DEVELOPING GENERIC AND SCALABLE FRAMEWORK FOR GEOGRAPHICALLY EXPLICIT INFECTIOUS DISEASE SIMULATING ABM

LOZA BEKALO SAPPA February, 2014

SUPERVISORS: Ir. P.W.M. Ellen-Wien Augustijn Dr. R. Raul Zurita-Milla

DEVELOPING GENERIC AND SCALABLE FRAMEWORK FOR GEOGRAPHICALLY EXPLICIT INFECTIOUS DISEASE SIMULATING ABM

LOZA BEKALO SAPPA Enschede, The Netherlands, February, 2014

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation. Specialization: [GFM]

SUPERVISORS: Ir. P.W.M. Ellen-Wien Augustijn Dr. R. Raul Zurita-Milla

THESIS ASSESSMENT BOARD: Prof.Dr. M.J. Kraak (Chair) Dr. Ir. A. Ligtenberg, (External Examiner, Wageningen UR)



DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Infectious diseases are one of the threats for the well-being and healthy life of human beings. Efficient approaches should be developed to gain understanding about outbreaks of infectious disease and analyze the impact of prevention and control method before their implementation. Agent based modeling (ABM) is one of the techniques used to analyze disease outbreaks by representing the detailed data of individuals and their contacts that have influence on the spread of epidemics. Despite their advantages, the computationally demanding nature of ABMs limits their applicability for simulation of large scale infectious diseases outbreaks.

This research is aimed at developing a generic and scalable framework that can simulate the outbreak of any types of infectious disease irrespective of their spatial scale. The study started by examining different modeling approaches to achieve generic-ness and scalability of the model. Three existing models were evaluated to identify suitable modeling techniques. After several investigations, the classical SEIR model, social network and parallel processing techniques are adopted to develop the model of this research.

Integrating the concepts of the selected modeling techniques, a conceptual model that can express the framework was developed. The conceptual model mainly relies on the concept of "divide and conquer", which decomposes the given spatial outbreak extent to sub simulation systems and conquer the outcomes of sub processes to represent the entire simulation. The sub simulation systems contain the detailed representations of individuals with their diversified interaction level and they are simulated in the parallel processing environment independent of each other. To represent the communication between sub processors, the movement of commuters from a given sub system to another was captured and handled by a reporter which has a responsibility of controlling these commuters. To achieve the overall process of the conceptual model, three sub models were developed namely an activity model, social interaction model and disease model. Beside this, the divide and conquer process was managed by the parallel processing model.

The framework was developed based on the conceptual model using RepastHPC modeling toolkit. The implementation was followed by a verification and evaluation of the framework. The model was implemented to simulate the outbreak of pertussis in two cities of the Netherlands namely Enschede and Hengelo. The verification and evaluation clearly state that the model is generic as well as scalable.

Keywords

Infectious diseases, Agent based modelling (ABM), Classic SEIR model, Social network, Parallel processing, RepastHPC

ACKNOWLEDGEMENTS

First and for most, I would like to glorify almighty God for his love and mercy in all my ways. Surely, without his protection and guidance, nothing would have been achieved. He has been faithful and unconditional helper throughout my journey. Glorified be his mighty name. Amen.

I am grateful to my first supervisor Ms Ellen-Wien Augustijn for her consistent encouragement and guidance during the research period. Her timely review of the research and outstanding comments enabled me to come up with this final volume. She has reshaped me not only in academic atmosphere, but also in my social life. I have learnt hardworking and diligence from her. THANK YOU VERY MUCH MY BELOVED "Mom".

I like to appreciate my second supervisor Dr. R. (Raul) Zurita-Milla for his constructive ideas during the research period.

I am thankful for Mr Bas Retsios who is brilliant programmer and amazing cooperative worker. His assistance and constructive comments in my implementation has irreplaceable contribution for finishing of this research. Thanks a lot. And I would like to acknowledge the repast HPC forum coordinators and especially for Mr Murphy, John T for fast and intelligent solutions.

I am indebted to The Netherlands Fellowship Programme-NFP for enabled me to have a quality education by covering all the expanses. Thanks to university of Twente, and University of Hawassa for granting me such a wonderful opportunity.

I like to say thank you for my beloved husband Mr Getachew, my mother Ms Amarech, my brother Mr Wolde Bekalo and all family members for their uncountable contributions. Without their prayer and help no achievement in my life could have been true.

Special appreciation goes to Mr Teshale Tadesse for his committed and all rounded contribution, during this difficult season. In the same way I like to extend my appreciations to my beloved friend and "sister" Frehiwot Melak Arega for her persistence helps in all my ways. Similarly many thanks for 'HABESHA' community members and ICF-Fellowship members for your kind contribution in my life. Love You All.

Loza Bekalo Sappa Enschede, The Netherlands, February 2014

TABLE OF CONTENTS

Acknowledgements iii 1. Introduction 1 11. Motivation and Problem Statements. 1 12. Research Identification 3 12.1. Research Objective 3 12.2. Research Objective 3 12.3. Incovation Aimed At 4 12.4. Related Work 44 1.5. Incovation Aimed At 4 1.4. Overview of the Following Chapters 5 14. Overview of the Following Chapters 5 2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.1. Social Interaction 88 2.2.1. Social Interaction 88 2.2.2. Commuting 9 2.3. Spreading Nature of Epidemics 10 2.4. Modeling Techniques. 11 2.4.1. Classic Epidemics 11 2.4.2. Social Network. 12 2.4.3. Agent Based Modelling. 13 2.5. Parallel Processing	Abst	tract	i
1. Introduction 1 1.1. Motivation and Problem Statements 1 1.2. Research Identification 3 1.2.1. Research Objective 3 1.2.2. Research Questions 3 1.2.3. Innovation Aimed At 4 1.4.4. Related Work 4 1.3. Project Setup 5 1.4. Overview of the Following Chapters 5 2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.1. Social Interaction 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Spreading Nature of Epidemics 10 2.4. Modeling Techniques 11 2.4. Social Network 12 2.4.3. Agent Based Modeling 13 2.5. Parallel Processing 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent Based Modeling 17 3.1. Daily Activity Model 20 31.4. <td>Ack</td> <td>nowledgements</td> <td>ii</td>	Ack	nowledgements	ii
1.1. Motivation and Problem Statements. 1 1.2. Research Identification 3 1.2.1. Research Questions 3 1.2.2. Research Questions 3 1.2.3. Innovation Aimed At 4 1.2.4. Related Work 4 1.2.5. Innovation Aimed At 4 1.2.4. Related Work 4 1.3. Project Stup 5 2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.2. Major Factors Related to the Outbreak of Infectious Diseases 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Immunity 9 2.4. Construing 9 2.5. Interaction 10 2.4. Modeling Techniques 11 2.4.1. Classic Epidemic Models 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.1.4. Strength and Limitations	1.	Introduction	1
1.2. Research Objective 3 1.2.1. Research Objective 3 1.2.2. Research Questions 3 1.2.3. Innovation Aimed At 4 1.2.4. Related Work 4 1.3. Project Setup 5 1.4. Overview of the Following Chapters 5 2. Background Concepts 7 2. Characteristics of Epidemics (Infectious Diseases) 7 2. Major Factors Related to the Outbreak of Infectious Diseases 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Immunity 9 2.3. Immunity 9 2.4. Addeling Techniques 10 2.4.1. Classic Epidemic Models 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1.1. Daily Activity Model 20 3.2.3. Huminity of Cographically-Explicit Agent-Based Modeling 17 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.1.4. Strength and Limitations of This Model 20<		1.1. Motivation and Problem Statements	1
1.2.1. Research Objective 3 1.2.2. Research Questions 3 1.2.3. Innovation Aimed At 4 1.2.4. Related Work 4 1.3. Project Scup 5 1.4. Overview of the Following Chapters 5 1.4. Overview of the Following Chapters 7 2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.1. Social Interaction 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Spreading Nature of Epidemics 10 2.4. Addeling Techniques 10 2.4. Classic Epidemic Models 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertusis in Enschede Region Using Agent-Based Modeling 19 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.1. Hierarchical Model 21 3.2.2.3. Regional Hierarchical Model 22 3.2.3. Municipality Lev		1.2. Research Identification	3
1.2.2. Research Questions 3 1.2.3. Innovation Aimed At 4 1.2.4. Related Work 4 1.3. Project Setup 5 1.4. Overview of the Following Chapters 5 2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.2. Major Factors Related to the Outbreak of Infectious Diseases 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Immunity 9 2.3. Innunity 9 2.3. Intronuity 9 2.4.1. Classic Epidemic Models 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling, 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 20 3.2.4. Strength and Limitations of this Model 20 3.2.4. Strength and Limitations of this Model 21 3.2.2. Regional Hierarchical Model 23 3.		1.2.1. Research Objective	3
1.2.3. Innovation Aimed At. 4 1.2.4. Related Work 4 1.3. Project Setup 5 1.4. Overview of the Following Chapters 5 2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.2. Major Factors Related to the Outbreak of Infectious Diseases 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Spreading Nature of Epidemics 10 2.4. Modeling Techniques 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling 12 2.4.4.2. Social Network 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.3. Disease Model 20 3.2.4. Strength and Limitations of this Model 20 3.2.3. Herarchical Model 20 3.2.4. Strength and Limitations of This Model 21 3.2.5. Parallel Processing of Geographically-Explicit Agent-Based Disease Diffusion M odels 21		1.2.2. Research Questions	3
1.2.4. Related Work 4 1.3. Project Setup 5 1.4. Overview of the Following Chapters 5 1.4. Overview of the Following Chapters 5 2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.2. Major Factors Related to the Outbreak of Infectious Diseases 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Immunity 9 2.3. Spreading Nature of Epidemics 10 2.4. Modeling Techniques 11 2.4.1. Classic Epidemic Models 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 20 3.1.4. Strength and Limitations of this Model 20 3.1.2. Social Interactioin Model		1.2.3. Innovation Aimed At	4
1.3. Project Setup 5 1.4. Overview of the Following Chapters 5 1.4. Overview of the Following Chapters 5 2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.1. A cotal Interaction 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Spreading Nature of Epidemics 10 2.4.1. Classic Epidemic Models 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling 11 2.4.4. Classic Epidemic Models 11 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.5. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels 21 3.2.1. Hierarchical Model 22 3.2.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23		1.2.4. Related Work	4
1.4. Overview of the Following Chapters 5 2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.2. Major Factors Related to the Outbreak of Infectious Diseases 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Immunity 9 2.3. Spreading Nature of Epidemics 10 2.4. Modeling Techniques 11 2.4.1. Classic Epidemic Models 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.1. Hierarchical Model 22 3.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3.1 IPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models		1.3. Project Setup	5
2. Background Concepts 7 2.1. Characteristics of Epidemics (Infectious Diseases) 7 2.2. Major Factors Related to the Outbreak of Infectious Diseases 8 2.2.1. Social Interaction 8 2.2.2. Commuting 9 2.3. Spreading Nature of Epidemics 10 2.4. Modeling Techniques 11 2.4.1. Classic Epidemic Models 11 2.4.2. Social Network. 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.1. Hierarchical Model 20 3.2.2. Regional Hierarchical Model 22 3.2.1. Hierarchical Model 22 3.2.2. Regional Hierarchical Model 23 </td <td></td> <td>1.4. Overview of the Following Chapters</td> <td>5</td>		1.4. Overview of the Following Chapters	5
21. Characteristics of Epidemics (Infectious Diseases) 7 22. Major Factors Related to the Outbreak of Infectious Diseases 8 2.1. Social Interaction 8 2.2.1. Social Interaction 9 2.2.3. Immunity 9 2.3. Spreading Nature of Epidemics 10 2.4. Modeling Techniques 11 2.4.1. Classic Epidemic Models 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.2. Regional Hierarchical Model 22 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3.1. Limitations and Strength of this Model 24 3.3.1. Limitations and Strength of this Model 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis </td <td>2.</td> <td>Background Concepts</td> <td>7</td>	2.	Background Concepts	7
22. Major Factors Related to the Outbreak of Infectious Diseases		2.1. Characteristics of Epidemics (Infectious Diseases)	7
2.2.1. Social Interaction		2.2. Major Factors Related to the Outbreak of Infectious Diseases	8
2.2.2. Commuting		2.2.1. Social Interaction	8
2.2.3. Immunity		2.2.2. Commuting	9
2.3. Spreading Nature of Epidemics 10 2.4. Modeling Techniques 11 2.4.1. Classic Epidemic Models 11 2.4.2. Social Network 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 19 3.1.3. Discase Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.1. Hierarchical Model 21 3.2.2. Regional Hierarchical Model 22 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 23 3.2.3. HPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. Purpose and Model Components 29 4.1.1. Purpose and Model Components 29 4.1.2 Stare Variables Agent behaviou		2.2.3. Immunity	9
2.4. Modeling Techniques		2.3. Spreading Nature of Epidemics	10
2.4.1. Classic Epidemic Models. 11 2.4.2. Social Network. 12 2.4.3. Agent Based Modelling. 13 2.5. Parallel Processing. 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 19 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.1. Hierarchical Model 20 3.2.2. Regional Hierarchical Model 22 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3.1. Limitations and Strength of this Model 25 3.4. Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. General Setup of the Model <t< td=""><td></td><td>2.4. Modeling Techniques</td><td> 11</td></t<>		2.4. Modeling Techniques	11
2.4.2. Social Network 12 2.4.3. Agent Based Modelling 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 19 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.1. Hierarchical Model 20 3.2.2. Regional Hierarchical Model 22 3.2.1. Hierarchical Model 23 3.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. General Setup of the Model 29 4.1.1. Purpose and Model Components 29		2.4.1. Classic Epidemic Models	11
2.4.3. Agent Based Modelling. 13 2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 19 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.1. Hierarchical Model 20 3.2.2. Regional Hierarchical Model 21 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 23 3.2.3. Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. General Setup of the Model 29 4.1.1. Purpose and Model Components 29 4.1.2 State Variables Agent behaviour and Scales 31		2.4.2. Social Network	12
2.5. Parallel Processing 14 3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 19 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.2. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels 21 3.2.1. Hierarchical Model 23 3.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3.1. Limitations and Strength of this Model 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. Purpose and Model Components 29 4.1.1. Purpose and Model Components 29 4.1.2 State Variables Agent behaviour and Scales 31		2.4.3. Agent Based Modelling	13
3. Evaluation of Existing Models 17 3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 19 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.2. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels 21 3.2.1. Hierarchical Model 22 3.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3. HPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. General Setup of the Model 29 4.1.1. Purpose and Model Components 29 4.1.2. State Variables. Agent behaviour and Scales 31		2.5. Parallel Processing	14
3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling 17 3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 19 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.2.2. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels 21 3.2.1. Hierarchical Model 22 3.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3. HPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1.1. Purpose and Model Components 29 4.1.2. State Variables Agent behaviour and Scales 31	3.	Evaluation of Existing Models	17
3.1.1. Daily Activity Model 19 3.1.2. Social Interaction Model 19 3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.2. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels 21 3.2.1. Hierarchical Model 22 3.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. Purpose and Model Components 29 4.1.2 State Variables. Agent behaviour and Scales 31	_	3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling	17
3.1.2. Social Interaction Model		3.1.1. Daily Activity Model	19
3.1.3. Disease Model 20 3.1.4. Strength and Limitations of this Model 20 3.2. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels 21 3.2.1. Hierarchical Model 22 3.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. General Setup of the Model 29 4.1. Purpose and Model Components 29 4.1.2 State Variables. Agent behaviour and Scales 31		3.1.2. Social Interaction Model	19
3.1.4. Strength and Limitations of this Model 20 3.2. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels 21 3.2.1. Hierarchical Model 22 3.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3. HPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. Purpose and Model Components 29 4.1.1. Purpose and Model Components 29 4.1.2. State Variables. Agent behaviour and Scales 31		3.1.3. Disease Model	20
 3.2. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels		3.1.4. Strength and Limitations of this Model	20
3.2.1. Hierarchical Model 22 3.2.2. Regional Hierarchical Model 23 3.2.3. Municipality Level Model 23 3.2.4. Strength and Limitations of This Model 24 3.3. HPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models 24 3.3.1. Limitations and Strength of this Model 25 3.4. Summary of Existing Models SWOT Analysis 26 4. Conceptual Model 29 4.1. General Setup of the Model 29 4.1.1. Purpose and Model Components 29 4.1.2. State Variables. Agent behaviour and Scales 31		3.2. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels	21
 3.2.2. Regional Hierarchical Model		3.2.1. Hierarchical Model	22
 3.2.3. Municipality Level Model		3.2.2. Regional Hierarchical Model	23
 3.2.4. Strength and Limitations of This Model		3.2.3. Municipality Level Model	23
 3.3. HPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models		3.2.4. Strength and Limitations of This Model	24
 3.3.1. Limitations and Strength of this Model		3.3. HPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models	24
 3.4. Summary of Existing Models SWOT Analysis		3.3.1. Limitations and Strength of this Model	25
 4. Conceptual Model		3.4. Summary of Existing Models SWOT Analysis	26
4.1. General Setup of the Model	4.	Conceptual Model	
4.1.1. Purpose and Model Components 29 4.1.2. State Variables. Agent behaviour and Scales 31		4.1. General Setup of the Model	29
4.1.2 State Variables Agent behaviour and Scales 31		4.1.1. Purpose and Model Components	29
1.1.2. Otate variables, rigent benaviour and bears		4.1.2. State Variables, Agent behaviour and Scales	31
4.1.3. Process Overview of the Sub Models		4.1.3. Process Overview of the Sub Models	32
4.2. Sub Models		4.2. Sub Models	33

	4.2.1. Activity Model	
	4.2.2. Modelling Social Interactions	
	4.2.3. Disease Model	
	4.3. Parallel Processing and Commuting Model	
	4.4. Summary	
5.	Data Preparation	
	5.1. Introduction	
	5.2. Population Census Data	
	5.3. Spatial Data	
	5.4. Vaccination and Commuters Data	
6.	Implementation	49
	6.1. Tools to Implement ABM	
	6.1.1. Fundamental Concepts of Repast	
	6.1.2. RepastHPC	
	6.2. Simulation Data Preprocessing	
	6.3. Implementing of the Framework	
	6.3.1. Synthetic Population Generation	
	6.3.2. Commuting	54
	6.3.3. Create Social Network	54
	6.3.4. Activity and Disease Model	
7.	Result and Discussion	
	7.1. Verification	
	7.1.1. Infection in Different Age Groups	
	7.1.2. The Infection in Different Time Steps	61
	7.1.3. Epidemic Curve	61
	7.1.4. Disease Transmission in Different Locations	63
	7.1.5. The Spatial Extent of the Outbreak	64
	7.2. Sensitivity Analysis	
	7.3. Scalability	
8.	Conclusion and Recommendation	
	8.1. Conclusion	
	8.2. Recommendation	
Re	ferences	69

LIST OF FIGURES

Figure 1:1 Workflow diagram of the research phases	6
Figure 2:1 Relationships between time periods for disease transmission. The bottom patient infected fir	rst
and the second patient followed and so on	7
Figure 2:2 Susceptibility and infectivity of individuals in relation to number of vaccination induced	10
Figure 2:3 Spatial diffusion natures of infectious diseases, adopted from Thrift et al. [28]	11
Figure 2:4 Undirected and directed graph representation between actors	12
Figure 2:5 Shared memory architecture of parallel processing adopted from Bhardwaj [64]	15
Figure 2:6 Distributed memory architecture of parallel processing adopted from Bhardwaj [64]	15
Figure 3:1 Class Diagram of Enschede Pertussis Model adopted from Abdulkareem [26]	18
Figure 3:2 Flow chart representation of daily activity model adopted from Abdulkareem [26]	19
Figure 3:3 Full and partial social interaction adopted fromAbdulkareem [26]	20
Figure 3:4 Structuring metapopulation modelling technique	21
Figure 3:5 Hierarchical decomposing of temporal and spatial for the Netherlands adopted from [25]	22
Figure 3:6 Flow of subpopulation between different hierarchical models	23
Figure 3:7 Hierarchical conceptual layers of HPABM Tang and Wang [24]	25
Figure 3:8 Graphs to show connection topology of super and sub models. Squares in A and circles in E	3
represent sub model and the colour shows the sub models under same super model adopted fromTang	5
and Wang [24]	25
Figure 4:1 Decomposing of the simulation spatial scale and allocating to different processors	29
Figure 4:2 UML diagram for the components of the model	30
Figure 4:3 The interaction between individual, urban level and vaccination zone	32
Figure 4:4 Interactions between sub models	33
Figure 4:5 Buildings included in each time step	34
Figure 4:6 Three types of social networks	35
Figure 4:7 Different stages of SEIR model adopted from adopted from Perez and Dragicevic [27] with	
modification	36
Figure 4:8 Interactions between three types of agents and reporter agent	40
Figure 4:9 Functional setup of reporter	40
Figure 5:1 Netherland, Overijssel , Hengelo and Enschede	43
Figure 5:2 Distributions of two types of family over neighbourhoods of Hengelo	44
Figure 5:3 Distributions of two types of family over neighbourhoods of Enschede	44
Figure 5:4 Distribution of Children over the neighbourhoods Hengelo	45
Figure 5:5 Distribution of Children over the neighbourhoods Enschede	45
Figure 5:6 Distribution of buildings in Hengelo and Enschede	46
Figure 5:7 Spatial distributions of schools both in Hengelo and Enschede	47
Figure 6:1 The workflow of implementation phases	52
Figure 6:2 Total numbers of individuals in different age group	53
Figure 7:1 Output registration in parallel processing	58
Figure 7:2 Information collected by the processor when infection occurs	59
Figure 7:3 Distribution of the infection over different age groups	60
Figure 7:4 Shows at the different ticks in which the infection occurred	61
Figure 7:5 Numbers of infections in each tick	62
Figure 7:6 The Accumulated Number of Infection	62
Figure 7:7 Disease transmissions in different activity place	63
Figure 7:8 The progress of disease both in Enschede and Hengelo in one year period of time	64

Figure 7:9 Infections of the outgoing commuters of Hengelo in Enschede	65
Figure 7:10 Number of individuals in different age group when the original data is double	66

LIST OF TABLES

Table 2:1 List of infectious disease and their approximated incubation and infectious period [7, 31, 32	2]8
Table 3:1 Summary of existing model analysis	27
Table 4:1 Template for creating schedule for different age group	34
Table 4:2 Numeric representations of susceptibility and infectivity levels	37
Table 5:1 The approximate values of various vaccination levels for different age groups	48
Table 5:2 Approximate values of commuters	48
Table 7:1 Methods to verify the functionality of sub models	58
Table 7:2 Changes in immunity level of teenagers and adults	65
Table 7:3 Infections during sensitivity analysis	65
Table 7:4 Time elapsed by each processor	66

1. Introduction

1.1. Motivation and Problem Statements

Infectious diseases are one of the threats for the well-being and healthy life of human beings. In 2009, 1.7 million people died because of tuberculosis and around 151,700 to 575,400 people died because of influenza [1, 2]. Measles and pertussis are also among the infectious diseases that cause severe mortality of children worldwide. The World Health Organization (WHO) estimated that in 2008, about 16 million cases of pertussis occurred and from these 95% were in developing countries [3]. For 2010, Simons et al. [4] estimated around 71,200 to 447,800 deaths caused by measles of which 36% occurred in Africa. Though different treatments have been imposed, infectious disease continues to be threats of the world.

Through different factors, infectious diseases may appear locally in a particular place and can spread to different levels such as city, region, country or even continental levels. The high contagious natures of infectious diseases and people movement and interaction in both local and worldwide extent are among the factors that facilitate the spread of infectious diseases. The movement of people from one pace to another may aim at various social activities including school, workplace, shopping, entertainment, visiting relatives, etc. The outbreak of an infectious disease can be controlled within a short time or it may persists for a long period of time, causing of serious problems like death, disability, etc.[5, 6].

Recognizing the severity of infectious diseases, various measurements have been under taken to prevent and control outbreaks. Some countries control import of animal products, restrict people movement from highly infected society to another, frequently check the health of citizens, periodically disseminate vaccination, etc. The Netherlands is one of the countries which conduct frequent check up on immigrants for tuberculosis and yellow fever. Besides imposing society oriented control methods, researchers proposed various techniques to understand the outbreak nature of epidemics and to evaluate control and prevention methods before being implemented [7].

Modelling is one of the techniques used to acquire understanding on the mechanisms underlying the spread of infectious diseases and to asses appropriate prevent and control methods. For better understanding, the contributions of models for disease diffusion can be explained in two ways: before and during the occurrence of the outbreak. Before the occurrence, models can be used to predict the spatial and temporal scales, identify society in high risk areas, forecast the expected severity level and related details of disease outbreaks by considering natural and manmade factors [8, 9]. During the outbreak, models can be used to evaluate the prevention and control methods before being implemented [7]. In general, models allow epidemiological researchers and policy makers to do "what-if" analyses with the purpose of understanding and assessing the disease outbreak behaviour under various conditions [10-12].

Epidemics can be modelled using different modelling techniques. SIR (Susceptible, Infected and Recovered) model is one of the oldest models that used to model the outbreak of epidemic by classify the society into three distinct groups called susceptible, infected and recovered[13]. SIR model uses differentiation equations to illustrate continues change in number of people in each groups. Contrary to its capacity to represent epidemic outbreak, SIR model fails to describe heterogeneity of individuals. SIR model assumes that all individuals have identical rates of disease causing contact [14]. In another words, SIR model not represents the diversified characteristic of individuals both in immunity level and social

interaction levels. To overcome the limitations of SIR model, several alternative models are developed and among them social network and agent based modelling (ABM) are the most widely used modelling approaches [15-17].

Social network modelling is type of modelling technique used to analyse the impact of social interaction on disease spread. It uses graphic representation concept where nodes are individuals and edges are the relationships between individuals[18]. Obviously, since a given person has not equal probability to meet all of people, some group of people like family, friends etc. are created. Though social network modelling assists interaction of defined groups, it cannot take into account the casual contacts like the contact that happen in shopping centres, public transportation etc.[19]. Agent based modelling is type of modelling method can handle the limitations of social network modelling.

Agent-based modelling (ABM) is a powerful modelling technique with ability to represent complex systems with detailed data representation of the system entities. Basic principles of ABM is that each individual is behaviourally and physiologically distinct because of environmental influence and they are influenced by nearby individuals[20]. When each individual is represented by its behaviour and the relationship with other individuals, realistic and complex phenomena can be represented. Agent based models consist of a population of individuals called agents, an environment, and a set of rules that control the interaction between agents and interaction of agents with the environment[15]. Modelling epidemics using an agent-based approach pursues the progression of a disease through each individual (consequently highly heterogeneous population can be represented), and tracks the contacts of each individual with others in the relevant social interaction and geographical areas.

Agent-based models represent a system in details that involves the description of each agent with respective behaviour, and their interaction with the environments. To represent details of individuals in a simulation, consumes significant amount of memory and time. Especially when the complexity of the system increases, the number as well as diversity of both agents and environments increases; it becomes an extremely computationally intensive process[21]. Although computing power is increasing rapidly, the high computational requirement of ABMs remains a limitation when modelling large systems.

To solve the high computational requirement of ABM, implementing hierarchy based parallel computing is recommended by several researchers [8, 22-24]. The hierarchy based parallel computing methods decompose an agent-based model into a set of sub-levels (comprised of agents and their spatially-explicit environments) and super levels of hierarchy. Sub levels function as computational units for parallel computing and are aggregated into a group of super levels that represent computing tasks. Applying hierarchy based parallel computing method allows ABM to efficiently utilize the computational resources (e.g. processing power, memory, and input/output capacity) [8, 24, 25].

By adopting different levels of innovation for infectious disease modelling, in this research a framework that uses the concepts of generic infectious disease modelling and agent based modelling is developed. To make the framework generic and scalable, existing models designed by Girmay [25], Abdulkareem [26], and Tang and Wang [24]are analysed and part of the methods are adopted. Additionally, to address computational problem of ABM, parallel computing software is developed using a programing language that can support simulation of large scale complex problem.

1.2. Research Identification

1.2.1. Research Objective

The main objective of this research is to develop a generic and scalable framework for spatially explicit ABM that supports simulation of the outbreak of infectious diseases. In this research, framework stands for a software developed using a particular programing language and generic and scalability indicate the capacity of the framework that simulates the outbreak of all types of infectious diseases in different spatial extent (e.g. town, city, country etc.). To achieve the main objective the following sub objectives are addressed.

- Understand characteristics and factors facilitate the outbreak of infectious diseases.
- Examine existing models.
- Determine suitable method to achieve scalability
- Design conceptual model
- Implement a framework using programming language
- Evaluate the generic and scalability of the framework

1.2.2. Research Questions

Sub objectives will be answered based on the respective questions mentioned below.

- 1. Understand characteristics and factors that facilitate the outbreak of infectious diseases
 - 1.1 What are the characteristics of infectious diseases that impact the outbreak?
 - 1.2 What are society related factors and how do these factors accelerate the outbreak of infectious diseases?
- 2. Examine existing models
 - 2.1 What are the methods implemented in existing models?
 - 2.2 What are strengths and limitations of existing models?
 - 2.3 Which components of the models are relevant for the model of this research?
- 3. Determine a suitable method/s to achieve scalability
 - 3.1 What is/are suitable approach/approaches to achieve scalability?
 - 3.2 How to implement scalability approach in ABM?
- 4. Design conceptual model
 - 4.1 What is the conceptual representation of the influences of disease and society related factors on the outbreak?
 - 4.2 How to represent scalability of spatial extent?
- 5. Implement a framework using programming language
 - 5.1 What is the appropriate programing language to design scalable ABM?
 - 5.2 How to represent scalable ABM in a framework?
- 6. Evaluate scalability of the framework
 - 6.1 Does the model represent the correct number of population and their spatially location?
 - 6.2 What spatial extent and number of population can be supported by a framework without affecting the performance of the model?

1.2.3. Innovation Aimed At

Emphasizing on boundary less spreading nature of infectious disease and the resource demanding nature of agent based model, this research aims at developing a scalable and generic framework. *Scalability* is to show capacity of the framework to support simulation of the outbreak of infectious diseases either in large or small area. *Generic-ness* indicates that the framework is not restricted to simulate specific types of infectious diseases. In general, the framework supports simulation of all types of infectious disease in different spatial extents.

1.2.4. Related Work

Because of its destructive nature and long history in societies, different researchers designed various types of models to understand and analyze the epidemic outbreaks. One of the most famous models is the stochastic model introduced by Kermack and McKendrick [13]. This stochastic model presents the general concepts of infectious disease transmission by using differential equations on distinct population groups called suspected, infected and recovered. Perez and Dragicevic [27] developed a contagious diseases spread agent based model by grouping human activities into stationary (school, work place and home) and mobile (transportation system) places. They implemented two algorithms that are responsible to control diseases transmission and agent behaviors such as interaction with other people, age etc. Borkowski et al. [22]designed a generic ABM application called discrete space scheduled walker (DSSW). DSSW allows the users to supply basic input parameters such as where (for group communication), who (agents), when (weekday or weekend), what (type of disease) and transportation line of a specific city to simulate the epidemic disease dynamics. Considering agent activities as individual and group activities that can be conducted either in weekday or weekday Abdulkareem [26] also developed an agent based model that simulates spread of Pertussis in Enschede, Netherlands.

Though the ABM is flexible and widely accepted, as the complexity of a given system increases, because of high computational resources demanding, ABM fails to support the simulation of such complex systems. To address this problem Tang and Wang [24] developed a generic framework called hierarchical parallel simulation framework for spatially-explicit agent based models (HPABM). The main logic behind HPABM is dividing the entire system into super and sub levels of hierarchy and implementing super levels in different processors to solve the computational delay of the model. Perumalla and Seal [8] designed a parallel discrete-event execution model that explains parallel execution of infectious disease models using a number of processors which are connected physically or logically using networks. To simulate the outbreak of pertussis in The Netherlands, Girmay [25] extended the model designed by Abdulkareem [26] using hierarchy modeling and metapopulation techniques. Girmay [25] divided the Netherlands into regions, municipalities and individual level and implement a sub model called commuting model disease model that controls agent movement from one hierarchy level to another and disease transmission among them.

By adopting different levels of innovation for infectious disease modeling, in this research a framework that uses the concepts of generic infectious disease modeling and agent based modeling is developed. To make the framework generic and scalable, existing method proposed by Girmay [25], Abdulkareem [26], and Tang and Wang [24] are used.

1.3. Project Setup

This section describes the work flow conducted to achieve the main objective of the research. The undertaken phases are classified into three main phases which include (1) knowledge building (2), developing of a conceptual model and lastly (3) implementation and evaluation of the framework. Each phase in turn consists of sub phases (Figure 1:1). The accomplished activities and outputs of each phase are discussed as follows.

Knowledge Building

This phase is aimed on acquiring knowledge from different literatures and existing models. The main focus is to understand:

- The characteristics of infectious diseases.
- Various modeling techniques (e.g. ABM)
- Suitable techniques to implement scalability in ABM.
- Programming language for implementations of scalable ABM
- Understand and evaluate existing models

Designing Conceptual Model

Using knowledge acquired, the conceptual model that describes the fundamental characteristics of generic and scalable infectious disease simulating ABM is designed. The model emphasize on three basic issues of the model. The firs explained theme is the basic characteristics of the ABM components including agent, environment, and interactions from the perspective of infectious disease simulation ABM; and followed by the processes that integrate the components to represent realist occurrences of the disease. Lastly, it explained the scalable representation of the model using parallel processing method.

Implementation and Evaluation Phase

This phase focuses on developing a framework based on the conceptual model and evaluating it using the real datasets. The implementation is using repastHPC modelling toolkit and to test the framework two cities of The Netherland called Hengelo and Enschede is used. The four datasets used for the evaluation are population census data, vaccination data, spatial data, and commuting data.

1.4. Overview of the Following Chapters

This section is to provide the general impression about chapters included in this document. The document consists of eight (including this chapter) chapters in which each of them have their own enlightenment and also integrated to the rest of the chapters. The broad presentations of the chapters are as follows:

Chapter two presents the literature reviews on the fundamental adopted concepts. The chapter starts by explaining the characteristics of epidemics that contribute the outbreak process; and it is follow by natural and manmade factors that facilitate disease spreading. The chapter also describes the implemented modelling techniques including agent based modelling, classic disease model, and social network analysis. Lastly, parallel processing techniques which used as a method to achieve scalability is presented.

Chapter three empathize on the analysing and evaluating existing model. Three exiting models are selected and the strength and weakness are identified. Chapter four presents the conceptual model of this research. The chapter start by describing the basic components of the model with their respective characteristics and sub model to integrate the components. The scalable representing of the model is also presented.

Chapter five describes the data processing phase which describes the basic datasets used for framework evaluation with their type and source they have taken. Chapter six follows by explaining the

implementation phase. Chapter six starts by describing background concept about the agent based model toolkits and emphases on repastHPC; the modelling toolkit used for implementation of the framework. The implementation chapter proceed with result and discussion chapter. This chapter explains the outputs of implementation and respective description. The last chapter is conclusion and recommendation chapter.



Figure 1:1 Workflow diagram of the research phases

2. Background Concepts

2.1. Characteristics of Epidemics (Infectious Diseases)

An uncommon large scale occurrence of an infectious disease in a community is called an epidemic [7]. An epidemic may originate from the community itself or can be introduced to the community from other places[28]. There are three essential requirements for an outbreak of infectious diseases which are: the presence of an infected person, an adequate number of susceptible people and an effective method of contact and transmission method[29]. The factor that activates the transmission of the infectious disease is the exposure of the susceptible person to infectious person. Different infectious diseases seek different levels of exposure extents. For instance, the transmission of smallpox needs very close contact between the susceptible and infectious persons, while measles, pertussis, and influenza require conversational proximity between people because these diseases diffuse through airborne droplets [10, 30].

When there is effective contact between an infectious and susceptible person, the initial outcome is the transmission of pathogenic (micro parasites) between them. After infiltration into the body of susceptible person, micro parasites may not manifest any symptoms until they invade most body parts of the person. Commonly, symptoms appear when most of the body parts are attacked by micro parasites. The time elapsed from exposure to the parasite and the appearance of the symptoms is called the *incubation period* [29]. Afterwards, since parasites reproduce the person has capability to transmit the disease to other susceptible individuals. For a couple of days, the person continues to infect those who have no immunity and can contact the person physical in contiguity distance. The time elapse when a given person has the capability of transmitting disease to susceptible individuals is called *infectious period*. Individuals who were infectious at some point can recover by developing natural immunity system to a particular infectious disease (Figure 2:1). Depending on the type of infectious disease, immunity induced naturally may persist for lifelong or may wane and the person may become susceptible again.



Infection

Figure 2:1 Relationships between time periods for disease transmission. The bottom patient infected first and the second patient followed and so on

Various infectious diseases have different time interval for the incubation and the infectious period [5, 7] (Table 2:1). In most cases, infected persons do not show symptoms during the incubation period. The symptoms start to appear after the incubation period when the person becomes infectious[17, 22]. The symptom levels can be categorized into first and second level symptoms. First level symptoms are common to many infectious diseases, which makes it difficult to differentiate the type of disease. The second level symptoms occur around the mid time of the infectious period and it is easier to distinguish the type of infectious disease. Second level symptom stage is also the period that a person could feel ill

and stay at home. This period is also the time for a person to contact health organization for medical treatment. In this stage the government has a chance to be informed about the disease either from health organizations or from those who see the major symptoms and inform the concerned body like municipality. From such information the government can decide and impose required prevention and/or control methods to the society.

Infectious diseases	Incubation period in days	Duration of infectious in days
Bordetella Pertussis	7-10	21-23
Measles	9-12	5-7
Smallpox	12-14	10
Rubella	17-20	14
Chicken pox	13-17	20-30

Table 2:1 List of infectious disease and their approximated incubation and infectious period [7, 31, 32]

2.2. Major Factors Related to the Outbreak of Infectious Diseases

Because of its contagious nature, infectious diseases can emerge in a given community and propagate to country level or even continents. The propagation of infectious diseases can be facilitated by both disease related factors (characteristics of infectious micro parasites, immunity level of a person etc.) and community related factors (social interactions, people mobility nature, cultural and geography) [5-7, 33, 34]. Among community related factors, social interactions and people mobility are considered as the main factors and from disease related factors immunity is an essential aspect.

2.2.1. Social Interaction

Social interaction states a particular forms of interaction in which the actions of a reference group (two or more individuals) affect an individual's preferences[35]. The reference group can be an individual's family, neighbours, peer, classmates, workers in office, friends, people in shopping places etc. Current social media technologies contribute easy and fast techniques for social interaction[36, 37]. People who stay far away from each other can interact through phone, email, Facebook etc. While physically adjacent individuals can interact face-to-face or closer like skin-to-skin contact.

Understanding the patterns of social contacts is a crucial determinant to analyse the spread of infectious diseases [30, 38]. Obviously contacts through telephone or email have no consequence on infectious disease transmission. The main type of interaction that facilitates the outbreak of an infectious disease is interaction that occurs between adjacent people in proximity distance. As number of adjacent contact increases for a particular person there are two major roles that can be played by the person. Firstly, the individual is at greater risk to being infected and, once infected, can transmit the disease to many others[39].

According to Read et al. [30], to analyse the outbreak of infectious disease through social interaction, it is not sufficient only to measure the number of contacts made by individuals. Understanding how often (how regular) each encounter is conducted during the infectious period and the time elapsed during interaction period are the key attributes needed to be considered. The nature and regularity of interaction between individuals is affected by many factors including culture, religion, age, intimacy and social context, and gender. For instance interactions among family members has different level and rate compared to the interactions occur between friends. Contacts occurring at home tend to be more stable and intimate while contact encounters in the workplace are predominantly irregular[30]. Furthermore, children mostly stay intimate to with children of their own age and adults are characterized by frequent interaction compared to children.

In general, understanding the contact patterns of individual is crucial for building a computational model of infectious disease outbreak. The important characteristics of the contact pattern that need to be emphasized are: (i) the topological structure of the contact place, (ii) number of contacts of a person, (iii) the clustering and presence of well-identified communities of people and (iv.) the frequency and duration of contacts. While social interactions yield local outbreaks and spread of a disease in single populations, multi scale human mobility is responsible for large scale spatial propagation[40]

2.2.2. Commuting

Commuting is repeated traveling of people between locations which mostly occur from residence to place of work or full-time study (school). In other words, individuals usually visit a limited number of places and predominantly they commute between home and work locations and possibly a few other locations[40, 41]. Mobility flows in very complex multi scale networks across several orders in intensity and spatiotemporal scales. Mobility can range from the long range intercontinental movement to the short range commuting within the society[42].

People mobility may initiate due to many factors and the commonly mentioned reason is the growth and expansion of cities or urbanization. Work opportunity, better life standard, and growth of the transportation, the establishment of businesses, industries and educational institutions in urban places are main features to attract people from other places. Despite its attractive features, mostly urban places have less housing facilities and those facilities are expensive. To balance the life system, people rent houses in nearby cities and conduct their daily activities in urban areas.

Mobility patterns can impact the outbreak of epidemics and it's among the major ingredients to be considered in the agent based modelling of infectious diseases [41, 42]. The reason to consider mobility is that it can create an opportunity for the interaction of infectious and susceptible people who live in different places[43, 44]. Traveling people can be susceptible or infectious and when they meet susceptible or infectious people at any social interactions places, they can infect or be infected by infectious disease which in turn leads to further spread of the disease.

2.2.3. Immunity

The human immunity system is a mechanism for defending against infections of pathogenic from causing the disease related consequences on human body. Immunity may result from vaccination or naturally induced after infection and recovery of a given disease. Immunity can affect both susceptibility (the state of likely to be infected) and infectivity (capability to infect the others) levels of a person depending on vaccination doses used [45].

Infectious diseases have various doses of vaccine and multiple doses are effective since each additional dose of vaccine seems to provide additional protection against the disease[45, 46]. As described in Figure 2:2, a given person who does not take vaccination (V) is fully susceptible (S) and highly infective (I_{highest}) in its infectious stage. When the person takes one dose of vaccine, though low immunity level is gained, the person still continues to be susceptible and infective; but with less susceptibility and infectivity level compared to those who did not get vaccinated. From the figure it is also clear that as the number of doses increases, the person gains more immunity and both susceptibility and infectivity decreases. The influence of the vaccination dose on the infectivity and susceptibility level of a person is mainly because each dose is capability to providing on additional level of immunity and the symptoms that cause the disease transmission such as coughing, sneezing etc. are decrease [45]. When the person is fully vaccinated, it becomes fully immune so that can be represented as recovered people. Though immunity can be gained

by vaccination it is also possible that vaccination-induced immunity can wane and lost through time and the person can be susceptible again (Figure 2:2) [47].



Figure 2:2 Susceptibility and infectivity of individuals in relation to number of vaccination induced

2.3. Spreading Nature of Epidemics

According to Thrift et al. [28]explanation, diseases spread in different forms including expansion, relocation or combination of both. Relocation is a type of diffusion which occurs when a disease appears in a given society and after sometime leaves an original place and moves to another place. On another hand, expansion diffusion refers to persistence and intensification of a disease in the original place and spread to nearby location through time. Expansion is a form of diffusion that can be described in two forms called contagious and hierarchical forms of diffusion (Figure 2:3). As illustrated in (Figure 2:3(c)) if we consider boxes represent distribution of people, contagious diffusion depend on close interaction between infectious and susceptible people. Increasing the distance between infectious and susceptible individuals can decrease the transmission probability of the infectious disease.

According to contagious diffusion, nearby society of the original area are affected before those far away. While hierarchical diffusion (Figure 2:3 (d)) explains transmission of disease from one place to another in an ordered sequence. For example, an infectious disease can propagate from crowded areas like cities to less crowed areas such as villages through commuting people without expanding to nearby cities. In most cases, the outbreak of an infectious disease has a hierarchical method of diffusion.



Figure 2:3 Spatial diffusion natures of infectious diseases, adopted from Thrift et al. [28]

2.4. Modeling Techniques

Modelling is a tool that helps to understand and analyse the real world objects and phenomena. In broad classification, modelling can provide two major benefits. (1)It assists the study and understanding of certain features of the real life systems and (2) enables control of objects or phenomena through facilitating the prediction of behavioural change under different conditions[48]. An epidemic outbreak is one of the real world phenomena that can be described using a model. Epidemic models mainly aim at (1) giving insight to understand the dynamic and complex spreading nature of epidemics through describing characteristics of infectious diseases including how they are transmitted, factors affecting the rate of transmission etc. [49] (2) Used to estimate and evaluate the prevention and/or control methods such as medical services, quarantine before being implemented in the society [7, 19, 49]. To utilize the advantage of modelling, various modelling techniques are proposed by researchers. Classic epidemic models, social network and agent based modelling are among the modelling techniques used to model epidemic outbreaks.

2.4.1. Classic Epidemic Models

Considering the common characteristics of an infectious disease as basic argumentation, researchers developed different epidemic models that can describe the diffusion nature of the diseases. One of the most famous models is the SIR model introduced by Kermack and McKendrick [13]. This model presents the general concepts of infectious disease transmission by using differential equations on distinct population groups called Suspected (S), Infected (I) and recovered (R).

Though the SIR model is able to capture most of the features of the epidemic processes, it does not consider the incubation period [49]. In other words, in the SIR model, individuals become infectious as soon as they are infected [50]. To model the incubation period, the SIR model is extended to a SIER (Susceptible, Exposed, Infectious and Recovered) model that embraces an exposed state of a person. In the SEIR model, initially susceptible individuals are considered as being exposed to the disease and the exposed individual incubates the disease for some times depending on the type of infectious disease. Even though exposed people have parasites inside their body, since these parasites are not abundant enough and still growing, those individuals have no ability to transmit the disease [5, 27, 33, 49]. After the incubation period, the status of exposed people changes into an infectious state. The infectious state is when people can transmit the disease to susceptible individuals. After a couple of days, infectious people loss their

ability to transmit a disease and are no longer infectious. After the infectious period the person gains immunity by natural interaction of the body to the disease.

Classic epidemic models are not restricted to SIR and SEIR model. Based on the aim of modelling, researchers can produce other types of classic models by remixing different groups of individuals. For example, to represent diseases whose infection does not confer immunity, the SIS (Susceptible, Infectious, Susceptible) model can be used[7]. SI, SIRS, SEIRS are also some type of epidemic models applied by different researchers. For this research, the SEIR model is chosen.

Despites their advantages on modelling epidemics, classic models neglects basic society related disease casing factors. The model assumes that population are "fully mixed" that infectious individuals have equally likelihood to spread the disease to other member group for which they belong[18]. The model does not consider the heterogeneity of individuals and social structure of the society so that understanding the spreading pattern of a given disease not possible. Furthermore classic model also assumes that all members of the community are equally susceptible to a given disease and complete immunity is conferred after the recovery [51-53]. Social network analysis and agent based model are types of modelling techniques to address the limitations of classic epidemic model and describe in succeeding sections.

2.4.2. Social Network

The spreading pattern of epidemics is determined not only by the properties of pathogen and immunity level of individuals. Social structure of population and the contact pattern between infectious and susceptible individuals also should be considered to model the outbreak of epidemics [54]. Social network analysis (SNA) is a technique used to understand and explain the effect of social structure on the spreading nature of diseases[55, 56]. SNA uses graph-theoretic which identifying nodes (actors) and the one or more types of edges (relations) between nodes to explain the social structure [56].

To analyze the effect of actors on each other, SNA provides attention to the nature of the relationship, particularly properties of being symmetric or asymmetric [39]. Symmetric nature describes whether a relationship between actors A and B implies a relationship between B and A while asymmetric is to explain actor A implies C but not vice versa or another way around. Symmetry and asymmetric relationships can be represented using undirected and directed graph respectively (Figure 2:4). In the case of an undirected graph, the content of the edge (relationship) flows in both directions while for a directed graph, the flow is only in one direction[56]. To choose the type of edge understanding the context is important.



Figure 2:4 Undirected and directed graph representation between actors

Social network modelling has been used to investigate the impact of social structures on disease outbreak. The major reason is the transmission nature of infectious disease which requires contact for a new infection to occur. However, social network modelling has two limitations. For large scale social network modelling it is difficult to obtain the social interaction of individuals from real data and more importantly, since it is based on social structures, representing casual contacts that happen in crowded areas like public transports, shopping centres etc. is difficult[19]. Furthermore, for computing type of modelling like ABM, producing the relationships between actors and storing them in the memory of computer demand a huge amount of time as well as memory space. According to Perrin and Ohsaki [19] single million-node network would take more than one day to generate, and a network ten times larger would take over four months. As the number of nodes to be connected increases it becomes clearly not practical.

On the contrary, agent based modelling approaches are suited to model emergent activities that happen in crowd area. In addition, to solve the memory demanding nature of network model, Perrin and Ohsaki [19] suggested parallel processing which deal on the generation and linkage of smaller sub networks to produce the entire network[57].

2.4.3. Agent Based Modelling

Agent based modeling (ABM) is a powerful modeling technique that provides an environment which can assist in the investigation of dynamics in complex systems [20, 24]. Agent-based models are capable of representing heterogeneous, randomness and irreducible interactions of very complicated systems[15]. The significant features of ABM that enable the investigation of complex systems are the agent (autonomous unit capable to make independent decision), the environment (the virtual world in which the agents act) and the interaction that occurs among agents and with the environment. As Andrew T. Crooks and Castle [58] explained, those substantial features provide ABMs with two basic characteristics: (1) it can capture emergent phenomena (unexpected and logically independent properties of the system) and (2) it provides a natural environment for the study of the systems.

ABM has a flexible computational platform that allows integration of ABM with other analytical and modeling approaches like geographic information system (GIS)[24]. GIS is a technology which has powerful ability for storing and retrieving spatial reference data. The integration of GIS and ABM facilitates the spatially explicit representation of referenced characteristics of both agents and the environments [20].

Being the cumulative outcomes of human interaction, the outbreak and spread of infectious diseases is a geographically complex system that can infect millions of people without being restricted to a boundary or a specific place. Agent based models are capable of considering the heterogeneity of individuals, environment and the stochastic essence of infectious disease transmission for infectious disease simulation. ABM also allows examining the temporal and spatial aspects of disease diffusion by facilitating tracking of individual's contact processes as well as people who are in contact with the person [59]. Furthermore, ABM signifies components of the real system and keeps track of individual behaviors over time[60].

Despite its advantages, to simulate complex geospatial processes, ABM is often computationally demanding and leads to a series of computational issues e.g. computing time and memory constraints[61]. Accordingly, ABM heavily relies on advances in computational science to hold these issues. Recently, the increasing availability of multi-core computer systems and the emergence of cyber infrastructure provide a considerable amount of computational resources that could enhance ABM and facilitate the resolution of the computational issues of ABM[24].

2.5. Parallel Processing

An epidemic outbreak is a complex process that results from the cumulative effects of many factors such as long and short distance movement of people, complex and nonlinear social interaction of people, transmissibility nature of infectious disease, etc. Representation of such complex processes using computational models like agent based models demand large amounts of resources like computer memory, simulation time etc. Commonly, simulating complex systems using only one standalone computer decreases the performance of the model as the number of entities of the complex system increases. To design a model not affected by the negative outcomes of the complex factors of an epidemic outbreak, various approached have been proposed. Scalable way of designing models is suggested by number of researchers.

In generic terms, scalability can be defined as the capability of a solution to a problem to work, when the size of the problem increases[21]. According to Rana and Stout [21] agent based modelling needs to be scalable when the total number of agents involved increases, when the size of the data (rules) the agents are operating on increases, and when the diversity of agents and complexity of the environment increases. Parallel computing is among the techniques suggested to achieve scalability of ABM.

Computers performance is continual improving to satisfy the demand for greater computational speed in multidisciplinary areas such as modelling motion of astronomical bodies, global weather forecasting, simulations of the large scale epidemic outbreak, etc. The improvement of computer may immanent from the faster hardware, efficient algorithms, processors operating speed, etc. According to Grama [62], the maximum efficiency level of those components could be achieved but still may continue to not fulfil the computational demanding natures of some disciplinary areas. One of the methods proposed to solve such problem is using multiple processors (CPU) to solve a particular task. In other words, imposing parallel computing processing elements that communicate and co-operate to solve a given problem in time period and accuracy[62].

Parallel computing is type of computation that decompose specific problem into small number of tasks that can be executed independently and communicate with each other to solve a large problem[63]. In this sense the main ideology behind parallel computing is divide- and- conquer technique that implement divided tasks in each of sub task in different processors and conquer the result to represent the entire problem.

Parallel computing environment has architectural paradigms to layout processors and memory they access. The two principal types architectures are shared memory multiprocessor and distributed memory multicomputer.

As the name indicates, shared memory architecture has global memory that accessed by all of the processors available in the system[62, 64]. Global memory of this system is place for all processors to access, modify and store the data they work on. Furthermore, global memory also facilitates the communication between processors. Shared memory architecture is easy to build as well as program. Since data communication conducted through global memory, it also does not have communication overhead when message transmission happens between processors. On another hand, shared memory architecture is limited to support scalability. The reason to not be scalable is that when number of processors added to the system, it introduces memory contention and more memory accessing traffic occur.



Figure 2:5 Shared memory architecture of parallel processing adopted from Bhardwaj [64]

The second type of parallel computing architecture is distributed memory system. In the case of distributed memory system, each processor has its own local memory to conduct task and data sharing with the rest of processors is through explicit internetwork connection[64, 65]. The network can be configured in different network topology such as tree, mesh, bus, star, etc. The main advantage of distributed memory system is memory size scales as number of CPU increase and each CPU has fast access to local memory without contention form another CPU[62]. Beside the scalability nature, this system is slightly difficult for programing and requires implementation of message passing technology to facilitate the communication between processors[66].



Figure 2:6 Distributed memory architecture of parallel processing adopted from Bhardwaj [64]

Besides specifying the architecture, parallel computing has additional issues need to be considered. Partitioning of data across the processor, mapping of data onto the processors, synchronization, scalability are some of concerns. The main concern of those issues is to distribute the tasks to processors through balancing the load and harmonize their output so that the processors are efficiently utilized and extensible when it is required[64]. Different process decomposition and mapping methods can be used to yield good performance on different processors for a given problem.

Data parallelism and process parallelism are the two prominent methods used to decompose a given problem in to sub tasks and distribute to processors. The data parallelism method divides a given dataset into number sub data groups and each of sub data group is executed on different processor using similar instruction. In another words, the implemented of same operations simultaneously on the various elements of a particular dataset is called data parallelism[64]. Commonly, data parallelism technique is acquired when the size of the problem is big. On another hand, process parallelism is when a particular process is combination of other several and diverse operations and each operation is executed on different processors. The outputs of the processors should be synchronized to generate the entire process[64, 65]. Process parallelism can be visualized using task graph where the nodes of the graph represent processes to be executed and edges stand for dependencies between the Processes. In this context dependency indicates an execution of a given task depending on all of the completion of preceding tasks.

Simulating the large-scale geospatial problems like infectious disease outbreak using ABM is computationally intensive process. Parallel computing technique is one of the methods used to design scalable infectious disease simulation ABM. Parallel computing is a technique suggested to solve the limitations of ABM using standalone computer on disease model simulation process.

3. EVALUATION OF EXISTING MODELS

Disease simulation models have been developed by a number of researchers [7, 17, 19, 24-27, 30, 67-70]. To design the framework of this research, three previous works are chosen to be evaluated. The main criterion to select these three existing models is the focus area of the models. The first model designed by Abdulkareem [26] has a detailed representation of individual's daily activities. The model uses a synthetic population generation method for the accurate representation of the population for whom the model is developed. In addition to this, the model is well implemented and tested for its accuracy. The second model, developed by Girmay [25], uses hierarchical modelling concepts to increase the scalability of the model developed by Abdulkareem [26], so that this model can provide understanding on developing scalable models. The last model is designed by Tang and Wang [24]. This model is a generic framework which intended to assist development of any types of scalable, spatially explicit ABM.

The evaluation was aimed at identify strengths, weaknesses; opportunities and threats (SWOT) of the existing models. This research adopts some of the basic idea of these models and extends to design a generic framework that enables simulation of the outbreak of any types of infectious diseases at various spatial scales. The sections of this chapter are organized as follows: section 3.1 describes the model developed by Abdulkareem [26], section 3.2 explains the model developed by Girmay [25] and section 3.3 enlightens the model developed by Tang and Wang [24] Section 3.4 summarizes the results of the SWOT analyses and compares the results.

3.1. Simulating the Spread of Pertussis in Enschede Region Using Agent-Based Modeling

This model was designed by Abdulkareem [26] to simulate the spread of pertussis in Enschede a city in the Netherlands. The model is based on an individual space-time activity-based model (ISTAM) which was designed by Yang and Atkinson [59]. ISTAM model was designed to simulate the transmission of infectious disease in Eemnes (the Netherlands). ISTAM used the concept of activity bundles (ABs) to represent interaction between people at a specific location. The underlying assumption of ABs is that being at the same xy location does not automatically mean people interact and are at risk of infection. Based on such assumption various level of people interaction were developed.

Abdulkareem [26] adopted the concept of AB from ISTAM and integrated it with the concepts of synthetic population generation technique and human activity pattern to model the spread of pertussis in Enschede, the Netherlands. Human activity pattern is a method that describes at what locations, at what times and how people pursue scheduled activities[71]. Synthetic population generation is a technique to create a population (agents) for the model based on empirical aggregated statistical information, using different datasets. The population resulting from the synthetic population technique is expected to be as precise and accurate as possible.

A conceptual model which incorporates the concept of activity bundles (AB), synthetic population and human activity patterns, was designed to describe the spread of pertussis in Enschede. The conceptual model is described two major stages of simulation. The first stage is the initialization and setup of the model and the second stage is the actual simulation.

The initialization and setup stage deal with building the population/agents with their required attributes, such as age, vaccination level, residence, workplace etc. and loading the environments (GIS) for the agent interactions. The population is generated using a synthetic population generation technique. The technique

used the census data of Enschede that contains information about the distribution of population in 69 city blocks of Enschede, spatial dataset that contains different types of buildings and vaccination data. Considering pertussis as a childhood disease, the population extracted from the census data for each blocks of the city includes the number of families (family with both parents or only mother or father) and the number of children per age groups (0 - 4, 5 - 9, 10 - 14 and 15 - 19). During the creation of agent, agents are assigned with attributes include age, family ID, and work status vaccination level, infection status, work status etc. Attributes distinguish a specific agent from the rest of the agents. To facilitate geographically explicit interaction of people, Enschede's spatial dataset that contains buildings (residential, school, workplace, etc.) is used. The building's attributes include building type (house, a school, a workplace etc.) and building address. Agents are assigned with type and address of buildings to assist them in identifying their activity place.



Figure 3:1 Class Diagram of Enschede Pertussis Model adopted from Abdulkareem [26]

After the agents are created and spatially distributed, they have to achieve their daily and weekly activities and interact with the other agents. Times of interaction are the moments that spread of disease may occur, because during activities infectious individuals and susceptible individuals meet. The simulation is in charge of facilitating agent interactions and disease spread among agents. To achieve its task the simulation consists of three sub models called daily activity model, social interaction model and disease model.

3.1.1. Daily Activity Model

The daily activity model is responsible for creating the activity pattern of agents. An agent has a weekday schedule (static) and a weekend schedule (dynamic). The main aim of scheduling is to allow multiple interaction of a given person with different groups of people through providing various activities in different place within a specified time. Abdulkareem divides human activities into two groups, individual activities and group activities. Individual activities are activities conducted by a single person (at a specific time) while group activities are carried out by two or more individuals. For instance, in the weekend, all family members can be part of the holy day activities engaging with other family groups. In the model, during weekdays, only individual activities are scheduled and in weekends, agents perform group activities. Activities for the complete week are stored in the activity table of the agent, and are copied (replicated) every following week.



Figure 3:2 Flow chart representation of daily activity model adopted from Abdulkareem [26]

3.1.2. Social Interaction Model

The social interaction model is in charge of managing the interaction levels between individuals that are in the same activity at the same time. The model assumes that in activity places, individuals may have full or partial social interaction. Full interaction occurs when individuals within a group have direct contact with the entire group members (Ex. family) and a partial interaction is when individuals within a group have contact with some of the group members but not all of them (Ex. student in a class group) (Figure 3:3). The level of interaction has a direct impact on the disease transmission. People with full interaction have a higher infectious probability compared to partially interacting people. Before disease transmission is checked by the disease model, the social interaction model identifies people's interaction levels.



Figure 3:3 Full and partial social interaction adopted fromAbdulkareem [26]

3.1.3. Disease Model

The disease model controls the transmission of the disease among individuals when they meet at activity places. The immunity level of individuals, distance, interaction level and duration of the contact between susceptible and infectious people are major attributes considered for the disease transmission. The disease model allows transmission when the distance between infectious and susceptible individuals is less than or equal to 1 meter; and the contact type is adjacent. The model also measures the duration of the contacts between individuals. When the duration is more than or equal to 1 hour, then the transfer of the disease is executed. If the contact duration is shorter, the frequency of contacts per day is computed at the end of the day. If the sum of the total time is equal to or more than one hour then the susceptible individual becomes infected. For partial interaction, if people are in the same activity place in close proximity distance for a period of time (1 hour or more), they get the disease.

3.1.4. Strength and Limitations of this Model

This model has strong points that can be adopted for the model of this research. First of all, the model tried to extract a population of the society with their respective detail information in a realistic and accurate manner using a synthetic population generation approach. The model also has individual as well as group activities that reflect realistic society interactions. The use of AB (activity bundle) enabled the model to reflect the detailed interaction level of individuals in a social interaction place. Furthermore, the model is fully implemented using Repast Symphony (agent based modelling and simulating toolkit) and tested for its accuracy based on the statistic data available.

The model also has some limitations. As the number of agents to be simulated increases, the performance of the model decreases. The reason for decreasing in performance is that it is not designed to support scalability and was implemented using a single computer. As described in chapter two, when number as well as diversity of agents and environment increases, because of ABM computationally resource demanding nature, simulating large scale system using standalone computer affect the performance. This model is implemented using Repast Symphony to be executed on a standalone computer which is not scalable. In addition to this, the temporal resolution of the model is also too detailed. Every simulation tick stand for 30 minutes of real time. Even though people do not change their activity places, or in cases without an infectious person in a group, it will still check the status of all agents during every tick. Furthermore, the model is a closed system; it fails to consider incoming and outgoing commuters from other cities. Since the disease can be introduced from another area through commuting people, not considering the commuters reduces the realistic ness of the model. Lastly, the model also does not represent vaccination or other interruptions to stop the diseases outbreak.

3.2. Performance and Scalability of Geographically-Explicit Agent-Based Disease Diffusion M odels

One of the models that is being reviewed in this research is a model developed by Girmay [25] called: "performance and scalability of geographically-explicit agent-based disease diffusion models". This model aims at simulating the outbreak of pertussis in the whole Netherlands by increasing performance and scalability of the model developed by Abdulkareem [26](see section 3.1). To increase the performance and scalability, Girmay [25] used two major modelling techniques called hierarchical modelling and metapopulation modelling.

Metapopulation modelling describes a particular population as a combination of subpopulations structured and localized in isolated discrete patches of a fragmented environment. Each subpopulation is connected to another subpopulation by some level of movement [72, 73]. For instance, as indicated in Figure 3:4 a population of a given city (bigger circle) can be described as a set of subpopulations structured per age group (smaller circle) which are allocated to a particular space of the city. Both city level (broken line) and subgroup level (sold line) can communication between themselves. Metapopulation modelling is used to aggregate individuals based on common characteristics to form groups with the same behaviour and activity. This modelling technique enables the user to model behaviour and activities of relatively big numbers of individuals/ people in a grouped manner. In other words metapopulation modelling can support scalable ABM modelling by permitting compact representation of people in group.



Figure 3:4 Structuring metapopulation modelling technique

The second adopted type of modelling is hierarchical modelling approach, which has the central idea of decomposing the whole system vertically into levels and horizontally into holons (elements at one level). According to hierarchical modelling, the higher levels are characterized by larger spatiotemporal extent and lower processes compared to the lower levels. Girmay [25] used hierarchical modelling concepts to decompose the spatial scale of the Netherlands to regional, municipality and individual level with different temporal scale for each spatial level (explain below in detail). In addition to the hierarchical modelling concept, Girmay [25] also applied a metapopulation modelling technique to enable scalable representation of ABM.

The regional level model and the municipality level models are built based on theories and assumptions of the metapopulation modelling approach. Girmay [25] assumed that families with children have the highest probability to be infected with pertussis. In accordance to this assumption the model used people aged between 0 to 55 years while individuals aged between 19 to 25 years were excluded. To generate a structured subpopulation, Girmay [25] used age as aggregation variable. Therefore, all individuals found within similar age categories are structured in the same subpopulation. For example, if there are 300 infants in one municipality, they will be grouped in one subgroup. In addition, agent level models are in accordance with the individual based Abdulkareem [26] model.

To represent the outbreak of the a disease, Girmay [25] developed several conceptual models. The main conceptual models are spatial hierarchy model, temporal hierarchy model, commuting model and disease model.

3.2.1. Hierarchical Model

The spatial hierarchical model is responsible for the decomposition of the spatial extent of the Netherlands. This extent is divided into three levels of hierarchy called regional, municipality and agent levels. The regional level contains the six regions of the Netherlands, the municipality level comprises all municipalities that are enclosed in a single region and agent level is a collection of all individuals and family within a municipality.

The temporal hierarchy model is in charge of allocating the interaction time interval of population in three spatial hierarchical levels. For the agent level, the model adopts the assumptions of Abdulkareem [26] which consider a 30 minutes interaction interval. For the municipality and regional level, the assumptions of Girmay is based on a distance decay function with states that people who live closer to each other interact more frequently that those who are far away and know each other[74, 75]. The municipality level assumes that people commute to municipality within the same region for daily activities in a way that leave their home municipality in the morning and come back in the evening. Due to this assumption the temporal extent of the municipality level is a 12 hour intervals (2 times per day). The last temporal scale is for the regional level. The model considers that the mobility of people in different regions occurs once a week, which mostly happens in the weekend. Based on this assumption the temporal extent assigned for the regional level is twice per week that is leaving on Friday evening and coming back to home region on Monday morning (Figure 3:5). Beside the hierarchical decompositions of the spatial and temporal extent, Girmay [25] assigned additional function for both regional and municipality level hierarchies.



Figure 3:5 Hierarchical decomposing of temporal and spatial for the Netherlands adopted from [25]

3.2.2. Regional Hierarchical Model

The regional level model has two main functionalities. The first purpose is to distribute commuters to their destination region and collecting them back to their home region when they finish their activity. The second function is to receive information from the municipality models. This information can be the number of infected people, infected commuters, or the total number of commuter per group and municipality. Commuter agent-groups are those agents who commute into other municipalities within the urban system for different reasons. For example: an agent-group can commute from their home municipality to work/ school and back from work/ school to their home.

To distribute and collect commuters, the regional level model cooperates with the commuting model. The commuting model is accountable to send the commuting subpopulation to their destination and collects them when the commuting time is completed. Commuter can be workers shoppers, students, visitors going for family (friend, holyday) visits, or all. Commuters are distributed and collected in groups but when they enter the individual level model they are decomposed and become individuals



Figure 3:6 Flow of subpopulation between different hierarchical models

To gather mormation about the occurrence of any mection within a given urban level, the disease model is integrated with the urban level model. The disease model inquires if there are infections and if infections occur, then the model directs the information to the urban systems. The urban system change disseminates this information to the other urban systems so that they change their behaviour for example, increase vaccination level of their population.

3.2.3. Municipality Level Model

The spatial extent of the municipality level model is the number of municipalities within the specific region. Population in each municipality is represented by an age structured population which consists of both commuter and non-commuter. The age structure is organized in 6 agent-groups that encapsulate four non-commuters groups (infants, children, teenagers, and adults) and 2 commuters (Teenager-student agents and adults). The daily activities of these agent-groups are not considered in the model so that agents are created and based on the metapopulation approach and positioned in the centre of the municipality. Each agent-group has a number of related attributes including age, health status, commuting status (commuter or non-commuter), and vaccination level and municipality code.

The health status attribute of the agents could be susceptible or infected, and each group consists of agents with different health status which could vary through time. In addition the agent groups are assigned with constant daily activities which differ between groups; for e.g., the activity of commuter is different from non-commuter. In general the municipality model is responsible for calculating the disease spread within the structured populations without identifying the interactions between individuals and the causes of the interactions. In the simulation, the agent-groups are represented by a single point symbols. To calculate the disease diffusion between municipalities and groups, the municipality level disease model compute the number of susceptible, infected, and recovered individuals. No birth or death rate is considered such that the number of population remains constant. In fixed rate, number of a given amount of people transferred from a specified group to another (susceptible, infective and recovered).

3.2.4. Strength and Limitations of This Model

As discussed in chapter two, using ABM to model large scale disease outbreaks with detailed spatiotemporal attributes and individual characteristics is computational resource demanding. To solve computational resource demanding problem of ABM, Girmay [25] utilized the advantages of two adopted modeling techniques(hierarchical and metapopulation modeling). Hierarchical modeling is used to decompose the outbreak extent (in the Netherlands) into various spatial levels (regional, municipality and individual) with respective temporal scale and simulated each of them independently. To facilitate the boundary crossing nature of disease, commuters who move from a given level to another are used. Moreover, metapopulation modeling technique assisted to characterize diversified activities of individuals in a compact form. Metapopulation technique is one of the methods to solve the computationally expensive nature of ABM in simulating large numbers of people and their interaction patterns.

Besides its strength, the model has some limitations. One of the limitations arises is because of metapopulation which structure subgroups based on the common characteristic of individuals. When such type of aggregation is implemented it is obvious that the unique and detailed data of members of subgroup is lost. Since most of the transmission behavior of infectious disease is individual behavior, the loss of detailed data can impact the realist representation of the outbreak. Additionally, though the movement of commuters as group is implemented, distribution member of groups as individual entity is not done. This limits the full evaluation of the model to analyze its performance and scalability.

3.3. HPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models

Understanding and analyzing the large scale geospatial problems using ABM requires large amount of computational resources like computer memory, high speed processor etc.[22, 23, 25]. HPABM aimed at designing a framework that enables spatially explicit ABM to have a high performance and scalability [24]. HPABM achieved its aim by applying a conceptual hierarchical decomposition of the ABM and permitting parallel computing to address computational challenges of ABM. In HPABM, the agent based model is decomposed into a number of sub models that contain sets of agents and their spatially explicit environment. Sub models in turn aggregated into super model which represent computing tasks in HPABM (Figure 3:7). Both sub and Super models have their own responsibility within the system. Super models are in charge of organizing sub models which are enclosed within the super model. Super models also facilitate information exchange between super models. Since each super model is implemented in one processor (one processor versus many super models and vice versa is also possible) information exchange between super models and vice versa is also possible) information exchange between agent-agent, agent – environment either with in sub model or across sub models.



Figure 3:7 Hierarchical conceptual layers of HPABM Tang and Wang [24]

To represent interaction between super models, a connection topology can be designed using directed graph in which each super model is a vertex and connection relationship are edges. Connection topology is aimed mainly to guide inter supermodel communication and task scheduling among processors. Beside super models and connection topology, sub models also can have connection topology. Sub models topology shows data communication induced by agent –agent and agent to environment among multiple sub models (Figure 3:8).



Figure 3:8 Graphs to show connection topology of super and sub models. Squares in A and circles in B represent sub model and the colour shows the sub models under same super model adopted from Tang and Wang [24]

3.3.1. Limitations and Strength of this Model

HPABM model has strengths that could solve the limitations of ABM. The model emphasizes on dividing the spatial extent of simulation into sub area that could help implement in computers computing parallel. The result of each computer is integrated to represent the entire system in ABM. This method can address the computational demanding natures of ABM by facilitating implementation using parallel
computing. HPABM is designed using C programing language to implement the model in each computer and used message passing interface technologies to allow the exchange of information between computers those simulating different supermodels.

HPABM requires some modification to simulate the outbreak of infectious diseases. Since the model is designed to enable scalability of any types of spatially explicit ABM model, the model need to be modified to simulate the infectious diseases outbreak. According to Tang and Wang [3] explanation, sub model represents agent and their interaction with the environment. Agents, environment and interaction levels are not clearly specified. And also the level of hierarchy classification is openly left for the people who want to extend the model. Since the model of our research is to simulate the outbreak of infectious disease, agents stands for population of the society and environment is for the geographical location where people conduct their daily activity. The interactions are daily activity interaction among individuals of the society.

3.4. Summary of Existing Models SWOT Analysis

To summarize the evaluation of existing models, four evaluation criteria are proposed. The result of evaluations are presented using table below (Table 3:1). The first analysed model was the model developed by Abdulkareem [26], which aimed at simulation the spread of pertussis in Enschede. The model integrated three sub models namely an activity, social interaction and a disease model with synthetic population to represent the realistic movement and interaction of individuals. Despites its strength on realistic representation of the society, the model emphasizes the city as an enclosed place that has no incoming and outgoing movement of people. In other words, the model does not include commuters coming to the city and going out which makes it. The model is also not scalable when a larger number of agents are introduced.

The second analysed model was developed by Girmay [25] which aimed at increasing the performance and scalability of the model developed by Abdulkareem [26]. This model uses hierarchical modelling and meta-population modelling techniques to increase scalability and performance of simulating the outbreak of pertussis in the whole of the Netherlands. The model divided the Netherlands into regions, municipalities, and individual levels and simulates the outbreak of the disease within a specified spatial extent, independent of each other. The communication between the levels is connected through commuters who move from a given spatial scale to another. Though the model is capable to represent scalability, because of metapopulation technique which represent sets of individual the disaggregation of agents from the higher aggregation level to the individual level was not implementable.

The last model is the model developed by Tang and Wang [24]. The work aims at designing a generic framework that enables the scalability of any types of ABM. The model uses the concept of hierarchical modelling and parallel processing to achieve scalability of ABM. Since the model is generic for every type of ABM, some modification should be done to simulate infectious disease using this framework.

At the end of existing model analysis, some characteristics of the model are identified to be adopted for this model. The method of integrating the activity model, social interactions model and disease model with synthetic population is used to generate the realistic society for the model. This model uses the concept of commuters and parallel processing concept to achieve scalability.

Main Characteristics	Sub Characteristics	1 st Model	2 nd Model	3 rd Model
	Types of Disease Simulated	Pertussis	Pertussis	None
Generic Model	Used Classic Infectious Disease Model	SIR	SIR	None
N. 1 11.	ABM	Yes	Yes	Yes
Modelling	Hierarchical Model	No	Yes	Yes
Approach/s	Metapopulation Model	No	Yes	No
	Social Networking	No	No	No
Prevention and	Pharmaceutical	No	No	No
control methods	Non-pharmaceutical	No	No	No
Scalability	Spatial	No	Yes	Yes
	Temporarily	No	Yes	No
Implementation	Parallel computing	No	No	Yes
	Programing language	Repast Symphony	Repast Symphony	C with MPI library
	Implementation	Full	Partial	Full

Table 3:1 Summary of existing model analysis

4. CONCEPTUAL MODEL

This section describes the conceptual model and its main focus is to create a generic and scalable model. The chapter is divided into three sub sections and each of them explains a fundamental component of the conceptual design. Section 4.1 describes the general setup of the model in which the purpose of its components is explained. Section 4.2 focuses on the sub models and section 4.3 explains parallel processing to achieve scalability of the infectious disease ABM.

Understanding the concept of social network and parallel processing described in chapter two is basic to understand the conceptual model.

4.1. General Setup of the Model

This subsection describes the general characteristics of the model including the purpose of the model, the model components, state variable and scales, and a process overview of the sub models

4.1.1. Purpose and Model Components

The main objective of this model is to design a scalable ABM that can simulate any type of infectious disease at various outbreak scales. Scalability is achieved by dividing the given simulation extent into sub areas and simulating sub areas independently using different processors (CPU) (Figure 4:1). To capture the overall outbreak, the result of each sub area is integrated by enabling the communication between them. The decomposition of spatial scale is mainly based on the number of people/agents available within the sub area. In other words, a sub area may represent different types of urban system which can be community, town, city, municipality or region which depends on number of available people to be modelled.



Figure 4:1 Decomposing of the simulation spatial scale and allocating to different processors

To continue designing the ABM, defining the building blocks of the agent based model, such as agents, environment, and time are crucial matters. In its basic characteristics, ABM is very sensitive to the initial conditions such as behavioural specifications of agents, the interaction rules of agents, etc.[58]. To generate accurate results, the simulation must be supplied with correct characteristics of agents and environments. In this model three types of agents called individual, municipality and health units are defined.

Since the main purpose of the model is to simulate infectious disease outbreaks, individuals stand for persons of the society who are affected or at risk of infection. As described in Figure 4:2, individuals can further be divided into commuters and non-commuters. Commuters are individuals who conduct their daily activity like school or work in a different urban system then the urban system they live in. While non-commutes live and work in the same urban system. Each individual should be aware of the home urban and for commuters they need to know their destination urban system.

Individuals also have social interaction with other agents. Social interactions stand for relationships that can occur in home with family members and in activity place like school or workplace with friends. Number and types of social interactions of an individual depend on their age and the place where they conduct their daily activity.



Figure 4:2 UML diagram for the components of the model

The municipality agents are abstract agents that are aware of the spatial area they manage and provided with a unique ID to distinguish them from the other municipality agents. Municipality agents are used to gather information about the total number of infections and illnesses within each urban system, exchange information about commuting agents with other municipality agents and initiate interventions from Health unit agents. Health unit agents are also abstract agents denoting the organizations in charge of vaccination. Their spatial scale may be the same as the area of the urban system or not.

Beside identification of agents, ABM also requires to determine the environments where agents interact with each other. In this research environment stand for the spatial environment that agents move and act on. As described in Figure 4:2, the model consists of three kinds of environment such as the buildings, the urban system and the vaccination zones. The buildings can be split further into three types of buildings,

homes, schools and business or workplaces. All types of building are represented as point locations while urban system and vaccination organization represented as abstract agents with unique ID of the urban system.

The environment that belongs to the Health unit is the vaccination zone. This zone defines are area for which the vaccination coverage is known and reported. Within this zone, the vaccination coverage is assumed to be uniform.

Furthermore, all environments in this model are static that no changes will occur during the simulation.

4.1.2. State Variables, Agent behaviour and Scales

All agents defined in this model have their own characteristics and functionality described as the way they act in the simulation environment. The aspects that are measured during the simulation are called the state variables. This is described by Grimm et al. [76] as "low-level variables that characterize the low-level entities of the model, i.e. individuals or habitats". The state variable of this simulation is the disease state of the individuals. Another important issue is the behaviour of the agents. This section describes the state variable of agents and the spatial they allowed to move and function.

Individuals

Individuals are persons with heterogeneous behaviour and each of them can be described by using a unique personal behaviour. Since the aim of the model is simulating disease outbreak, they can be explained using characteristics that have impact on the transmission of the disease. For instance, an agent/ person can be labelled by micro level data including (but no restricted only to this) age, disease status (exposed, infectious, or recovered), immunity, susceptibility (the state of likely to be infected) and infectivity (capability to infect the others) levels of the person, etc. Some of the attributes are interdependent; like susceptibility and infectivity are the results of immunity level. As the immunity level of the person increases, both susceptibility and infective level decrease. Immunity level in turn depends on the level of vaccination consumed by a given person. As described in chapter two, increasing the vaccination level provides higher immunity. Immunity levels can be induced either from vaccination or naturally from recovering from a given infection. For some of the diseases immunity induced stay lifelong while for the others it wanes and turns the person susceptible status.

Depending on the type of the individuals (whether commuter or no-commuter), an individual has a defined spatial scale to move and function. Both types of individuals have a house, activity place (school or workplace based on the age of an agent), and a home urban system. When an agent is a commuter, the person has an additional spatial extent to move and act which is called the destination urban system. The destination urban system contains the activity place of the person.

In general, individual type of agent has state variable called disease status which can have the status of being exposed, infections and recovered, in determined by number of factors. The factors include vaccination, immunity, susceptibility, infectivity, age and the social interaction the person conduct with others. The social interaction of the person can be defined on the spatial scale. For instance house determines the interaction of individuals with respective family.

Municipality agent

A municipality agent is an abstract agent that represents the spatial extent of a sub area of the modelling extent. This agent has three major functionalities including: sending commuters to destination urban system, counting the number of infectious people, sending the information about individuals to the respective Health unit agent. To assist the vaccination organization, the municipality agent has to record

which agent is infected and where the agent was infected. The age, house address and the place of infections are the information transmitted to the Health unit agent. Since the spatial extent of the vaccination zone may or may not coincide with area of the urban system, each urban system has to collect the house address of the individual agent to identify the vaccination zone which is in charge of controlling that area (Figure 4:3). The functional scale of the municipality agent is obviously that it functions within its spatial extent and communicate with other municipality agents and Health units agents.



Figure 4:3 The interaction between individual, urban level and vaccination zone

Health unit agent

Health unit agent is one of abstract type of agent in which the decision of vaccination zone is transferred to individual agents. When conducting vaccination is needed, the agent checks if the government or health organization has announced an extra vaccination campaign. When this is the case, the immunity level of a certain group of agents (e.g. pregnant woman and young parents) will be shifted to the next level. Full immunity for certain diseases (like Pertussis) only is obtained after a sequence of a number of vaccinations. For instance, if a person had only one single vaccination, this person will need a second and possibly a third and fourth vaccination to reach complete immunity. The immunity of a given person depends on its vaccination level and because of waning it may be lost.

The vaccination zone may also decide to close activity places like school or workplaces. When closing activity place is decided, the agent deactivates the working status of the activity place. This is indirectly informing activity model that for individuals who are scheduled to go to that activity place, the work status of the place is not activated. When this happens the activity model should not allow the agent to go to that activity.

4.1.3. Process Overview of the Sub Models

In the previous sections of this chapter the main setup of model was described. To simulate the disease spread, agents should be scheduled to conduct their activities. Particularly the interactions that occur between individual agents are a fundamental requirement to track the spread of the disease. To achieve this purpose, three sub models were designed including: activity model, interaction model and a disease model. The execution of the models is sequential as well as iterative. For instance, as described in Figure

4:4, the execution of the disease model must be followed by the activity and interaction model and each full process is repeated in every given time interval.



Figure 4:4 Interactions between sub models

Each sub model has its own functionality. The main purpose of the activity model is to move agents to new locations according to their activity plan. The purpose of the interaction model is to identify which agents are in contact during a specific time step and what type of interaction occurs. Lastly, the disease model is to determine if the disease status of the given individual agent will change. This change in disease status of an individual may reflect the status of being exposed, infectious or recovered. Additionally, the disease model also controls the waning level of a person's immunity through time. Additional elaboration on each sub model is included in the following sub section.

4.2. Sub Models

In section 4.1.3 the overview sub model is described. In the following sections, the detailed explanation of every sub model will be given; including the functionality of each sub models is described. Sub section 4.2.1 describes about the activity model. Section 4.2.2 deal at the social interaction model and lastly, section 4.2.3 describes about the disease model.

4.2.1. Activity Model

The purpose of the activity model is to generate a schedule that guides the daily activity of individuals and moves them to the next location when the previous task is accomplished. The schedule comprises a list of intended daily activities and time allocated to accomplish each activity. The schedule can be static (repeat itself over a long period of time) or dynamic (frequently changing). Commonly, weekday schedules are static (the same for all weekdays) while weekend schedules are dynamic (person can have different schedule for every weekend). Planning the schedule is determined by a number of factors such as age of a person, available activity places, opening and closing hours of the activity place, working status of activity place (closed or working)etc. For this research, agents are assigned with a fixed schedule for weekdays and are assumed to stay at home in weekends.

Age group and activity places are among the factors that impact the schedules. Some assumptions are made regarding the activities. The age groups included in the model are divided in to four age groups including infants (0-3 years), children (4-12), teenagers (13-18) and adults (19-55). Besides the age groups, identifying the activity place where people spent most of their time should be identified. In reality the interaction of individuals happens in different places like home, schools, workplaces, and shopping

canters, public transportation etc. But not all activities places have an equal impact on facilitating disease transmission. For example, even though going for shopping provides a probability to contact a number of people, since the time spent is short and no intimate interactions (like kissing and hugging) take place, the likelihood of disease transmission in shopping places is not significant[40]. Belik et al. [40], also mentioned that family, school and workplaces are the places where people spent most of their time and have intimate interactions. Modelling those activity places can reflect realist disease outbreaks. Based on this consideration, the interaction places represented in this model are home, schools (kindergarten, primary school and secondary school), and workplace (business or professional work).

In addition to the activity places, the schedule also needs a time interval during which activities stated in the schedule are executed. If we consider only the activity places mentioned above, starting from mid night, a day can be divided into three blocks of 8 hours. The first block represents the time interval starting from 12 o'clock early in the morning (24.00 hours) till 8 o'clock (8.00 hours). This is the time people stay at home. The second block begins at 8 o'clock (8.00 hours) and ends at 4 o'clock on afternoon (16.00 hours). Second block is the time period that all individuals (agents) stay at activity place (school, workplace). The third block is starting from 4 o'clock (16.00 hours) up to 12 o'clock in mid night (24.00 hours). Based on the age of the person, the schedule contains activity and allocated time. The template of the schedule is as mentioned in the table below (Table 4:1).

Day	Age (Years)	Time allocation			
		1 – 8 o'clock	8 – 16 o'clock	16 – 24 o'clock	
	0-3	Home	Kindergarten	Home	
Weekdays	4 - 12	Home	Primary School	Home	
weekdays	13-18	Home	Secondary	Home	
			School		
	>=19	Home	Work place	Home	
Weekends	All age groups	Home	Home	Home	

Table 4:1 Template for creating schedule for different age group

In summary, as described in Figure 4:5, in the design of the simulation, the scheduling of the activities into three blocks has a consequence for the buildings included in each time step. As all people are presumed to be at home during time steps 1 and 3 during these time steps, the simulation will only run the interaction and the disease model for buildings that represent home locations. Alternatively, as during time step 2 all people are at work or school. Moreover, Activity model has to check the opening, closing hour and working status (closed or working) of activity place before allowing the agent to move. This is because the activity place may be not function based on the decision from vaccination zone.



Figure 4:5 Buildings included in each time step

4.2.2. Modelling Social Interactions

Besides personal characteristics of individuals, transmission of infectious diseases also relies on the interaction between people/agents. To represent the interactions among people, different approaches including social network modelling are used. For this research, three main types of social interactions are defined and they represent family, colleagues and schoolmates. Family relationship denotes any group of people who live together in the same house (not restricted to blood relationship). Colleagues are people who work together. Lastly, schoolmates are agents who attend their education in the same school. In the most simplified case, an agent must have at least family relationship and based on the age, the agent can have colleagues or schoolmate relationships.

Schoolmate and colleague can be further classified into sub relationships. Schoolmate relationships can be split into two types of relationship namely classmates and intimate-friends. The classmate relationship stands for student who attend the same class while intimate-friends are student who spend most of their time together when they are in school. Intimate-friends may be from the same or different classes. Colleague also contains two types of relationships called same-department and same-office relationship. The reason behind the further classification is to handle the impact of different contact levels on the transmission of disease. Obviously, the possibility of a particular worker infecting an office mate is much higher than worker in another office.

As described in chapter two, social network can be represented using graph in which people/ agents are the vertexes of the graph and edges are the relationship between them. The edge can be either directed or undirected graph. The specification of the edge type depends on the aim of the model. In our case, we assume that in a particular group each individual can affect and being affected by other. Thus, the social network representation between agents has directed graph representation (Figure 4:6).



Figure 4:6 Three types of social networks

In Figure 4:6, three types of relationships are defined, using the colours red, green and blue. The green colour indicates that all agents connect via green lines are living in the same house and have intimate interaction. Likewise, the blue and green colours indicate the colleagues and schoolmate relationships that occur between members of different families.

4.2.3. Disease Model

The disease model is one of the sub models and it is in charge of controlling the spread of the infectious disease over the population. The model has three basic responsibilities. Firstly, it monitors the progress of a disease within a given person/agent. Secondly, it controls the transmission of the disease among individuals. And lastly, it controls the immunity level of individuals that are lost by waning. Based on the task to be achieved, the disease model is further divided into two models and the functionality of the models is explained below.

Within an individual progress controller

When an infected person coughs sneezes or even breathes, tiny droplets containing micro parasites of an infectious disease are released into the air and can be inhaled by person who is in contact with the infectious person. As described in chapter two, if the person is fully immune, the inhaled micro parasites cause no effect, otherwise the micro parasites can start to disseminate within the body of the inhaled person. The time elapsed in dissemination is called incubation period and in this period the person will not manifest any symptoms or can infect other people. After the incubation period, the person gains the ability of infect others and continues to be infectious for a couple of days. The time period that a person remains infectious is called infectious period. Both incubation and infectious period are different for different infectious diseases. Recovering from the disease also enable the person to become fully immune.

To model the processes that happen within the body of an individual, a simple arithmetic operation is formulated. The arithmetic operation considers the number of days elapsing from the incubation and infectious period. The calculation starts from the day when a person is exposed to the disease and adds a number of days pass as incubation period to change the status of the person in to infectious. The recovered status in turn is calculated by adding a number of days to the infectious period, after which the status is changed to infectious. In this model the following formulas are used to represent the progress of disease with in the host (Figure 4:7).

$$Pic = De + Dic$$
(1)

$$Pif = Pic + Dif$$
(2)

Where *Pic* denotes the incubation period and De stands for the first day that an individual is exposed to the micro parasites for the first time and Dic is the number of days that have to elapse before the exposed individual becomes infectious. Pif stands for the infectious period with Pic representing the last day of incubation period and Dif stands for the number of days that the person remains infecting other people. Dic and Dif take different values for different diseases.



Figure 4:7 Different stages of SEIR model adopted from adopted from Perez and Dragicevic [27] with modification

Between hosts transmission model

Since the model considers the structure of social network, between individuals investigation of disease spread is limited to the social network where a person appears at a given time and place.

Transmission of a particular infectious disease from infectious to susceptible agent can be influenced by a number of factors. According to Barrett et al. [17] the number of infectious people within a group, susceptibility and infectivity of susceptible and infectious people respectively, transmissibility (probability of transferability of micro parasites), the contact duration etc. are factors that can impact transmission of disease within a group. Distance of contact is the physical distance between two individuals that are in contact with each other. The duration of an interaction is the amount of the time that individuals are spending while interacting. This time is important, as its length will decide if it is enough for a disease to transfer from an individual to another[17, 26].

The transmission is expressed using probabilistic equation adopted from Barrett et al. [17](equation 4). To express susceptibility and infectivity some assumptions are considered. The assumption is based on number of vaccination consumed by the particular person. For instance fully vaccinated person considered as 0% susceptible and not infective; while not vaccinate person is 100% susceptible and 100% infective. The susceptibility and infectivity with respective of vaccination level are summarized in (Table 4:2).

$$P_i = 1 - \exp\left(\tau \sum_{r \in \mathbb{R}} N_r \ln(1 - rs_i \rho)\right)$$
(3)

Where pi stands for the probability that a given susceptible person i is infected, τ is the time duration elapsed when the person is stay in contact with infectious person, R is the set of infectivities of the infected individuals at the location, Nr is the number of infectious people with infectivity r, si is the susceptibility level of i, and ρ is the transmissibility.

Vaccination dose	Range of	Susceptibility	Infectivity	
	Immunity			
No Vaccinated	0 - 25%	Highest (100%)	Highest (100%)	
One dose	25-50%	High (75%)	High (75%)	
Two doses	50 - 75%	Medium (50%)	Medium (50%)	
Three doses	75-100%	Low (25%)	Low (25%)	
Four doses	100%	No (0%)	No (0%)	

Table 4:2 Numeric representations of susceptibility and infectivity levels

Immunity of a particular person is also a fundamental concern during transmission of a disease. In reality, immunity can be induced either from vaccination or naturally after recovering from an infection. For some diseases, immunity will stay lifelong and because of that the person may not be susceptible again for that disease. Immunity also can wane through time and turn the person to susceptible status. In meanwhile, if the person receives a vaccination in the waning period, the person can re-boost immunity.

The disease model is in charge of managing the waning level of immunity of individual who are fully immune at some time either because of vaccination or recovery of the disease.

An arithmetic operation is used to represent the waning of immunity. According to Hethcote [45], for immunity that can wane, there is a time period that has to pass when the full immunity fade and return the person to susceptible. For instance, if no vaccination is taken, immunity induced from vaccination of pertussis can persist only for seven to twenty years; so that after those years the person is susceptible to the disease. Based on this concept decay rate can be represented using 1/decay-period (e.g. for pertussis, if 20 years is considered, the decay rate is 1/ 20*365 where 365 indicate number of days in one year). In that sense the person loses 1/ 20*365 times the current immunity per a day. Different diseases have various decaying period. In general, to represent waning of immunity the following formula is used.

$$Inew = ICurrent \left(1 - \frac{1}{\text{decay_period * 365}}\right)$$
(4)

Where Inew stands for new immunity level calculated from current immunity ICurrent of the person

4.3. Parallel Processing and Commuting Model

Parallel processing method is a technique used in this research to increase scalability of infectious disease simulating ABM. The main concept behind parallel processing is to decompose a given problem into sub processes that can be executed independently using simultaneously running processors and communicate with each other to produce a solution for the entire problem. The efficiency of parallelization depends on fundamental functional and technical factors such as specifying architectural layout of parallelism, defining the extent of sub processes in a way of load balancing between processors, enable communication between sub processes and synchronizing the communication. Any work using parallel processing should consider those factors and from infectious disease simulating view they are explains as follows.

In chapter two, some of the methods that assist the specification of parallelization factors are described. Shared memory and distributed memory systems are among the methods used to describe parallelization architectural layout of parallel processing. The conceptual model design of this research is based on the concept of distributed memory architecture. The reason to choose such architecture is its scalable nature when a number of processors are introduced to the computing environment. Beside architectural layout specification, processors also required to be allocated with a task in a manner of not being over or less loaded compared to the others. In other words, task distribution must balance the load between processors so that the time delay caused by waiting for the result of another processor will be reduced. From the two ways of task distribution, namely data parallelization and process based parallelization, which are explained in chapter two, this research emphasize on the data parallelization technique.

In accordance with the explanation of Rana and Stout [21], the issue of agent based model scalability arises when there is an increase in the number and/or diversity of basic components (agents, interaction, environment or all of them). Specifically for infectious disease simulating ABM, scalability is needed when there is increase in number and/or diversity of population, social interaction between individuals and the range and diversification of simulation environment.

Obviously, from the above description scalability of infectious disease ABM is desirable when the amount of data for a simulation increases. Based on this fact, it is clear that the data parallelization method is the appropriate method to decompose the simulation. This method divides the dataset into sub datasets and distributes them over processors so that similar operation will be conducted on sub datasets. While execution is undertaken, communication between processes can be conducted using internetwork connection between processors.

The Urban system encapsulates its respective population, buildings, and an abstract agent called municipality agent that facilitate the communication between municipality agent of other urban systems and the Health unit agent. The communication could be sending commuters to the destination processor, collecting number of infections and informing Health unit agent. Municipality agent should be created for each of the urban system and the Health unit agent should be created in for vaccination zones.

The transfer of commuters needs the exchange of information between the municipality agents. This can be achieved in different ways e.g. by physically moving the agents or by creating a copy of the agent in the other processor and based on the time step either run the agent itself or the copy of the agent. The physical movement of agents from one processor to another processor is a realistic way of representation but when the number of commuters and the destination processors of commuters increase, it introduces a lot of communication between processors. Increasing of communication may result in a computation delay of the overall process through leading the communication suspension between processor. For this simulation the second option (creating copy of commuters in destination urban system) was selected and implemented in the following way:

- Create copies of commuter agents
- Create an abstract link between original and copy agent to transfer the updated information from the original to the copy agent
- Create a reporter agent for each processor that collects information about disease status of commuter agents.
- Create copy of commuter destination processor reporter agent in home processor of commuter
- When there is a change in disease status of a commuter agent, the reporter agent of the destination processor transferring information to its copy agent in sender processor
- Copy of the reporter agent informs the original agent of commuter in sender processor to change its status.
- Original agent updates its states and the status of copied agent in the receiver processors through the abstract link between them.

The main point to not forget in this level is that the steps illustrate above are the task of the municipality agent in facilitating the communication between other municipality agents. For scalability purpose the task of municipality agent is further decomposed to creation of reporter agent and the communication between original and copied agents. Further elaboration of the steps is described in more detail as follows.

The commuting process desires communications between commuter sender processor and receiving processors/s. Initially, the sender processor has to identify commuter agents, their destination processor; and produce their copy with similar current status of the original agents. The copied agents will be sent to the destination processor and after receiving the copied agents, the destination processor/s is expected to allocate the activity place for each agent. When an agent is allocated to an activity place, it has to join the social network of the place which could be schoolmate or colleague relationship. Since the copied agent must have the same status as the original agent, any attempt in the destination processor to change the status of a given copied agent should be denied and only the original agent in the home processor is authorized to change the status. To capture the status change that may happen in the destination process, the synchronizing method called reporter agent is created.



Figure 4:8 Interactions between three types of agents and reporter agent

As indicated in Figure 4:9, reporter is created for each of the processor and when a given processor receives commuters, it sends the copy of reporter to the processor that sends the commuters. When the update on commuter is needed, the original reporter informs the copied reporter; and copied reporter in turn informs the original agent of the commuters in home processor about the attempted change. After receiving the information, original agent change its status and also update the copied agent.



Figure 4:9 Functional setup of reporter

Commuting agents are duplicated and may be available in two or more processors. Unless proper controlling mechanism is applied, at the same time agents have a chance to conduct different activities on different processors. Scheduling is one of the techniques to solve such problems. The original agent always stays in the home processor and is used to check disease transmission within a family during the first and third time step of the day. In other words, original agents stay inactive during time step 1 and 3 of the activity table, copied agent will be active at work time and deactivated on the rest of the time. This

means, that the original agents are used to check disease transmission that happens within family whereas copied agents are used during disease transmission check-up in the activity place.

As described in Figure 4:8, besides commuting agent management, reporter agent is also in charge of controlling the information exchange that occurs between vaccination zone and the urban system. The copy of all reporters from each urban system should be in contact with the vaccination zone to inform about the infection and illness in a given urban system. On the another hand, vaccination zone that manage the given urban system also need to send the copy of Health unit agent to urban system so that urban system will be aware of the decision of vaccination zone. Since the boundary of a particular vaccination zone may not coincide with the boundary of the urban system, an urban system has a chance to have more than one Health unit agent. Based on this concept is it clear that a given urban system may get different decision form various vaccination zone.

4.4. Summary

This chapter explains the conceptual model to achieve generic and scalable infectious disease ABM. It start by describing the basic components of the model with respective characteristics and followed by explaining the spatial scale of agents of the model to move and be functional. The model also describes sub models used to allow realistic social interaction among agents and track disease transmission between them. The model introduced prevention and control method with vaccination organization and required communication to implement to the society.

Finally the model explained the technique to achieve scalability of ABM model called parallel processing. Parallel processing used distributed memory architecture and data parallelization method to achieve scalability. Furthermore to synchronize the communication between processes happening independently in different processors, the synchronizer called reporter is introduced. The main purpose of introducing reporter is to reduce communication dependency of processors so that communication delay between them can be reduced.

To test the correctness of the conceptual model, the model is implementing to simulate the outbreak of pertussis in the Netherlands. Though the model is capable to simulate for the entire county, because of time limitation two cities of Netherlands called Enschede and Hengelo are chosen. The implementation precede with data preparation which described in following chapter and implementation using repastHPC modelling toolkit is discussed in chapter six. The evaluation and recommendations are also included in the last chapter.

5. DATA PREPARATION

5.1. Introduction

In chapter four, the conceptual model was presented to describe the outbreak of infectious diseases in various spatial scales. Based on the conceptual model, a framework was developed using RepastHPC and the details of the implementation will be described in chapter six. To test the framework, pertussis will be simulated for two municipalities in the Netherlands. The motivation to choose the Netherlands is the availability of large studies conducted by different researchers and the availability of data sets required for the implementation. The main data sources are the official websites and documented material written by different governmental organizations.

The Netherlands is a European country which is administratively divided into fourteen provinces and 408 municipalities. The provinces represent the administrative layer in between the central government and the local municipalities. The Netherlands is among the European countries which have relatively many inhabitants per municipality. In 2009, the average number of inhabitants per municipality increased from 40 to nearly 41 thousand. Overijssel is one of the fourteen provinces of the Netherlands with the capital city of Zwolle. Overijssel consists of 25 municipalities which include Enschede and Hengelo (See Figure 5:1). The conceptual model of this research can be implemented for the whole Netherlands, but, because of time limitation only the outbreak of pertussis in Hengelo and Enschede was implemented.



Figure 5:1 Netherland, Overijssel, Hengelo and Enschede

This section describes the preparation of the data sets required for the evaluation of the framework. Four major datasets are needed for both Hengelo and Enschede which includes: population census data, buildings, vaccination coverage and commuting data. Sections 4.1, describes the population census dataset and in section 4.2, the building dataset will be presented. In section 4.3 vaccination and commuters data are explained.

5.2. Population Census Data

One of the dataset used is the population census data of the two municipalities. The census data was used to extract the number of children per age group and the number of households. This dataset is downloaded from the official websites of the two municipalities (<u>http://www.enschede.buurtmonitor.nl/</u> and <u>http://www.hengelo.buurtmonitor.nl/</u>). The websites presents data in aggregate form for the population distributed in the neighbourhood of the municipality. For instance it contains the total population, population per ethnic group, population in various age groups, number of families with and without children, families with two parents and only one parent, people in different economy levels, etc.



Figure 5:2 Distributions of two types of family over neighbourhoods of Hengelo



Figure 5:3 Distributions of two types of family over neighbourhoods of Enschede

Assuming infectious disease as diseases that mostly occur in children, households with children and children in different age groups are extracted from the census dataset. The households can have one or two parents; and children are grouped to four age groups called infants (0-3), child (4-12), teenager (13-18) and adults (19 - 55). It is assumed that adults in the age group of 19 - 24 are students cannot have children while those who are in the age group of 25 - 55 can have children and be households of a given house.



Total Number of Children per Neighborhood 1600 1400 1200 1000 800 600 400 200 0 t Zwering Laares Oikos Boswinkel/ De. Marssteden Bentveld/ Bultserve de Eekmaat Deppenbroek Dorp Lonneker Eilermarke Goorseveld Het Brunink Schreurserve Stadsveld-Zuid Stroinkslanden NW Ribbelt/. Twekkelo /eldkamp Getfert-. Walhof/ Roessingh Wesselerbrink ZO Zuid-Esmarke Cromhoffsbleek/ Horstlanden/ Roombeek/

Figure 5:4 Distribution of Children over the neighbourhoods Hengelo

Figure 5:5 Distribution of Children over the neighbourhoods Enschede

The datasets are available for different years but for this research the dataset of 2009 is chosen. The reason to choosing the data of 2009 is that compared to the dataset of another year, this data has complete list of information for each neighbourhoods. The data set has been downloaded from the website and was organized in CSV (Comma separated vector) file format.

5.3. Spatial Data

Agent based modelling requires environments to allow spatially explicit interactions of agents. Environments can be like buildings, roads, etc. The representation of the environments also needs to be loaded into the computer. For this model, since the model deal with infectious disease outbreaks, the

considered simulation environments are the activity place layer where people meet and the spatial dataset for the boundary of the municipality where they live. Within the activity place layer (buildings) the interaction places identified are kindergarten, primary school, secondary school, work place and houses. To bring the required data sets we have collected two major spatial data sets called city neighbourhoods and building shape files.



Figure 5:6 Distribution of buildings in Hengelo and Enschede

The first spatial dataset is the polygon Shape file that includes the boundary and neighbourhood divisions of each municipality. This dataset is downloaded from the ITC website and includes about 73 and 63 neighbourhoods for Enschede and Hengelo respectively. Since population census data contains the distribution of population over the neighbourhood, considering neighbourhood in spatial dataset allows for a realistic distribution of individuals over the respective neighbourhoods.

The second spatial data set is the buildings layer. This dataset contains the polygonal representations of the buildings, which lacks the type specification and neighbourhood. The type specification can be residential, shop, school, workplace, etc. and identifying their type allows allocation of individuals to correct type of house for their daily activities. For instance, identifying residential buildings allows a correct allocation of house types as home location for family members of a given house. Furthermore, identification of school type allows computing the nearest school for children. Additionally, identifying the neighbourhood of the buildings also facilitate reasonably accurate distribution of individuals to their house in the neighbourhood where they are. To capture these two fundamental attributes of each building, two further processes are conducted.

To provide building with corresponding neighbourhood codes, a spatial layer joining process is conducted between the boundary and building dataset using ArcGIS. To make the processes easier, the building dataset which was in polygon form, was converted to a point file. The joining process allows the buildings that fall in the polygon of the neighbourhood to have all attributes of the neighbourhood. Since only the code of the neighbourhood is needed for the buildings, the rest of the attributes are removed and only the Id of the neighbourhood is maintained in the information of the buildings. Using this process the neighbourhood of the buildings were identified, but also the type of building is required.

To identify the types of buildings, Google Earth search engine is used. The search engine is used to find the location of activity places (kindergartens, primary and secondary schools, industrial, shopping centres and workplace) in Google map. Longitude and latitude of the buildings are extracted from Google Earth and searched in the Shape file of the buildings. Some of the extractions from Google Earth did not coincide with the exacted building, so the comparison of the shape of buildings in the Shape file and Google Earth was undertaken to decide the type of the building. Schools are searched by typing their names on the search engine, but for the identification of houses and workplace another assumption was made. The assumption is that mostly workplaces like industrial areas are reserved and commonly residential are not available there. Based on this assumption, industrial areas are identified from Google map and reserved as workplace. Removing schools and workplaces from the original data, the rest of the data was considered to be residential.



Figure 5:7 Spatial distributions of schools both in Hengelo and Enschede

5.4. Vaccination and Commuters Data

The last datasets used are the vaccination and commuting data. Finding the exact data for those datasets is one of the challenges of this research. To approximate the data, some official documents were reviewed and one of the documents was the RIVM Immunization Program [77]. This document lists the number of disease and respective vaccination schedule. Pertussis is one of the diseases in which periodical vaccination is disseminated and has four step of vaccination. According to the document, children are getting their first vaccination while they are at the age of 6-9 weeks and in the time of 3, 4 and 14 months additional four vaccination level is provided. The second stage of vaccination is introduced while they are at age of

four and it is stated that that 96% of children at this age are fully vaccinated. The third and the fourth level of vaccinations are conducted when children are at age of nine and twelve years. The data about teenagers and adults is not clearly mentioned. Since vaccination stops at the age of 12, for older age groups some approximation was made. According to Hethcote [45], the immunity for pertussis wanes in a time period of 7 to 20 years and can bring the person to be susceptible. Unless the person is infected and recovers, the time of waning starting from the teenager's time, can lead to susceptible when the person reaches an age of 19 till 32. But, if the person has a chance of being infected, the person can boost and retain full immunity level. Because of data limitations, all adults are considered as susceptible.

Age Groups	Level of vaccination (%)					
	None First Second Third Fourth				Fourth	
Infants (0-3)	1	1	1	1	96	
Children(3-12)	5	5	5	5	80	
Teenagers(13-18)	55	5	5	5	30	

Table 5:1 The approximate values of various vaccination levels for different age groups

Commuting is also one of the dataset displayed by simple approximation. The dataset expected to contain two types of commuters namely workers and students. Both of the cities should have incoming and outgoing commutes. The approximated values are mention in table below.

Type of commuters	Hengelo	Enschede
Workers	1000	1000
Students	250	250
TI11 50 A	1 C	

Table 5:2 Approximate values of commuters

The implementation of the framework is discussed in chapter six and the verification and evaluated of the framework is explained in chapter seven. The verification and validation are based on the datasets explained in this chapter.

6. IMPLEMENTATION

In chapter four, a conceptual model was designed and chapter five explained the data preparation. In this chapter the parallel computing software development phase is discussed. Implementation is conducted by using RepastHPC. This section is organized into sections and sub sections. 6.1 describes the programing languages used for agent based modeling and their limitation. Furthermore the section describes Repast HPC and the reason why it was chosen for the implementation of the model. In section 6.2 the detailed implementation of the model is discussed.

6.1. Tools to Implement ABM

Agent based modelling is a new and rapidly developing modelling techniques. Thanks to thoroughly improving technologies, a number of simulation toolkits (software) have been developing to implement ABM which includes Swarm, NetLogo, MASON, Recursive Porous Agent Simulation Toolkit (Repast), etc.[78, 79]. Despite the availability of toolkits, selecting the most suitable one is important point. Using evaluation criteria researches tried to specify the weakness and strengths of simulation frameworks. According to Tobias and Hofmann [79], Repast is the winner based on three aspects. The first feature of this language is its flexibility to support heterogeneity of both agents as well as environment; secondly it is open source code software so the functioning of the toolkit is visible and can be adapted or even extended. Lastly Repast uses Java as the implementation language and since, Java is a widely used powerful independent platform "write once, run anywhere", it gives the advantage of running simulations on computers with different operating systems including WindowsTM, MacTM, LinuxTM, UnixTM, etc[26].

RepastS toolkit provides libraries that enable the developer to create, run, display and collect data from agent based simulations [80]. In accordance to the explanation of Crooks [80], the availability of a basic functions library and point-and-click interface allow programmers that are using Repast to spend more time on specifying the characteristics of models like agents behaviour, interactions rather than dealing with the language itself.

6.1.1. Fundamental Concepts of Repast

Repast has built in libraries to implement basic building blocks of agent based models. *Context* is a data structure in RepastS which handles agents by providing an abstract environment for them at a given simulation time. Though contexts can work as bucket to contain all agents, context does not provide relationship between agents. The relationship between agents is created by the data structure called *projection*[81]. Projections are added to a context to allow interaction between agents whether in spatial or network form.

There are four types of projection in Repast symphony called *continuous space*, *grid*, *network* and *GIs projection*[81, 82]. Continues space is the real space in which an agent's location is represented by floating point coordinates. For instance, a given agent can be at a point 2132.133, and 43276.521 and another may be at 67431.20, 0.045. Since this projection use floating points using longitude and latitude of real system agent location can be loaded for simulation. In contrast to continuous space, grid projection uses discrete integer values to represent the location of agents but for the rest of its characteristics it has the same functionality as the continuous space projection.

Network is another type of projection available in Repast symphony that can be used as a tool to work with social networks or graphic connection between agents. A network projection allows agents to have network link relationships by adding agents as nodes and their relationship as edges. An edge could be friendship, a family relationship, or a colleague's relationship etc. A given agent can have more than one relationship either with the same agent or with other agents.

The last type of projection is the GIs projection which support is based on the spatial geometry like polygon, point, etc. In the GIs projection, agents are stored by type in layers within the geometry and agents of the same type must have the same geometry. When agents are queried spatially, all the agents within a particular area are retrieved. To load a GIs layer as an environment of the model, the user has to link the ABM model to the Java Topology Suite (JTS). The role of JTS is to create and import geometry objects from GIs [26]. A context can have more than one projection to represent diversified and complex interaction that occurs within agents and environment.

Another significant element of Repast is that it allows programmers to create nested timing models, in which a very complex hierarchy of simulation events can occur[83]. Agent based models are scheduled in two basic ways called synchronous and asynchronous [80]. According to Crooks [80] synchronous assumes that the behaviour of all agents change simultaneously (not in an ordered manner). In this case ordering of the agents has no influence on the simulation but conflicts can occur when agents compete over limited resources. The synchronous way of modelling updates or/and changes agents in turn by observing the reality left by the previous agent. In the last case the conflicts that could occur in asynchronous modelling are solved.

Repast adopts the asynchronous method of scheduling and represented as a discrete event simulator whose unit of time is known as a 'tick'. The tick exists only as a catch on for ordering and suspending the execution of events when it is required. For instance, if event A is scheduled to happen at tick one and event B is to occur at tick two, and event C is scheduled at tick three, then event B will be executed after A and before C. In general, ticks are just a way to order the execution of events relative to each other.

Most ABM simulating toolkits, including Repast, have the limitation that they cannot support simulation of a model on more than one computer. In other words, those toolkits are built on a single-CPU computer architectures and thus, are unable to effectively exploit increasingly available computational resources often based on parallel computing architectures[24]. According to Gasser and Kakugawa [83], the reason why Repast cannot allow simulation in parallel computing environments is that it has no communication (messaging) facility to send and receive message to/from another computers. Recognizing this limitation, researchers of the Argonne National Laboratory [84] extended Repast symphony to RepastHPC which has the required libraries to assist simulation of models on two or more computers. RepastHPC inherits the fundamental concepts of repast like context, projection and scheduling. The details of RepastHPC are described in the succeeding section.

6.1.2. RepastHPC

RepastHPC (repast for high performance computing) is an agent based modelling and simulation toolkit that was designed to be implemented on high performance distributed computing environments [85]. The computing environment is expected to contain highly performing computers like large clustered computers, supercomputer and etc.[82]. RepastHPC is implemented using C++ Programing language instead of Java which was used for RepastS.

The main objective of RepastHPC is to extend the fundamental features of Repast symphony and enable parallel computing for ABM. The concept of packaging is the main idea introduced to achieve parallel processing. *Packaging* is the process of packing states of agents within a given processor, sending them via the networks to a receiver processor and unpacking the agents when they reach the destination processor.

This process is initiated when agents are 'requested' from another process to the process that generated the agent. The generator processor will answer a request by providing a copy of the original agent. If the agents is previously transferred, the copy of this agent in the other processor is updated with the borrowed agents' current information [82, 86].

The agents transferred from other processors are called non-local agents to the receiver processor and local agents for the original processor. Each processor is responsible for their local agents and can interact with agents from other processes but are not allowed to change the status of non-local agents. The reason for this is that copies are passive, so reading information from them is possible, but changes made to the copies are not transmitted back to the original. To update original agents with a change in another processor, agents must be supplied with an appropriate projection like a network projection. Projections will manage the changes to be automatically applied to original agent. The developed model is not responsible for the synchronization process; the only requirement of the programmer is proving the code for packaging pattern and adding appropriate projection to the system.

To help processors in distinguishing local and non-local agents from one another, both in context and processors, agents have a unique ID. The agent ID is generated to include a combination of four different variables including: the ID of the agent's creating processor, type of agent, the generation stage of the agent, and the ID of the processor where the agent currently available. For instance, an agent with ID of [8, 7, 3, 5] indicates that the agent was created on processor 7, it is of type '3', was the ninth agent created on that process (numbers begin with 0, so agent '8' is the ninth one), and is currently on process 5.

Since repastHPC is new and under development modelling toolkit, some features are not yet fully included. One of that features that are lacking is the built in library to load GIs data layers like polygon, point etc.

6.2. Simulation Data Preprocessing

Before starting the implementation using RepastHPC, some of the dataset required pre-processing. As discussed in chapter five, the spatial datasets are in the form of GIS geometric data called polygon Shape files. Since RepastHPC has no built in library that allows simulations using GIS layers, the data should be changed into the format that could be loaded in the simulation environment of RepastHPC. Projection in RepastHPC is a fundamental concept that determines the structure and simulation environment.

The projections that are supported by RepastHPC are network, continuous space and grid projections. Contrary to the grid projection, which uses discrete value, continues space uses floating point values to represent a given location using either 2D or 3D. Since the continuous space projection uses floating point data, the polygon shape files can be changed into point Shape files which have X and Y coordinate and be represented using the continuous space projection. Points of a particular polygon are extracted using ArcGIS. Further, the longitude and latitudinal value of the point is calculated using ArcGIS function called ADD XY function. The model also works on social network analysis, so that the network projection is also included. Using both continuous and network projection the basic components of the model are loaded into the simulation environment and the steps are discussed in the following section.

To allocate children to the closest kindergarten and primary school for infants and children respectively, the family should know the nearest school. To enable this, after getting the X and Y coordinates of houses, the nearest kindergarten and primary school is distinguished and each house is allocated to its nearest school. Thus, infants and children will be given the nearest school when they are assigned to houses.

6.3. Implementing of the Framework

Implementation of the framework is divided into four stages called: synthetic population generation, social interaction model, activity model, commuting and disease model (Figure 6:1). The whole process is similarly applied to both cities (Enschede and Hengelo).



Figure 6:1 The workflow of implementation phases

6.3.1. Synthetic Population Generation

As indicated previously, generating a synthetic population is an important element in agent based modelling and the accurate representation of agents (number and characteristics) can impact the outcome of the model. Loading the complete list of data for each individual person of the true population is not possible for reasons. Firstly, most of the time full datasets are not available due to high acquisition cost. Secondly, loading the complete population into the model demands a large amount of computing

Resources, time and power influence the performance level of the model. To solve such problems, researchers suggested techniques of generating a subset of population that can represent the population of the society accurate and efficiently. Population synthesis techniques is one of the methods used to generate such a subset of population [87].

Population synthesis technique is an algorithm that uses statistical data of the whole population in terms of the demographic, social and economic properties, to produce a complete list of a population's members, each with associated attribute data, including the geographic location of each member[88]. Mostly, the statistical datasets are in tabular form. The datasets are integrated using iterative proportional fitting, which assists estimation of the proportion of a given dataset with a desired combination of other

datasets. For example, to generate the synthetic population for simulation of infectious disease outbreak, different datasets can be combined through iterative proportional fitting to generate agents with attributes required for the simulation like age, family members, classmates, house, workplace etc.

Creating a synthetic population starts with selecting each household from the population census dataset of households and randomly assigning a house from the buildings dataset. During the allocation, each household is assigned a unique ID, family ID and age. Unique ID enables to identify a person from the rest of people while family ID helps to group people in the same family. In other words, people with the same family ID belong to the same family. Since each house is allocated a nearest kindergarten and primary school during the data pre-processing stage, when a particular person is assigned to a house, the person also inherits the nearest schools as an attribute. The reason for this is that, later children will inherit a nearest school from the household. The age of head of households the head of the household (mother) has a particular age ranging between 25 and 55 years old. Multiple households can be allocated to one building, for example in the case of apartment buildings. Individuals that are workers are also randomly assigned a workplace from the same city where they live. The processes iterates till all households have an allocated house. When the family has two parents a second head of household is created with the same family ID and house but different attributes like age etc.

When all heads of household are assigned houses, children are generated according to the census dataset. Before children are assigned to a particular household, two major checks are conducted. Firstly, the age of the mother must be checked. For instance, when the mother has an age of 25 she cannot have a child with age of 19. Secondly, the system has to ensure that all households have children. For this reason the first iteration of the process allocates all households with children. The second and following iterations assign children to certain families until the number of children matches the average number of children per family. During the allocation of families, based on age, children will be assigned to a school. If the child is in age group of infants, it will be allocated to the nearest kindergarten and if the child is in the age group 4-12, this child will inherit the nearest primary school from the household. For teenagers, a secondary school from the same city where it lives will be assigned randomly. In this stage children are assigned with Unique ID, age, family ID and house and their activity place. Since the data is organized in age groups, age is extracted from the census data. A unique ID is randomly created for every agent.

As shown in Figure 6:2, both Hengelo and Enschede have individuals in different age groups. In the diagram, city 1 stands for Enschede while city 0 is for Hengelo.

```
-----
City 1 Number of Individuals per Age Group
Infants -> (0-3 years)
                       = 6887
Children -> (4-12 years)
                       = 15617
Teenagers -> (13-18 years) = 10687
Adults -> (25-55 years)
                       = 36154
City O Number of Individuals per Age Group
Infants -> (0-3 years)
                       = 2874
Children -> (4-12 years)
                       = 8895
Teenagers -> (13-18 years) = 6940
Adults -> (25-55 years)
                       = 22295
```



Immunity level of an individual is one of the most important attributes that determines the infection probability of a person. Immunity level has a relationship with susceptibility as well as infectivity of a given person. As immunity of the person increases, susceptibility and infectivity deceases and if immunity is low the person has a high susceptibility and infectivity characteristic (see chapter 2 for more explanations). Immunity level in turn is determined by the vaccination level and the waning rate. For this implementation, waning rate is not implemented; only vaccination level of an individual is used to present immunity level.

To generate susceptibility and infectivity levels of a person using synthetic population techniques, the total number of children in different age groups is multiplied with statistical data of vaccination. The vaccination data include percentage values of each age group of children with the respective vaccination level. The output is the number of children in a certain vaccination level. The next step is selecting the specified number of children from the whole set of children and assigning the value's corresponding level of infectivity and susceptibility to the child.

For example, let us consider a particular municipality with 100 teenagers. From this group of teenagers, 10% has been vaccinated only at the first level of vaccination while 20% has a vaccination at the second level, 30% at the third level, 40% has been vaccinated at the fourth level and 50% has not taken any vaccination. Based on the procedure we followed, 10 teenagers have both susceptibility and infectivity of 75% while 20, 30 and 40 teenagers have 50%, 25% and 0% of susceptibility and infectivity respectively. Lastly, 50% of teenagers who did not have any vaccination at all are 100% susceptibility as well as infectivity. 0% susceptibility and infectivity describes that the person is fully vaccinated and immune enough to resist the infection. At this stage, individuals are already knows by their unique ID, age, family members, house they live, infectivity and susceptibility level. The next step is the identification of commuting and non-commuting people. Before assigning individuals a vaccination level, all agents are initially created as susceptible and their status is updated based on the data. Adults are considered as susceptible that so that they remain fully susceptible and infective for entire simulation.

6.3.2. Commuting

The commuting process is initiated by identifying the commuter individuals. Commuter identification depends on the commuter data that encapsulates both students and workers. Since initially all agents are allocated with an activity place from the same municipality where they live, the activity place of commuters should be changed to an activity place in the destination municipality. The allocation of activity places is random, so that each commuter is given a corresponding location in the destination Municipality randomly.

After allocating activity places, commuter agents are copied and the copies are sent to the destination address. When destination address receives the commuters, it sends the copy of reporter to sender processor. The commuting process is implemented using the package concept of RepastHPC. This process leads to a change in number of population in each municipality. Number of outgoing commuters is reduced from the original number and the number of incoming commuter is added to the total population.

6.3.3. Create Social Network

The model contains five types of social network edges namely connections between family members, connection between workers in the same department, links between officemates, connections between classmates and intimate friends. Family social network is created between individuals who have the same family ID. The creation of other type of social network edge is based on some assumptions. Initially all agents with the same activity place are collected and for children in the age interval of 0-18, average of 30

students are grouped to be their classmates and for each students four students are randomly joined to form intimate friendships. On the another hand, for adults who are in the age group 25 to 55, similarly agents with the same activity place are collected and an average of 20 individuals is collected to be from the same department. Among them, 10 agents are connected to form an office relationship.

6.3.4. Activity and Disease Model

This phase is to create a schedule that allows individuals to interact with different types of people at different time and place. The schedule is based on time block discussed in the conceptual model which divides a given day in three blocks. In the first and third blocks agents stay at home and in block two they are at their workplace. In block one and three the disease transmission between family members is checked. In block two disease transmission is checked in workplaces and schools and kindergarten (and homes for people that are not workers or too young to go to school). Disease spreading in activity place is influenced by two type of social interaction. In the case of workplace a particular individual may get infected from officemates or department relationship.

The basic point to not forget in this implementation is that the processes described above are executed by each processor separately and parallel. The result of each processor is registered separately. The following section describes the results of implementation and discussion about the results is included.

7. RESULT AND DISCUSSION

In chapter six, the implementation of the conceptual framework which described in chapter four was discussed. In this chapter, first, the implemented model will be verified. During the verification the model is tested to see if the model performs according to the specifications in the conceptual model. The setup and the results of the verification will be discussed in section 7.1. After the verification, a test is performed to check the sensitivity of the model to changes in immunity levels. The results of the sensitivity tests will be discussed in section 7.2. To proof that the framework is indeed scalable, the model will have to be checked on its scalability. These checks will be discussed in section 7.3.

7.1. Verification

In order to perform verification, a number of simulation runs had to be performed. In this research it was decided to perform two runs with two different starting points for the infection. In the first run, the origin of the disease (initial infection) was in Enschede, in the second run the initial infection was in Hengelo. This setup allows checking if the commuting is implemented correctly in both directions.

To verify the simulation it was decided to run the simulation for the duration of 1 year (total of 1085 ticks). This period was chosen because it allows for a reasonable period for the disease to spread. It was decided to start each simulation with two initial infections that were randomly chosen. This means that the initial infection can be in any age group and can be a commuter or non-commuter. Only the municipality in which the initial infection occurred was controlled.

Individuals in various age groups have different levels of susceptibility based on various factors. In the case of this model, vaccination is the factor that mainly affects the susceptibility. According to the data presented in chapter five, most children are fully vaccinated. As the percentage of vaccination coverage per region is known, the susceptibility of children can be calculated. For adults, it is much more difficult to determine the level of susceptibility, because of waning of immunity and robust of immunity after natural infection. Although this does not reflect reality, for the test runs it was assumed that all adults are susceptible. This will lead to an overestimation of the number of disease cases, but also creates a simulation with a complete spread pattern, allowing for a good analysis of the runs.

Before conducting the verification analysis, it is important to determine which factors should be analyzed. As explained in chapter 4, the model has three sub models: activity model, the interaction model and the disease model. As already described, the activity model determines which activity agents should perform at a given time step (where the agent is), the interaction model determines with which other agents interaction occurs, and the disease model controls the spread of the infection. Besides these three models, the parallel processing using different processors and interaction between the processors (commuting) should be checked during the verification. Testing the functionality of parallel processing and commuting is very important because it is essential for the scalability of the model.

An overview of the verification checks that will be performed, and the sub models they relate to, can be found in Table 7:1.

Sub Models	Are different age groups infected?	Is there infection in different time steps?	Is epidemic curve correct?	Does the disease transfer in different locations?	Spread from Hengelo to Enschede?	Spread from Enschede to Hengelo
Activity model						
Interaction Model						
Disease model						
Parallel						
processing and						
commuting						

Table 7:1 Methods to verify the functionality of sub models

Functioning of the activity model and the interaction model will be checked together as it is difficult to identify between the activity that is being performed and the interactions with agents at this location. The verification will check if infection occurs in all three different time steps (during the night, day and evening). The expectation is that transfer of infection should take place mainly during the day and evening time step. For the activity model, it should also be confirmed that the disease spreads during different types of activities (while in school, at work or at home). And the verification will also proof that the disease spreads in all different age groups (related to their different activities and corresponding interactions). As the disease status is the state variable, the functioning of this model will be done during all verification checks. To proof that the parallel processing works according to the specifications, it will be checked if the disease spreads form Enschede to Hengelo and vice versa, and special attention will be paid to the role that commuters play in this spread.

Before discussing the results of the verification, the execution and output file structure of the processors should be clear. It is important to remember that each processor undertakes the execution of three sub models and the action of the processors is independent of each other. A given processor is responsible to register the occurrence of infection inside their boundary. This means that the number of output files for each simulation run is equal to the number of processors used. This system is illustrated in Figure 7:1. The fact that each processor produces an individual output files which can be used during the verification by analyzing the combined results, or the individual results of each processor.



Figure 7:1 Output registration in parallel processing

People who can be infected are two types. The first type of people is non-commuters those who live and work in same city while the second types are commuters those who work and living in different cities. Commuters have a chance of being infected in the city they live in, or in the city where they conduct their activities. If they are infected in the morning or evening, since they are at their home and their infection will registered by the city where they live. Otherwise, if they get infected while they are in their activity place, the processor were they commuted will count their infection.

AgentID	age	houseX	houseY	tick	infection_placeX	infection_placeY
20190	28	258304	468677	1	257234	470206
20191	32	258304	468677	1	257264	470201
24637	50	256610	467966	1	255335	471867
9910	40	254983	470725	2	255292	471876
27268	36	255382	468375	2	255292	471876
30950	26	263044	470463	2	255292	471876
19378	42	259209	468748	2	255292	471876
22534	29	257461	468600	2	255292	471876
28846	45	254310	470924	2	255292	471876
12540	41	256083	472179	2	255292	471876
968	37	258944	471756	2	255292	471876
12014	37	255685	471953	2	255292	471876
45785	12	254983	470725	66	472004	259765
22535	27	257461	468600	66	255366	471876
37088	16	258944	471756	66	259297	472395
28845	51	254310	470924	66	255335	471867
12013	41	255685	471953	66	255335	471867

Figure 7:2 Information collected by the processor when infection occurs

An example of an output file is shown in Figure 7:2. The information gathered by the registering processor includes agentID, age, tick of infection, house X and Y coordinates, and infection place X and Y coordinates. This makes it possible to analysis the place where the infection occurred.

7.1.1. Infection in Different Age Groups

Different age groups can have various types and numbers of contacts with family or people in their activity place. Confirming that agents from all groups get infected provides important information for the functionality of the activity model, the social interaction model and the disease model. The social interaction model facilitates the interaction among people in groups and the disease model controls disease transmission between people. The Figure 7:3 shows the total number of infections per age group infected in both Hengelo and Enschede, for the test run starting the infection in Enschede.



Figure 7:3 Distribution of the infection over different age groups

As can be seen in Figure 7:3, there is a clear difference in the number of infections over the different age groups. The number of infections for babies and young children (below the age of 4) is the lowest. Although babies are only partially vaccinated and have a relatively high susceptibility, they have very few social interactions, explaining the low number of disease cases. The group of children going to primary school is the second group that can be identified. This group has a very high immunity level (Low susceptibility) but has a lot more interactions (school and friends). This explains the higher number of disease cases.

In the teenage group, immunity starts to wane and their susceptibility is higher than that of children. This leads to a higher number of disease occurrences compared to children in primary school.

The group of young adults (18-24) has a lower level of infection compared to the teenagers. This group is modeled as workers (not going to school) but are also not parents yet. This group has a relatively low number of interactions with disease carriers, explaining the lower number of infections compared to the adults (parents with children).

Pertussis is often regarded to be a children's disease and the high number of adult infections may seem contradictory to this, but in this simulation, all adults were regarded as fully susceptible. This explains the high number of adult infections. In reality it is plausible that in reality their susceptibility is much lower. When the correct immunity level is implemented, the number of disease cases in this group will drop considerable.

7.1.2. The Infection in Different Time Steps

As indicated in chapter four, the simulation time (tick) is based on the classification of a given day into three blocks called evening, day and night. These blocks are represented using three consecutive ticks of simulation time. The occurrence of disease transmission in those three ticks reveals that individuals are contacting each other both at home and activity place and disease is transmitted.



Figure 7:4 Shows at the different ticks in which the infection occurred

In total 15237 infections took place in this run. Two agents were infected in the first tick to start the infection. As indicated in

Figure 7:4, no further infection took place in tick 1 (night) and all other infections occurred during the day (tick 2) and evening (tick 3). A total of 7747 infections took place during the day; this is 50.8 % of the total infections. The other half, 7487 infections (49,1%) took place in the evening.

7.1.3. Epidemic Curve

An epidemic curve is a plot of the numbers of incident cases of a disease outbreak plotted over time. The shape of the curve gives important information on the propagation of the disease. Based on the simulation output, two types of curves can be plotted: the number of new infections during a particular tick, the total number of accumulated infected agents per tick. The simulation does at this moment not register when the infected agent recovers; this makes it difficult to plot a regular epidemic curve.

The number of infections per tick is shown in Figure 7:5 below. This plot shows the number of new infections for every tick. Initially the number of new infections per tick is small. From approximately tick 600 onwards, a pattern starts to emerge with high numbers of new infections in a number of equally spaces peaks. Until tick 1000, these peaks grow in intensity (become higher) after tick 1000, their intensity declines.

Peaks in this graph represent a fundamental characteristic of epidemics called incubation and infectious period pattern. In reality, infectious individual continue to be infecting the others till the end of their infectious period however, exposed people are not yet infectious. In the case of pertussis, the average period in which people are infections is between days 8 to 21, after being infected. From this point, the infection process is less frequent in the time interval when infectious people recover and exposed people
change their status to infectious and start infecting. This is one of the reasons for the peaks in the simulation output. In Figure 7:5, the intervals with fewer infections are the time when existing infectious individuals recover and exposed people are in incubation period.



Figure 7:5 Numbers of infections in each tick



Figure 7:6 The Accumulated Number of Infection

Normally the accumulated number of infections shows a gentle slope in the beginning of the spread, to develop into a much steeper curve in the middle of the epidemic and to flatten out in the last phase of the spread. In the curve of the simulated spread (Figure 7:6), the first two phases can be recognized but the last step does not occur yet. Perhaps the duration of the simulation should have been longer. The curve also shows a stepwise line, relating to the peaks of infection seem in the previous curve (Figure 7:5)

7.1.4. Disease Transmission in Different Locations

The transmission of disease can also be checked using the location where the transmission occurs. The locations analysed in this model are home, workplace, secondary school, primary school and kindergartens. Compared to the rest of location, home is the place where people mostly spent their time and it has intimate interaction. Because of such interaction characteristics, home undertake largest infection rate. As shown in Figure 7:4, home location covers averagely 50.8 % of overall infection and workplaces utilize only 49.1%.



Figure 7:7 Disease transmissions in different activity place

Figure 7:7shows the results for individuals who live in Hengelo. The primary schools in Hengelo contribute to a relatively high number of infections. A total of 999 of the 19869 infections take place in a primary school (5%), however no infections in Enschede occur in primary schools. For the secondary schools we see that disease transfer occurs in both Hengelo and Enschede. In total, 2821 infections originate from a secondary school. This is 14% of which take place in Enschede. From the diagram it is also clear that, since students for primary schools are not commuters, the primary school map (the left) show that the infections of primary school students occur only in Hengelo. On another hand, for secondary school and workplaces, individual infected both in Hengelo and Enschede because some of them are commuters and have chance to be infected in Enschede.

7.1.5. The Spatial Extent of the Outbreak

At different time intervals, the outbreak of the disease shows different spatial patterns in both the original municipality and in the newly introduces place (other municipality). The outbreak of the disease within the given city is managed by one processors that controlling them. The disease spreading across the boundary from a given city to another is either due to incoming commuters or outgoing commuters. Through time, progress in the disease is expected to intensify both in original and newly introduced places. Plotting a time series of this progression is a way of confirming that the parallel computing works as expected.

To proof this reality, two random infections are introduced in Enschede and progression of the disease over time is displayed in Figure 7:8. The outbreak is captured for one year and this time frame is four groups representing three month blocks to analyse disease spread over time. The map shows that for the first three months the disease intensified only in Enschede and is diffused to Hengelo in the time period of six to nine months. After introduction to Hengelo, the disease continues to intensify in both cities



Infection in Hengelo and Enschede (both)

Figure 7:8 The progress of disease both in Enschede and Hengelo in one year period of time

By conduction further processing it is possible to identify which type of commuters (outgoing or incoming commuters) introduced the disease to Hengelo. To analyse this, the file registered by the processor that manages the infection that occurs in Enschede is examined. Since the data includes the house address of infected people, it is possible to display infected people based on their home address Figure 7:9 This figure reveals that infection of people who come from Hengelo is registered in the time interval of six to none months. But from Figure 7:8, it is clear that the disease is introduced to Hengelo in the time frame of three to six month. This explains that, since outgoing commuters of Hengelo (incoming commuters of Enschede) are infected in the time period of six to nine months, the disease in Hengelo is introduced by outgoing commuters of Enschede (incoming commuters of Hengelo).



Infection in Hengelo and Enschede (Enschede)

Figure 7:9 Infections of the outgoing commuters of Hengelo in Enschede

7.2. Sensitivity Analysis

To analyse the uncertainty in the outputs, sensitivity analysis is conducted by changing the vaccination input data. Initially the implementation is based on the data explained in chapter five but for the sensitivity analysis; the immunity of teenagers and adults is changed to 50% (Table 7:2).

Age group for immunity change	Vaccination level				
	None	First	Second	Third	Fourth
Teenagers	50%	0%	0%	0%	50%
Adults	50%	0%	0%	0%	50%

Table 7:2 Changes in immunity level of teenagers and adults

The process is conducted independently for each both teenagers and adults and the outcome is explained in Table 7:3.

Sensitivity	Total Number of Infected and (%)			Total number	Tick		
analysis					infections		
Change in	Infants	Children	Teenagers	Adults		Initial	Hengelo
vaccination data							Arrival
Normal	42	781	1426	12993	15242	1	269
	(0.3%)	(5.1%)	(9.4%)	(85.2%)			
Teenagers	0.5	4.5	6.6	88.4	4987	1	471
Adults	0.4	4	8.3	87.3	1664	1	740

Table 7:3 Infections during sensitivity analysis

The other important change shown in the table is that the time when the disease introduces to Hengelo. In the normal data simulation around, the disease is spread to Hengelo around the end of third months, but in the case of change in teenagers and adult, the disease reached Hengelo around the begin of fifth and eighth month respectively.

7.3. Scalability

As discussed previously, scalability is required when the number or diversity of agents, environments and or interactions increases. In this research, scalability is tested for increases in the environment (spatial extent) of the outbreak and increasing the number of agents. To conduct scalability for increasing the number of agents, the household datasets and children per age group are doubled and a new simulation is undertaken (Figure 7:10). The outcome reveals that, the time consumed by the original data is doubled (See Table 7:4). To test scalability by increasing spatial extent, two ideal cities are created by exchanging the longitude and latitude of both Hengelo and Enschede cities and two more processors are introduced. This is because increasing processors enables sharing the task of one processor over other processor so that each of them performs in a given time and simulation are same for both increasing agent and introducing two new cities, the second case simulated in less amount of time. This is because of that the data are distributed to two more processor and so the processes become fast.



Simulation data	Time elapsed (minutes)					
	First processor	Second processor	Third processor	Fourth processor		
Two Cities (Original)	9.3	9.4	No	No		
Two Cities (Doubled)	19.5	19.6	No	No		
Four Cities	12.3	12.2	12.3	12.3		

Figure 7:10 Number of individuals in different age group when the original data is double

Table 7:4 Time elapsed by each processor

From this all analysis it is undoubtedly that the model is functionally correct and can be extended to simulate more extended outbreaks.

8. CONCLUSION AND RECOMMENDATION

8.1. Conclusion

The general objective of this study was to create an agent-based framework that was suitable for all types of infectious diseases and was scalable. To achieve this main objective, sub objective were posed and the answers for the questions are described as follows. The first sub objective was to identify disease and human related factors that can influence the outbreak of an infectious disease and to develop a method to model the outbreak process. The incubation time and the infectious period are defined disease related factors while the immunity level of individuals and the social structure of the society are considered as human related factor that impact the outbreak of infectious diseases. Besides ABMs, the classical SEIR model and social network methods were selected to represent the social structure of the society.

The second sub objective was to analyse existing models and identify a suitable method to achieve scalability. Three existing models were analysed and respective strengths and weakness were recognized. At the end of existing model analysis, some characteristics of the model are identified to be adopted for this model. The method of integrating the activity model, social interactions model and disease model with synthetic population is used to generate the realistic society for the model. This model uses the concept of commuters and parallel processing to achieve scalability.

Based on the concepts acquired during the answering of first and second question, the conceptual model of this research was developed. To generate a framework for any type of infectious disease the general SEIR model was chosen as the most generic disease model. The SEIR model was integrated with a social network and an agent-based model to simulate disease outbreaks. To make the model scalable, the parallel computing method was implemented. A conceptual model was created based on three sub modes: an activity model, interaction model, and disease model. The sub models were linked to a synthetic population builder, able to generate populations based on statistical information.

The framework was created using the conceptual model and using RepastHPC modelling toolkit; which is a new Repast extension and suitable for simulation models on multiple processors. In order to make the parallel computing in RepastHPC work, a model was developed to create a copy of commuting agents in the processor they commute to, and runs either the original agent or the copy depending on the time of day. RepastHPC does not allow for copies to chance the original agents. This problem was solved via a reporter agent, which has the ability to collect information from copy agents, and transfer this information to a reporter agent in the home processor. This agent can transfer the changes to the original agent.

The model was verified via analysis of a number of test runs, proving the functionality of the model in relation to disease spread and parallel computing. Verification was very successful proofing that both the parallel computing and the sub models enabling disease spread worked appropriately. A sensitivity analysis has proven that the model is sensitive to the immunity level of the agents.

In order to test the scalability extra runs were conducted in which the number of agents was increased and the performance of the model was checked.

8.2. Recommendation

Despites its strengths, this research has also some limitations and further studies should consider them to achieve a far better model. The first and obvious limitation which is clearly stated in conceptual model is that the model concentrates only on part of social activities such as home, schools and workplaces. In reality people's interactions are complex and not restricted to such activities only. The next implementation has to consider extended activities like shopping centres, entertainment places, interaction in public transportation area etc. Though their interaction time is short they can contribute for highly contagious disease like influenza. Additionally, the sensitivity analysis has shown that the model is sensitive to the immunity level of the adults. Data should be collected, or methods should be developed to more accurately estimate the immunity level of adults. After doing this, the output of the model could be compared to actual disease outbreaks. This method also needs to be extended for commuters. Currently, the number of commuters is randomly chosen but to have a realistic representation the correct percentage of commuters should be collected and used.

RepastHPC

From the implementation and verification sections, it is also clear that RepastHPC is a very promising platform (toolkit) to implement parallel processing for agent based modelling. However, it is not an easy toolkit because of the lack of tutorials and missing documentation on the required updated functionalities. It took a lot of time to implement the code and fix the problems faced with RepastHPC. Help provided from developers via the user forum and example solutions (code) from experienced developers contributed to solving a lot of problems. Despite the help, there are still the problems of documentation and introducing supportive tutorial should be considered and improved.

Parallel processing

It is undoubtedly that parallel processing is a suitable method to achieve scalability. But to utilize the advantages of parallel processing, some fundamental technical issues must be given appropriate attention. If the load of the simulation task that is distributed over the processors is not balanced, a processor may finish an execution faster than other processors and this may lead to a delay in communication between them. Furthermore, since processor are executing independent of each other, if one of the processor faces an error, the other processor continues execution till the time when they need the result of the other processor. Proper synchronization and continues check-up method should be implemented to solve such problems.

Disease Model

Within the given time frame, it was not possible to test if the implemented SEIR model would also allow running for a different disease. This test could have proven that the model is generic. In order to test this, users would have to change the time periods of incubation, and infectious period of the respective disease to be modelled. Thought, the conceptual model includes waning model, it is not included in the implementation Because of this, no long runs (larger number of years) could be conducted. In case this would be implemented the disease should be re-initiated at certain time intervals and the long term disease patterns could be evaluated. Currently the health unit and vaccination zones are also not implemented yet. Agents that become ill do not stop their activities.

Model flexibility and Complexity

The model is currently simulating three fixed time steps. People normally have multiple activities (possibly at different locations) over de period of day and evening. This is at this moment not possible. Furthermore the model should also be extended to include the weekend activities.

References

- 1. WHO, Tuberculosis Global Facts. 2010.
- 2. Dawood, F.S., et al., *Estimated global mortality associated with the first 12 months of 2009 pandemic influenza* A H1N1 virus circulation: a modelling study. The Lancet infectious diseases, 2012. **12**(9): p. 687-695.
- 3. Organization, W.H., *Weekly epidemiological record: relevé épidémiologique hebdomadaire*, 2009, World Health OrganizationGenebra, Suica, CH.
- 4. Simons, E., et al., Assessment of the 2010 global measles mortality reduction goal: results from a model of surveillance data. The Lancet, 2012. **379**(9832): p. 2173-2178.
- 5. Keeling, M.J. and P. Rohani, *Modeling infectious diseases in humans and animals*. 2008: Princeton University Press.
- 6. Colizza, V., et al., *Epidemic modeling in complex realities*. Comptes rendus biologies, 2007. **330**(4): p. 364-374.
- 7. Hethcote, H., *Three Basic Epidemiological Models*, in *Applied Mathematical Ecology*, S. Levin, T. Hallam, and L. Gross, Editors. 1989, Springer Berlin Heidelberg. p. 119-144.
- 8. Perumalla, K.S. and S.K. Seal. Reversible parallel discrete-event execution of large-scale epidemic outbreak models. in Proceedings of the 2010 IEEE Workshop on Principles of Advanced and Distributed Simulation. 2010. IEEE Computer Society.
- 9. Organization, W.H., Using climate to predict infectious disease outbreaks: A review. Geneva: World Health Organization, 2004.
- 10. Epstein, J.M., *Toward a containment strategy for smallpox bioterror: an individual-based computational approach.* 2004: Brookings Institution Press.
- 11. Ferguson, N.M., et al., *Planning for smallpox outbreaks*. Nature, 2003. **425**(6959): p. 681-685.
- 12. Taylor, N., Review of the use of models in informing disease control policy development and adjustment. 2003.
- Kermack, W.O. and A.G. McKendrick, *Contributions to the mathematical theory of epidemics. II. The problem of endemicity.* Proceedings of the Royal Society of London. Series A, 1932. 138(834): p. 55-83.
- 14. Bansal, S., B.T. Grenfell, and L.A. Meyers, *When individual behaviour matters: homogeneous and network models in epidemiology*. Journal of The Royal Society Interface, 2007. **4**(16): p. 879-891.
- Bonabeau, E., Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences of the United States of America, 2002. 99(Suppl 3): p. 7280-7287.
- 16. Crooks, A., C. Castle, and M. Batty, *Key challenges in agent-based modelling for geo-spatial simulation*. Computers, Environment and Urban Systems, 2008. **32**(6): p. 417-430.
- 17. Barrett, C.L., et al. *EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks.* in *Proceedings of the 2008 ACM/IEEE conference on Supercomputing.* 2008. IEEE Press.
- 18. Newman, M.E., Spread of epidemic disease on networks. Physical review E, 2002. 66(1): p. 016128.
- 19. Perrin, D. and H. Ohsaki. A parallel approach to social network generation and agent-based epidemic simulation. in Proceedings of the Ninth Australasian Symposium on Parallel and Distributed Computing-Volume 118. 2011. Australian Computer Society, Inc.
- 20. Gimblett, H.R., Integrating geographic information systems and agent-based technologies for modeling and simulating social and ecological phenomena. 2002: p. 1-20.
- 21. Rana, O.F. and K. Stout. What is scalability in multi-agent systems? in Proceedings of the fourth international conference on Autonomous agents. 2000. ACM.
- 22. Borkowski, M., B. Podaima, and R. McLeod, *Epidemic modeling with discrete-space scheduled walkers:* extensions and research opportunities. BMC Public Health, 2009. **9**(Suppl 1): p. S14.
- Aaby, B.G., K.S. Perumalla, and S.K. Seal. *Efficient simulation of agent-based models on multi-gpu and multi-core clusters.* in *Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques.* 2010. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

- 24. Tang, W. and S. Wang, HPABM: A Hierarchical Parallel Simulation Framework for Spatially-explicit Agent-based Models. Transactions in GIS, 2009. **13**(3): p. 315-333.
- 25. Girmay, N.H., Performance And Scalability Of Geographically-Explicit Agent-Based Disease Diffusion Models, 2012.
- 26. Abdulkareem, S.A., Simulating the spread of pertussis in enschede region using agent based modelling, 2010.
- 27. Perez, L. and S. Dragicevic, *An agent-based approach for modeling dynamics of contagious disease spread*. International Journal of Health Geographics, 2009. **8**(1).
- 28. Thrift, N., et al., *Spatial diffusion an historical geography of epidemics in an island community*. New Zealand Geographer, 1982. **38**(2): p. 92-92.
- 29. Evans, A.S. and R.A. Kaslow, Viral infections of humans: epidemiology and control. 1997: Springer.
- 30. Read, J.M., K.T. Eames, and W.J. Edmunds, *Dynamic social networks and the implications for the spread of infectious disease*. Journal of The Royal Society Interface, 2008. **5**(26): p. 1001-1007.
- Keeling, M.J. and B.T. Grenfell, Understanding the persistence of measles: reconciling theory, simulation and observation. Proceedings of the Royal Society of London. Series B: Biological Sciences, 2002.
 269(1489): p. 335-343.
- 32. Hethcote, H.W., *Simulations of pertussis epidemiology in the United States: effects of adult booster vaccinations.* Mathematical Biosciences, 1999. **158**(1): p. 47-73.
- 33. Haggett, P., The geographical structure of epidemics. 2000: Oxford University Press.
- 34. Yang, Z., A.-X. Cui, and T. Zhou, *Impact of heterogeneous human activities on epidemic spreading*. Physica A: Statistical Mechanics and its Applications, 2011. **390**(23): p. 4543-4548.
- 35. Scheinkman, J.A., *Social interactions*. The New Palgrave Dictionary of Economics, 2008. 2.
- 36. Kaplan, A.M. and M. Haenlein, *Users of the world, unite! The challenges and opportunities of Social Media.* Business horizons, 2010. **53**(1): p. 59-68.
- 37. Viswanath, B., et al. On the evolution of user interaction in facebook. in Proceedings of the 2nd ACM workshop on Online social networks. 2009. ACM.
- 38. Stehlé, J., et al., *Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees.* BMC medicine, 2011. **9**(1): p. 87.
- 39. Keeling, M.J. and K.T. Eames, *Networks and epidemic models*. Journal of The Royal Society Interface, 2005. **2**(4): p. 295-307.
- 40. Belik, V., T. Geisel, and D. Brockmann, *Natural human mobility patterns and spatial spread of infectious diseases*. Physical Review X, 2011. **1**(1): p. 011001.
- 41. Gonzalez, M.C., C.A. Hidalgo, and A.-L. Barabasi, *Understanding individual human mobility patterns*. Nature, 2008. **453**(7196): p. 779-782.
- 42. Balcan, D., et al., *Multiscale mobility networks and the spatial spreading of infectious diseases*. Proceedings of the National Academy of Sciences, 2009. **106**(51): p. 21484-21489.
- 43. Cliff, A. and P. Haggett, *Time, travel and infection*. British Medical Bulletin, 2004. 69(1): p. 87-99.
- 44. Dickinson, R.E., *The geography of commuting*. 1960.
- Hethcote, H.W., An age-structured model for pertussis transmission. Mathematical Biosciences, 1997. 145(2): p. 89-136.
- Mandell, L.A., et al., Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clinical infectious diseases, 2007.
 44(Supplement 2): p. S27-S72.
- 47. Rouderfer, V., N.G. Becker, and H.W. Hethcote, *Waning immunity and its effects on vaccination schedules*. Mathematical Biosciences, 1994. **124**(1): p. 59-82.
- 48. Voinov, A., Systems science and modeling for ecological economics. 2008: Academic Press.
- 49. Lim, K.-H., The SEIR Epidemic Model. 2004.
- 50. Lloyd, A.L. and V.A.A. Jansen, *Spatiotemporal dynamics of epidemics: synchrony in metapopulation models.* Mathematical Biosciences, 2004. **188**(1–2): p. 1-16.
- 51. Khalil, K.M., et al., *An agent-based modeling for pandemic influenza in Egypt*, in *Handbook on Decision Making*. 2012, Springer. p. 205-218.
- 52. Carley, K.M., et al., *BioWar: scalable agent-based model of bioattacks.* Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2006. **36**(2): p. 252-265.
- 53. Connell, R., P. Dawson, and A. Skvortsov, *Comparison of an agent-based model of disease propagation with the generalised SIR epidemic model*, 2009, DTIC Document.
- 54. Easley, D. and J. Kleinberg, *Epidemics*. 2010.

- 55. Brandes, U., *A faster algorithm for betweenness centrality**. Journal of Mathematical Sociology, 2001. **25**(2): p. 163-177.
- 56. Holzhauer, S., *Developing a Social Network Analysis and Visualization Module for Repast Models.* 2010: kassel university press GmbH.
- 57. Claude, B., D. Perrin, and H.J. Ruskin. *Considerations for a social and geographical framework for agentbased Epidemics.* in *Computational Aspects of Social Networks, 2009. CASON'09. International Conference on.* 2009. IEEE.
- 58. Andrew T. Crooks and C.J.E. Castle, *Principles and Concepts of Agent-Based Modelling for Developing Geospatial Simulations* 2006.
- 59. Yang, Y. and P. Atkinson, *Individual space-time activity-based model: a model for simulation of airborne infectious disease transmission by activity bundle simulation*. Environment and Planning B: Planning and Design, 2008. **35**(1): p. 80 99.
- 60. Barnes, D. and D. Chu, *Agent-Based Modeling*, in *Introduction to Modeling for Biosciences*. 2010, Springer London. p. 15-77.
- 61. Parunak, H.V.D., R. Savit, and R.L. Riolo. Agent-based modeling vs. equation-based modeling: A case study and users' guide. in Multi-Agent Systems and Agent-Based Simulation. 1998. Springer.
- 62. Grama, A., Introduction to parallel computing. 2003: Pearson Education.
- 63. Geist, A., *PVM: Parallel virtual machine: a users' guide and tutorial for networked parallel computing.* 1994: MIT press.
- 64. Bhardwaj, D., Parallel Computing, 2002.
- 65. Mohamed Iskandarani and A. Srinivasan, Introduction To Parallel Computing, 2008.
- 66. Wilkinson, B. and C.M. Allen, *Parallel programming*. Vol. 999. 1999: Prentice hall New Jersey.
- 67. Yang, Y., P. Atkinson, and D. Ettema, *Simulation of Infectious Disease Transmission within a City: the Application of ISTAM to Eemnes by Role-based AB Simulation*. Journal of The Royal Society Interface, 2008. **5**: p. 759 - 772.
- 68. Darwish, M., Linking an Existing ABM for Disease Diffusion with a Social Network Analysis and Visualization Module, 2012.
- 69. Eubank, S., et al., *Modelling disease outbreaks in realistic urban social networks*. Nature, 2004. p. 180 184.
- 70. Zheng, D.-F., et al., *Epidemics and dimensionality in hierarchical networks*. Physica A: Statistical Mechanics and its Applications, 2005. **352**(2–4): p. 659-668.
- 71. Ettema, D. and H. Timmermans, *Activity-based approaches to travel analysis:*[selected papers presented at the workshop, May 1995]. 1997.
- 72. Hanski, I., Metapopulation ecology. 1999: Oxford University Press.
- 73. Colizza, V. and A. Vespignani, *Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations.* Journal of theoretical biology, 2008. **251**(3): p. 450-467.
- 74. Rengert, G.F., A.R. Piquero, and P.R. Jones, *Distance decay reexamined*. Criminology, 1999. **37**(2): p. 427-446.
- 75. Jain, P.K., W. Huang, and M.A. El-Sayed, On the universal scaling behavior of the distance decay of plasmon coupling in metal nanoparticle pairs: a plasmon ruler equation. Nano Letters, 2007. 7(7): p. 2080-2088.
- 76. Grimm, V., et al., *The ODD protocol: a review and first update*. Ecological Modelling, 2010. **221**(23): p. 2760-2768.
- 77. RIVM Immunisation Programme, The national immunisation programme in the netherlands. 2010.
- 78. Macal, C.M. and M.J. North. *Tutorial on agent-based modeling and simulation*. in *Proceedings of the 37th conference on Winter simulation*. 2005. Winter Simulation Conference.
- 79. Tobias, R. and C. Hofmann, *Evaluation of free Java-libraries for social-scientific agent based simulation*. Journal of Artificial Societies and Social Simulation, 2004. **7**(1).
- 80. Crooks, A.T., The repast simulation/modelling system for geospatial simulation. 2007.
- 81. Howe, T.R., et al., *Containing agents: contexts, projections, and agents*, 2006, Argonne National Laboratory (ANL).
- 82. Collier, N., Repast HPC manual, 2013.
- 83. Gasser, L. and K. Kakugawa. MACE3J: fast flexible distributed simulation of large, large-grain multi-agent systems. in Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2. 2002. ACM.
- 84. Team, R.D.; Available from: <u>http://repast.sourceforge.net/repast_hpc.html</u>.

- 85. Collier, N. RepastHPC Tutorial. 2013; Available from: http://repast.sourceforge.net/hpc_tutorial/TOC.html.
- 86. Murphy, J.T., *Computational social science and high performance computing: a case study of a simple model at large scales.* Argonne National Laboratory, Argonne, Illinois, USA, Tech. Rep, 2011.
- 87. Beckman, R.J., K.A. Baggerly, and M.D. McKay, *Creating synthetic baseline populations*. Transportation Research Part A: Policy and Practice, 1996. **30**(6): p. 415-429.
- 88. Ryan, J., H. Maoh, and P. Kanaroglou, *Population synthesis: Comparing the major techniques using a small, complete population of firms.* Geographical Analysis, 2009. **41**(2): p. 181-203.