.2 Comparison between Scn2 and Scn3

In Scn3, at the end of simulation, the maximum and minimum degree centrality of a given agent is 2 and 1, respectively. With the presence of SN (in Scn3), as shown in Figure 4.7 C, the fraction of CEU in the population increased dramatically – rising from 6% in Scn2 to 61% in Scn3. With the exposure to ECig use from connected

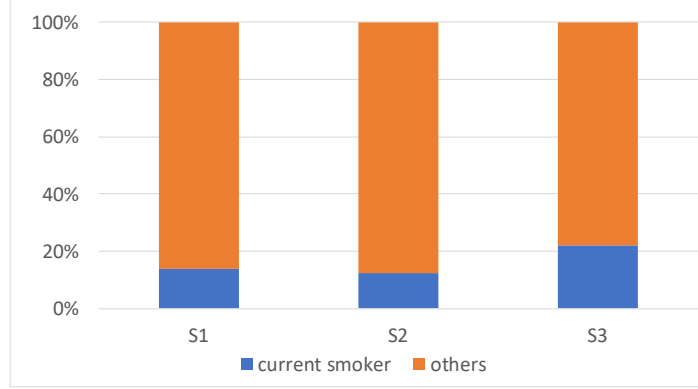


Figure 4.8: Fraction of CS and others

Panels A, B and C depict the fraction of CS in S1, S2, and S3, respectively. Blue and orange labels CS and others (including NS and FS), respectively.

individuals or neighbors, people tend to initiate ECig use. The increased portion of CEU are mostly from the agents who were NS\NEU. In Scn2, the fraction of CEU among CS and that among FS are similar, with value of 15% and 14% in Figure 4.7 E and 4.7 F, respectively, which are considerably larger than fraction of CEU among NS, as shown in Figure 4.7G. In Scn3, the SN significantly increased the fraction of CEU among CS and FS, with the value of 86% and 65%, respectively in Figure 4.7H and Figure 4.7I, respectively, while the fraction of CEU among NS does not show obvious increase due to SN, compared with that of Scn2.

4.3.3 Sensitivity Analysis

Results from the SA on the rate of ECigUC and the rate of ECigUR suggests that the PCEU and prevalence of former ECig user can substantially change the PCS, as shown in Figure 4.5 and 4.6. Figure 4.5 demonstrates that when ECR is increased from 0.2 to 1.0 – holding invariant the value of ERR – PCS are gradually increased, and PCEU decreases. The results in Figure 4.5 suggest that although incidence of SI is reduced by the lower PCEU, the decreased rate of SC and elevation in SR due to the decreased PCEU compensates for the decrease in the rate of SI. Similarly, the change in the rate of ECigUR also influences PCEU. Holding constant the rate of ECigUC, an increase in the rate of ECigUR generally increases the PCEU, but lowers the PCS, with a possible exception at the lowest levels of the rate of ECigUR. Results in Figure 4.6, the line from the rate of ECigUR of 0.2 having the lowest PCEU, reflect that agents were more likely to remain as CS.

4.4 Discussion

From the results in the three scenarios, the model demonstrates that ECig use and SN encourage agents to uptake ECig, therefore, shape population-level smoking behavior. Although the use of ECig increases the rate of SI, the combined effect of the increase in the rate of SC and the decrease in the rate of SR results a considerably large decline in PCS and increase in PFS. The results of SA further shows the PCS is sensitive to

the ECig use behavior change. The outputs of the model largely depend on the feedback between smoking and ECig use, and interactions between agents. First, we assumed the rate of ECigUI of CS, FS and NS are in a declining order, specifically, the CS has highest rate of ECigUI compared with other smoking category. Second, if an individual is CS, being a CEU increases the probability of quitting smoking and staying in FS state, which means they have a relatively higher probability of using ECig as a SC tool. We assumed the ECig use helps greatly in SI for NS. Given the model results, fraction of CEU among NS is considerably lower, compared with CS and FS. Accordingly, as a combined result of the rate of SI, the rate of SC and rate of SR, the PCS is decreased due to ECig use. Furthermore, we assumed gender effect as divisors in the rate of ECigUI. Thus, the model behaves a relatively stronger influence from ECig use on smoking behavior. The effect of SN is modeled as a multiplier to the rate of ECigUI, which generates more CEU during the simulation.

Despite the fine resolution of the model, there are some limitations. First, the model is highly sensitive to the use of ECig, however, the model has no good assumption on the rate of ECigUC and the rate of ECigUR. Second, at this resolution, the model cannot capture the smoking episodes, dynamics of nicotine metabolism, allowing the model to analyze whether ECig use helps in relieving nicotine cravings at a fine-grained level as SC tool. Finally, the model assumes the effect of SN in a relatively simple way.

Allowing for capturing dynamics in a richer context, such as those associated with a network and individual characteristics, the ABM is suitable for combining the simulated population-level results with the aggregate data considered here by aggregating individual outcomes with the cross-sectional data via model calibration process. The cross-sectional data considered here to calibrate the model results are summary statistics from individual-level survey data. Such relatively coarse resolution data constraints the model’s ability to reliably capture some dynamics at a fine-grained level, including smoking episodes, and dynamics of nicotine metabolism, and significantly limits resolution even at coarser levels – such as , with regards to broad changes in smoking behaviour over time. Notably, social network data and data associated with the rates of smoking and ECig use cessation and relapse would be particularly valuable to additionally incorporate for grounding the dynamics of smoking and ECig use, and for further analyzing decision-making effects from social network contacts.

Although with some limitations, the model outcomes can provide some explicable understanding of the consequences of the complex feedback between smoking and ECig use at the individual level, then allow us to analyze population-level smoking behavior. With the support of additional data sources, the model could be extended to simulate smoking and ECig use dynamics at a finer resolution – such as capturing the build-up of nicotine tolerance due to e-cigarette exposure – and used as a learning tool to advance health insights on smoking and ECig use behavior. Additionally, the model is also a useful tool for examining how SN influences smoking and ECig use, particularly among adolescents.

CHAPTER 5

MULTI-SCALE SIMULATION MODELING FOR PREVENTION AND PUBLIC HEALTH MANAGEMENT OF DIABETES IN PREGNANCY AND SEQUELAE

This chapter is based on the manuscript of "Multi-Scale Simulation Modeling for Prevention and Public Health Management of Diabetes in Pregnancy and Sequelae" by Yang Qin, Louise Freebairn, Jo-An Atkinson, Weicheng Qian, Anahita Safarishahrbiari, and Nathaniel Osgood, published in Proceedings of the 2019 Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation. The contribution of each author is presented in Section 1.5 of Chapter 1 [17].

While combining the ABM with cross-sectional data assists in grounding the model, as illustrated in Chapter 4, richer data availability can considerably increase the level of sophistication of the model that can be evidenced, which can extend the model resolution to advance insights into the dynamics of the underlying complex system, and provide additional information for grounding model behavior. The multi-scale hybrid model of DIP described in this chapter incorporates a sophisticated model structure that draws on building blocks characterized using SDM, ABM, and DES systems simulation techniques. The model is informed by both longitudinal individual-level data and cross-sectional data via calibration to minimize the discrepancy between empirical data and corresponding model outputs. In addition to grounding model behavior, the availability of the statistics from longitudinal individual-level data allows the model to extend its structure to capture the effects of inter-generational transfer of risk for GDM and T2DM and occurrence of diabetes amongst the the offspring.

5.1 Introduction

Gestational diabetes mellitus (GDM) is an increasing public health priority in the Australian Capital Territory (ACT), particularly on account of its impact on the risk of Type 2 Diabetes (T2DM) across the population [11, 12]. The increase of GDM is associated with increasing prevalence of risk factors including advanced maternal age [15], obesity [16], and sedentary behavior, growing GDM risk factors in those with family history of diabetes, and a growing number of residents whose ethnic background has traditionally been subject to elevated rates [11].

Mathematical models characterizing diabetes progression, glucose hemostasis, pancreatic physiology and complications related to diabetes have been built by many researchers [54, 55]. De Gaetano et al. [56] formulated a model representing the pancreatic islet compensation process, related to insulin resistance, beta-cell mass and glycemia (G) of a diabetic individual. Hardy et al. [58] proposed a model, characterizing mechanisms of anti-diabetic intervention and the corresponding impact on glucose homeostasis. Lehmann and Deutsch [83] modeled the physiology underlying the interaction between insulin sensitivity (K_{xgl}) and G of an individual with Type 1 diabetes (T1DM).

Health simulation models commonly apply one of three types of modeling techniques: system dynamics modeling (SDM), agent-based modeling (ABM) and discrete event simulation (DES). SDM captures and describes complex patterns of feedback and accumulation by solving sets of differential equations. While SDM can be applied at different scales [44], it is most commonly applied at the aggregated level, and its core components include the accumulation of elements (stocks), rate (flows), causal loops involving stocks (feedback), and delays [42, 43]. By contrast, ABM simulates complex social dynamics by characterizing emergent system behavior as the result of within-environment interactions between individual elements in a system that are referred to as agents. ABM readily captures heterogeneous characteristics of agents, including agent history, situated decision making, structured interaction between agents typically evolving along with multiple aspects of states and transitions and aggregation of individual outcomes [44, 45]. DES characterizes individual-level, resource-limited progression through structured workflows which often associated with service delivery, queuing processes, waiting for times and lists and resource utilization [40].

Previous studies examining the health burden of GDM and its risk factors have predominantly relied upon cohort studies, administrative data or clinical trials [18, 19, 20]. While filling a key set of research needs, given the dynamically complex nature of the interactions including feedback, accumulations, delays, heterogeneity, and interacting factors across many levels, it is difficult to use such studies to answer “what-if” question related with the risk factors and effects of interventions, particularly counter-factual whose outcomes have not yet been observed. Given the long time scales involved, cost, logistics, and ethical concerns, clinical trial studies may not be feasible for providing a timely evaluation of novel portfolios of clinical-level and population-level interventions (PLI).

In this work, we built a multi-scale hybrid model in AnyLogic (version 8.3.3) including SDM, ABM, and DES, to describe the dynamics of glycemic regulation (DGR), weight status and pregnancy, and to evaluate impacts of the interventions on DGR. While leaving most aspects of examination of model health findings to other forthcoming contributions, this paper introduces the design and structure of the model, provides illustrations of some of the types of interventions that the model can capture and simulation outputs.

The structure of the remainder of the paper is as follows: Model overview section describes the model structure and the simulation description. The next section briefly discusses model calibration and assumption. The model formulation section then describes the statecharts, DGR, weight dynamics, interventions, service delivery and offspring outcomes by hyperglycemia. Part 3 and 4 illustrate and discuss some of the model

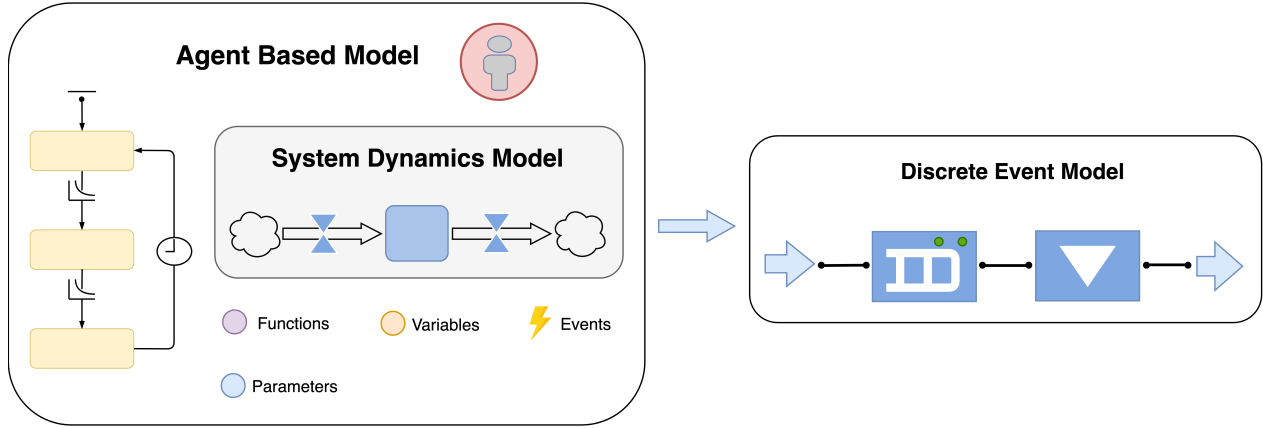


Figure 5.1: Illustration of hybrid model structure

outputs and limitations of the model.

5.2 Methods

5.2.1 Model Overview

Figure 5.1 describes the hybrid model structure of ABM, SDM, and DES. The `Person` class of the ABM includes the individual-level characteristics such as evolving states, actions that change them, and the rules to trigger those actions (all captured in statecharts), parameters and functions. The ABM further represents the family structure, weight at birth and evolution over the adult life course, individual history, inter-generational family context, pregnancy and diabetes classification, and implementation of the PLI. The SDM describing the DGR forms a sub-model encapsulated in the `Person` class. By encapsulating this SDM in the ABM structure, the model can capture side-by-side both individual characteristics and their evolution and continuous dynamics of the glucose-insulin system. The clinical service pathway for pregnant women in the ACT is described by a shared (global) DES, building on top of the ABM. The model will be discussed in its essentials in the following sections. Added elements of detail, the technical description associated with the model, are listed in the supplementary material, <https://www.cs.usask.ca/faculty/ndo885/GDM-ACT>, other material will be available later.

The model simulates a population of 200,000 female agents, each an instance of `Person` class. During the simulation, the agents can become pregnant, thereby experiencing the risk of GDM, and subsequently give birth, influencing the weight status and DGR of their descendants. The second-generation agents also have their life-course shaped thoroughly by model dynamics. The information available for the descendants is, therefore, richer than in the initial population. Thus, the simulation requires a burn time of 60 years.

5.2.2 Model Formulation

We discuss here several aspects of model formulation, particularly concentrating on statecharts, which encapsulate a discrete set of collectively exhaustive and mutually exclusive (lowest-level) states with respect to particular concerns, the actions by which the individual transitions between such states, and the rules under which such actions take place.

Pregnancy statechart (Figure 5.2) indicates whether an agent is pregnant or not, and their transitions through different stages of pregnancy. Female agents with ages between 15 and 50 can transit between the `notPregnant`, `planPregnant`, and `pregnant` states. Agents in the `pregnant` hierarchical state will be in one of three substates, corresponding to trimesters of pregnancy. `notPregnant` agents will either be in the `fertile` or `PostPartum` state. `PostPartum` agent will either be in the `breastFeeding` or `notBreastFeeding` state. Of these, two of seven state transitions are memoryless transitions driven by a hazard rate (henceforth known as rate transitions), `becomePregnant` and `leaveBreastFeeding`; the hazard rate for `becomePregnant` is an age-ethnicity-specific fertility rate [84]. While the others are timeout transitions triggered after a specified residence time. The timeout transition `birthTransition` is particularly notable, as it introduces a new agent into the model.

Population statechart separates the population into three categories, initial female population, female descendant and male descendant. Agents are initialized with different ages [85], and assigned their ethnicity according to ACT demographic information taken from the 2011 Australian Census and National Health Survey. Type of ethnicity includes Australian Born, Australasian Diabetes in Pregnancy Society at risk group (ADIPS) [86], Aboriginal and Torres Strait Islander (ATSI) and Other. All male descendants are excluded during simulation, and female agents leave the model upon reaching age 50. The female agents aged less than 50 leave the model by the age-specific death rate [87].

Dysglycemia classification statechart (Figure 5.3) divides the `G` of an agent into four categories: `T1DM`, `NormoglycemicAndIGR`, `T2DM` and `GDM` states, according to clinical classification categories. Agents can occupy one of four states, and switch states by checking whether the `G` of agents exceed the threshold of each state (known as condition transition). Reflecting the fact that residence in the `GDM` state is only an option during pregnancy, pregnancy status is also considered.

The `GDM` agents will either be in `NormoglycemicAndIGR` or `T2DM` state after pregnancy. Thresholds for `T2DM` and `GDM` state are denoted as G_{T2DM} and G_{gdm} , which are calculated by $C_{T2DM} \times G_t$ and $G_t \times C_{T2DM} \times C_{gdm}$, respectively, where C_{T2DM} , C_{gdm} and G_t are calibrated and equal to 1.636, 0.642 and 5.504, respectively.

DGR, the interaction between beta-cells, `G` and `KxgI`, is represented as an SDM based on the ordinary differential equation models of diabetes progression by De Gaetano et al. [56,57] and Hardy et al. [58]. To improve model scalability, a cyclic timeout event with time interval (Δt) is used to solve the compartmental equations in SDM [56,57,58]. Another cyclic timeout event with a time interval of 5 and 30 days during pregnant and non-pregnant periods, respectively, updates `G` and `KxgI` using the Newton-Raphson method and other components in SDM. Parameter and function details are listed in the supplementary material.

To capture dynamics of `KxgI` in different trimesters of pregnancy, postpartum and different weight status,

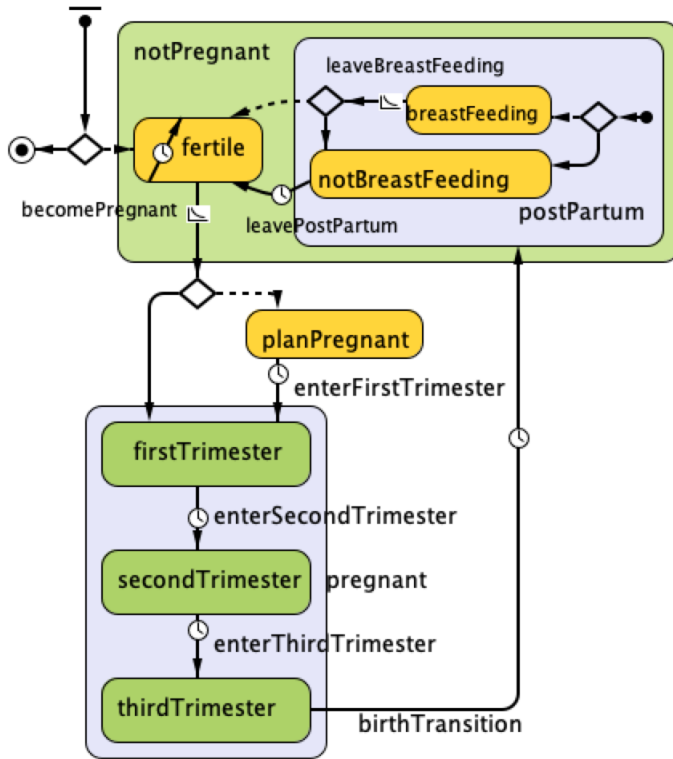


Figure 5.2: Pregnancy statechart

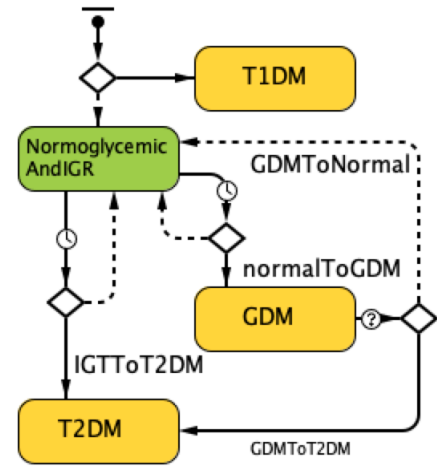


Figure 5.3: Dysglycemia classification statechart

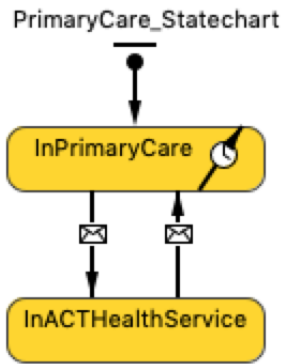


Figure 5.4: Primary care and ACT health service statechart

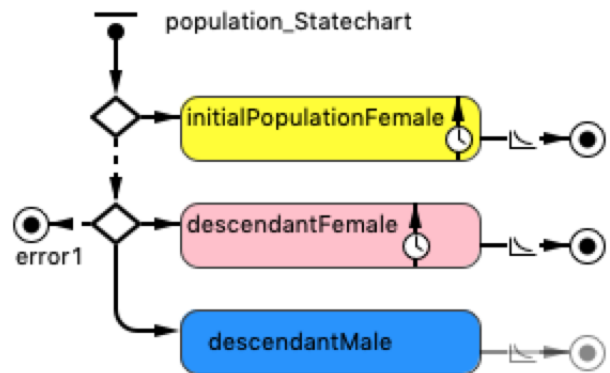


Figure 5.5: Population statechart

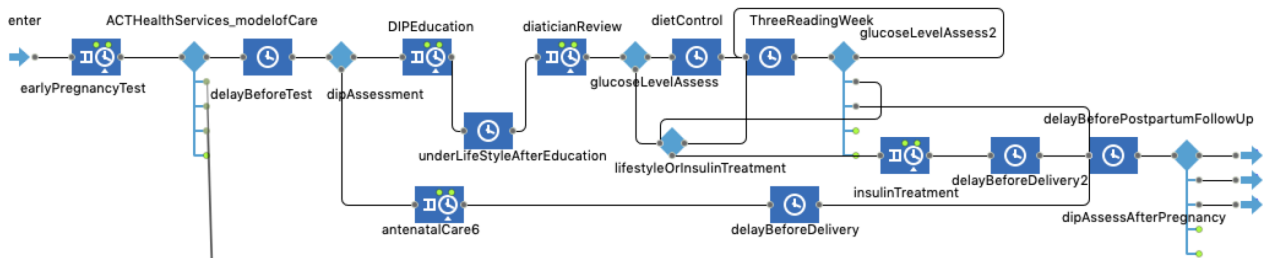


Figure 5.6: DES of the ACT health clinical service pathway

respectively, we modified the (exogenous) equations giving K_{xgI} over time introduced by De Gaetano et al. [56, 57], Function `linearInterpInPregnancy()`, `InsulinSensitivityInPostPartum()` and `insulinSensitivityBasedWeight()`, calculate the K_{xgI} in different trimesters of pregnancy, postpartum and different weight status, respectively. `linearInterpInPregnancy()` assumes that K_{xgI} is declining during pregnancy. By this assumption, the function is formulated by linearly interpolating the K_{xgI} at beginning, at three month and at the end of pregnancy proposed by Catalano [88] using the `LinearInterpolator` of `org.apache.commons.math3` library. The model assumes the diminished K_{xgI} in pregnancy will gradually recover during the postpartum period to the value it would have held absent the pregnancy. Therefore, `InsulinSensitivityInPostPartum()` takes time after delivery, the more the time close to $\tau_{recovery}$, the more it close to the K_{xgI} whose value is equal to the point on the K_{xgI} trajectory as if the agent is not pregnancy. The model further assumes the K_{xgI} of overweight and obese agents would decline faster over age than that of agents with normal weight. Based on this assumption, in `insulinSensitivityBasedWeight()`, we modified the `KxgI0` and `t1` in the equation of De Gaetano et al, by changing `KxgI0` and `t1` with value of BMI and weight category, respectively. `t1` for overweight and obese agents were represented as `t1_overweight` and `t1_obese`, and `KxgI0` for different BMI was calculated by table function `tf_BMI2KxgI()` [89]. In addition to the modification of equations giving K_{xgI} , we modified the equation giving the spontaneous recovery rate of the pancreas (τ_η) for ADIPS. While ADIPS represents agents from a recognized high-risk group with respect to GDM, the empirical data revealed that the ADIPS group actually had a higher proportion of healthy weight agents than that were present in the other groups, indicating that weight as a risk factor did not fully account for the higher risk levels. Therefore, to capture the high incidence of GDM of ADIPS, the model assumes the τ_η of ADIPS declines faster than that of other ethnic groups. Furthermore, to investigate effects on various types of intervention on the DGR, we incorporated the mechanism introduced by Hardy et al. [58] for the impact of lifestyle change (LC), metformin treatment (MT) and insulin treatment (IT) on K_{xgI} . Elements of interventions making use of the LC, IT and MT are discussed in the next sections.

Weight dynamics are characterized as a continuous variable of BMI value, and a variable of Z-score of a BMI distribution (BMID), representing the position of BMI within the age group (AG) specific BMID. Upon entry to adulthood, agents are assigned a BMI value based on an AG specific BMID introduced by Hayes et al. [90], and its corresponding Z-score calculated by the BMI and mean of the BMID. Hayes et al. [90] reported that the BMID of the population within AG moves toward higher BMI value through their life course. Applying an identical Z-score into the BMID of different AGs may position the agents into different weight categories. Therefore, for simplicity, the Z-score of agents are assumed to stay the same as they age, unless intervention or pregnancy [19] changes their BMI value and assigns a new Z-score to them. When an agent transfers from one AG to another, the BMI value of next AG of the agent will be calculated by applying the Z-score to the BMID of next AG in an event with a cyclic timeout of 10 years. As a continuous variable, another cyclic timeout event with an interval of 1 year is used to make the BMI of current AG change towards the BMI of next AG gradually. With this BMI-Z-score mechanism, we also captured BMI

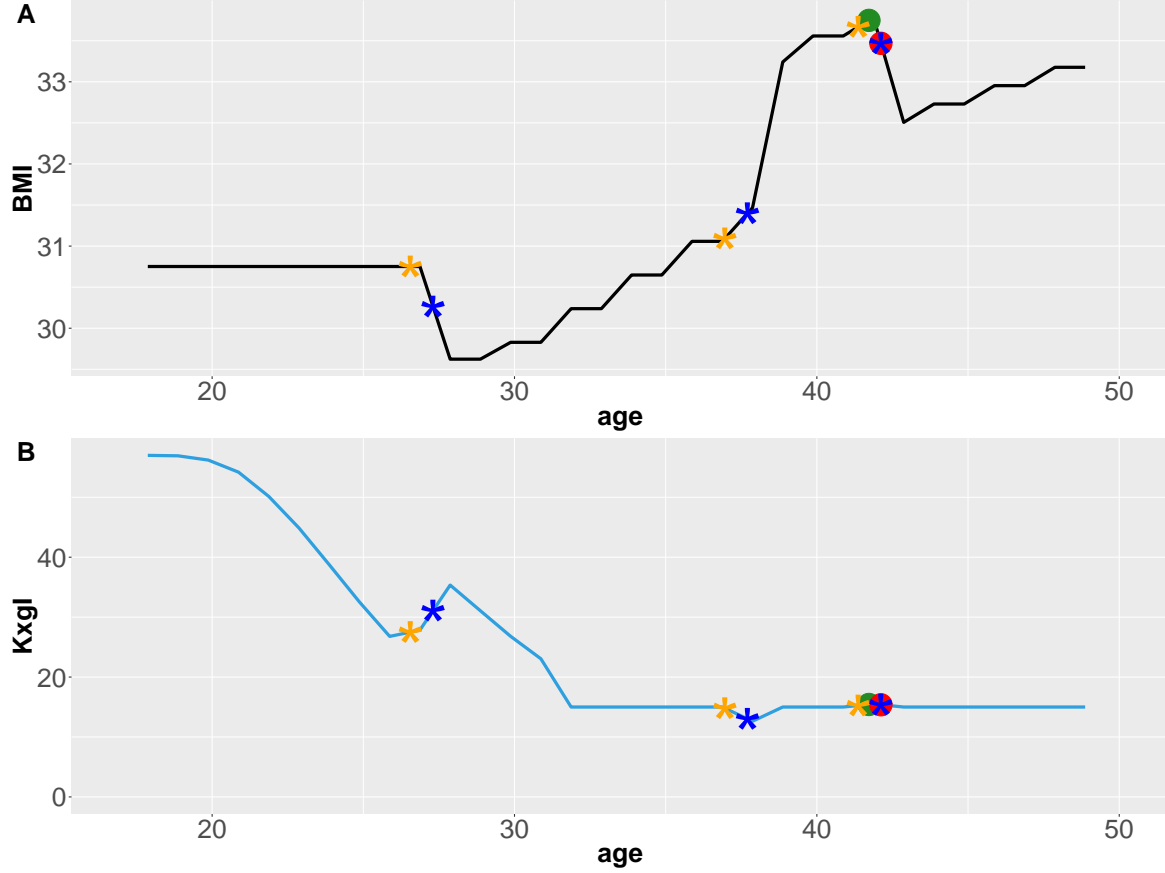


Figure 5.7: Individual trajectory of BMI change (A) and K_{xgl} change (B) over age without PLIs and Services. Green and red dots are the start and end of the GDM period, respectively. Orange and blue stars are the beginning and the end of the pregnancy, respectively.

change following pregnancy; due to space considerations, the interested reader is referred to the supplemental material. Time-Varying weight distribution is required in light of the simulation burn time. We, therefore, employed importance sampling using an alternative BMID of female adults aged 25 to 64 years in 1980 and 2000 [91] and AG specific BMID in 1995 and 2008 [90], to estimate the AG specific BMID in 1980.

ACT clinical service pathway, *Services*, is modeled using DES, as shown in Figure 5.6. And in *Person*, a statechart reflects type of health care that an agent is currently being delivered, which is separated as the *InPrimaryCare* state reflecting that a non-pregnant woman is receiving usual health care services through a general practitioner, and the *InACTHealthService* state reflecting that a pregnant woman is moving through the *Services*, as shown in Figure 5.4. The DES and statechart not only models the effects of LC and IT in reducing the risk of progression to T2DM after delivery and implemented PLIs in the *InPrimaryCare*, but also leave room for investigating resource use and costs associated with service provision. The blocks (i.e. *dipAssessment*, *antenatalCare*, *dieticianReview* and *lifestyleOrInsulinTreatment*) in the *Services* form a sequence of operations, providing pregnant agents the DIP assessment test, education of LC, IT for the agents with DIP, and further deliver postpartum checks to agents.

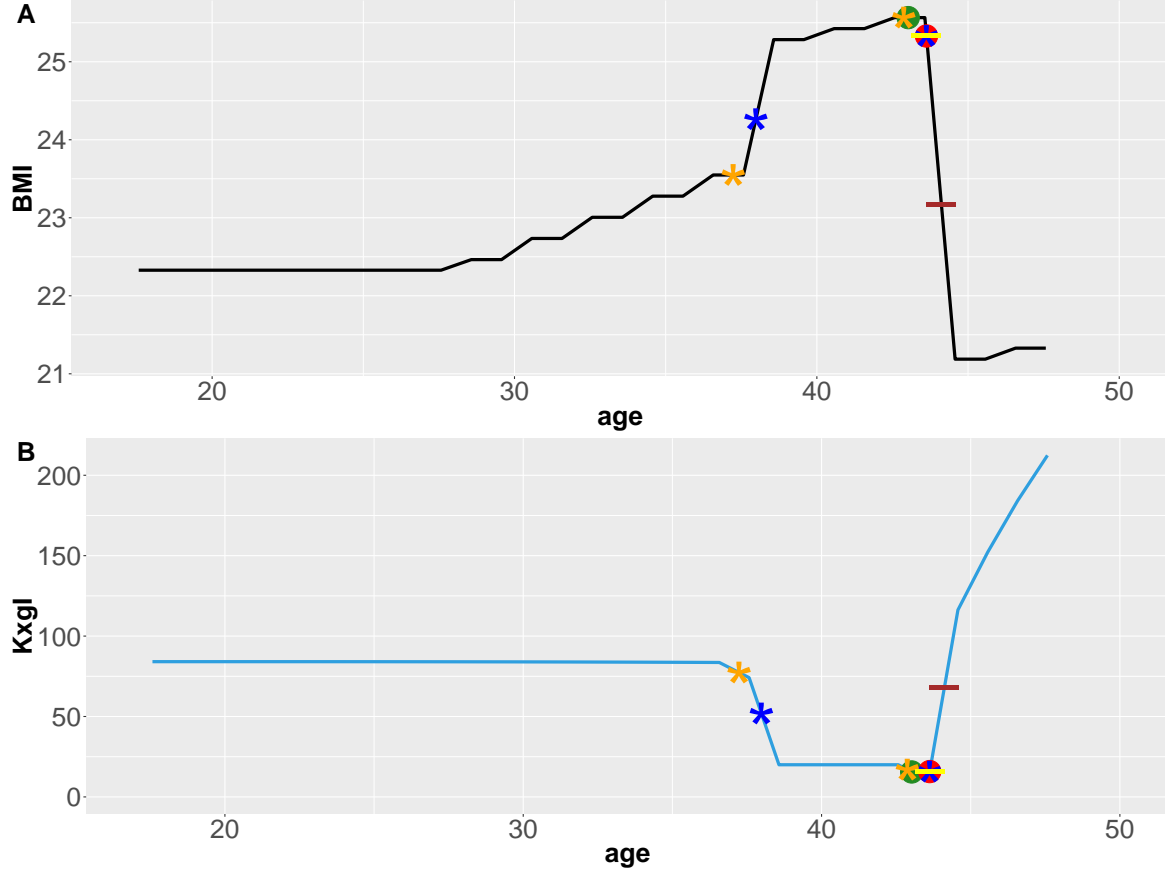


Figure 5.8: Individual trajectory of BMI change (A) and K_{xgl} change (B) over age with DRI. Color Labels of GDM and pregnancy are the same as Figure 5.7. Yellow and brown bars are the start and the end of the DRI, respectively.

Details of the DES are described as follows. The agent enters the `Services` at beginning of their pregnancy, then receives an early pregnancy assessment test at `earlyPregnancyTest` block, then waits at `delayBeforeTest` block for the DIP tests in `dipAssessment` block. The DIP test time varies with individuals. Test time is at 8-12 weeks gestation if the agent is considered as a high-risk group (e.g., obese, ADIPS or whose age is over 35), but the test time of non-high-risk agent is at 26 - 28 weeks gestation. At `dipAssessment`, the agent who passed DIP test will be referred to `antenatalCare6` block for standard antenatal care and remain there until delivery. However, if the agent failed the test, the agent first moves to the `DPIEducation` block, then starting 1 week LC in `underLifestyleAfterEducation` block, then the agents are referred to `dieticianReview` to check G again. In `dieticianReview`, if the G does not exceed the G_{gdm} , the agent continues LC in `dietControl` block for 1 week, then the agent's G is monitored and checked in `ThreeReadingWeek` and `glucoseLevelAssess2` blocks every week until delivery. During this time, if the G in any week is higher than the G_{gdm} , the agent is referred to `lifestyleOrInsulinTreatment` block, otherwise, the agent goes to `delayBeforePostpartumFollowUp` block for postpartum check. If G of the agent exceeds the G_{gdm} in `dieticianReview`, the agent will directly be referred to `lifestyleOrInsulinTreatment`, determining whether the agent will continue LC or start IT. The model assumes

that if the G is higher than G_{T2DM} , the agent will start and continue IT until delivery, otherwise the agent will go back to `ThreeReadingWeek` and `glucoseLevelAssess2`. At six week after delivery, the agent receives postpartum follow-up DIP assessment in `dipAssessAfterPregnancy` block determining follow-up interventions, then leave the `Services`. In the mean time, the state of agent is transferred from `InACTHealthService` to `InPrimaryCare`. The follow-up interventions includes only LC for agents whose G is higher than G_{gdm} but not yet developed into T2DM, combined LC, IT and MT for agents with T2DM or T1DM.

PLI includes consideration of a public health messaging and mobile app support intervention (PHMMASI), health professional support intervention (HPSI), diet review intervention (DRI), and public health messaging and support intervention (PHMSI). The difference between the `Services` and the PLI is that PLIs are initiated during the non-pregnant period, while the `Services` is triggered during pregnancy. All PLIs share a similar mechanism of taking optional LC and BMI reducing. Specifically, overweight and obese agents reduce their BMI, drawing the extent of that reduction from a normal distribution, while the normal weight agents keep their BMI invariant. The interventions are variants of each other with respect to who they are the target, the intervention triggering time, and the length and strength of adherence of the LC. In detail, the agents with age between 20-35 take the PHMMASI and retake it according to certain probability [92]. The HPSI takes place at the `planPregnant` state, and works on women with risk factors, e.g., BMI > 28, age > 30, ADIPS ethnicity [93]. DRI and PHMSI both take place between pregnancies and target women who had DIP in previous pregnancies and on women who have given birth, respectively. For the HPSI and PHMSI, the adherence and length of LC are flexible, whereas the agents who are subject to DRI take mandatory, lifelong, strongly adherent LC.

The outcomes for baby and mother including birth weight (e.g., macrosomia), type of birth (e.g., Caesarean section), NICU admission, and shoulder injury, are triggered in `birthTransition` of the pregnancy statechart. The probability of occurrence of baby outcomes was calculated based on the study by The HAPO Study Cooperative Research Group [94]. Furthermore, information of the mother is passed on to the new child, including DIP status, age, weight status, and ethnicity. The mother's DIP status influences the K_{xgI} of the child by multiplying a coefficient to the K_{xgI} calculated by the modified equations giving K_{xgI} over time.

5.3 Model Development Process

The multi-scale hybrid simulation model of DIP was developed in the joint sessions with domain experts across fields. The team worked collaboratively to discussed the model scope, the risk factors for the growing epidemic of GDM, the dynamics of diabetes progression, and the mechanisms of selected interventions. The model development process went through three big categories: design of the concept, implementation, and parameter calibration. The validation and calibration process made the model capable of exploring and producing results for demonstration. The model was previously built, modified, extended, and simplified by the supervisor and a series of students, including Ph.D. students Weicheng Qian and Allen Mclean, and M.Sc.

student Anahita Safarishahrbiari. Nathaniel Osgood helped with the manuscript, supervised and advised on the entire study. Allen McLean initiated the project and collected background information and broad structure of the model, particularly the DES representing treatment and exploring the resource allocation of the public health service pathway, which is up to version 16 of the model. Weicheng Qian, Nathaniel Osgood, and Geoff McDonald constructed the key system dynamic casual loop diagram depicting dynamic relationships including beta-cell compensation and dry-out, dynamics of glycemia regulation. Weicheng Qian overhauled and rebuilt core logic among Glucose-Insulin-Beta Cell (replacing discrete state charts to SDM) from version 17 to version 19 of the model (due to their adoption of minor versions, they did in total 23 minor versions) and independent concept proving model (7 major versions). The key SDM was further refined and reused until the current version of the model (version 107); Anahita Safarishahrbiari implemented fast system dynamic solver applying Newton-Raphson numeric method, constructed the mechanism of insulin resistance changing with pregnancy and treatment, which can be referred to version 37 of the model, and initiated LaTeX based documentation for the model archiving the source and usage of the parameters, which can be accessed from <https://www.cs.usask.ca/faculty/ndo885/GDM-ACT>. Yang Qin (myself) contributed to the model development in the following aspects:

- I extensively debugged the fast system dynamic solver, which had numerical error under the condition of insulin sensitivity being extremely low. With this issue, the model was unable to simulate the agents in an overweight or obese state for a long time, and stop model running by giving not a number (NaN) error.
- Along with Weicheng Qian, we profiled the model via JVisualVM to scale up the model performance. Using 2.2 GHz Intel Core i7 and 16GB RAM, the un-optimized model (version 37), having a memory consumption of 15073.28M, simulated 10000 agents over 40 years for 5676.49 seconds per realization, and was unable to run with 40000 population. After performance optimization, the model takes 27.56 seconds to run with 10000 agents over 40 years using 1146.88M memory size. The model can simulate 200,000 agents over 93 years for 23.25min per realization.
- I calibrated the parameters of the model using AnyLogic optimizer. The calibration experiment ran 2000 iteration with 100,000 female agents to reach a minimal objective, 61.2. During model calibration, I further identified the missing model assumption about the spontaneous recovery rate of the pancreas for ADIPS.
- I contributed to refine and elaborate model mechanism. For modeling BMI dynamics, I added (continuous) BMI evolution as age cohorts shifting, and estimated historical age-specific BMI patterns through importance sampling. For modeling disease mechanisms, I refined the expression of offspring outcomes based on the maternal hyperglycemia status, and modified equation of calculating insulin sensitivity over time by dependent on weight status and interventions, and elaborated the equations giving the spontaneous recovery rate of the pancreatic beta cells for diabetes high-risk ethnicity groups.

- I innovated and redesigned the model by separating model logic and presentation-data collection to modularize complex hybrid models using active-agent class in AnyLogic.
- Along with Louise Freebairn and Nathaniel Osgood, I implemented five population-level experiments, including aspects of a public health messaging and mobile app support intervention, health professional support intervention, diet review intervention, and public health messaging and support intervention. The interventions shed light on the effect of lifestyle change and weight reduce on improving glycemia regulation, and further supports decision making and following projects.
- Along with Louise Freebairn, I implemented an DES of ACT health service pathway.
- I am one of two co-contributors, providing aggregate population level results and individual-level trajectories, towards an interactive dashboard for exploration and demonstration by the ACT Minister of Health.

5.4 Model Calibration

To estimate poorly- or non-measured parameters and to support the projection of status quo future incidence of DIP using model outputs, A baseline model without interventions was calibrated against the following historical data: the incidence of DIP of each ethnicity in ACT from 2008-2016, the prevalence of macrosomia by DIP status of in ACT from 2010-2016, and number of diabetes by age 30 of offspring by their mother’s diabetic status (e.g., GDM, T2DM) and birth weight from population-wide administrative data from Saskatchewan, Canada. Calibrating the model against the empirical data of number of diabetes of offspring by age 30 allowed the model to capture the effects of inter-generational transfer of risk for GDM and T2DM which has been recognized from multi-generational epidemiological studies [13, 95], and occurrence of later-life diabetes. Birth weight (e.g., macrosomia) can serve as an important marker of control of G in utero, and epidemiological studies suggest that it further influences the tendency towards GDM. The details of the calibrated parameters are listed in Appendix A.

In addition to calibrating the parameters of the model, the missing model assumption about the spontaneous recovery rate of the pancreas (τ_η) for ADIPS was identified during this process. As a high-risk group, agents from the ADIPS group show a higher incidence of GDM than other ethnic groups. However, the empirical data revealed that the ADIPS group had a higher proportion of healthy weight agents than the other ethnic groups, indicating that weight as a risk factor can not fully explain the elevated incidence of GDM. Therefore, the assumption that the τ_η of ADIPS declines faster than that of other ethnic groups were incorporated into the model for capturing the high incidence of GDM of the ADIPS group.

The objective function of the calibration is shown in Equation 5.1, where d_m and d_h , and n denote the model output, empirical data and count of empirical data point. The calibration experiment ran 2000 iteration with 100,000 female agents to reach a minimal objective, 61.2. The results of the calibration are

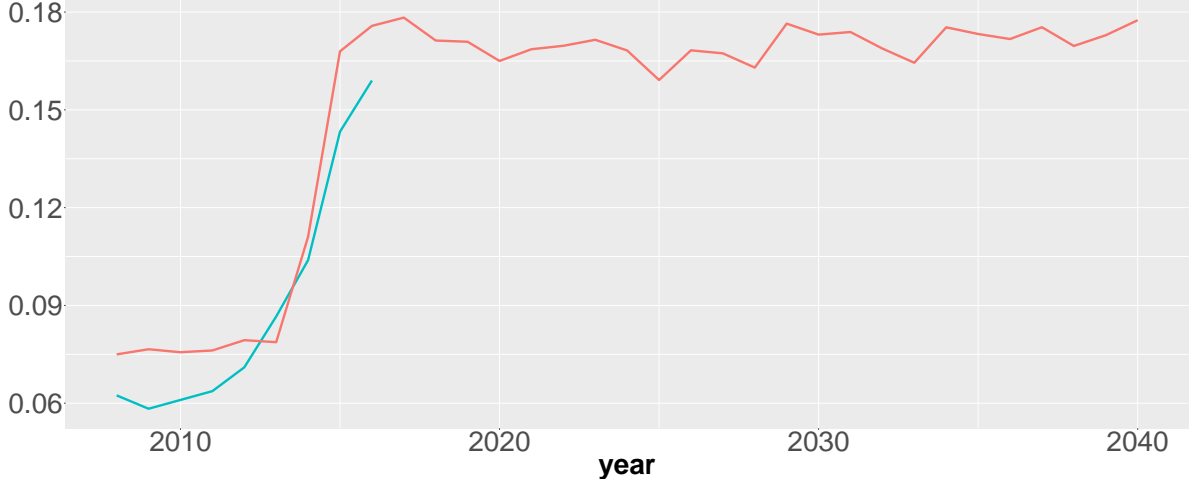


Figure 5.9: Model simulated incidence of DIP (red line), and empirical incidence of DIP in ACT from 2008 to 2016 (green line)

shown in Figure 5.9.

$$\sum_{i=1}^n \left(\frac{d_{mi} - d_{hi}}{(d_{mi} + d_{hi})/2} \right)^2 \quad (5.1)$$

As the empirical data shown in Figure 5.9, the incidence of DIP started to increase dramatically in 2015, due to the threshold for diagnosing DIP was modified after 2015. Instead of calibrating the diagnosing threshold used before and after 2015 separately, the threshold for diagnosing DIP after 2015 was calculated using variable `incidenceOfDIPIncrease`. The `incidenceOfDIPIncrease` represents the fraction increase was required to make the model generated the incidence of DIP close to the empirical data. With the fraction increase of incidence of DIP, the threshold for diagnosing GDM after 2015 can be calculated by following steps: (1) The G of all agents were stored in a list and sorted by descending order. (2) The number of DIP cases after 2015 is equal to the production of `incidenceOfDIPIncrease` and incidence of DIP and the number of birth of 2013. (3) The new diagnosis threshold after 2015 is equal to the G represented by the position index equal to the number of DIP cases.

5.5 Performance Optimization

To scale up the hybrid multi-scale model of DIP, the following model configuration modifications were made. The Garbage First Garbage Collector(G1GC) was switched to ParallelGC (OldParallel) given the consideration that the throughput of processing simulation events is of our favor, rather than the low latency on visualizing animation or responding to user interactions. We further made shared constant agent fields as the statistics fields of the Java class to reduce memory footprint and class instantiation time. We further moved the agent level event scheduler out to the main class to reduce the memory cost for the instantiation of an `EventOriginator`. The agent level visualizations, e.g., plots and histograms, whose presentation are stored in arrays, were also removed for saving memory and time for updating the visualization components. Using 2.2

GHz Intel Core i7 and 16GB RAM, the un-optimized model, having a memory consumption of 15073.28M, simulated 10000 agents over 40 years for 5676.49 seconds per realization, and was unable to run with 40000 population. After performance optimization, the model takes 27.56 seconds to run with 10000 agents over 40 years using 1146.88M memory size. The model can simulate 200,000 agents over 93 years for 23.25min per realization.

5.6 Logic-Presentation Separation

Separating model logic and presentation is a key to reduce complexity, decouple, and encapsulate models from both the classical software engineering perspective and the modern best practice of simulation model development perspective. This section focuses on discussing a novel set of guidelines that can help modularized complex hybrid models and improve logic-presentation separation rules that have been successfully coined in well-know software architectures such as Model-View-Control (MVC) model and frontend-backend web design. The exploratory nature of model development requires modelers to balance quick and light implementation and flexibility and robustness. The active-agent based development environment as AnyLogic has implemented, provides a tangible approach that being able to predict the potential growth of a stylized model, to balance quick implement and flexibility and robustness.

In an active-agent based development environment, an active-agent class is a basic module of the model encapsulates internal logic which can be reused and further inherited for an extension. The relationship can be abstracted as containment and messaging. To coin common patterns of messaging, the concept of environment and network is then introduce to automatically establish messaging channels (named as `connections`) among agents within the same parent environment and connected following classic patterns that is parameterizable built-in network types, such as small-world network, scale-free network, distance-based network, and furthermore as an overlay of GIS environment. Similarly, for DES models each module such as `Seize`, `Delay`, `Release`, `Process` are active agents that connects to previous and next module in a pipeline via agent-connections. As a further step for decoupling logic from visualization, rather than directly put visualization components, such as collecting datasets, plots and collecting events, transform function inside the agent which producing data itself. One could create a new active agent class dedicated to visualizing a certain type of data, such that the agent who produces data can be easily paired with a group of "data consumer".

The `Person` agent of previously built model includes the parameters, variables, distribution, datasets, functions, and events related with weight dynamcis, diabetes progression dynamics and interventions. In model refactoring process, the modules of weight dynamics, diabetes progression dynamics (SDM), and interventions were separated from agent `Person` and placed into three agents, `WeightDynamics`, `DiabetesDynamics` and `Interventions`, while the statecharts of population, pregnancy, and dysglycemia classification are kept in agent `Person`. A single agent of each of the three agent type is instantiated in agent `Person`, as shown

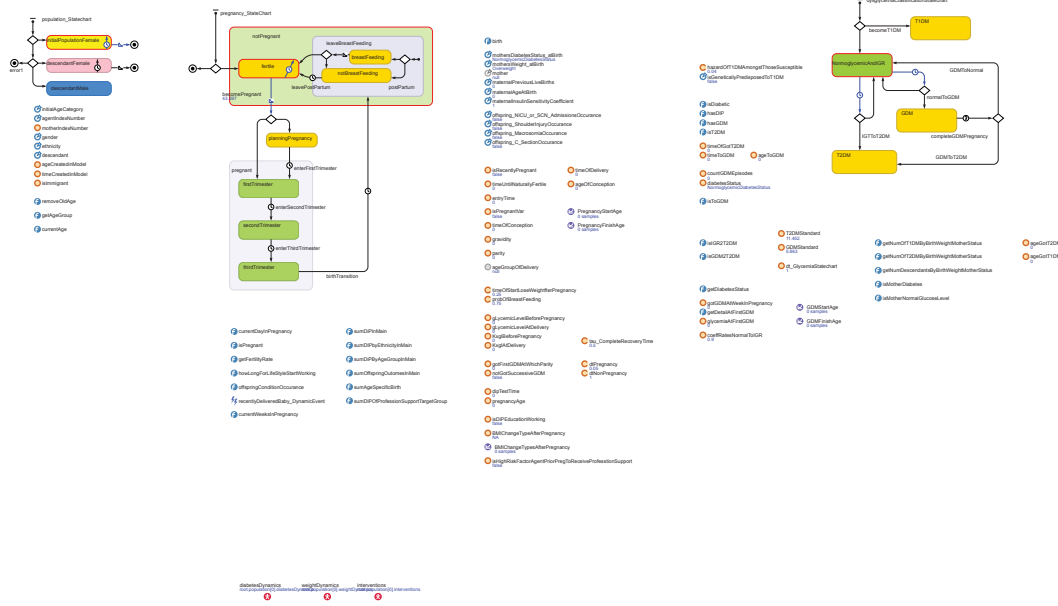


Figure 5.10: Agent Person after model refactoring

in Figure 5.10. With this model hierarchy, agent **Person** is the upper level agent to agent **WeightDynamics**, **DiabetesDynamics** and **Interventions**. After model refactoring, agent **Person** allowed flexible pulg-in of other sub-modules, and has a clear boundary of each components based on their logic. In a similar manner, visualization and population level statistics are also placed in separate agents and instantiated a single agent in agent **Main**, e.g., aggregated results related to interventions, diabetes dynamcis, pregnancy, offspring and weight dynamcis were placed in agent **statisticsForIntervention**, **statisticsForDiabetesDynamics**, **statisticsForDiabetesAndPregnancy**, **statisticsForOffspring**, and **statisticsForWeightAndPopulation**, respectively.

5.7 Results

5.7.1 Individual Trajectory

We show here several scenarios that demonstrate the functioning of the model at the level of an individual's health history. Figure 5.7 A and B show the individual trajectories of BMI changes and K_{xgI} over age without the PLIs and the **Services**, respectively. Figure 5.7 A illustrates the agent entered adulthood with a BMI of 30.75, and her BMI reduced over one BMI unit after the first and third pregnancy and three BMI units after

the second pregnancy. Other than pregnancy, Figure 5.7 A also demonstrates continuous BMI change over age. The K_{xgI} remained constant under the age of 18 but declined over time according to the value of BMI after the age of 18. The agent developed GDM at the third pregnancy, as shown by the dots in Figure 5.7. Furthermore, Figure 5.7 B reflects the decreasing K_{xgI} during pregnancy and recovery in postpartum. We can see from Figure 5.7 that K_{xgI} is declining in parallel to, and, in fact, in response to the increase in BMI.

From Figure 5.8, we can see that the agent increased over one BMI unit after the first pregnancy, and K_{xgI} was decreased in response to this BMI increase. In the second pregnancy, the agent developed GDM but retained their BMI and corresponding Z-score after pregnancy. But the DRI reduced that agent's BMI and significantly increased K_{xgI} from 20 to 116 at the end of BMI reduction period (6 months), following which the K_{xgI} continued to increase due to strong adherence in LC, as shown by the bar labels in Figure 5.8 A and B.

Figure 5.11 demonstrates the individual trajectory under PHMSI. we can see that the agent had the PHMSI, including both BMI reduce and two years of strong adherence to LC. A second pregnancy, the agent showed a strong increase of K_{xgI} , due to LC and BMI decrease. Furthermore, with no BMI change after age 42.9, the K_{xgI} decreased after the LC stopped, as shown in Figure 5.11

5.7.2 Population-Level Outcomes

In addition to the individual trajectories discussed in the previous section, population-level results generated by the model under four scenarios, by taking the average results of 36 simulation run, are presented to demonstrate the functioning of the model at the aggregating individual trajectories and capturing the effect of various interventions. The scenarios are as follows : (1) In scenario 1 (Scn1), the baseline scenario, the model assumes that no agent can receive the interventions during simulation (2) Model in scenario (Scn2) posits that 80% of agents take PHMMASI with BMI reduces and 50% of agents who have taken PHMMASI can retake this intervention. (3) For scenario 3 (Scn3), all agents in the model can take PHMMASI once with BMI reduces, and a lifestyle change in addition to BMI reduces. (4) In scenario 4 (Scn4), the model assumes that 80% of agents take HPSI with only BMI reduces. Figure 5.12 illustrates BMI reduces and K_{xgI} improvement of individuals via interventions can be a key driver to decrease the prevalence of DIP. The degree of BMI reduces during interventions directly affect the prevalence of DIP, as shown in Figure 5.12, model in Scn2 yields a lower prevalence of DIP than the model in Scn4.

As the glycemic level is driven by the SDM component in `Person` class, the prevalence of DIP also reflects the average beta-cell mass, glycemic level, and K_{xgI} , as shown in Figure 5.14, 5.15, and 5.16, respectively. The simulated outcomes from the model in Scn1, the highest prevalence of DIP among the four scenarios, as shown in Figure 5.12, can be substantiated by results in Figure 5.14 and 5.16 of the highest glycemia and lowest K_{xgI} . Notably, the model in Scn1 generating the highest level of beta-mass reflects the model assumption about the function of increasing beta-cell mass to compensate the insulin sensitivity decrease due to high glycemia level.

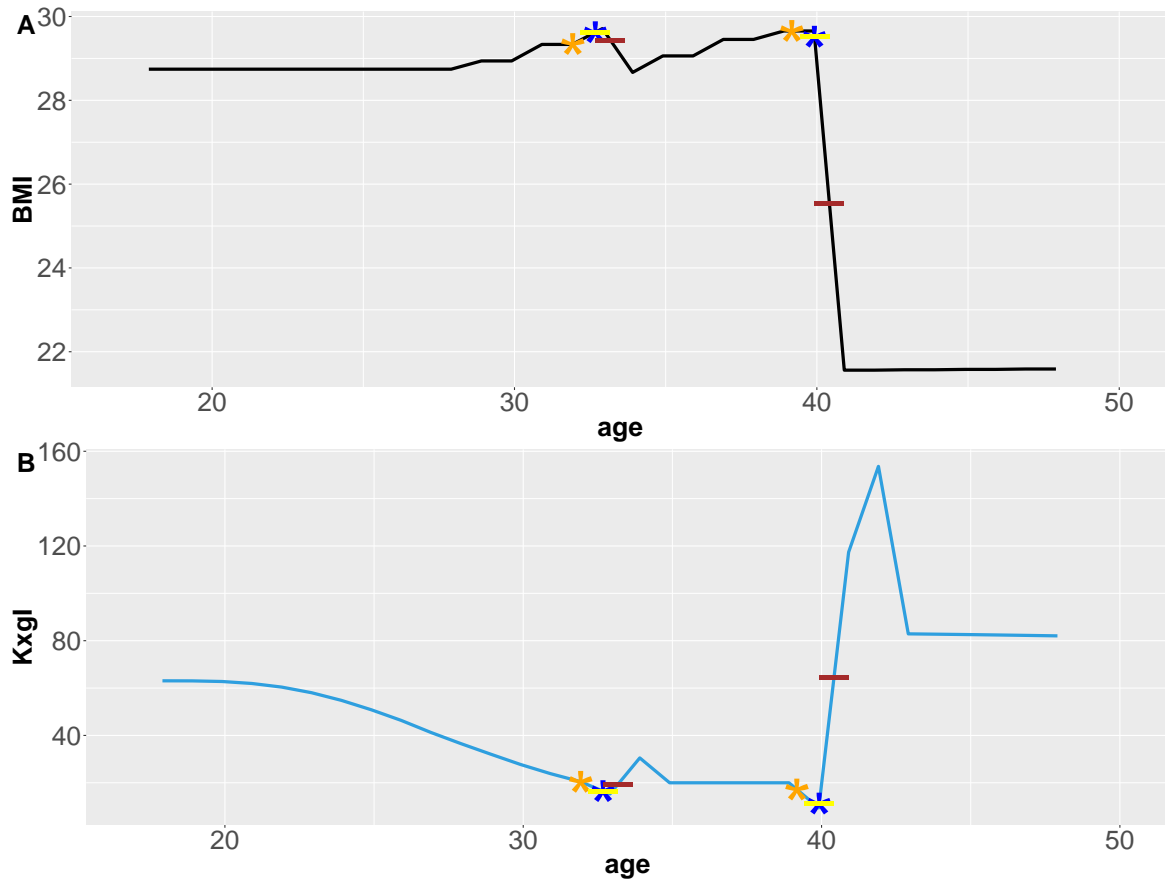


Figure 5.11: Individual trajectory of BMI change (A) and K_{xgl} change (B) over age under PHMSI. Color Labels of GDM and pregnancy are same with Figure 5.7. Yellow and brown bars are the start and the end of the PHMSI, respectively.

The model further demonstrates its potential for examining various risk factors at the individual level and their effects on population-level outcomes. The results in Figure 5.13 compares the prevalence of DIP generated by the model under the assumption that all children are overweight or obese and that from a model in `Scn1`. Figure 5.13 shows that obese and overweight at childhood is a risk factor that can increase the prevalence of DIP.

5.8 Discussion

This paper has described a novel multi-scale model that utilizes three types of system simulation methods to provide a versatile, powerful and general platform for examining interventions to address the growing epidemic of GDM and T2DM in the ACT. The model achieves such versatility by virtue of maintaining a core underlying physiological representation that captures the common generative pathways mediating diverse needs in the model, to capture effects of lifestyle and clinical interventions, to capture clinical categorization, to represent the effects of each of pregnancy, aging and BMI change, and the longer term effects of one

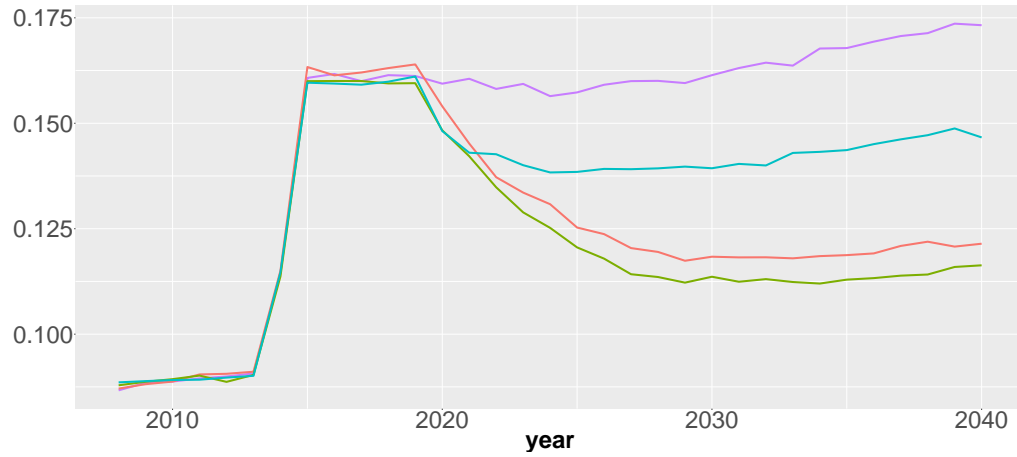


Figure 5.12: Prevalence of DIP generated by the model in Scn1, Scn4, Scn3, and Scn2, are labelled as purple, blue, red, and green line, respectively

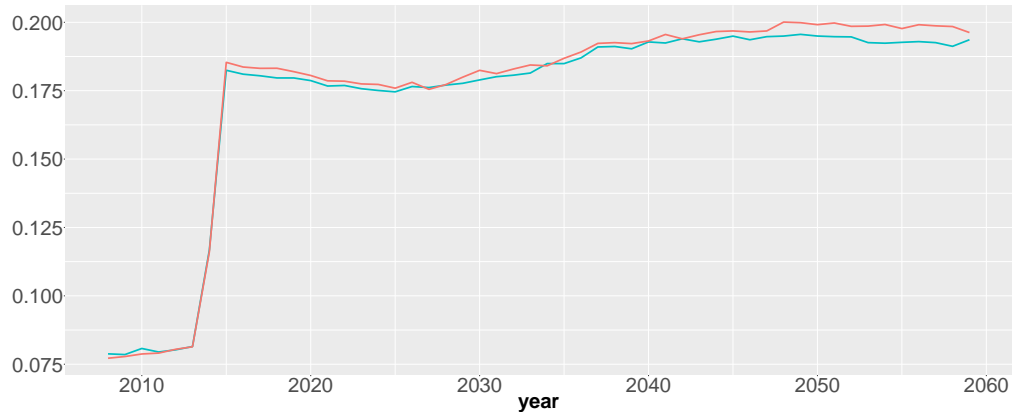


Figure 5.13: Prevalence of DIP from the model in Scn1 and the scenario assuming all children are either overweight or obese, are labelled as green and red line, respectively

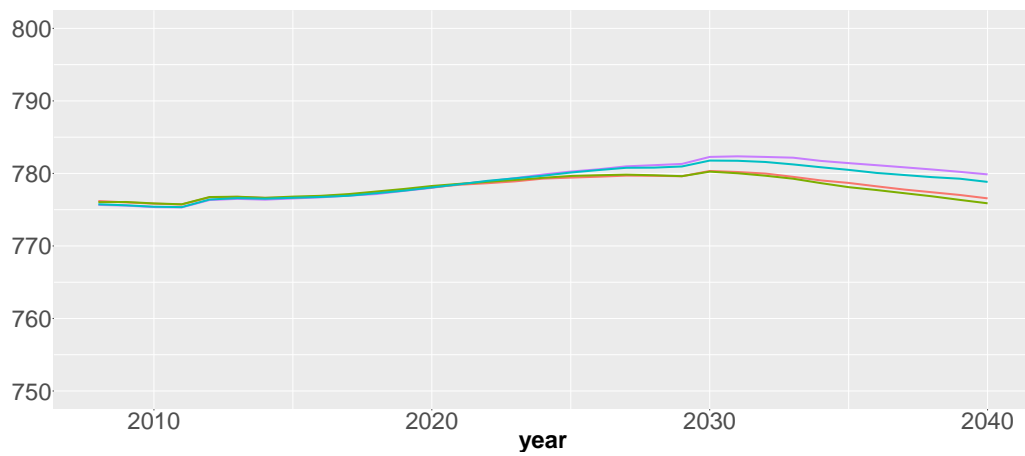


Figure 5.14: Average of beta-cell mass generated by the model in Scn1, Scn4, Scn3, and Scn2, are labelled as purple, blue, red, and green line, respectively

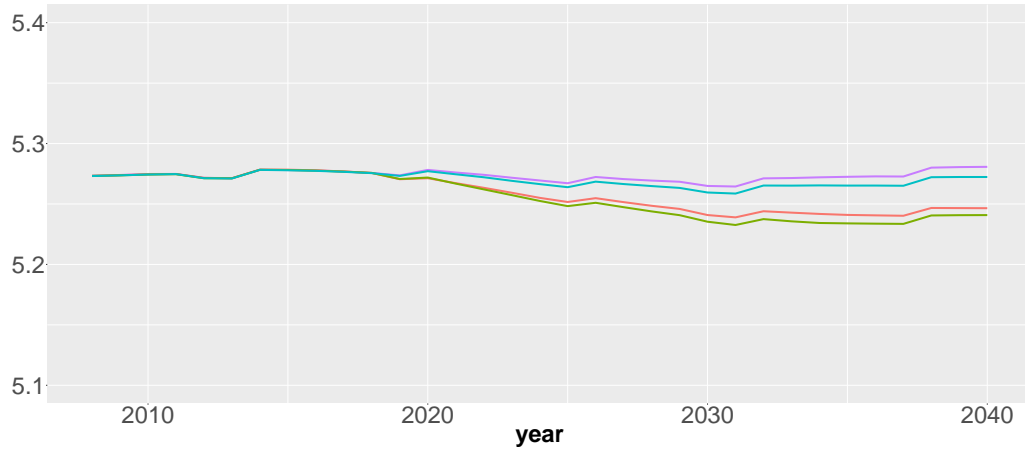


Figure 5.15: Average of glycemia generated by the model in Scn1, Scn4, Scn3, and Scn2, are labelled as purple, blue, red, and green line, respectively

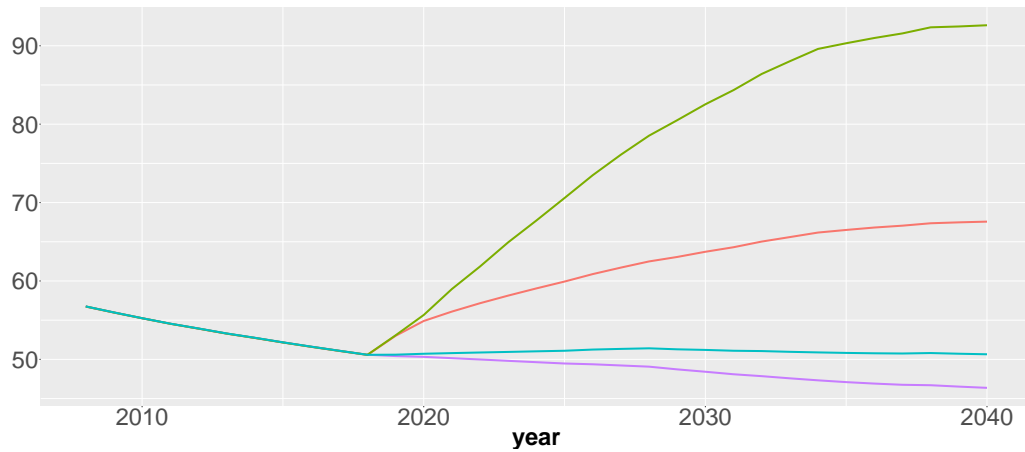


Figure 5.16: Average of KxgI generated by the model in Scn1, Scn4, Scn3, and Scn2, are labelled as purple, blue, red, and green line, respectively

pregnancy (via beta-cell mass and function) on later pregnancies and subsequent material risk of T2DM, and outcomes of interest. Such a representation can also flexibly capture the impacts of maternal status on the offspring.

A high level of heterogeneity at the individual level, e.g., family context, risk factors for diabetes and life course trajectories motivated the use of ABM as the core component of this hybrid model. The ABM permits a high-resolution representation of relevant dynamics of individual objects and further allows the implementation of finely targeted interventions. Compared to ABM, SDM simulates a system in a more abstract and general way. The high level of abstraction of DGR makes it a suitable candidate for SDM. The Services can be described as a sequence of operations, DES, therefore, was selected to model the Services, and to study the resource allocation and effect of clinical interventions.

While empirical models of necessity represent simplifications of processes in the world, the model here includes a requisite degree of detail to capture a remarkably broad set of factors. Nonetheless, they remain

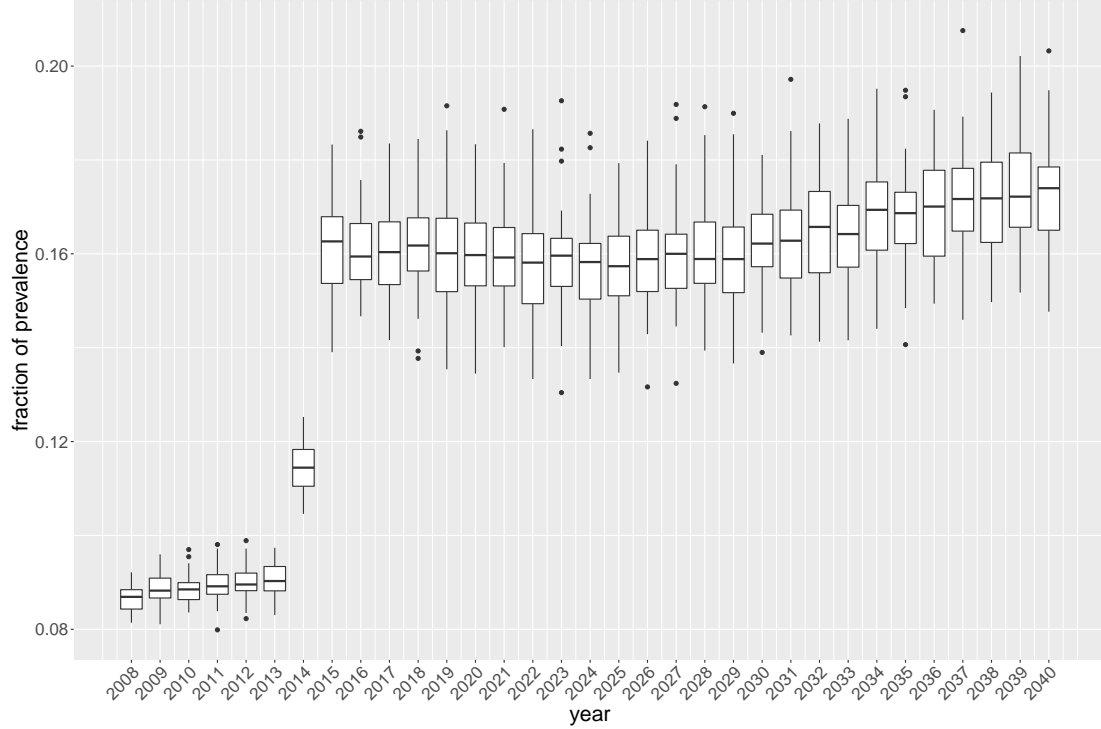


Figure 5.17: Prevalence of DIP from the model in Scn1 showing uncertainty of model stochastics

important limitations in the model that are ripe for addressing. These notably include a lack of detail with regards to childhood dynamics (including weight change), neglect to social network effects on behavior, and an overly simplistic representation of changes in K_{xgI} and behavior change. Extensions of the model to capture such effects, and to capture cost and resource components of scenarios, remain an important priority.

The multi-scale hybrid model of DIP presented in this chapter – which interweaves ABM, SDM, and DES elements – was informed by longitudinal and cross-sectional empirical data from the ACT and Saskatchewan Ministry of Health. While this model is the most sophisticated dynamic model in the thesis, it is notable that the model interfaces with the data in a relatively simple way – via model calibration. Similar to the ABM in Chapter 4, the cross-sectional data over time considered here – the incidence of DIP in each ethnicity – allows the model to address the growing epidemic of GDM and T2DM at a high level. The rich longitudinal empirical data available for this model, including the prevalence of macrosomia by DIP status and number of diabetes by age 30 of offspring according to their mother’s diabetic status (e.g., GDM, T2DM) and birth weight, provides additional information for extending model mechanism to explore the interaction of risk factors and simulate the inter-generational transfers of risk for GDM and T2DM and occurrence of later-life diabetes in the offspring. A high-resolution representation of dynamics of glycemic regulation and heterogeneity at the individual level, as a hybrid model, permits the model to effectively utilize various types of data in grounding model behavior. The work in this chapter demonstrates that availability of rich data can support powerfully enriching evidence-informed model structure. The resulting dynamic model with sophisticated

structure further increases the ability of the model to utilize empirical data.

The model here simulates the interaction of risk factors, coupled dynamics of the glycemia-beta-cell-insulin system and insulin resistance, and the impacts of a diverse portfolio of interventions at a requisite degree of detail. There are ways this model could be used to address the growing epidemic of GDM and T2DM in the ACT and support public health policy and decision making. The use of a hybrid model capturing a remarkably broad set of factors provides a way to examine and estimate the system-wide outcomes for counter-factual situations and to elevate understanding of the dynamics of implemented interventions in a complex system. With the DES component built atop the ABM building block, the model can be employed to investigate delays, bottlenecks in and optimization of resource use in public health service delivery – including by addressing the greater demands imposed on public health services due to earlier diagnostic screening.

CHAPTER 6

PARTICLE FILTER APPLIED TO SYSTEM DYNAMICS MODEL FOR MOSQUITO POPULATION SURVEILLANCE

In this chapter, the machine learning technique of particle filtering was applied to an SDM of mosquito population dynamics to predict the adult mosquito population. The model of DIP in Chapter 5 has the most sophisticated model structure in this thesis, but combines with empirical data in a relatively simple and traditional way – via model calibration. While having a simpler model structure than the hybrid dynamic model in Chapter 5, the SDM of mosquito development examined here both uses a more sophisticated means of grounding the model in data – via the Sequential Monte Carlo approach of Particle Filtering – and requires a more complex way of relating the empirical data to corresponding model outputs, due to the critical mediating influence of the probability of capturing a *Culex* mosquito adult, a quantity dependent on time-varying environmental factors. The empirical data – including weather-related factors and mosquito-related time series – are used as direct observations to support the particle filtering in recurrently regrounding the latent state of the dynamic model. The count of *Culex* mosquito adults being captured in mosquito traps, the number of such traps employed, in conjunction with the estimated resulting probability of capturing a *Culex* mosquito adult per trap night and model generated *Culex* adult mosquito population, were used in the likelihood function of the particle filtering model.

6.1 Introduction

West Nile virus (WNV) infection is historically one of the leading causes of mosquito-borne disease (MBD) in Saskatchewan, and across Canada [21]. WNV maintains its enzootic transmission cycle primarily through bird biting mosquitoes and WNV infected birds, and typically spreads to humans and other mammals – importantly and devastatingly including horses – as the feeding preference of mosquito shifts [23]. In many regions – including Saskatchewan – the bridge vector for WNV between birds and humans consists predominantly of mosquitoes in genus *Culex* [22]. *Culex pipiens* and *Culex restuans*, and *Culex tarsalis* are dominant vectors in eastern Canada and in western Canada, respectively [25]. *Culex tarsalis* amplifies the spread of infections as feeding preference shifts from bird to human during American robins (*Turdus migratorius*) migration period [23]. Within Saskatchewan, it is believed that control of *Culex* mosquitoes can greatly reduce the burden of WNV infection. Further afield, *Culex pipiens* is associated with the highest number of humans

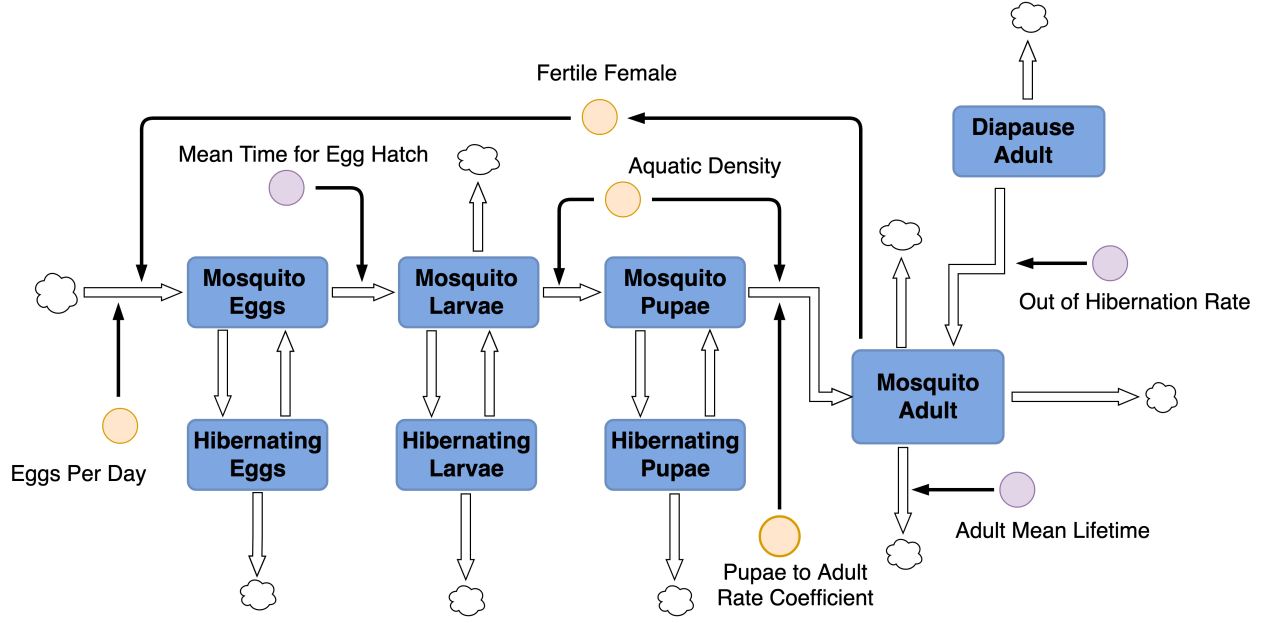


Figure 6.1: SDM of *Culex* mosquito development

cases in WNV infection in northeast and north-central United States [23,24]. Despite the very high burden of WNV historically experienced in jurisdictions such as Texas and Saskatchewan, researchers have argued that needed information or data about the abundance of mosquitoes is not currently sufficient for public health entities to effectively control the outbreaks of MBD [26].

Previous studies estimated the dynamics of abundance of *Culex* mosquitoes on the basis of counting the number of *Culex* mosquito adults being captured per trap set in the local environment [96,97,98]. While there are two primary types of traps employed in Saskatchewan, the most widely used trap to capture *Culex* mosquitoes [96] is the CDC light trap, which attracts mosquitoes using CO_2 . One important limitation of trap-based mosquito monitoring and control is that the traps can only estimate the relative abundance of mosquitoes after a change in the mosquito population, and delay responses to the outbreaks of MBD [21]. Furthermore, the number of mosquito adults being captured in a given trap relies heavily on the number of nights that the traps are set, on weather conditions during that period, and the landscape of the trap location [96].

Many researchers have investigated predictive modeling of mosquito abundance to forecast outbreaks of MBD [26,99]. Within such contributions, weather-based factors play a notably important role in constructing the population abundance model of mosquitoes. Peper et al. [26] constructed a statistical model of weather variables for the detection of WNV in a mosquito pool. Specifically, the model indicated the detection of WNV was negatively associated with wind speed, but positively related to the week of a year, visibility, humidity and dew point. Yu et al. [99] developed a temperature-dependent dynamic model for mosquito life cycle – development, mortality, and diapause – to predict the abundance of mosquitoes in a single season. Rosà et al. [100] proposed a linear mixed model to simulate the dynamics of *Culex pipiens* population using local land

use and environmental drivers, e.g., temperature and precipitation. In model results, the early occurrence of warm temperatures in a year increases the length of the mosquito season and decreases mosquito population later in the year, whereas early precipitation is associated with a shortening of the mosquito season and with an increased magnitude of mosquito abundance. Cong [101] developed a probability generating model using an SDM – an ancestor of that employed in this chapter – to simulate a mosquito population, and suggested that the environmental variables were key drivers for variations of the probability that a mosquito will be trapped. Using a synthetic datasets, the Markov Chain Monte Carlo (MCMC) method was employed to analyze the effect of the measuring frequency of mosquito observations on model performance, with results suggesting the value of daily measurements for accurately estimating the impact of weather-based factors, given the timespans of data available.

Machine learning techniques have been applied by many researchers in modeling mosquito distributions or classifying mosquito species [102, 103]. Wieland et al. [102] employed a support vector machine (SVM) model to build a relationship between the distribution of a particular mosquito species and weather data. Fruha et al. [103] evaluated the performance of SVM, random forest, logistic regression and decision tree models in predicting the distribution of particular mosquito species by weather data, and concluded combining models yielded better performance than does a single model.

Due to the marked variability in mosquito population dynamics year-to-year, and the strong limitations of trap-based estimation of mosquito abundance, in this chapter, the machine learning technique of particle filtering was applied to an SDM of mosquitoes development to predict adult mosquito population using various direct observations – including weather-related factors – and mosquito-related time series; in contrast to previous applications of particle filtering models to health-related topics, this model included a likelihood function critically dependent on time-varying factors (here, weather-related observations).

The structure of the remainder of the chapter is as follows: The methods section describes empirical data used for model parameterization, calibration and validation, an overview of the SDM of mosquito population dynamics, modeling for mosquito control, the use of linear regression and model calibration to estimate the probability that a mosquito will be captured by a trap, description of the state equation model, formulation of the particle filtering adaptation to the model, model calibration, and model scenarios. The results section characterizes outputs of the particle filtering model. The discussion section characterizes the model findings, outcomes, limitations, and notes prospective lines of future work for the model.

6.2 Empirical Data

Daily weather records from 2010 to 2015 from the city of Saskatoon [104] were employed as exogenous factors driving the dynamics of the SDM of *Culex* mosquito development, as well as in the particle filtering model as exogenous factors determining the estimated probability of capturing a *Culex* mosquito adult in a CDC light trap. The weather data employed includes temperature, precipitation, humidity, and wind speed levels,

with the SDM model using just the first two of those, while the likelihood function used all four. Data from 2011 to 2015 provided by the Mosquito Control Program of the City of Saskatoon, including the number of *Culex* mosquito adults being captured by the CDC light traps deployed by the city, the number of the traps deployed for a given time period (typically day), and records of mosquito control efforts, were incorporated in the particle filtering model. The empirical number of *Culex* mosquito adults being captured informed the particle filtering model, and provided the basis for that technique to recurrently estimate (via sampling) the latent state of the system. The number of traps, in conjunction with the estimated probability of capturing a *Culex* mosquito adult and model generated *Culex* adult mosquito population, were used in the likelihood function of the particle filtering model. VectoBac is an effective mosquito larvicide applied in standing water for the control of mosquito populations [105]. Three types of VectoBac records – including the number of sites visited for applying VectoBac, the number of sites with mosquito larvae among the visited sites, and VectoBac amount in kilograms, are employed in the SDM to simulate the effect of VectoBac on reducing the population of mosquito larvae.

6.3 Methods

6.3.1 Overview of the System Dynamics Model of Mosquito Population

The *Culex* mosquito development is formulated as an SDM previously built by the supervisor and modified, extended, and simplified by a series of students, including M.Sc. students Wenyi An [106] and (separately) Peibo Cong [101], undergraduates Curtis Theoret, Michael Richards, Chin Wang, and student – subsequently staff member – Karen Yee. The work described in this chapter built the particle filter upon partial particle filter structure contributed by Wang and Cong. Reflecting the need to simulate the life cycle of *Culex* mosquitoes, the SDM includes the following stocks to represent distinct stages of the mosquito lifecycle: *Mosquito_Eggs* (ME), *Mosquito_Larvae* (ML), *Mosquito_Pupae* (MP), *Hibernating_Eggs* (HE), *Hibernating_Larvae* (HL), *Hibernating_Pupae* (HP), *Mosquito_Adult* (MA), *Diapause_Adult* (DA), as shown in Figure 6.1. Reflecting the fact that the development of *Culex* mosquitoes from larvae to adult is affected by several environmental factors, the SDM consequently employs daily weather-related variables – including temperature, humidity, wind speed, and precipitation – to allow for accurate simulation of the dynamics of the population of *Culex* mosquito adults. The inflows and outflows for a given stock specify the rates of flow increasing and decreasing the value of that stock, as quantified by the time unit of the model, respectively. The relationship between a state variable (termed a stock in System Dynamics) stock and its corresponding flows is represented by the series of first-order ordinary differential equations shown in Equation 6.1. The system of equations applying this within the specific context of the SDM applied here is shown in Equation

6.7.

$$S(t) = \int_0^t (inflow(x) - outflow(x))dx + S_0 \quad (6.1)$$

Amongst many other endogenous factors, the SDM outputs the population of *Culex* mosquito adults in the Saskatoon area, and the daily count of *Culex* mosquitoes that the model estimates to be captured by traps across the city. This latter quantity is a product of the model generated *Culex* adult mosquito population and an estimated capturing probability varying according to the environmental factors. The model time unit is 1 day, and the duration of the model time horizon is 6 years – from the beginning of 2010 to the end of 2015. Because of uncertainty regarding the initial state, a burn-in period of 1 year is used to allow some measure of balance in model state, so as to better estimate the value of stocks and parameters over the continuous simulation. The subsections below discuss different regions of the model.

6.3.2 Model of Mosquito Control (VectoBac)

As shown in Figure 6.1, the mosquito larvae mature into mosquito pupae, which in turn develop into the adult mosquito population. The application of VectoBac to standing water ponds provides effective control of mosquito larvae [107], which can strongly reduce the population of larvae – and, in coming weeks, of adult mosquitoes. In light of the widespread application of VectoBac within the area represented by the model, to more precisely simulate the dynamics of the mosquito population, the SDM incorporates a representation of the effect of larvaciding. The effect on larvae of an application VectoBac on a given day is characterized in the SDM using the following equation (Equation 6.2):

$$\frac{n_s}{n_t} \times \left(1 - e^{-\alpha \times \frac{V_t}{n_s}}\right) \quad (6.2)$$

Specifically, equation 6.2 specifies the fraction of mosquito larvae assumed to remain after applying VectoBac, where n_s is the count of sites at which VectoBac was applied on that day, n_t is the count of all possible such sites, α is a coefficient characterizing the effectiveness of VectoBac, and V_t represents the total VectoBac (in kg) applied on that day; by extension, $\frac{V_t}{n_s}$ represents the per-site average amount of VectoBac applied on that day. By multiplying Equation 6.2 by the stock representing mosquito larvae, the number of larvae controlled by the VectoBac are essentially subtracted from the stock.

6.3.3 Probability of Capturing a Mosquito by CDC Light Trap

Weather-based factors play an important role in governing the number of *Culex* mosquito adults being captured by the CDC light traps. This occurs through two primary pathways: By affecting the underlying dynamics of the mosquito population, and by affecting the probability that a given mosquito will be trapped. Speaking with respect to both such pathways, higher temperature increases the maturation rate of *Culex* – and thus eventually the abundance of mosquito adults – but also elevates the activity of existing *Culex*

mosquito adults [21]. Precipitation is another important driver affecting both pathways for *Culex* mosquitoes. Soverow et al. reported that precipitation is positively associated with an increase in the population of *Culex* mosquitoes [108], presumably in part due to availability of egg-laying sites. However, count of mosquitoes being trapped is not linearly associated with precipitation [96]; a higher count of mosquitoes being captured is observed with higher precipitation prior to the week of trapping. It is also well-known that precipitation affects the degree to which mosquitoes take wing – and, by extension, the probability that they are captured by a trap. This model characterized the effect of precipitation on the probability that a given mosquito would be trapped via two variables – one quantifying the precipitation level, and the other the square of precipitation level. By contrast to the first two factors, the influence of wind speed on measured mosquito counts is characterized by this model as operating purely through probability of entrapment. Specifically, wind speed influences the behavior of *Culex* mosquito adults, e.g., host-seeking activities and flight direction, and is negatively related to the measured abundance of *Culex* mosquito adults [96].

The model assumes that the weather-based factors independently influence the probability that a trap will successfully capturing a given *Culex* mosquito adult in the course of the day, and that probability p for a given day is characterized in the model by Equation 6.3:

$$\ln(p) = \beta_0 + \beta_T \times T + \beta_W \times W + \beta_{P1} \times P + \beta_{P2} \times P^2 \quad (6.3)$$

where T , W , P and P^2 represent the weather variable for that day for temperature, wind speed, precipitation, and square of precipitation, respectively, and β_0 , β_T , β_W , β_{P1} and β_{P2} denote the intercept term, and coefficients for temperature, wind speed, precipitation and square of precipitation, respectively.

Linear regression based estimation of the weather coefficients

This model used two approaches to estimate the coefficients β_T , β_W , β_{P1} , and β_{P2} : Via regression and (separately) calibration; the first approach is covered in this section. The empirical data provides the number of *Culex* mosquito captured per trap C_t , which is treated for the derivation as equal to the product of the capturing probability p_t at day t and the population m_t of *Culex* mosquitoes at day t . For the derivation via regression, the model assumes m_t and m_{t-1} are approximately equal, yielding Equation 6.4. On the basis of this, the coefficients for weather variables are characterized as a linear regression model, as shown in Equation 6.5. The results of linear regression model are shown in Appendix B.

$$\frac{C_t}{C_{t-1}} \approx \frac{p_t}{p_{t-1}}, \quad (6.4)$$

$$\ln\left(\frac{C_t}{C_{t-1}}\right) = \beta_T \times (T_t - T_{t-1}) + \beta_W \times (W_t - W_{t-1}) + \beta_{P1} \times (P_t - P_{t-1}) + \beta_{P2} \times (P_t^2 - P_{t-1}^2) \quad (6.5)$$

Model calibration based estimation of the weather coefficients

Model calibration was also employed to estimate the weather coefficients (β_T , β_W , β_{P1} and β_{P2}) in Equation 6.3. The calibration used the values of the weather coefficients estimated via linear regression as initial

estimates for the β_T , β_W , β_{P1} and β_{P2} being calibrated. Within this process the values of β_T , β_W , β_{P1} and β_{P2} were tuned to search for a best combination of parameter value to yield a minimized difference between the predicted number of *Culex* mosquitoes being captured and empirical data. The calibrated β_T , β_W , β_{P1} and β_{P2} , denoted in calibration-related code as `betaTemp`, `betaWindSpeed`, `betaPrecipitation`, and `betaPrecipitationSquare`, and their corresponding calibrated values are listed in Appendix C.

6.3.4 State Equation Model

The model for mosquito population dynamics consists of the following state variables: `Mosquito_Eggs` (E), `Mosquito_Larvae` (L), `Mosquito_Pupae` (P), `Hibernating_Eggs` (HE), `Hibernating_Larvae` (HL), `Hibernating_Pupae` (HP), `Mosquito_Adult` (A), and `Diapause_Adult` (DA). Auxiliary variables (termed dynamic variables in AnyLogic) and parameters are incorporated to characterize the rates for inflow and outflow of the stocks. Figure 6.1 illustrates the stocks and corresponding inflows and outflows used in the SDM, and Equation 6.7 to 6.14 are the compartmental equations used to characterize their relationship. Parameters and notation used in the SDM and thesis are listed in Appendix D.

In this model, the rate for mosquito eggs being laid (λ_e) of a female mosquito adult is affected by the following three factors: The dynamic variable c_e represents the count of batch of eggs that a female mosquito adult can lay per day – with an assumption of approximately 200 eggs being laid per batch – the parameter t_{eg} refers to the egg gestation period; and f is favourability for egg laying influenced by precipitation and water height (the parameter `waterHeightForMaxEggLaying`). λ_e is correspondingly calculated by Equation 6.6:

$$\lambda_e = \frac{c_e \times 200}{t_{eg}} \times f \quad (6.6)$$

λ_{ih} represents the rate for entering the diapause state, which is affected by temperature and hours of sunlight; for simplicity, this is treated as being identical for mosquito eggs, larvae and pupae. Conversely, parameter λ_{oh} is the rate for diapaused mosquito adults as well as eggs, larvae, and pupae, to leave diapause. Parameter t_{eh} represents the mean time for an egg to hatch. Both temperature and the density of mosquito larvae and pupae in a particular aquatic environment influence both the mosquito larvae death rate (λ_{ad}) and the development time t of a mosquito from larvae to pupae and pupae to adult. The flows of stock *A* are determined by the following parameters: Parameter t_m represents the mean lifetime of mosquito adults; λ_d is mosquito diapause rate based on the day of a year, and γ_f is the rate of mosquito adults dying due to exposure to low temperatures.

$$\frac{dE}{dt} = \frac{1}{2}A\lambda_e f + HE\lambda_{oh} - \frac{E}{t_{eh}} - E\lambda_{ih} \quad (6.7)$$

$$\frac{dHE}{dt} = E\lambda_{ih} - HE\lambda_{oh} - \frac{1}{100}HE \quad (6.8)$$

$$\frac{dL}{dt} = \frac{E}{t_{eh}} + HL\lambda_{oh} - L\lambda_{ih} - L\lambda_{ad} - \frac{L}{t} \quad (6.9)$$

$$\frac{dHL}{dt} = L\lambda_{ih} - HL\lambda_{oh} - \frac{1}{100}HL \quad (6.10)$$

$$\frac{dP}{dt} = \frac{L}{t} + HP\lambda_{oh} - P\lambda_{ih} - \frac{P}{t}\gamma \quad (6.11)$$

$$\frac{dHP}{dt} = P\lambda_{ih} - HP\lambda_{oh} - \frac{1}{100}HP \quad (6.12)$$

$$\frac{dA}{dt} = \frac{P}{t}\gamma + DA\lambda_{oh} - \frac{A}{t_m} - A\gamma_f - A\lambda_d \quad (6.13)$$

$$\frac{DA}{dt} = -\frac{5}{1000}DA - DA\lambda_{oh} \quad (6.14)$$

6.3.5 Particle Filtering Model

The model for mosquito population dynamics at time t can be represented by the system state vectors $x_t = f_t(x_{t-1})$ (based on the state equations above) and observation vectors y_t , where f_t is a Markov process of the model. At time t , each particle characterizes the states of the model, including E, L, P, HE, HL, HP, A, and DA. Furthermore, dynamic parameters c_e and γ , and stock `minTemperatureAdultOutOfDiapause` are associated with stochastic processes put in place within the model to include the randomness over time required for useful particle filtering. The primary objective of the particle filtering model is to obtain the discrete random measure $X_t = \{x_{0:t}^{(m)}, w_{t-1}^{(m)}\}$, given X_{t-1} and the observation vector y_t . Generating particles $x_t^{(m)}$ and updating their corresponding weights $w_t^{(m)}$ are achieved by importance sampling methods. After updating weights, the distribution of value of states across the particles X_{t-1} reflects X_t . Resampling, a scheme that eliminates particles with small weights and replicates particles with large weights, takes place when the effective particle size falls below a predefined threshold k . The likelihood function employed in the model posits a Poisson distributed error, around a mean value (λ) calculated as follows:

$$\lambda = n_t \times MP_t \times p_t \quad (6.15)$$

where t denotes the day, n_t , MP_t and p_t are the empirical number of traps being set, the simulated mosquito population from the SDM, and the estimated probability of capturing a *Culex* mosquito adult per trap, respectively.

6.4 Model Calibration

To estimate poorly- or non-measured parameters, the model was calibrated so as to minimize an objective function given by the discrepancy (Equation 6.16) between the predicted number of *Culex* mosquitoes being captured and empirical data. The discrepancy was computed in a calibration scenario for the model with

800 particles. The calibration experiment was set to run 25000 iterations (each associated with a single realization), conducted with a fixed seed value of 1. The calibration experiment yielded a starting objective function value of 336.84 and a minimal objective value of 110.287 following parameter optimization. The calibrated parameters and their values are listed in Appendix C. The parameters optimized within calibration play an important role in shaping the development of *Culex* mosquitoes from eggs to adults, and estimating the effects of four weather-related factors – precipitation, temperature, wind speed, and relative humidity – on the probability of capturing a mosquito.

6.5 Model Scenarios

This work examined the performance of particle filtering under two scenarios. The model in Scenario 1 (Scn1) employed all of the calibrated parameters listed in Appendix C. By contrast, the model in Scenario 2 (Scn2) used the weather coefficients from the linear regression model shown in the Table B.1 of Appendix B rather than the calibrated weather coefficients. As in Scn1, mosquito development related parameters were estimated by model calibration (`waterHeightForMaxEggLaying`, `outOfHibernationRate`, `AdultMeanLifeTime`, `EggGestationPeriod`, and `MeanTimeForEggHatch`, given in Appendix C). In order to assess the impacts of particle filtering, two reference scenarios were also employed. In Scenario 3 (Scn3), the model used the all of the calibrated parameters listed in Appendix C, but with particle filtering mechanism turned off. Finally, Scenario 4 (Scn4) was used to characterize a scenario having the same parameter setting as Scn2, but without particle filtering.

6.6 Results

The discrepancy between the particle filtering-based predictions and corresponding empirical observations is quantified using Equation 6.16 [38], where x_{ij}^P is the output represented by the sampled particle j at observation i , and x_i^E is correspondingly empirical observation. T_f and T^* represents the final observation time, and the time t up to and including the final point at which the weights of particles were updated. For the calculation of discrepancy, n ($n = 10,000$) particles were sampled according to their weights in accordance with the dictates of importance sampling.

$$\text{discrepancy} = \frac{\sum_{i=T^*+1}^{T_f} \left(\frac{\sum_{j=1}^n (x_{ij}^P - x_i^E)^2}{n} \right)}{T_f - T^*} \quad (6.16)$$

This work also yielded a set of 2D-histograms sampling from both posterior distributions (as shown in Figures 6.2a, 6.2b, 6.2c, 6.2d, and 6.2e), and prior distributions (Figures 6.3a, 6.3b, 6.3c, 6.3d, and 6.3e). These figures illustrate the distribution of number of *Culex* mosquitoes being captured by the traps as predicted by the sampled particles regrounded by the particle filtering model in Scn1, and visually compare those results with empirical data (as shown in red). Analogous comparisons between model generated distributions

Table 6.1: Discrepancy between the model estimate of the count of *Culex* mosquitoes being captured and the empirical data in Scn1

Year	Discrepancy of prior distribution	Discrepancy of posterior distribution
2011	8354.02	70.36
2012	41.82	30.51
2013	140.95	101.91
2014	249.88	232.63
2015	166.81	146.992

for Scn2 and empirical data are shown for posterior distributions in Figure 6.4a, 6.4b, 6.4c, 6.4d, and 6.4e, and for prior distributions in Figure 6.5a, 6.5b, 6.5c, 6.5d, and 6.5e.

The results presented in the 2D-histograms were generated by an underlying model using 10,000 particles, where the plots were generated by sampling 12,500 particles from the distribution of captured *Culex* mosquitoes in traps according to their weight, in accordance with the principles of importance sampling. Figure 6.3d and Figure 6.3e shows the superposition of the value of empirical data with the corresponding estimated prior distribution of model-estimated mosquitoes traps; these plots demonstrate that the model generally underestimate the number of *Culex* mosquitoes being trapped in 2014 and 2015. The posterior distribution of 2015 in Scn1, as shown in Figure 6.2e shows that the posterior distribution sampled from the model somewhat underestimates the number of *Culex* mosquitoes being trapped around day 226 of the year (mid-August), but the difference between the model outputs and empirical data is decreased in the posterior distribution compared to the prior distribution. Figure 6.2d shows the posterior distribution for the model in Scn1 is not able to effectively reground the model to adequately match the rapid escalation and the following fast decline of the reports of *Culex* mosquito population occurring on around August 18 in 2014 (day 230 of the year). However, both the prior and posterior distribution for the model of 2014, as shown in Figure 6.3d and 6.2d, respectively, is close to the empirical data until early August. The prior distribution of the counts of *Culex* mosquitoes being trapped in 2012, as shown in Figure 6.3b, can be recognized by the substantial visible overlap between the prior distribution and empirical data. As shown in Figure 6.2c and Figure 6.3c, the particle filtering model in Scn1 captures the peak of the empirical number of reported *Culex* mosquitoes captured during the summer of 2013.

The results for Scn2 – which features linear regression estimates for the weather-related coefficients contributing to the probability that a mosquito will be trapped – the model outputs of 2012, 2013, and 2015 show similar patterns with the model results of Scn1. However, in contrast to the posterior distribution for

Table 6.2: Discrepancy between the model estimated number of *Culex* mosquitoes being captured and the empirical data in Scn2

Year	Discrepancy of prior distribution	Discrepancy of posterior distribution
2011	11816.13	47.93
2012	45.29	30.1
2013	146.22	104.3
2014	268.92	247.43
2015	175.35	141.65

2014 in Scn1, as shown in Figure 6.5d, the model in Scn2 underestimates the rapid escalation of the reports of the *Culex* mosquito population occurring on around August 18 in 2014, but captures slightly better than Scn1 the decline of count of trapped *Culex* mosquitoes after that escalation. Figure 6.4b shows the particle filtering model captures the two peaks of the empirical number of reported *Culex* mosquitoes captured during the summer of 2012. As shown in Figure 6.4a, around early August, the posterior distribution for the model in Scn2 substantially superposes the value of empirical data.

Both Table 6.1 and 6.2 demonstrate that the particle filtering method decreases the discrepancy associated with the posterior distribution for each year after the regrounding of model states. The results further show that the model makes the most accurate estimates the number of *Culex* mosquitoes being captured in 2012. In both Scn1 and Scn2, the model suffers from its largest discrepancy in 2014, compared with the discrepancies of other years. Furthermore, the discrepancy of 2014 in Scn2 is substantially smaller than that for Scn1. Again focusing on the posterior distribution, the average of the discrepancy from 2011 to 2015 of Scn1 and Scn2 are 116.48 and 114.28, respectively.

In contrast to the results above – which focus on estimated counts of reported (i.e., trapped) mosquito populations, the particle filtered model can also estimate the size of the underlying mosquito population. Figures 6.6a, 6.6b, 6.6c, 6.6d, and 6.6e demonstrate the distribution of *Culex* mosquito population across all particles for each of 2011, 2012, 2013, 2014, and 2015, respectively. Figure 6.6c shows a pattern for 2013 that the *Culex* mosquito population had a sharp decrease around day 230 (August 18), and then increased dramatically in the next 10 days. This results demonstrates the particle filtering regrounds the model state by incoming observations, and measurement bias – here, particularly associated with estimates of capturing probability – can significantly influence the the behavior of the particle filtered model.

Figure 6.7a to Figure 6.7e and Figure 6.8a and Figure 6.7e shows the count of *Culex* mosquitoes estimated as being trapped generated by the model in Scn3 and Scn4, respectively. Compared with the SDM with particle

Table 6.3: Discrepancy between the model estimated number of *Culex* mosquitoes being captured and the empirical data in Scn3 and Scn4, respectively

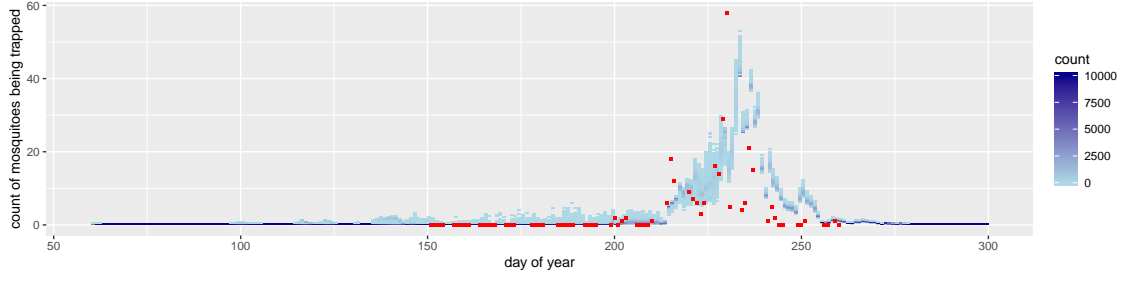
Year	Discrepancy of model in Scn3	Discrepancy of model in Scn4
2011	4259.26	3962.13
2012	330063.85	314106.57
2013	2777842.6	2457183.64
2014	44661488.9	47712305.97
2015	23215026498.085	22089962994.27

filtering, e.g. Scn3 and Scn4, the results in Scn3 and Scn4 demonstrate that the model without particle filtering yields a considerably larger discrepancy between the model outputs and empirical data.

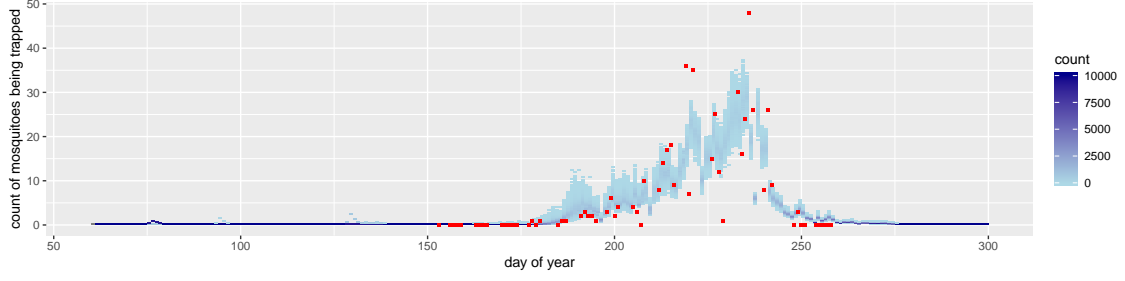
6.7 Discussion

This chapter has described applying the particle filtering method to an SDM to estimate underlying states of the model. The particle filtering method allows the SDM to simulate the population of *Culex* mosquitoes with considerably more accuracy, by virtue of recurrently regrounding the states and stochastically evolving parameters of the model based on incoming empirical observations. The 2D histograms in Section 6.6 demonstrate that by regrounding the states and stochastically evolving parameters of the SDM model by the particle filtering method, the technique is able to capture much of the fluctuation of mosquitoes population, despite ongoing the evolution of environmental factors that shape both mosquito population dynamics and probability of mosquito entrapment.

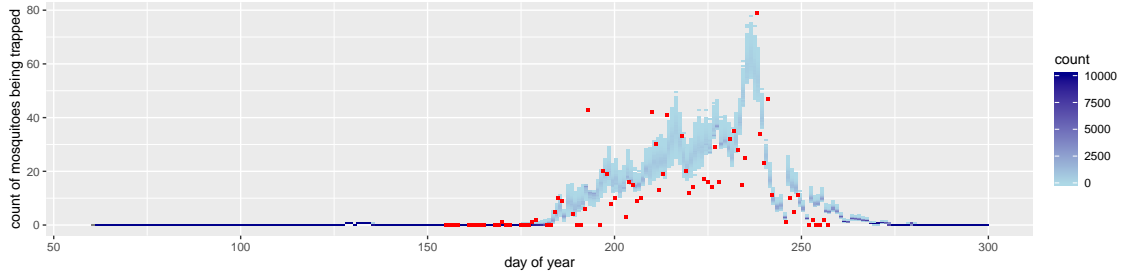
The dynamics of the *Culex* mosquito population depend heavily on environmental factors. The warm temperature and increased precipitation during summer play an important role in increasing the *Culex* mosquito population, via increased egg-laying rate, lower diapause rate, and reductions in maturation time for mosquitoes development from larvae to pupae to adults. The increased population of *Culex* mosquitoes further contribute critically to the peak in the number of mosquitoes being captured during summer in empirical data. Beyond influencing the mosquito abundance itself, because the weather factors also strongly impact the probability that a given mosquito will be trapped, such weather factors are tied up with parameter uncertainties associated with the multifactorial dynamics of the number of mosquitoes being captured by traps. Within the work laid out in this chapter, we pursued two ways of estimating weather related coefficients ($\beta_T, \beta_W, \beta_{P1}$, and β_{P2}) capturing at a relatively high temporal resolution how weather-related factors influence the capturing probability. Firstly, I used a linear regression model to infer such coefficients,



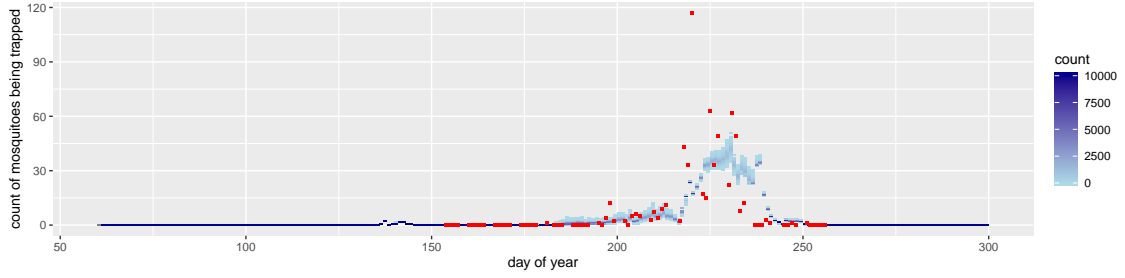
(a) 2011



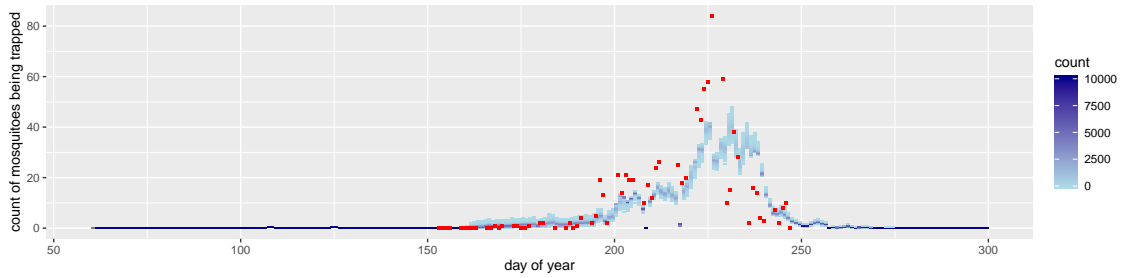
(b) 2012



(c) 2013



(d) 2014



(e) 2015

Figure 6.2: Posterior distribution of the number of *Culex* mosquitoes being trapped in Scn1

Blue and red represents the posterior distribution of the count of *Culex* mosquitoes being trapped estimated by the particle filter model in Scn1 and the empirical count of *Culex* mosquitoes being trapped, respectively.

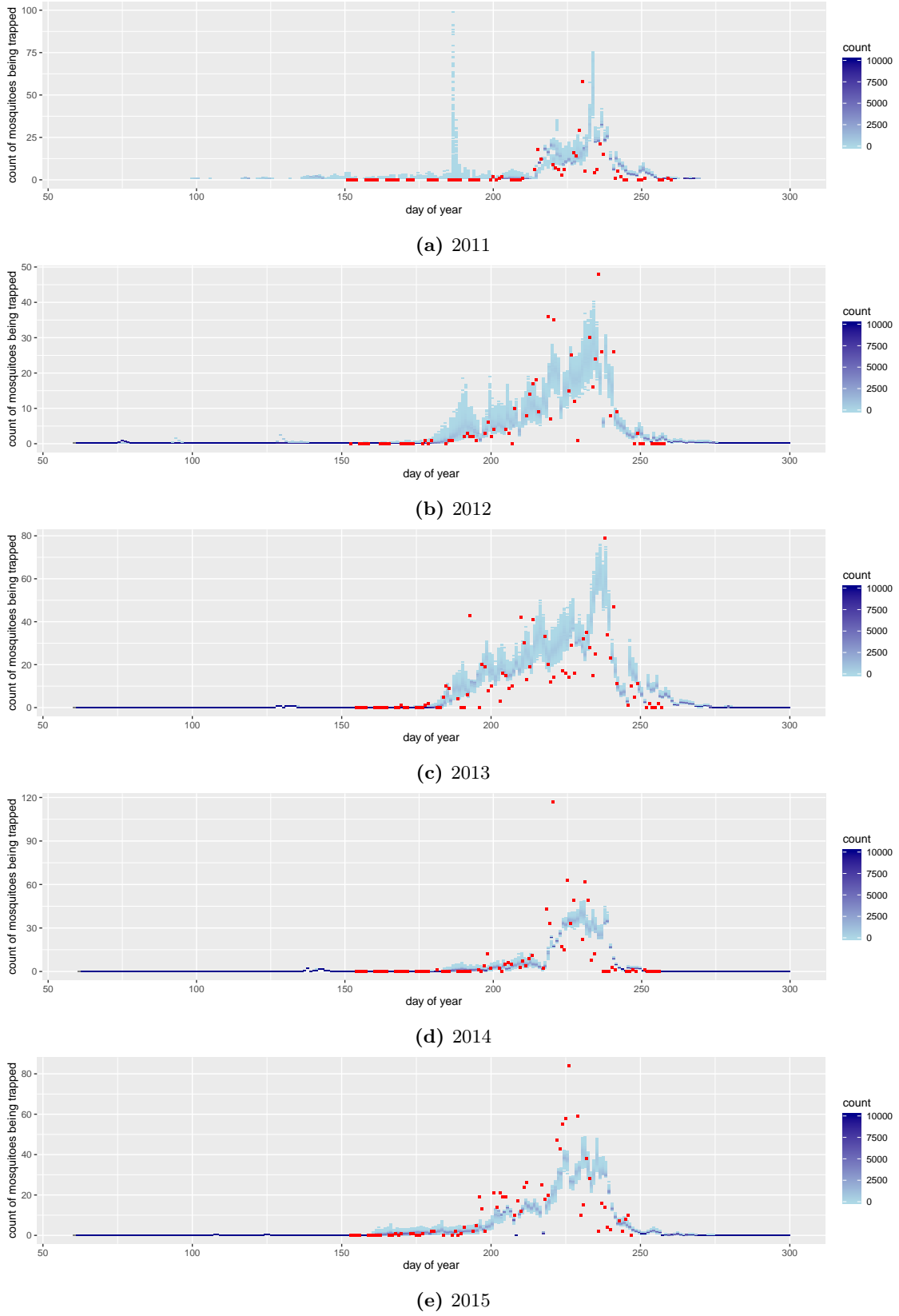


Figure 6.3: Prior distribution of the number of *Culex* mosquitoes being trapped in Scn1

Blue and red represents the prior distribution of the count of *Culex* mosquitoes being trapped estimated by the particle filter model in Scn1 and the empirical count of *Culex* mosquitoes being trapped, respectively.

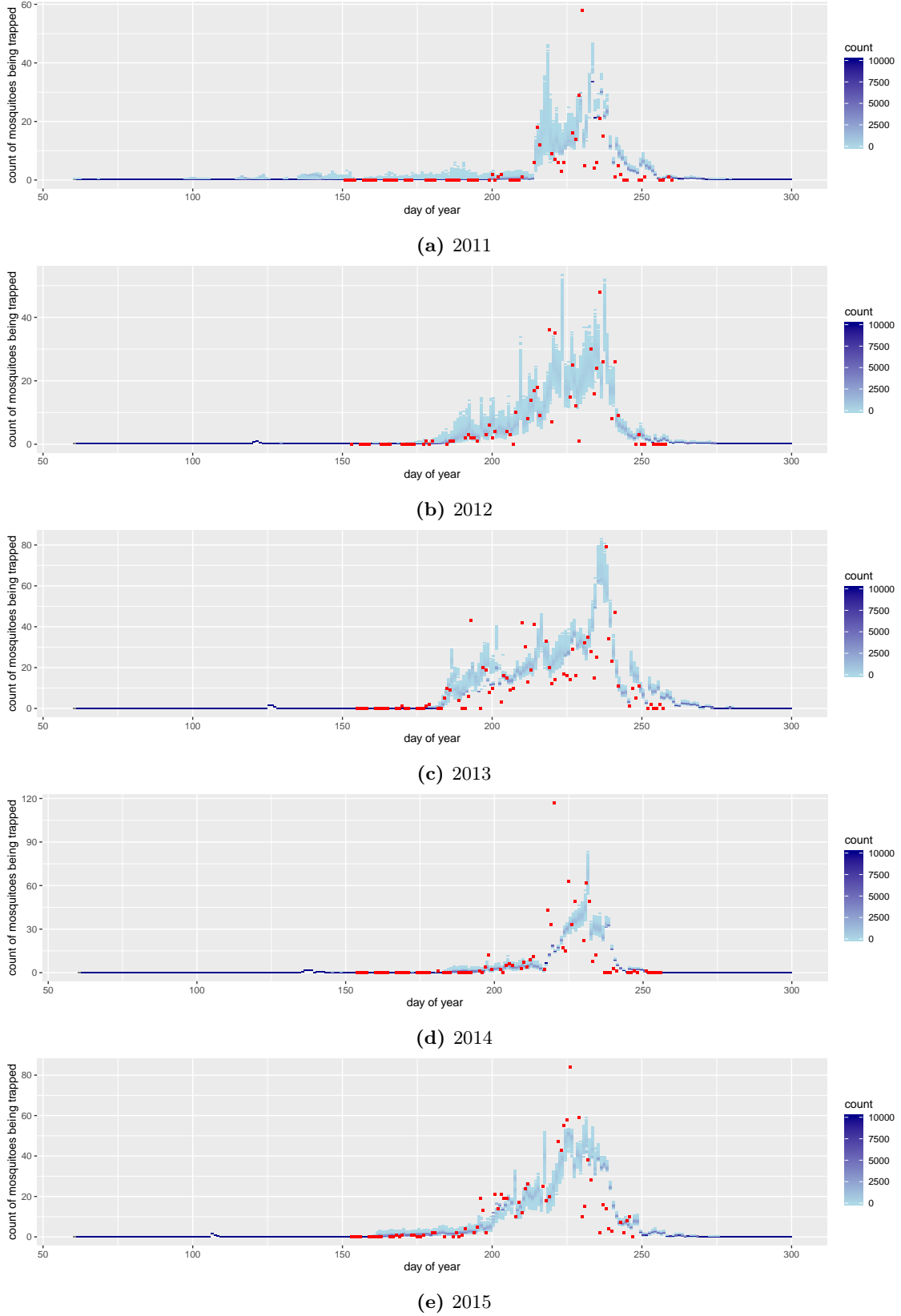
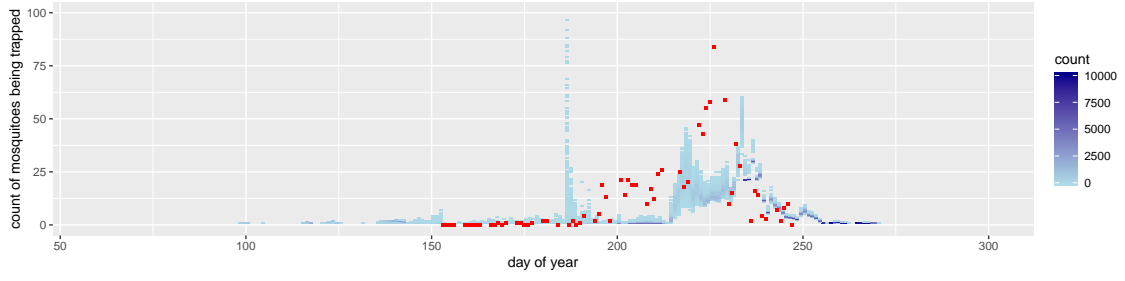
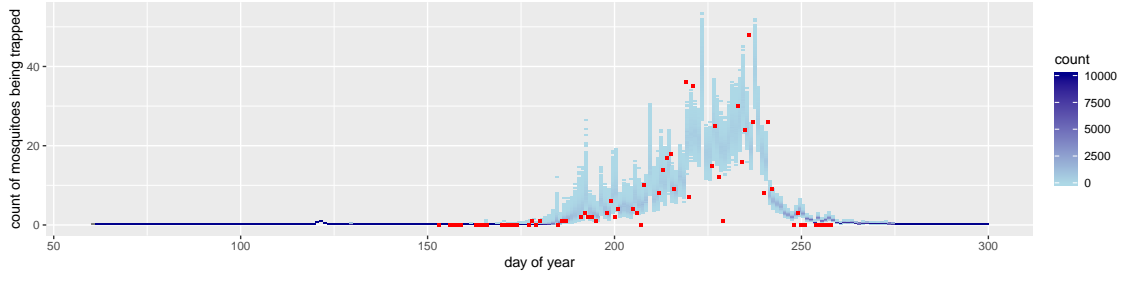


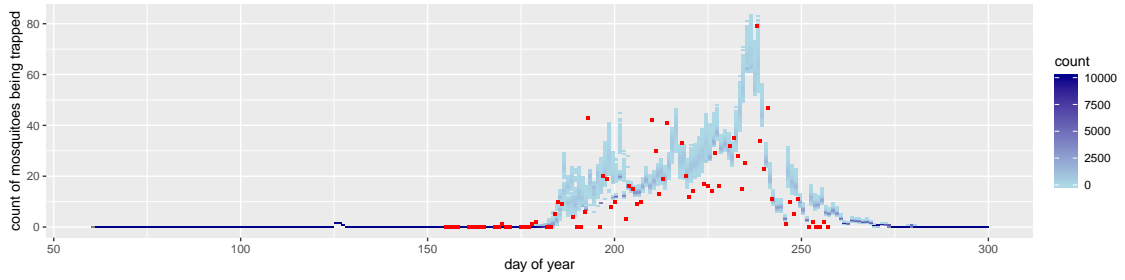
Figure 6.4: Posterior distribution of the number of *Culex* mosquitoes being trapped in Scn2
 Blue and red represents the posterior distribution of the count of *Culex* mosquitoes being trapped estimated by the particle filter model in Scn2 and the empirical count of *Culex* mosquitoes being trapped, respectively.



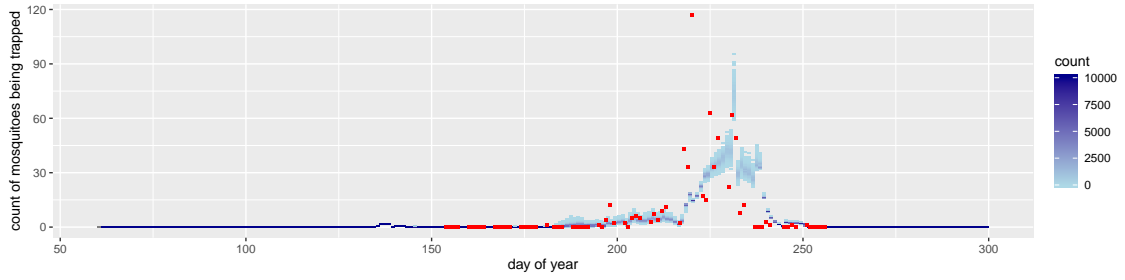
(a) 2011



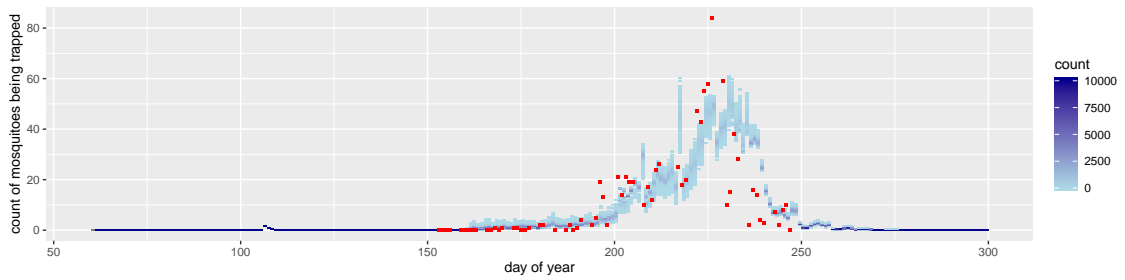
(b) 2012



(c) 2013

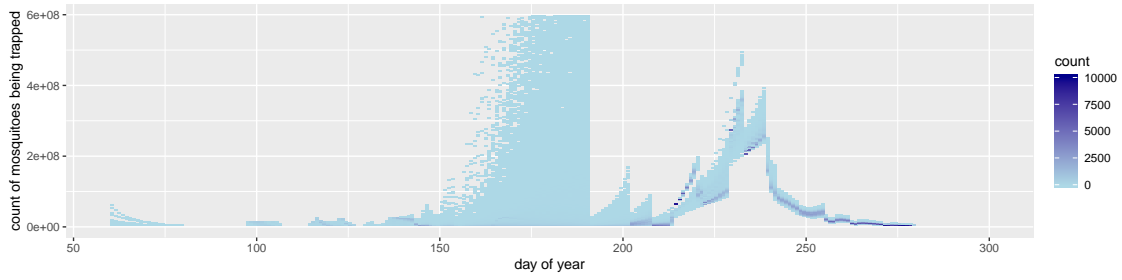


(d) 2014

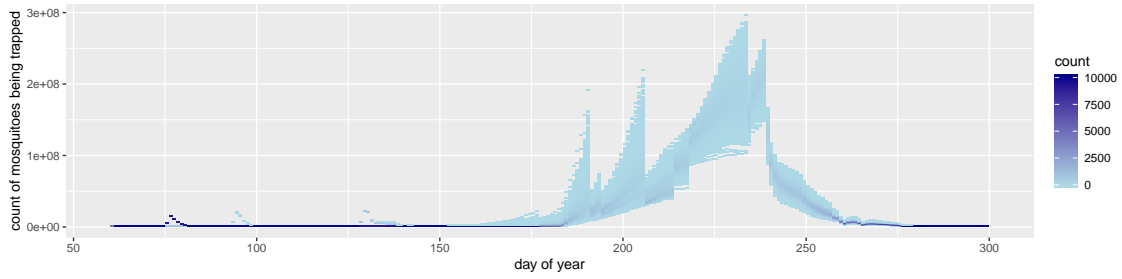


(e) 2015

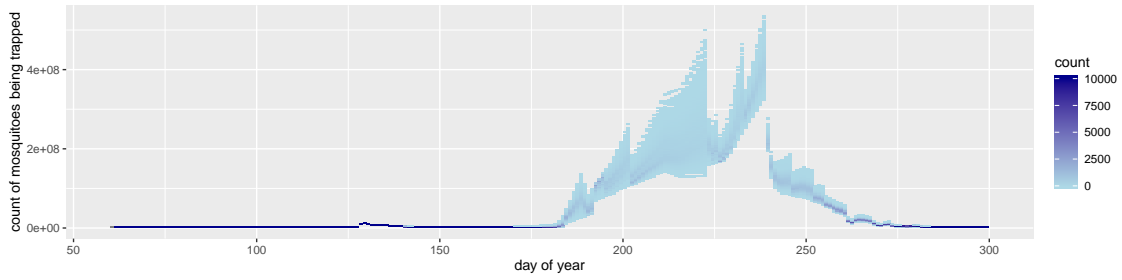
Figure 6.5: Prior distribution of the number of *Culex* mosquitoes being trapped in Scn2
Blue and red represents the prior distribution of the count of *Culex* mosquitoes being trapped estimated by the particle filter model in Scn2 and the empirical count of *Culex* mosquitoes being trapped, respectively.



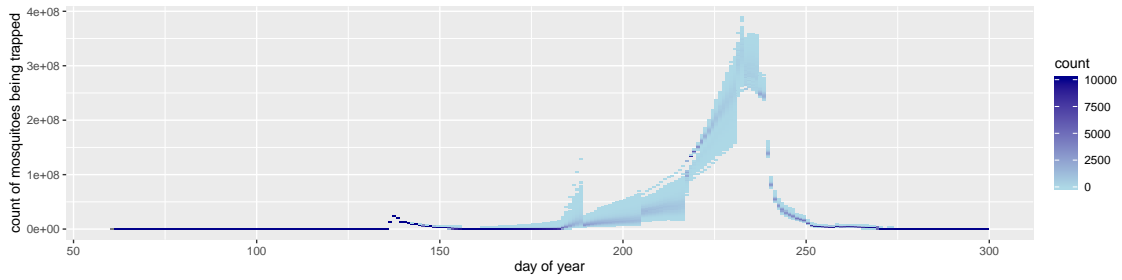
(a) 2011



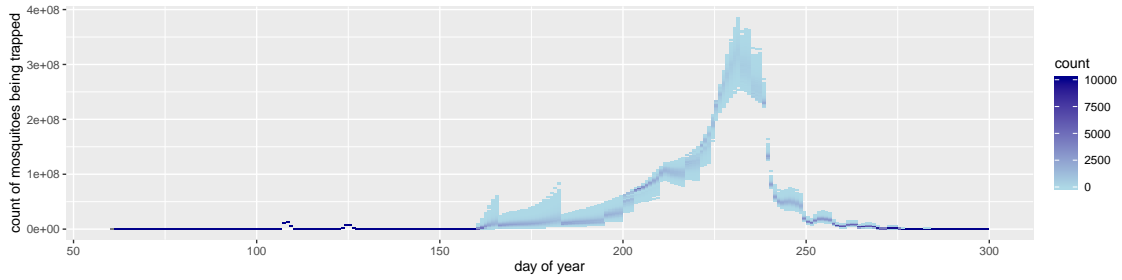
(b) 2012



(c) 2013

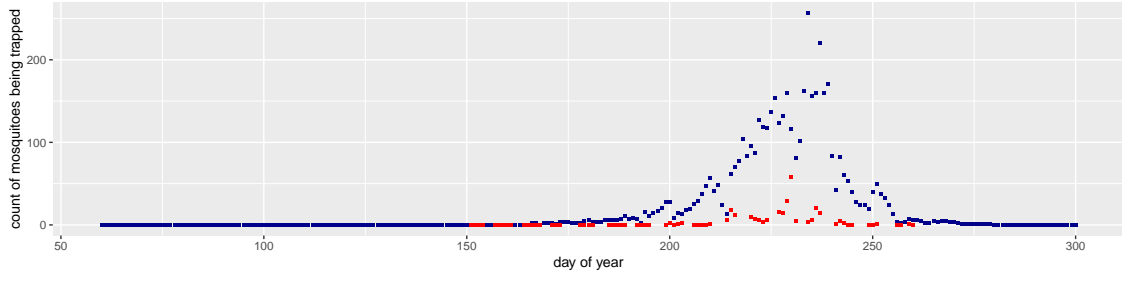


(d) 2014

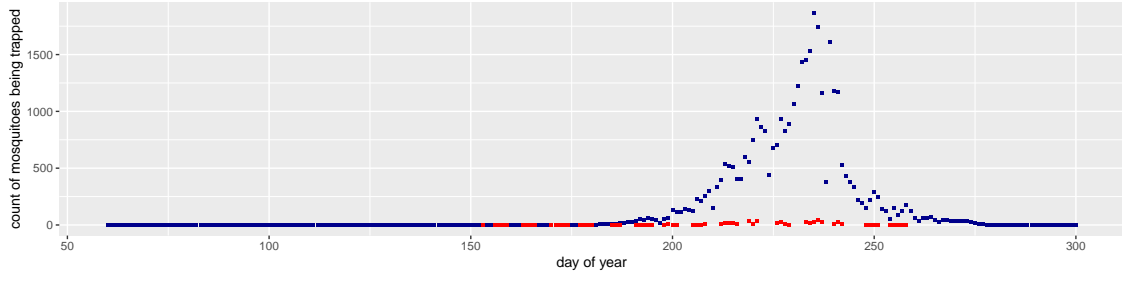


(e) 2015

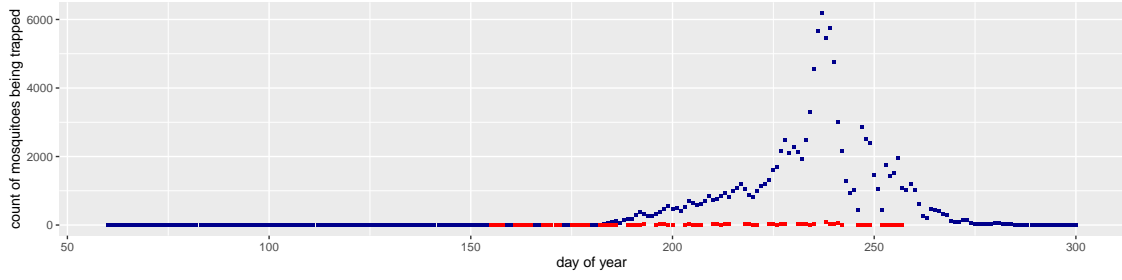
Figure 6.6: Distribution of *Culex* adult mosquito population in Scn1



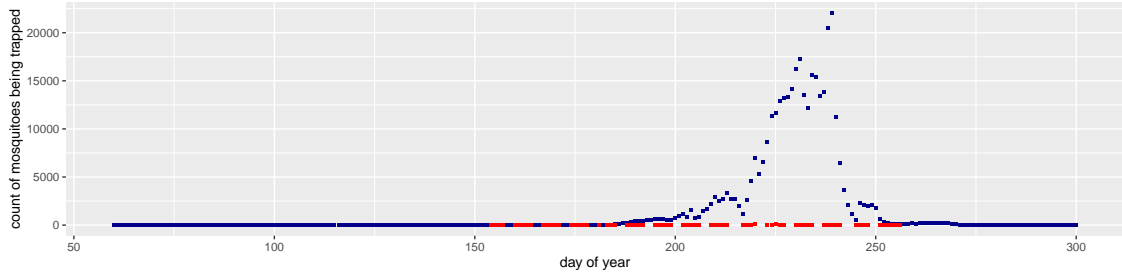
(a) 2011



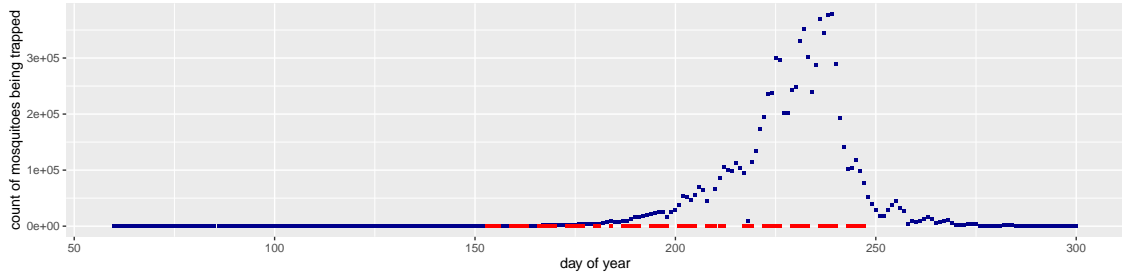
(b) 2012



(c) 2013



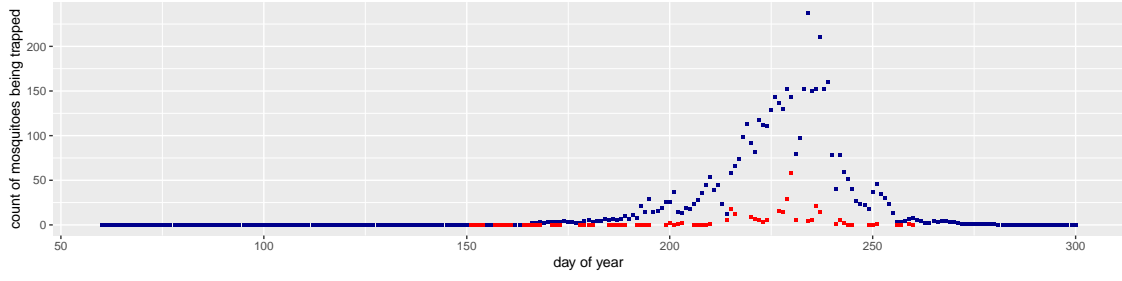
(d) 2014



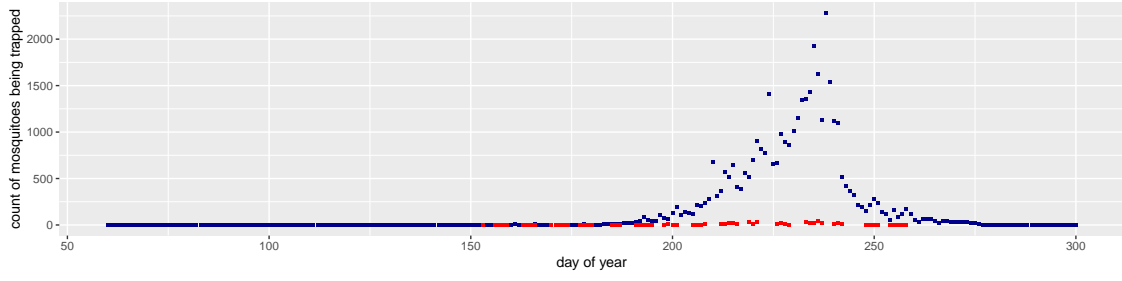
(e) 2015

Figure 6.7: Count of *Culex* mosquitoes being trapped in Scn3

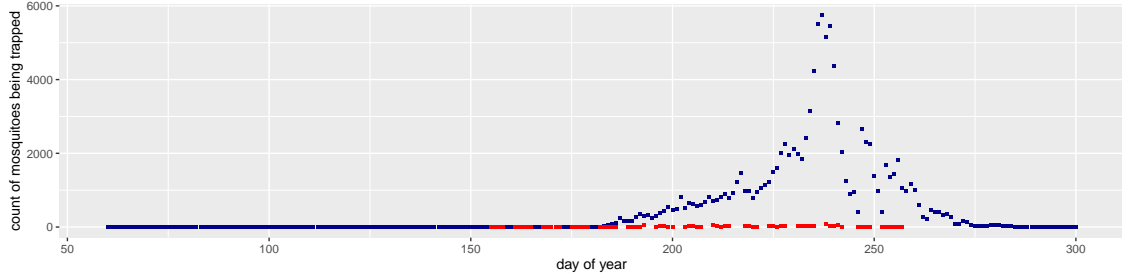
Blue and red represents the number of *Culex* mosquitoes being trapped estimated by the SDM in Scn3 and the empirical count of *Culex* mosquitoes being trapped, respectively.



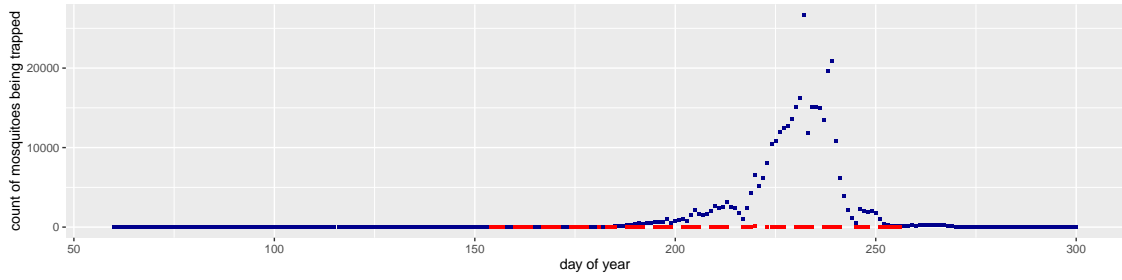
(a) 2011



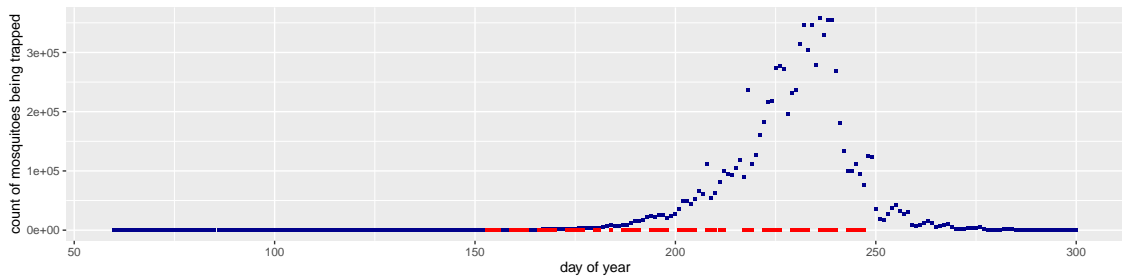
(b) 2012



(c) 2013



(d) 2014



(e) 2015

Figure 6.8: Count of *Culex* mosquitoes being trapped in Scn4

Blue and red represents the number of *Culex* mosquitoes being trapped generated by the SDM in Scn4 and the empirical count of *Culex* mosquitoes being trapped, respectively.

using near day-resolution resolution data. However, linear regression is insufficient to deal with other parameter uncertainties, e.g., static parameters associated with the dynamics of mosquito lifecycle. To address key elements of model parameter uncertainty, this work made an innovative methodological contribution by extensive use of both model calibration and particle filtering to deal with model uncertainties – here, particularly those associated with mosquito abundance and the dynamics of the number of mosquitoes being captured, in the context of the complex relationship between environmental factors and the probability of trapping a given mosquito. Used in this joint fashion, dynamic model calibration can serve a powerful role by efficiently estimating the value of static parameters and addressing the parameter uncertainties regarding model outputs associated with bias in the measurement process – areas not addressed by particle filtering – while particle filtering is used to estimate the latent state of the model. Kalman Filtering yields more accurate estimates of the parameters of a model having noisy simulation data than does unassisted calibration, by flexibly dealing with the system nonlinearities and measurement bias associated with the model [109]. The model parameter estimates obtained from calibration allow for effective operation of particle filtering by using an appropriate treatment of measurement bias.

Particle filtering assumes an underlying Markov process that cannot be directly observed and is applied to state-space models with continuous states, and recurrently regrounds the states and stochastically evolving parameters of the model based on incoming empirical observations. The empirical data compared with model outputs – here, count of mosquitoes being captured by traps on successive days – has obvious time-series characteristics, and each data point is typically individually insufficient to conclusively identify the unobserved state at a given time. But combined with particle filtering using the observations together with the dynamic model, the data are able to give insights into the model latent state that reflects the abundance of mosquitoes at different points of their lifecycle over continuous time. While the underlying SDM is not the most sophisticated dynamic model structure examined in this thesis, the particle filtered model exhibits a complex interface with data not only by virtue of its capacity to estimate model latent state, but also because the observed data relates to model outputs in a way that is mediated by the probability of capturing a given *Culex* mosquito adult, which differs dramatically between different days due to time-varying environmental factors. Because of such variation and the fact that the SDM simulates the *Culex* mosquito population, the empirical count of mosquitoes being captured by traps cannot itself be directly used to parameterize the model, and it is necessary to reason explicitly about the probability of capturing a *Culex* mosquito adult per trap to combine the model with the corresponding empirical time series.

The SDM examined here simulates the development of *Culex* mosquitoes with a requisite degree of detail for the problem at hand, capturing a broad set of factors affecting the life cycle of *Culex* mosquitoes, such as weather, maturation phases, and diapause. The joint use of particle filtering and model calibration allows the SDM to simultaneously reground its underlying state and estimate parameter uncertainties, so as to more effectively match simulation output with empirical data.

The calibration experiment employed the model with 800 particles to run 25000 iterations with a fix

seed value of 1. Within this process, the values of model parameters were tuned to search for the best combination of parameter value to yield a minimized discrepancy between the predicted number of *Culex* mosquitoes being trapped and empirical data. The particles represent the underlying distribution of the model states. Stochastic processes incorporated into the model via dynamic parameters and the stock provides the randomness over time required for particle filtering. The large size of particle used in each realization can offer notable benefit in characterizing the randomness and stochastic process associated with the underlying states of the model, and may thus play an important role in minimizing discrepancy associated with the posterior distribution during the model calibration.

Despite the strength of the approach taken here has yielded strong benefits, there are some important limitations to that approach and the accompanying model. These include the fact that the representation of mosquito diapause is overly simple, the likelihood function used in particle filtering exhibits difficulty in providing a suitably wide distribution of the outputs represented by the particles. Importantly, the parameter uncertainties were estimated by model calibration, which is a relatively simplistic way to deal with unknown parameters. Extensions of the current model to address such limitations, particularly accurate inference regarding poorly-measured static parameters by making use of more sophisticated sampling techniques such as PMCMC, remain an important priority for future work. In addition to the number of *Culex* mosquitoes being captured, the dynamics of the model can also be compared with data on mosquito larval populations so as to better ground model states; this can offer particular benefit, given that such measurements are governed by different values of weather-related factors than are measurements of mosquito adults – and may thus offer effective regrounding even during weather patterns that suppress accurate measurements of mosquitoes on the wing. As *Culex* mosquitoes are the predominantly bridge vector for the high-burden pathogen WNV, the model can be further extended with a representation of spread of WNV infection. This could both provide health insights and open up the possibility using additional data sets for grounding the model, e.g., those involving reported counts of infected horses through clinical and lab test results, hospitalization, rates of severe neurological outcomes and death rates in the population, and historical reporting of dead birds.

CHAPTER 7

CONCLUSION AND FUTURE WORK

Collecting reliable data through traditionally available methods is encountering increasing difficulty in reaching potential respondents and – in some spheres – becoming increasingly expensive. Public health researchers are increasingly aware of the need to leverage the tools of both systems and data science. While traditionally apply separately, this thesis includes elements of an emerging discipline of Systems-Data Science (or systems aware data science) that links dynamic models suitable for investigating counter-factual regimes, including those involving policy and intervention questions, and investigations probing different pathways – including by drawing on statistical modeling and machine learning techniques such as explored in this thesis. These elements are captured in different ways within the various studies included in this thesis.

Smoking is one of the foremost public health threats, and effective surveillance tools – such as those combining increasingly available high-velocity, high-variety data with models over time, such as HMMs – are key to informing policy making. Smartphone sensor data is viable, yet an effective classification algorithm and platform for broad-scale smoking identification remains elusive. ECig use has increased dramatically – particularly in the vulnerable teen years – and has an inconclusive effect on smoking behavior change. GDM is a challenging problem for public health in the ACT, Saskatchewan and in a growing number of jurisdictions worldwide, particularly on account of its influences on future risk of diabetes. Given the interaction of complex risk factors across many levels in ecological context that are controlled tightly for in Randomized Clinical Trials, and the cost, logistics, and ethical challenges associated with such trials, such studies offer incomplete – and sometimes misleading – solutions for evaluating interventions, especially those at the population level [110]. WNV infection is one of the leading causes of MBD and maintains its transmission through mosquitoes of the *Culex* genus. The estimation of the abundance of mosquitoes is of key importance in public health messaging and for effective control of outbreaks of MBD. At the same time, effective prediction of mosquito dynamics is complicated by the marked variability in mosquito population dynamics year-to-year, the pronounced effects of weather-related factors on both mosquito population dynamics and the probability of detection of a given mosquito, and the strong limitations of and labour-intensive character of trap-based estimation of mosquito abundance.

To address such challenges, this thesis has contributed four models demonstrating different levels of sophistication by which dynamic simulation models can be combined with data – with an emphasis on engagement with big data sources, and discussing the characteristics and availability of data in shaping model structure

and the ways in which dynamic model structure shapes the ability to utilize data. These studies show how advanced analytics can aid dynamic model parameterization, calibration, filtering, classification and inference concerning latent state, and temporal projection. Specifically, beyond a common use of parameterization, the applications illustrate use of dynamics models with diverse forms of calibration and filtering techniques – with the ABM of smoking and ECig using predominantly cross-sectional data, a hybrid model of DIP making additional use of longitudinal data, categorical classification with HMMs being used to classify smoking intervals according to categorical classification across time series of smartphone sensor data, and the particle filter being used to estimate the evolving continuous latent state of an SDM of *Culex* mosquito development using daily mosquito trap and weather data, even in the context of highly data-dependent observation functions.

Dynamic model structure affects the ways that the data can be used – for example, by supporting or making infeasible effective use of individual-level longitudinal data in calibration or model validation. That structure further determines the level of sophistication of the model interacting with data – for example, by opening or restricting opportunities for use of certain machine learning algorithms. The characteristics and availability of data also plays an important role in shaping model structure. Machine learning classifiers such as HHMs require a large volume of data associated with categorical underlying states, and assume a memoryless transition between the latent states according time-invariant probabilities, and independence of successive observations emitted from a given state. Therefore, smartphone-based sensor data with self-reported categorical smoking status can be well-suited for informing multivariate HMMs in identifying the unobserved state sequence. In contrast to HMMs, dynamic models allow for capturing continuous states, state-varying time transitions and richer context, and in general support use of cross-sectional data points for the calibration of model results. Individual-level models further support memoryful states, both continuous and discrete heterogeneity, and use of history statistics from individual-level longitudinal data in model calibration. An ABM and (separately) multi-scale hybrid simulation model were constructed to explore the potential for examining the causal structure of a system and interventions using data and dynamic models. The ABM of smoking and ECig use demonstrates the use of cross-sectional data via model calibration. The hybrid model of DIP further illustrates the capacity to exploit rich data availability – including both cross-sectional data and longitudinal data. It further supports increasing the level of sophistication of the model structure – including as building blocks SDM, ABM, and DES – and extending model resolution to advance insights regarding the dynamics of a complex system. The final case study applied particle filtering with an SDM of mosquito development using detailed time-series data. While using a simpler model structure than the model of DIP, the use of particle filtering offers a far richer capacity to make use of data – such as in estimating latent state across the system, and the value of time-varying parameters. This application of particle filtering offers several distinct features. The first is a complex likelihood function to deal with a textured relationship between empirical data and underlying latent model state, where that relationship depends centrally on factors that vary dramatically over time. Secondly, the model extensively jointly uses particle filtering and automated model calibration to deal with the uncertainties associated with the dynamics

of the mosquito population and the dynamics of the number of mosquitoes being captured.

7.1 Solutions

The solutions to the problems encountered in this thesis are briefly described as follows.

- **Identifying Smoking from Smartphone Sensor Data and Multivariate Hidden Markov Models**

Windowed aggregation (every 30 seconds in our case) can effectively reduce noise as well as support extraction of derived features while considering time dependency, taking into account temporal context, rather than simply treating the sensor data of each feature as a Markov process. Here, for high-velocity measurements such as accelerometer data, GPS data and Wi-Fi readings, we successfully extracted derived features such as average and standard deviation for accelerometer data, count of GPS readings drawn from satellite sources, maximum Received Signal Strength Indicator (RSSI), and count of unique media access control (MAC) address from Wi-Fi data, respectively.

- **Multi-Scale Simulation Modeling for Prevention and Public Health Management of Diabetes in Pregnancy and Sequelae**

To scale up the hybrid multi-scale model of DIP, the following model configuration modifications were made. The Garbage First Garbage Collector(G1GC) was switched to ParallelGC (OldParallel) given the consideration that the throughput of processing simulation events is in our favor, rather than the low latency on visualizing animation or responding to user interactions. We further made shared constant agent fields as the statistics fields of the Java class to reduce memory footprint and class instantiation time. We further moved the agent level event scheduler out to the main class to reduce the memory cost for the instantiation of an `EventOriginator`. In the process of calibrating a multi-scale model of DIP, missing assumptions and constraints were identified. Making use of other longitudinal data in addition to the data about the incidence of DIP of each ethnicity, we constrained the outcomes of the model with regards to intergenerational outcomes, particularly those involving the occurrence of macrosomia and later-life diabetes outcomes by age 30 of offspring according to their mother's glycemic status during pregnancy. Investigation revealed that while the risk group identified by the Australasian Diabetes in Pregnancy Society bears the burden of higher incidence rates of diabetes in pregnancy than are seen in the Australian Born population, these elevated rates persist despite lower levels of an important risk factor – overweight and obesity. Based on consultation with members of the expert advisory panel, we incorporated additional assumptions in the form of ethnic-specific differences in β -cell mass and function and/or insulin sensitivity.

- **Particle Filter Applied to System Dynamics Modeling for Mosquito Population Surveillance**

Particle filtering is employed and combined with SDM to reground the state of the model and estimate the uncertain parameters associated with stochastics, in light of incoming observations. We applied linear regression to relate key detection-related parameters – here, related to capturing adult mosquitoes – with weather-related factors. Model calibration was combined with particle filtering for optimizing values of parameters for minimizing the discrepancy coming out of the particle filtering. Both calibration and particle filtering were used to deal with the uncertainties associated with the dynamics of the mosquito population and the dynamics of the number of mosquitoes being captured associated with measurement bias; the use of calibration was particularly notable both in terms of likely being the first use of automated calibration with particle filters for health, and in terms of markedly reducing the discrepancy between model outputs and observed mosquito counts.

7.2 Summary of Findings

7.2.1 Identifying Smoking from Smartphone Sensor Data and Multivariate Hidden Markov Models

Multivariate HMMs informed by binned time-series of transformed smartphone sensor data were employed to classify smoking and non-smoking intervals. In a supervised-learning approach, the smartphone-based Wi-Fi, GPS, and accelerometer time-series sensor data with self-reported smoking periods contributed to estimation of the parameters of multivariate HMMs. The results demonstrate that multivariate HMMs have higher accuracy in classifying states than single feature HMMs due to the use of additional types of sensor data. Both improving data conditioning achieved by the transformations of noisy raw data and tailoring training set and test sets close to the entire smoking cycle serve important roles in improving classifier performance. In light of the central use of supervised learning in this work, the accuracy of self-reporting of ground truth data also significantly affects the predictive accuracy of the multivariate HMMs.

7.2.2 Effect of E-cigarette Use and Social Network on Smoking Behavior Change: An Agent-Based Model of E-cigarette and Cigarette Interaction

We built an ABM of conventional smoking and ECig use to better study the impact of ECig use on smoking behavior change in Canada. We further explored the effects of proximity-based social network influences on smoking and ECig use initiation. Simulation outcomes suggest that ECig use results in substantially lower prevalence of current smokers, and social network increases the prevalence of current ECig users. With calibration of the model to longitudinal data, the ABM of smoking and ECig use demonstrates a substantial potential for providing a deeper understanding of the complex system involved and influences of social network. Outcomes of the simulation suggest that behaviors of ECig use, initiation, cessation, and relapse may notably alter smoking behavior. Further extending the ABM to a multi-scale model involving

dynamics of nicotine tolerance and craving could help to greatly deepen understanding the dynamics of nicotine addiction. Capturing such dynamics could be of particular importance in light of the far longer exposure intervals common in e-cigarette use in some subpopulations [111], and the diversity of nicotine levels employed by e-cigarette users.

7.2.3 Multi-Scale Simulation Modeling for Prevention and Public Health Management of Diabetes in Pregnancy and Sequelae

A multi-scale hybrid model employing ABM, SDM, and DES was built as a versatile and general platform to explore the interaction of risk factors, coupled dynamics of the glycemia-beta-cell-insulin and insulin resistance system, and interventions to address the growing epidemic of GDM and T2DM, employing empirical data from the ACT and Saskatchewan Ministry of Health. The model demonstrates ability to examine health outcomes resulting from a complex system, and the mechanisms for and impact of interventions at the physiological, health service and population levels.

7.2.4 Particle Filter Applied to System Dynamics Modeling for Mosquito Population Surveillance

The particle filtering method was applied to a previously developed SDM of *Culex* mosquito development to predict *Culex* mosquito adult populations using various direct observations – including weather-related factors – and mosquito-related time series from mosquito traps. Model calibration combined with particle filtering allows the model to optimize values of parameters and markedly reduces the discrepancy coming out of the particle filtering. The results demonstrated that particle filtering can reground the dynamic model in light of real-time incoming observations, and support the model in accurately estimating the underlying state and evolving parameters of that model, while calibration can work with particle filtering with high effectiveness to markedly improve model accuracy by estimating the value of static parameters.

7.3 Contributions

The contributions made in this thesis focus on the following areas. First, the work presented in Chapter 3 serves as one of the first contributions using multivariate HMMs informed by multiple types of time-series smartphone sensor data to identifying smoking and non-smoking intervals. In this work, we presented a novel way of fusing various types of smartphone sensor data – including Wi-Fi, GPS, and accelerometer time series – after critical transformations of the raw data. Second, the ABM of smoking and ECig use presented in Chapter 4 incorporated the influence of the social network on smoking and ECig use initiation and simulated complex interactions between smoking and ECig use. Chapter 5 proposed a cutting edge work of developing a multi-scale hybrid model consisting of an SDM module describing the underlying physiological regulation

of glycemia based on beta-cell dynamics and insulin resistance nested in an open population ABM enclosing the previous SDM module representing the population dynamics – including birth, death, weight change, and age cohort shifting – and a DES module describing the operational flow of health services. Combined with three modules taking advantage of various modeling methods, we achieved a performant model of this complex system via a hybrid modeling approach by linking various types of data in the context of judicious performance optimizations. The final line of work contributed an investigation applying particle filtering on an SDM for mosquito population surveillance using various direct observations. Furthermore, this work contributes to the first investigation of combining both model calibration and particle filtering for optimizing values of parameters to minimize the discrepancy coming out of the particle filtering. The combination of model calibration and particle filtering is a novel method to deal with the uncertainties associated with the dynamics of an SDM and measurement bias, and proved strikingly effective in improving dynamic model accuracy.

7.4 Limitation and Future Work

In this thesis, true to the vision of Systems-Data Science we presented the combination of counterfactual-capable dynamic simulation models with machine learning techniques and time series smartphone sensor data, longitudinal data and daily environmental data. Despite fine granularity data and fine resolution of the models that capture a requisite degree of detail, there are some limitations in the models and approaches. Certain directions of future work can be conducted to enhance the performance of models presented in this thesis.

Although the multivariate HMMs examined in Chapter 3 are informed by high granularity data for each participant, incorporating a larger participant pool could help in investigating the influence of tailoring the dataset to an entire smoking cycle on accurate classification of smoking interval. As future work, we will investigate the performance of multivariate HMMs making use of joint distribution of three observations using empirical distribution estimated by using linear interpolated Kernel-Density-Estimation. Another interesting point of future work for investigation is whether the empirical distribution of one participant can be reused on other participants; finally, this work suggests the importance of incorporating added measures to ensure data quality in future studies (such as by asking participants to report their sense of the accuracy of their contributed data, secure in the knowledge that such statements will not prevent receipt of participant payments).

For the ABM of smoking and ECig use, supplementing the model with an evidenced estimate of the rate of ECig use cessation and the rate of ECig use relapse will make the model less sensitive to the use of ECig, which helps to elevate the accuracy of model simulation of smoking and ECig use behavior. This work further suggests the strong value of investigating broader changes to model design, particularly to capture the effect of smoking episodes and dynamics of nicotine metabolism at the fine-grained level. By adding these

mechanisms, the resulting model could support analysis of the effect of ECig use on not just relieving nicotine cravings but also on elevating nicotine tolerance as a smoking cessation tool, and smoking behavior change from the nicotine craving satisfied by ECig instead of smoking. The results also suggest that social network play an important role in shaping smoking and ECig use behavior; therefore, an extension of the current model with a more theoretically grounded and evidenced social network assumptions remains an important priority to simulate effects of social network on the behaviors.

An urgent sphere of extension for the current DIP model presented in Chapter 5 consists of connecting the model to a rich endogenous representation of early weight change, its impact on child birth characteristics and in-utero environment, and its impact on the underlying physiological regulation of glycemia. Especially in light of the strong evidence base recognizing the role of social support in weight change [112], it is further important for future work to tap the model’s capacity to support ready extension for incorporation of social network effects on behavior – for example, social network influences on starting and maintaining lifestyle change and other interventions. Another limitation in the model relates to the simplistic current representation of the dynamics of K_{xgI} and the stylized manner in which it is currently affected by a pregnancy, weight change, and interventions. By extending the current mechanism of K_{xgI} change to a finer resolution, the model will have the potential to simulate the DGR in a more detailed way. Furthermore, the model leaves room for readily incorporating cost and resource utilization into DES to investigate the optimization of resource use during service delivery, and resource- and cost-based tradeoffs between interventions.

Enhancing the sophistication of the model characterization of the probability of *Culex* mosquitoes being captured in traps would be an important priority in future work. As discussed in Chapter 6, weather-based factors directly influence the trapping success, but estimation of such parameters is challenging, due especially to the impacts of such weather-related factors on both mosquito dynamics and detection probability. Extension of the current particle filtering into a Particle Markov chain Monte Carlo (PMCMC) model is an attractive means of making estimating the weather coefficients. An accurate estimate of the weather coefficients could support the model in characterization of mosquito dynamics despite small trap counts due to environmental conditions, reflecting the fact that mosquito activities are decreased due to windy and rainy weather while the population of mosquitoes may remain large. Another limitation remaining in the model is that the representation of mosquito diapause is overly simple. Extending the dynamics of mosquito diapause may be important in estimating the population of mosquitoes, particularly in late spring and early summer. Given the critical role of *Culex* mosquitoes in spreading WNV, it is important to extend the SDM of mosquito population with a representation of spread of WNV infection, which would both provide insight into WNV-related policy, and open up additional data sets for grounding the model – such as those associated with human cases, dead birds, and horses.

7.5 Conclusion

The studies in this thesis demonstrate different levels of sophistication by which dynamic models can be combined with data, the characteristics and availability of data in shaping model structure and scope. The dynamic model structure directly affects the way that data can be used, including by shaping the capacity to represent the underlying data generating process, by supporting a capacity to compare against certain types of data (e.g., longitudinal data), and by the capacity to support analytic approaches that make more effective use of data and predictions with high accuracy. Combining dynamic model with analysis methods such as particle filtering enable the model to be appropriately matched to highly variable data sources, and such methods can be fruitfully and computationally tractably combined with more traditional techniques for use of models with data, such as model calibration and parameterization. Within this study, we employed multivariate HMMs informed by time-series smartphone sensor data with self-labeled smoking intervals to classify smoking and non-smoking intervals. The results demonstrated that, despite their simplicity, such multivariate HMMs can have notable accuracy in classifying such states using multiple types of smartphone sensor data. Richer dynamic simulation models, e.g., SDM, ABM, and DES, support representation of state-varying behaviour and richer contexts; individual-based models can be informed by longitudinal data allowing for capturing memoryful states, and often support more accurate projections and richer investigations of intervention tradeoffs. Particle filtering was applied on a compartmental model to support the model to make an accurate projection, and the results demonstrated the effectiveness of particle filtering in regrounding the states and evolving parameters of the model based on incoming empirical observations, despite highly variable and environment dependent data observation processes. Collectively, the work demonstrates that the appropriate use of advanced analysis methods in parameterization, calibration, and filtering can play an important role in yielding model outcomes with high accuracy and versatility.

REFERENCES

- [1] Center for Disease Control and Prevention, “How Tobacco Smoke Causes Disease,” in *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General*, 2010, pp. 1 – 16.
- [2] U.S. Department of Health and Human Services, *The Health Consequences of Smoking-50 Years of Progress: A Report of the Surgeon General*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014.
- [3] Y. Qin, W. Qian, N. Shojaati, and N. Osgood, “Identifying smoking from smartphone sensor data and multivariate hidden markov models,” in *Social, Cultural, and Behavioral Modeling*, D. Lee, Y.-R. Lin, N. Osgood, and R. Thomson, Eds. Cham: Springer International Publishing, 2017, pp. 230–235.
- [4] C. Pisinger and M. Døssing, “A systematic review of health effects of electronic cigarettes,” *Preventive Medicine*, vol. 69, pp. 248 – 260, 2014.
- [5] B. Demick, “A high-tech approach to getting a nicotine fix,” *Los Angeles Times*, Apr 2019. [Online]. Available: <https://www.latimes.com/archives/la-xpm-2009-apr-25-fg-china-cigarettes25-story.html>
- [6] O. Rom, A. Pecorelli, G. Valacchi, and A. Z. Reznick, “Are e-cigarettes a safe and good alternative to cigarette smoking?” *Annals of the New York Academy of Sciences*, vol. 1340, no. 1, pp. 65–74, Dec 2014.
- [7] Y. Qin, R. Edjoc, and N. D. Osgood, “Effect of e-cigarette use and social network on smoking behavior change: An agent-based model of e-cigarette and cigarette interaction,” in *Social, Cultural, and Behavioral Modeling*, R. Thomson, H. Bisgin, C. Dancy, and A. Hyder, Eds. Cham: Springer International Publishing, 2019, pp. 245–255.
- [8] R. Axtell, S. Durlauf, J. M. Epstein, R. Hammond, B. Klemens, J. Parker, Z. Song, T. Valente, and H. P. Young, “Social influences and smoking behaviour final report to the American Legacy Foundation,” Center on Social and Economic Dynamics(CSED) Economic Studies Program, The Brookings Institution, Washington, DC, Report, February 2006.
- [9] Y.-L. Zhuang, S. E. Cummins, J. Y. Sun, and S.-H. Zhu, “Long-term e-cigarette use and smoking cessation: a longitudinal study with US population,” *Tobacco Control*, vol. 25, no. Suppl 1, pp. i90–i95, Oct 2016.
- [10] S.-H. Zhu, Y.-L. Zhuang, S. Wong, S. E. Cummins, and G. J. Tedeschi, “E-cigarette use and associated changes in population smoking cessation: evidence from US current population surveys,” *BMJ*, vol. 358, 2017.
- [11] A. Ferrara, “Increasing prevalence of gestational diabetes mellitus: a public health perspective.” *Diabetes Care*, vol. 30, no. Supplement 2, pp. S141–S146, 2007.
- [12] H. Lee, H. C. Jang, H. K. Park, B. E. Metzger, and N. H. Cho, “Prevalence of type 2 diabetes among women with a previous history of gestational diabetes mellitus,” *Diabetes Research and Clinical Practice*, vol. 81, no. 1, pp. 124 – 129, 2008.

- [13] D. Dabelea, R. L. Hanson, R. S. Lindsay, D. J. Pettitt, G. Imperatore, M. M. Gabir, J. Roumain, P. H. Bennett, and W. C. Knowler, "Intrauterine exposure to diabetes conveys risks for type 2 diabetes and obesity: a study of discordant sibships," *Diabetes*, vol. 49, no. 12, pp. 2208–2211, 2000.
- [14] N. Osgood, R. F. Dyck, and W. Grassmann, "The inter- and intragenerational impact of gestational diabetes on the epidemic of type 2 diabetes," *American journal of public health*, vol. 101, pp. 173–9, 01 2011.
- [15] T. T. Lao, L.-F. Ho, B. C. Chan, and W.-C. Leung, "Maternal age and prevalence of gestational diabetes mellitus," *Diabetes Care*, vol. 29, no. 4, pp. 948–949, 2006.
- [16] C. Athukorala, A. R. Rumbold, K. J. Willson, and C. A. Crowther, "The risk of adverse pregnancy outcomes in women who are overweight or obese," *BMC Pregnancy and Childbirth*, vol. 10, no. 1, p. 56, Sep 2010.
- [17] Y. Qin, L. Freebairn, J.-A. Atkinson, W. Qian, A. Safarishahrbijari, and N. D. Osgood, "Multi-scale simulation modeling for prevention and public health management of diabetes in pregnancy and sequelae," in *Social, Cultural, and Behavioral Modeling*, R. Thomson, H. Bisgin, C. Dancy, and A. Hyder, Eds. Cham: Springer International Publishing, 2019, pp. 256–265.
- [18] X. Xiong, L. Saunders, F. Wang, and N. Demianczuk, "Gestational diabetes mellitus: prevalence, risk factors, maternal and infant outcomes," *International Journal of Gynecology & Obstetrics*, vol. 75, no. 3, pp. 221 – 228, 2001.
- [19] C. R. Knight-Agarwal, L. T. Williams, D. Davis, R. Davey, T. Cochrane, H. Zhang, and P. Rickwood, "Association of BMI and interpregnancy BMI change with birth outcomes in an Australian obstetric population: a retrospective cohort study," *BMJ Open*, vol. 6, no. 5, p. e010667, May 2016.
- [20] D. S. Feig, B. Zinman, X. Wang, and J. E. Hux, "Risk of development of diabetes mellitus after diagnosis of gestational diabetes," *CMAJ*, vol. 179, no. 3, pp. 229–234, 2008.
- [21] N. Ogden, L. Lindsay, A. Ludwig, A. Morse, H. Zheng, and H. Zhu, "Weather-based forecasting of mosquito-borne disease outbreaks in Canada," *Canada Communicable Disease Report*, no. 5, pp. 127–132, May.
- [22] B. V. Giordano, S. Kaur, and F. F. Hunter, "West Nile virus in Ontario, Canada: A twelve-year analysis of human case prevalence, mosquito surveillance, and climate data," *PLOS ONE*, vol. 12, no. 8, pp. 1–15, 08 2017.
- [23] A. M. Kilpatrick, L. D. Kramer, M. J. Jones, P. P. Marra, and P. Daszak, "West Nile virus epidemics in north america are driven by shifts in mosquito feeding behavior," *PLOS Biology*, vol. 4, no. 4, 02 2006.
- [24] K. A. Bernard, J. G. Maffei, S. A. Jones, E. B. Kauffman, G. D. Ebel, A. P. Dupuis, K. A. Ngo, D. C. Nicholas, D. M. Young, P.-Y. Shi, V. L. Kulasekera, M. Eidson, D. J. White, W. B. Stone, L. D. Kramer, and NY State West Nile Virus Surveillance Team, "West Nile virus infection in birds and mosquitoes, New York State, 2000," *Emerging Infectious Diseases*, vol. 7, no. 4, pp. 679–685, Aug 2001.
- [25] A. Ludwig, H. Zheng, L. Vrbova, M. Drebot, M. Iranpour, and L. Lindsay, "Increased risk of endemic mosquito-borne diseases in Canada due to climate change," *Canada Communicable Disease Report*, vol. 45, pp. 91–97, 04 2019.
- [26] S. T. Peper, D. E. Dawson, N. Dacko, K. Athanasiou, J. Hunter, F. Loko, S. Almas, G. E. Sorensen, K. N. Urban, A. N. Wilson-Fallon, K. M. Haydett, H. S. Greenberg, A. G. Gibson, and S. M. Presley, "Predictive modeling for West Nile virus and mosquito surveillance in Lubbock, Texas," *Journal of the American Mosquito Control Association*, vol. 34, no. 1, pp. 18–24, 2018.
- [27] Statistics Canada, "Canadian tobacco, alcohol and drugs survey," *Health Canada*, Oct 2018. [Online]. Available: <https://www.canada.ca/en/health-canada/services/canadian-tobacco-alcohol-drugs-survey.html>

- [28] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” in *Proceedings of the IEEE*, vol. 77, no. 2, 1989, pp. 257–286.
- [29] R. L. Walter Zucchini, Iain L. MacDonald, *Hidden Markov Models for Time Series An Introduction Using R, Second Edition*, 2nd ed. New York: Chapman and Hall/CRC, 2016, no. 398 pages.
- [30] C. Jared and S. Højsgaard, “Hidden semi markov models for multiple observation sequences: The mhsmm package for R,” *Journal of Statistical Software*, 01 2011.
- [31] Wikipedia, “Baum welch algorithm,” May 2019. [Online]. Available: https://en.wikipedia.org/wiki/Baum%E2%80%93Welch_algorithm
- [32] L. E. Baum and T. Petrie, “Statistical inference for probabilistic functions of finite state markov chains,” *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.
- [33] G. D. Forney, “The viterbi algorithm,” *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, March 1973.
- [34] S. Thrun, “Particle filters in robotics,” in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI’02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 511–518.
- [35] P. M. Djuric, J. H. Kotecha, Jianqui Zhang, Yufei Huang, T. Ghirmai, M. F. Bugallo, and J. Miguez, “Particle filtering,” *IEEE Signal Processing Magazine*, vol. 20, no. 5, pp. 19–38, Sep. 2003.
- [36] A. Doucet and A. M. Johansen, “A tutorial on particle filtering and smoothing: fifteen years later,” in *The Oxford Handbook of Nonlinear Filtering*, D. Crisan and B. Rozovsky, Eds. Oxford University Press, 2009.
- [37] A. Safarishahrbiari, “Particle filtering in compartmental projection models,” Master’s thesis, University of Saskatchewan, January 2019.
- [38] A. Safarishahrbiari, A. Teyhouee, C. Waldner, J. Liu, and N. D. Osgood, “Predictive accuracy of particle filtering in dynamic models supporting outbreak projections,” *BMC Infectious Diseases*, vol. 17, no. 1, p. 648, Sep 2017.
- [39] P. E. Plsek and T. Greenhalgh, “The challenge of complexity in health care,” *BMJ*, vol. 323, no. 7313, pp. 625–628, Sep 2001.
- [40] D. A. Marshall, L. Burgos-Liz, M. J. IJzerman, N. D. Osgood, W. V. Padula, M. K. Higashi, P. K. Wong, K. S. Pasupathy, and W. Crown, “Applying dynamic simulation modeling methods in health care delivery research—the simulate checklist: Report of the ISPOR simulation modeling emerging good practices task force,” *Value in Health*, vol. 18, no. 1, pp. 5 – 16, 2015.
- [41] A. Maria, “Introduction to modeling and simulation,” in *Proceedings of the 29th Conference on Winter Simulation*. IEEE Computer Society, 01 1997, pp. 7–13.
- [42] J. B. Homer and G. B. Hirsch, “System dynamics modeling for public health: Background and opportunities,” *American journal of public health*, vol. 96, no. 3, pp. 452–458, Mar 2006.
- [43] L. K. Kreuger, W. Qian, N. Osgood, and K. Choi, “Agile design meets hybrid models: Using modularity to enhance hybrid model design and use,” in *2016 Winter Simulation Conference (WSC)*, Dec 2016, pp. 1428–1438.
- [44] N. Osgood, “Using traditional and agent based toolset for system dynamics: Present tradeoffs and future evolution,” in *Proceedings of the 25th International Conference of the System Dynamics Society*, 07 2007, p. 19.
- [45] D. A. Luke and K. A. Stamatakis, “Systems science methods in public health: Dynamics, networks, and agents,” *Annual Review of Public Health*, vol. 33, no. 1, pp. 357–376, Apr 2012.

- [46] A. M. El-Sayed, P. Scarborough, L. Seemann, and S. Galea, "Social network analysis and agent-based modeling in social epidemiology," *Epidemiologic Perspectives & Innovations*, vol. 9, no. 1, p. 1, Feb 2012.
- [47] R. A. Nianogo and O. A. Arah, "Agent-based modeling of noncommunicable diseases: A systematic review," *American Journal of Public Health*, vol. 105, no. 3, pp. e20–e31, 2015, PMID: 25602871.
- [48] C. M. Jenkins and S. V. Rice, "Resource modeling in discrete-event simulation environments: A fifty-year perspective," in *Proceedings of the 2009 Winter Simulation Conference*. IEEE, Dec 2009, pp. 755–766.
- [49] B. Milstein, A. Jones, J. Homer, D. Murphy, J. Essien, and D. Seville, "Charting plausible futures for diabetes prevalence in the United States: A role for system dynamics simulation modeling," *Preventing chronic disease*, vol. 4, p. A52, 08 2007.
- [50] A. P. Jones, J. B. Homer, D. L. Murphy, J. D. K. Essien, B. Milstein, and D. A. Seville, "Understanding diabetes population dynamics through simulation modeling and experimentation," *American Journal of Public Health*, vol. 96, no. 3, pp. 488–494, 2006, PMID: 16449587.
- [51] A. Gao, N. Osgood, Y. Jiang, and R. F. Dyck, "Projecting prevalence, costs and evaluating simulated interventions for diabetic end stage renal disease in a Canadian population of aboriginal and non-aboriginal people: An agent based approach," *BMC Nephrology*, vol. 18, 12 2017.
- [52] A. Gao, N. D. Osgood, W. An, and R. F. Dyck, "A tripartite hybrid model architecture for investigating health and cost impacts and intervention tradeoffs for diabetic end-stage renal disease," in *Proceedings of the 2014 Winter Simulation Conference*. IEEE Press, 12 2014, pp. 1676–1687.
- [53] W. Grassmann, J. Zhang, R. Dyck, and N. Osgood, "A system simulation model for type 2 diabetes in the Saskatoon Health Region," in *Proceedings of the 29th International conference of the System Dynamics Society*, St. Gallen, Switzerland, 2012, p. 18.
- [54] I. Ajmera, M. Swat, C. Laibe, and V. C. Le Novère, "The impact of mathematical modeling on the understanding of diabetes and related complications," *Pharmacometrics & Systems Pharmacology*, vol. 2, pp. 1–14, July 2013.
- [55] B. Topp, K. Promislow, G. Devries, R. M. Miura, and D. T. Finegood, "A model of β -cell mass, insulin, and glucose kinetics: Pathways to diabetes," *Journal of Theoretical Biology*, vol. 206, no. 4, pp. 605 – 619, 2000.
- [56] A. De Gaetano, T. Hardy, B. Beck, E. Abu-Raddad, P. Palumbo, J. Bue-Valleskey, and N. Pørksen, "Mathematical models of diabetes progression," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 295, no. 6, pp. E1462–E1479, Dec 2008.
- [57] A. DeGaetano, S. Panunzi, P. Palumbo, C. Gaz, and T. Hardy, "Data-driven modeling of diabetes progression," in *Lecture Notes in Bioengineering*. Springer, Berlin, Heidelberg, 2014, pp. 165–186.
- [58] T. Hardy, E. Raddad, N. Pørksen, and A. De Gaetano, "Evaluation of a mathematical model of diabetes progression against observations in the diabetes prevention program," *American journal of physiology. Endocrinology and metabolism*, vol. 303, pp. E200–12, 05 2012.
- [59] E. Sazonov, P. Lopez-Meyer, and S. Tiffany, "A wearable sensor system for monitoring cigarette smoking," *Journal of studies on alcohol and drugs*, vol. 74, no. 6, pp. 956–964, 2013.
- [60] P. M. Scholl and K. van Laerhoven, "A feasibility study of wrist-worn accelerometer based detection of smoking habits," in *2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, 2012, pp. 886–891.
- [61] B. U. Toreyin, Y. Dedeoglu, and A. E. Cetin, "Contour based smoke detection in video using wavelets," in *2006 14th European Signal Processing Conference*. IEEE, 2006, pp. 1–5.

- [62] P. Lopez-Meyer, S. Tiffany, Y. Patil, and E. Sazonov, "Monitoring of cigarette smoking using wearable sensors and support vector machines," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1867–1872, July 2013.
- [63] S. T. Cherng, J. Tam, P. J. Christine, and R. Meza, "Modeling the effects of e-cigarettes on smoking behavior: Implications for future adult smoking prevalence," *Epidemiology*, vol. 27, no. 6, pp. 819–826, Nov 2016.
- [64] D. Chao, H. Hashimoto, and N. Kondo, "Dynamic impact of social stratification and social influence on smoking prevalence by gender: An agent-based model," *Social Science & Medicine*, vol. 147, pp. 280 – 287, 2015.
- [65] World Health Organization, "Tobacco factsheet," p. 339, 2016. [Online]. Available: <http://www.who.int/mediacentre/factsheets/fs339/>
- [66] Community Preventive Services Task Force, "Reducing tobacco use and secondhand smoke exposure: mobile phone-based cessation interventions," 2013. [Online]. Available: <https://www.thecommunityguide.org/findings/tobacco-use-and-secondhand-smoke-exposure-mobile-phone-based-cessation-interventions>
- [67] M. Raja, "Diagnostic methods for detection of cotinine level in tobacco users: A review," *Journal of Clinical and Diagnostic Research*, vol. 10, no. 3, pp. 4–6, 2016.
- [68] S. E. Meredith, A. Robinson, P. Erb, C. A. Spieler, N. Klugman, P. Dutta, and J. Dallery, "A mobile-phone-based breath carbon monoxide meter to detect cigarette smoking," *Nicotine and Tobacco Research*, vol. 16, no. 6, pp. 766–773, 2014.
- [69] A. A. Ali, S. M. Hossain, K. Hovsepian, M. M. Rahman, K. Plarre, and S. Kumar, "mpuff: Automated detection of cigarette smoking puffs from respiration measurements," in *2012 ACM/IEEE 11th International Conference on Information Processing in Sensor Networks (IPSN)*, April 2012, pp. 269–280.
- [70] M. Hashemian, D. Knowles, J. Calver, W. Qian, M. C. Bullock, S. Bell, R. L. Mandryk, N. Osgood, and K. G. Stanley, "iEpi: an end to end solution for collecting, conditioning and utilizing epidemiologically relevant data," in *Proceedings of the 2nd ACM international workshop on Pervasive Wireless Healthcare*, ACM. New York, NY, USA: Association for Computing Machinery, 2012, pp. 3–8.
- [71] W. Qian, K. G. Stanley, and N. D. Osgood, "The impact of spatial resolution and representation on human mobility predictability," in *Web and Wireless Geographical Information Systems*, S. H. L. Liang, X. Wang, and C. Claramunt, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 25–40.
- [72] Ethica Data, 2017. [Online]. Available: <https://www.ethicadata.com/>
- [73] P. Chaturvedi, A. Mishra, S. Datta, S. Sinukumar, P. Joshi, and A. Garg, "Harmful effects of nicotine," *Indian Journal of Medical and Paediatric Oncology*, vol. 36, no. 1, p. 24, 2015.
- [74] PopulationPyramid.net, "Population pyramids of the world from 1950 to 2100," 2017. [Online]. Available: <https://www.populationpyramid.net/canada/2017/>
- [75] T. R. Holford, D. T. Levy, L. A. McKay, L. Clarke, B. Racine, R. Meza, S. Land, J. Jeon, and E. J. Feuer, "Patterns of birth cohort-specific smoking histories, 1965–2009," *American Journal of Preventive Medicine*, vol. 46, no. 2, pp. e31 – e37, 2014.
- [76] T. A. Wills, J. D. Sargent, F. X. Gibbons, I. Pagano, and R. Schweitzer, "E-cigarette use is differentially related to smoking onset among lower risk adolescents," *Tobacco Control*, vol. 26, no. 5, pp. 534–539, Aug 2016.
- [77] A. M. Leventhal, D. R. Strong, M. G. Kirkpatrick, J. B. Unger, S. Sussman, N. R. Riggs, M. D. Stone, R. Khoddam, J. M. Samet, and J. Audrain-McGovern, "Association of electronic cigarette use with initiation of combustible tobacco product smoking in early adolescence," *JAMA*, vol. 314, no. 7, pp. 700–707, 08 2015.

- [78] H. McRobbie, C. Bullen, J. Hartmann-Boyce, and P. Hajek, “Electronic cigarettes for smoking cessation and reduction,” *Cochrane Database of Systematic Reviews*, no. 12, Dec 2014.
- [79] R. Polosa, P. Caponnetto, J. B. Morjaria, G. Papale, D. Campagna, and C. Russo, “Effect of an electronic nicotine delivery device (e-cigarette) on smoking reduction and cessation: a prospective 6-month pilot study,” *BMC Public Health*, vol. 11, no. 1, Oct 2011.
- [80] J. L. Reid, V. L. Rynard, C. D. Czoli, and D. Hammond, “Who is using e-cigarettes in Canada? nationally representative data on the prevalence of e-cigarette use among Canadians,” *Preventive Medicine*, vol. 81, pp. 180 – 183, 2015.
- [81] Statistic Canada, “Fertility: Overview, 2012 to 2016,” June 2018. [Online]. Available: <https://www150.statcan.gc.ca/n1/pub/91-209-x/2018001/article/54956-eng.html>
- [82] Statistics Canada, “Deaths and mortality rates, by age group,” January. [Online]. Available: <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310071001>
- [83] E. Lehmann and T. Deutsch, “A physiological model of glucose-insulin interaction in type 1 diabetes mellitus,” *Journal of Biomedical Engineering*, vol. 14, no. 3, pp. 235 – 242, 1992, annual Scientific Meeting.
- [84] *ACT Maternal Perinatal Data Collection, Australian Capital Territory. Health Directorate & 2011 Census of Population and Housing.* Australian Bureau of Statistics, 2011.
- [85] PopulationPyramid.net. [Online]. Available: <https://www.populationpyramid.net/australia/1978/>
- [86] A. Nankervis, H. McIntyre, G. Moses, R. and Ross, L. Callaway, C. Porter, W. Jeffries, C. Boorman, B. De Vries, and A. McElduff, “ADIPS consensus guidelines for the testing and diagnosis of gestational diabetes mellitus in Australia,” June 2014.
- [87] Australian Bureau of Statistics, “Deaths, year of occurrence, age at death, age-specific death rates, sex, states, territories and Australia,” February 2019. [Online]. Available: http://stat.data.abs.gov.au/Index.aspx?DataSetCode=DEATHS_AGESPECIFIC_\OCCURENCEYEAR
- [88] P. M. Catalano, “Trying to understand gestational diabetes,” *Diabetic Medicine*, vol. 31, no. 3, pp. 273–281, Feb 2014.
- [89] S. Panunzi, A. De Gaetano, and G. Mingrone, “Advantages of the single delay model for the assessment of insulin sensitivity from the intravenous glucose tolerance test,” *Theoretical Biology and Medical Modelling*, vol. 7, no. 1, Mar 2010.
- [90] A. Hayes, E. Gearon, K. Backholer, A. Bauman, and A. Peeters, “Age-specific changes in BMI and BMI distribution among Australian adults using cross-sectional surveys from 1980 to 2008,” *International Journal of Obesity*, vol. 39, no. 8, pp. 1209–1216, Apr 2015.
- [91] H. L. Walls, R. Wolfe, M. M. Haby, D. J. Magliano, M. de Courten, C. M. Reid, J. J. McNeil, J. Shaw, and A. Peeters, “Trends in BMI of urban Australian adults, 1980–2000,” *Public Health Nutrition*, vol. 13, no. 5, pp. 631–638, 2010.
- [92] G. F. Mateo, E. Granado-Font, C. Ferré-Grau, and X. Montaña-Carreras, “Mobile phone apps to promote weight loss and increase physical activity: A systematic review and meta-analysis,” *Journal of Medical Internet Research*, vol. 17, no. 11, p. e253, Nov 2015.
- [93] C. S. Weisman, M. M. Hillemeier, D. S. Downs, M. E. Feinberg, C. H. Chuang, J. J. Botti, and A.-M. Dyer, “Improving women’s preconceptional health: Long-term effects of the strong healthy women behavior change intervention in the central Pennsylvania Women’s Health Study,” *Women’s Health Issues*, vol. 21, no. 4, pp. 265 – 271, 2011.
- [94] The HAPO Study Cooperative Research Group, “Hyperglycemia and adverse pregnancy outcomes,” *Obstetrical & Gynecological Survey*, vol. 63, no. 10, pp. 615–616, Oct 2008.

- [95] P. W. Franks, H. C. Looker, S. Kobes, L. Touger, P. A. Tataranni, R. L. Hanson, and W. C. Knowler, "Gestational glucose tolerance and risk of type 2 diabetes in young Pima Indian offspring," *Diabetes*, vol. 55, no. 2, pp. 460–465, 2006.
- [96] S. Karki, G. L. Hamer, T. K. Anderson, T. L. Goldberg, U. D. Kitron, B. L. Krebs, E. D. Walker, and M. O. Ruiz, "Effect of trapping methods, weather, and landscape on estimates of the *Culex* vector mosquito abundance," *Environmental Health Insights*, vol. 10, p. EHL.S33384, Jan 2016.
- [97] D. L. Kline, "Traps and trapping techniques for adult mosquito control," *Journal of the American Mosquito Control Association*, vol. 22, no. 3, pp. 490 – 496, 2006.
- [98] H. E. Brown, M. Paladini, R. A. Cook, D. Kline, D. Barnard, and D. Fish, "Effectiveness of mosquito traps in measuring species abundance and composition," *Journal of Medical Entomology*, vol. 45, no. 3, pp. 517–521, 10 2014.
- [99] D. Yu, N. Madras, and H. Zhu, "Temperature-driven population abundance model for *Culex pipiens* and *Culex restuans* (Diptera: Culicidae)," *Journal of Theoretical Biology*, vol. 443, pp. 28 – 38, 2018.
- [100] R. Rosà, G. Marini, L. Bolzoni, M. Neteler, M. Metz, L. Delucchi, E. A. Chadwick, L. Balbo, A. Mosca, M. Giacobini, L. Bertolotti, and A. Rizzoli, "Early warning of West Nile virus mosquito vector: climate and land use models successfully explain phenology and abundance of *Culex pipiens* mosquitoes in north-western Italy," *Parasites & Vectors*, vol. 7, no. 1, p. 269, Jun 2014.
- [101] P. Cong, "Modeling mosquito activity built on mosquito population dynamics: A simulation study," Master's thesis, University of Saskatchewan, September 2017.
- [102] R. Wieland, A. Kerkow, L. Früh, H. Kampen, and D. Walther, "Automated feature selection for a machine learning approach toward modeling a mosquito distribution," *Ecological Modelling*, vol. 352, pp. 108 – 112, 2017.
- [103] L. Früh, H. Kampen, A. Kerkow, G. A. Schaub, D. Walther, and R. Wieland, "Modelling the potential distribution of an invasive mosquito species: comparative evaluation of four machine learning methods and their combinations," *Ecological Modelling*, vol. 388, pp. 136 – 144, 2018.
- [104] Canadian Centre for Climate Services, "Historical climate data," April 2019. [Online]. Available: <http://climate.weather.gc.ca/>
- [105] A. Djènontin, C. Pennetier, B. Zogo, K. B. Soukou, M. Ole-Sangba, M. Akogbéto, F. Chandre, R. Yadav, and V. Corbel, "Field efficacy of Vectobac GR as a mosquito larvicide for the control of Anopheline and Culicine mosquitoes in natural habitats in Benin, West Africa," *PLOS ONE*, vol. 9, no. 2, pp. 1–7, 02 2014.
- [106] W. An, "Decision trees for dynamic decision making and system dynamics modelling calibration and expansion," Master's thesis, University of Saskatchewan, September 2014.
- [107] CLARKE, "Vectobac larvicide." [Online]. Available: <https://www.clarke.com/vectobac>
- [108] J. E. Soverow, G. A. Wellenius, D. N. Fisman, and M. A. Mittleman, "Infectious disease in a warming world: How weather influenced West Nile virus in the United States (2001–2005)," *Environmental Health Perspectives*, vol. 117, no. 7, pp. 1049–1052, Jul 2009.
- [109] D. W. Peterson, "Hypothesis, estimation, and validation of dynamic social models : energy demand modeling," Ph.D. dissertation, Massachusetts Institute of Technology, 1975.
- [110] D. A. Luke and K. A. Stamatakis, "Systems science methods in public health: Dynamics, networks, and agents," *Annual review of public health*, vol. 33, no. 1, pp. 357–376, 04 2012.
- [111] N. Osgood and E. Penz, "Feeding the habit: Insights into cigarette & e-cigarette use and interaction via big data," Respiratory Research Center, University of Saskatchewan, Talk, 2019.

- [112] E. Johnson, T. Carson, O. Affuso, C. Hardy, and M. Baskin, "Relationship between social support and body mass index among overweight and obese African American women in the rural deep South, 2011-2013," *Preventing Chronic Disease*, vol. 11, p. E224, 2014.

APPENDIX A

CALIBRATED PARAMETERS OF THE HYBRID MULTI-SCALE MODEL OF DIABETES IN PREGNANCY

Parameters	Value	Unit	Description
thresholdCoefficientToT2DM	1.636	unit	Coefficient of diagnosis criteria for T2DM.
coeffRatesToGDM	0.642	unit	Coefficient of diagnosis criteria for GDM.
fastingBloodGlucoseThresholdBefore2015	5.504	mM	Glycemia value used as diagnosis threshold.
tEthaEnd	0.341	mo^{-2}	Value of pancreatic reserve η at the end of the period of observation
offspringKxgIGlycemiaCoefficient	0.508	unit	Coefficient for offspring's <i>KxgI</i> impairment
incidenceOfDIPIncrease	1.5	unit	Fraction increase of incidence of DIP after diagnosis threshold change after year 2015
coefficientTEthaEndADiPS	0.354	unit	Coefficient for tEthaEnd of ADiPS ethnicity group.

Table A.1: Calibrated parameters of the hybrid multi-scale model of diabetes in pregnancy

APPENDIX B

LINEAR REGRESSION ON WEATHER-RELATED VARIABLES FOR THE PROBABILITY OF A *Culex* MOSQUITO BEING CAPTURED BY CDC LIGHT TRAP

Notation	Value	Description
β_0	-18.421	Intercept
β_t	0.055	Coefficient for temperature
β_w	-0.002	Coefficient wind speed
β_{p1}	0.063	Coefficient for precipitation
β_{p2}	-0.0012	coefficient for precipitation square

Table B.1: Coefficients of weather-related variables for the probability of a *Culex* mosquito being captured by CDC light trap

Coefficients	Estimate	Std. Error	$P_r(> t)$	2.5% CI	97.5% CI
β_t	0.055	0.042	0.183	-0.026	0.137
β_w	-0.002	0.02	0.922	-0.041	0.037
β_{p1}	0.063	0.045	0.162	-0.026	0.152
β_{p2}	-0.0012	0.001	0.219	-0.003	0.0007

Table B.2: Summary of the linear regression on weather-related variables for the probability of a *Culex* mosquito being captured by CDC light trap

APPENDIX C

CALIBRATED PARAMETERS OF THE PARTICLE FILTERING MODEL APPLIED TO THE SDM OF *Culex* MOSQUITO DEVELOPMENT

Parameters	Value	Unit	Description
waterHeightForMaxEggLaying	0.55	mm	Water height for favorability of egg laying
outOfHibernationRate	0.76	per day	Rate for mosquito adults out of diapause state
AdultMeanLifeTime	13.67	day	Adult mean life time
EggGestationPeriod	1.41	day	Egg gestation period
MeanTimeForEggHatch	1.82	day	Mean time for egg being hatched
betaTemp	0.06	unit	β_t , coefficient for temperature variable for determining trap capture probability
betaWindSpeed	-0.004	unit	β_w , coefficient for temperature variable for determining trap capture probability
betaPrecipitation	0.033	unit	β_{p1} , coefficient for temperature variable for determining trap capture probability
betaPrecipitationSquare	-0.005	unit	β_{p2} , coefficient for temperature variable for determining trap capture probability

Table C.1: Calibrated parameters of the particle filtering model applied to SDM of *Culex* mosquito population

APPENDIX D

PARAMETERS OF THE SDM OF *Culex* MOSQUITO DEVELOPMENT

Parameter name	Notation	Value	Unit
initialStandingWaterMm		100	mm
startingAquaticDensity		0.346	unit
initialEggDensity		0.637	unit
waterHeightForMaxEggLaying		calibrated	mm
minTempForOutOfDiapausePreAdult		17.834	mm
durationForLowPassPrecip		25	day
outOfHibernationRate	α	calibrated	per day
minTempForIntoDiapausePreAdult		16.285	C^o
adultMeanLifeTime	t_m	calibrated	day
minumAcceptableEffectiveSampleSizeFraction		0.25	unit
minInitialMinTemperatureAdultOf0fDiapause		0.0	C^o
maxInitialMinTemperatureAdultOf0fDiapause		15.0	C^o
daylightHoursForDiapause		15.4	hour
beta0	β_0	-18.421	unit
betaTemp	β_T	calibrated	unit
betaWind	β_W	calibrated	unit
betaPrecipitation	β_{P1}	calibrated	unit
betaPrecipitationSquare	β_{P2}	calibrated	unit
eggGestationPeriod	t_{eg}	calibrated	day
eggLayingTempCoeffPerDegreeCMin		0.005	unit
eggLayingTempCoeffPerDegreeCMax		0.4	unit
cRandomWalkStdDev		0.5	unit
meanTimeForEggHatch	t_{eh}	calibrated	day
maxTemperatureCoefficientForEggLayingRate		4.0	unit
pupaeToAdultRateCoeffMin		0.005	unit
pupaeToAdultRateCoeffMax		0.7	unit
stdPupaeToAdult		0.8	unit

Table D.1: Parameters of the particle filtering model applied to SDM of *Culex* mosquito population