A GLOBAL STOCHASTIC MODELING FRAMEWORK TO SIMULATE AND VISUALIZE EPIDEMICS

Saratchandra Indrakanti

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

May 2012

APPROVED:

Armin R Mikler, Major Professor
Chetan Tiwari, Committee Member
Xiaohui Yuan, Committee Member
Rajeev Azad, Committee Member
Barrett Bryant, Chair of the Department of Computer Science and Engineering
Costas Tsatsoulis, Dean of the College of Engineering
James D. Meernik, Acting Dean of the Toulouse Graduate School

Indrakanti, Saratchandra. A global stochastic modeling framework to simulate and visualize epidemics. Master of Science (Computer Science), May 2012, 72 pp., 9 tables, 26 illustrations, 28 titles.

This thesis proposes a framework to simulate and visualize the spread of an infectious disease in a population of a region such as a county. As real-world populations have a non-homogeneous demographic and spatial distribution, this framework models the spread of an infectious disease based on population of and geographic distance between census blocks; social behavioral parameters for demographic groups. The population is stratified into demographic groups in individual census blocks using census data. Infection spread is modeled by means of local and global contacts generated between groups of population in census blocks. The strength and likelihood of the contacts are based on population, geographic distance and social behavioral parameters of the groups involved. The disease dynamics are represented on a geographic map of the region using a heat map representation, where the intensity of infection is mapped to a color scale.

This framework provides a tool for public health personnel and epidemiologists to run what-if analyses on disease spread in specific populations and plan for epidemic response. By the means of demographic stratification of population and incorporation of geographic distance and social behavioral parameters into the modeling of the outbreak, this framework takes into account non-homogeneity in demographic mix and spatial distribution of the population. Generation of contacts per population group instead of individuals contributes to lowering computational overhead. Heat map representation of the intensity of infection provides an intuitive way to visualize the disease dynamics.

ACKNOWLEDGMENTS

I am very thankful to my advisor, family, friends and colleagues in the lab for all the support and encouragement I have received through the course of my Masters. In particular, I am grateful to my major professor, Dr. Armin Mikler for introducing me to the field of computational epidemiology and constantly guiding me through the course of thesis work. I would not have completed this work if not for the guidance and motivation I received from him. I really appreciate all the time Dr. Mikler spent discussing challenging ideas with me. I would like to thank all of my committee for the time they spent on my thesis. I am also thankful to all the members of CERL for all always being available to give me valuable suggestions and feedback. I would like to thank Tamara Schneider for suggesting and helping me with GeoTools and Census data. I would like to thank my parents for always encouraging me to pursue my goals. They always inspired and supported me to aspire for higher goals. Last but not the least; I would like to thank all my family friends for the invaluable support and motivation I received.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

An epidemic is the occurrence in a community or region of cases of an illness, specific health-related behavior, or health-related events clearly in excess of normal expectancy as defined by Merrill and Timmreck [21]. Though it is not a requirement for an epidemic to be contagious, most epidemics that are caused by infectious agents are. Propagated epidemics arise from the disease being transmitted from an infected individual to a non-infected one. Disease transmission can occur either by direct or indirect modes of transmission. Direct transmission is the direct transfer of the disease-causing agent through direct physical contact or direct person-to-person contact. Indirect transmission occurs when disease-causing agents are carried by some intermediate entities such as droplets or dust particles in case of airborne diseases, drinking water or water-bodies in case of waterborne diseases, or fomites such as utensils, clothing or shared items in case of vehicle-borne diseases [21]. In the context of a localized area, epidemics occur when the rate of transmission of the infection is exponential, resulting in a significant raise in the number of infected people in the area.

History is strewn with epidemics of varying scales, ranging from local outbreaks to pandemics that spread to various continents, causing significant human and monetary losses. The Black Death caused by Bubonic plague during 1338 1351, that spread over Europe and Asia with an estimated death toll of 100 million [17] and the 1918 flu pandemic that took an estimated 75 million lives worldwide [25] are some of the most devastating pandemics in recorded human history. The 2009 flu pandemic that took a toll of about 14,286 worldwide during 2009-2010 [4] is a reminder of the threat posed by epidemics in the modern world, and the importance of planning and prevention of disease spread in the context of globalization and international travel [12] [15].

Merrill and Timmreck [21] defined epidemiology as the study of the distribution and determinants of health-related states or events in human population and the application of this study to prevention and control of health problems. Epidemiologists study the occurrence,

frequency and pattern of diseases in specific populations, identifying individuals and populations at risk of contracting a disease, among others. Epidemiologists work with public-health professionals towards developing public-health policy and response plans that contribute to preventing and controlling disease spread. What-if analyses on epidemic outbreaks aid the processes of planning the response for an impending epidemic outbreak and decision making in the case of an epidemic.

The ever-changing demographics of communities and transportation options influence the pattern and rate of disease spread in a given population. The characteristics of disease spread vary not only for different geographic regions but also for the same region with changing times due to variation including changes in demographics and advancements in transportation infrastructure among others. An Influenza epidemic in a region today would evolve rather differently from one in the same region a century ago owing to a varied population distribution, improved awareness and altered interpersonal interaction patterns. With the prevalence of emerging and re-emerging diseases that have differing infection characteristics, developing epidemic response and public-health policies based on historic information is not reliable [24] [14]. Additionally, historic data on epidemics often suffers from under-reporting and inaccuracies due to insufficient training and infrastructural issues. The lack of accurate data on the basis of which decisions and plans can be made, calls for infection spread models and simulations that predict the pattern, frequency of the spread of a disease.

Various disease spread models have been developed to study epidemic outbreaks in a population [24] [22] [23]. Mathematical models based on compartmental models of epidemics help to predict infection spread in homogenous populations by means of stochastic or deterministic equations. In order to model the spread of infectious diseases in a population, it is important to consider the non-homogeneity and spatial distribution of the population of a region; identify risk groups for the disease among the various demographics and the social behavior of the participating demographic groups [24]. Models based on the cellular automata paradigm make an effort towards capturing the spatial distribution of a disease [23]. More complex computational models like agent based models recognize the non-homogeneity in

the population [22].

By the means of simulation, the dynamics of an epidemic in the desired population can be studied, while taking into account the characteristics and features of the geographic region and its population. As the disease spread is influenced by a variety of real life constraints and involves a large number of stake holders, simulating it can be complex. This necessitates the use of computational tools and methodologies to perform simulations. As a consequence of the large numbers of individuals and inter-personal interactions involved, the simulation may generate sizeable amounts of data that have to be analyzed. Methods to visualize the data greatly facilitate the process of studying and analyzing the simulation [22] [14]. An intuitive way to depict the progression of an epidemic in a geographic region is to represent it on a geographic map. With the assistance of color codes to denote intensity of infection in a sub-region, the spatial and temporal dynamics of disease spread in a region can be effectively studied.

1.1. Overview of Thesis

As part of this thesis, a framework to simulate an epidemic in the population of a region and visualize the disease dynamics has been developed. Census data provided by the US Census Bureau is used to capture the spatial distribution of population in the region. The population is classified into demographic groups on the basis of disease-specific risk groups, such as age or gender. Disease spread is simulated by the means of contacts between sub-populations. The contact model is based on a global stochastic field simulation paradigm where stochastic infectious contacts are generated between sub-populations of different census blocks on the basis of population and geographic distance between census blocks [24]. The simulation is depicted on a geographic map of the region to provide visualization of the disease dynamics in order to aid the analysis of the epidemic. A heat map representation with a color scheme that maps a color to the intensity of infection at any time step in a census block is used for the visualization.

The focus of this thesis is on the modeling of propagated epidemics. Epidemics resulting from infectious diseases that spread through transmission of the pathogen from

one person to another by the means of interactions, either physical or via a medium are modeled and analyzed. The SEIR compartmental model of epidemiology is used to classify the population into compartments based on infection state and formulate rules for disease transmission.

Disease spread is simulated by generating local and global infectious contacts between sub-populations in individual census blocks. Infectious contacts are generated randomly between the infected populations of every demographic group in a census block and susceptible populations of the region. Demographic groups to be contacted while making infectious contacts and their locations are selected randomly with the help of social behavioral constraints of participating groups, population and distance between census blocks they are located in.

Disease dynamics are depicted using a heat map representation. At the end of every time step of the simulation, intensity of infection in each census block is shown on the regions geographic map using a color scheme. Geometry of the region as provided by the census data is used to generate its geographic map. Methods from the GeoTools code library [5] are used to render the map and depict the color scheme of the heat map representation.

The simulator has been built using an object oriented methodology and written in Java. The census data is retrieved from Postgres databases using PostgreSQL queries. Java Database Connectivity (JDBC) technology is used for database connectivity from the Java classes. Methods from GeoTools [5], a Java based open source code library are used for the visualization of disease dynamics.

1.2. Contribution

The epidemic outbreak simulator developed as part of this thesis, provides a modeling tool for public health professionals and epidemiologists to study and interpret disease dynamics for a propagated epidemic in a region. By the means of what-if analyses on disease spread in a population, public health personnel can fine-tune public health policies and plan response in case of an outbreak. The simulator helps predict disease spread patterns, estimate epidemic characteristics, and study impact of the epidemic on sub-regions and sub-groups of populations. The incorporation of spatial distribution of population and social behavior of

demographic groups into the model enables more realistic modeling of the epidemic. Heat-map representation of the disease spread and intensity of infection on a geographic map of the region provides an intuitive method to analyze the large amount of data produced by the simulation. In summary, this thesis aims at providing a stochastic method to simulate an epidemic in a region taking into account the non-homogeneity and spatial distribution of its population, and an intuitive way to visualize and analyze the disease dynamics.

CHAPTER 2

BACKGROUND

2.1. Epidemic Theory

An epidemic is defined as an unusually high occurrence of a disease or illness in a population or area [3]. The term outbreak, which is often used synonymously with epidemic, refers to an epidemic confined to a localized area. An ongoing, usual, or constant presence of a disease in a community or among a group of people is referred to as an endemic, while a pandemic is an epidemic affecting or attacking the population of an extensive region, country or continent [3] [21].

The contemporary usage of the term epidemic is made to describe a relative excess of diseases under a wide variety of conditions. These include both communicable infectious diseases like influenza or cholera and non-communicable diseases such as breast cancer, or physical conditions such as obesity [18]. Infectious-disease epidemics can be classified into common-source epidemics and propagated epidemics based on how they spread. Common source epidemics arise from a specific source such as contaminated food, while propagated epidemics, such as tuberculosis or influenza arise from infection transmitted from one person to another. A mixed epidemic occurs when inter-personal contacts among hosts of common-source epidemics leads to a propagated epidemic [21] [18].

Epidemiology is the study of the distribution and determinants of health-related states or events in human populations and the application of this study to the prevention and control of health problems. The history of epidemiology dates back to Hippocrates, when he made observations regarding occurrence of diseases. Daniel Bernoullis mathematical model to evaluate effectiveness of techniques of variolation against smallpox, William Farrs fitting of quarterly smallpox data to a normal curve were some of the early applications of mathematics in epidemiology [19]. John Snows investigation of the 1854 London cholera outbreak is considered one of the most important contributions to the field of epidemiology [21] [19]. Various infectious disease spread models using different approaches and modeling paradigms

have been developed since.

Disease-causing agents and hosts are important elements of epidemics caused by communicable diseases. Agents, such as bacteria or viruses are the cause of the disease, while hosts such as humans are the organisms that harbor the disease. Infectious-disease epidemics are associated with timelines which govern the temporal nature of an epidemic. The time period between a hosts exposure to a disease causing agent and the host becoming infectious, i.e. transmitting the infection is called the latent period or the period of latency. The duration for which an infected host can transmit the disease-causing agent to a susceptible host is referred to as the infectious period. The disease-causing pathogen is associated with a value of infectivity, which refers to the ability of a pathogen to establish an infection in a host population.

2.2. Disease Compartments

At any time during the course of an epidemic, there are people who are at different levels of exposure to the disease. While some of them are yet to contract the disease, some of them may be infectious, actively transmitting the infection, while others are recovered and attain immunity to the disease. The host population is classified into different groups or compartments with respect to the state of infection they are in and the nature of the disease involved.

One of the conventional models of epidemiology is the SIR model where the population is classified into susceptible, infectious and recovered compartments. Individuals who are at risk of contracting the infection are placed in the susceptible compartment. Individuals who contracted the infection and actively transmit it are classified as infectious, while individuals who either recovered and gained immunity or succumb to the disease are placed in the recovered (or removed) compartment. These compartments are modified depending on the disease being studied. For diseases like influenza, where there is a period of latency involved, an additional latent (or exposed) compartment is used. In the case of certain diseases such as common cold, against which the hosts do not gain immunity, the host population is classified into either susceptible or infectious [11] [7].

2.3. Epidemic Modeling

Various attempts have been made at modeling epidemics through the years. These range from mathematical models that use simple deterministic equations to agent-based computational models that make use of heavy computational resources. Some of the significant and relevant epidemic-modeling approaches are reviewed below:

2.3.1. Mathematical Models

Mathematical models have been used in modeling epidemics since the early 20^{th} century. Many of them are based on time-dependent differential equation systems. Deterministic equations were used to model transition of hosts between disease compartments in the initial models. Later on, binomial distributions and other probabilistic aspects were made use of to represent disease spread bringing a stochastic dimension to epidemic modeling [7]. Mathematical models can be classified into deterministic and stochastic models, based on the inclusion of probabilistic functions into the model.

2.3.1.1. *Deterministic Models.* Deterministic models use differential equations, which govern the movements of population between disease compartments, to model the spread of an epidemic. The classic SIR model, which is based on the Kermack-McKendrick threshold theorem [10], assumes population equilibrium and ignores changes to population due to migration, births or deaths owing to the comparatively shorter duration of the epidemic. The following differential equations govern the population dynamics of the model:

$$\begin{aligned} \frac{\mathrm{d}\,S}{\mathrm{d}\,t} &= -\beta I S \\ \frac{\mathrm{d}\,I}{\mathrm{d}\,t} &= \beta I S - \upsilon I \\ \frac{\mathrm{d}\,R}{\mathrm{d}\,t} &= \upsilon I \\ \frac{\mathrm{d}\,S}{\mathrm{d}\,t} + \frac{\mathrm{d}\,I}{\mathrm{d}\,t} + \frac{\mathrm{d}\,R}{\mathrm{d}\,t} &= 0 \end{aligned} \tag{1}$$

Where $S(t)$, $I(t)$, $R(t)$ represent susceptible, infectious and recovered (or removed) compartments respectively at time t, β is the infection rate and υ is the rate of recovery.

Variations of this model include the SEIR model where latent period of the disease is considered and the SIRS model where the recovered hosts can contract infection again, among others. These models assume a homogeneous population mix, and do not consider the spatial distribution of the population [24].

2.3.1.2. *Stochastic Models.* Factors such as the environment or demographics contribute to the inherent variability that exists in the system for which the disease spread is being modeled [9]. In the last few decades, the need to incorporate stochasticity into epidemic models to represent variability in the system has been well recognized [20]. A stochastic model is expressed as a stochastic process with a collection of random variables. The solution of a stochastic model is a probability distribution for each of the random variables [9]. Stochastic models were discussed in literature in the early 20^{th} century, alongside deterministic ones. Early stochastic models were based on discrete-time systems and chain binomial models.

The Reed-Frost model [8] is one of the well-known chain binomial models. The number of susceptible individuals in the population at the end of a time interval is represented as a binomial distribution in the Reed-Frost model. During the time interval $(t, t+1)$, I_t infectives infect S_t susceptibles, where each susceptible individual infects an infective individual with a probability p. Given S_t and I_t, the number of susceptibles at time $t+1$, S_{t+1} is binomially distributed with index S_t and mean $S_t(1-p)^{I_t}$.

There have been various attempts at producing variations of the continuous time stochastic SIR model, by adding stochastic terms to the deterministic SIR model, such as:

$$\frac{\mathrm{d}\, E\{I(t+dt)\}}{\mathrm{d}\, t} = \left(\frac{\alpha E[S(t)]}{n} - \gamma\right) E[I(t)] + \frac{a}{n} cov\{S(t), I(t)\} \tag{2}$$

where,

$$cov\{S(t), I(t)\} = E\left[\left(S(t) - E[S(t)]\right)\left(I(t) - E[I(t)]\right)\right] \tag{3}$$

It can be seen that in a deterministic case, the expected values $E[S(t)]$ and $E[I(t)]$ are equal to $S(t)$ and $I(t)$ respectively and the effect of covariance term is nullified. However,

when stochasticity is incorporated into the model and the variables are based on probability distributions, the covariance term comes into play, affecting the final outcome.

Though in many systems deterministic solution turns out to be a good approximation to the stochastic mean in case of modeling an outbreak in a large population, it is widely accepted that fluctuations due to stochasticity do not always average out, thus leaving an effect on the final outcome [20]. By introducing probability metrics, stochastic models make an attempt towards a closer real life modeling of epidemics [7].

2.3.2. Computational Models

Epidemiological studies often involve large populations and make use of large datasets to model epidemic outbreaks. The use of stochastic models that take into account non-homogeneity and spatial distribution of population, make it imperative to use computational resources to model epidemics. Over the past few decades, various computational modeling paradigms have been applied to model epidemics. Stochastic models based on Markov chain processes and Monte Carlo techniques, cellular automata modeling paradigm and agent-based models are some of the significant computational models in literature

2.3.2.1. *Cellular Automata.* A cellular automaton is defined in Wolfram Mathworld as a collection of cells on a grid of specified shape that evolves through a number of discrete time steps according to a set of rules based on the states of neighboring cells [27]. Cellular automata have been used for several decades for computational modeling in the sciences [26].

Cellular automata mainly depend on the dimensionality of the grid on which they are modeled. One, two and three dimensional grids have been used in modeling. Each cell in a one dimensional cellular automaton has a left and a right neighbor. Cells in two-dimensional automata usually have four or eight cells in their neighborhood depending on von Neumann neighborhood or Moore neighborhood respectively. Each cell is associated with a state, and its state in the next generation is computed according to fixed rules using the states of cells in its neighborhood [6].

In modeling epidemics, two-dimensional cellular automata with each cell representing an individual and state of the cell corresponding to the individuals disease-state (i.e. suscep-

tible, infectious or recovered) have been traditionally used. Classic cellular automata suffer from neighborhood saturation, thus limiting the spread of the disease. Additionally, spread of infection in an extended neighborhood in a single time-step is not possible [23].

2.3.2.2. *Agent-Based Models.* There has been a significant amount of work on using agent-based models in computational epidemiology. In agent-based models, each individual in the population of interest is represented as an autonomous agent. Interactions between individual agents are modeled based on specific behavioral rules. An attempt is made to model the epidemic more realistically by simulating the actions of each individual involved. Consequently, agent-based models are computationally expensive and demand extensive computational resources. Systems that make use of agent-based models for large populations require high performance computing resources [24].

Some of the significant agent-based epidemic modeling efforts include BioWar [13] which is a citywide multi-agent system that analyzes disease spread based on interactions from social, health and professional networks and EpiSims [16] which uses data from TRANSIMS, which is a simulation tool for transportation systems, among others. Agent-based systems have been used on a smaller scale in studies such as analyzing tuberculosis outbreaks in homeless shelters by Mikler et al.[22].

Agent-based models have the ability to capture the real-world mixing patterns, spatial distribution and non-homogeneity of population, but are computationally expensive. They can lead to cumulative modeling errors as a result of the choice of parameters used in modeling individual behaviors.

2.3.2.3. *Global Stochastic Field Simulation.* Each of the computational modeling paradigms mentioned above have specific strengths and weaknesses. Hybrid models that are based on more than one of the above paradigms have been developed to leverage the positives of different approaches. Global stochastic field simulation (GSFS) proposed by Mikler et al. [24] is one such model.

GSFS paradigm is a hybrid of agent-based simulation and cellular automata. The geographic region is represented as a filed, which is an overlay of the spatial distribution of

population into cells of a fixed number of individual. Unlike pure agent-based models statistics of disease states for each cell are maintained. The interactions between geographic regions are stochastic and occur globally instead of a well-defined neighborhood.

CHAPTER 3

METHODOLOGY

Prior planning and preparedness for an impending epidemic outbreak requires public health professionals and epidemiologists to study and analyze the disease dynamics in the host population. Simulations of the epidemic outbreak provide a platform to perform these studies and analyses. Techniques to visualize the epidemic propagation and disease dynamics greatly help in effective and elegant interpretation of the simulations. In this section, a framework for the simulation of an epidemic outbreak and visualization of disease dynamics in the host population is introduced.

This framework must facilitate analysis of the characteristics of an outbreak, in order to aid the decision making process of public health professionals. Duration of the outbreak, numbers of infected people, sub populations more likely to be infected and their locations, pattern of disease spread and spatial distribution of infection at any given time are some of the characteristics public health professionals are interested in. Design choices pertaining to the following framework components were made so as to cater to the above requirements:

- Geographic layout and distribution of population
- Disease transmission model
- Contact model
- Social behavior model
- Visualization

As part of this thesis, a simulator that models an epidemic outbreak in a region such as a county has been developed. Census data have been used to classify population into demographic groups relevant to the disease. Global stochastic contacts that are sensitive to the spatial distribution of population are generated to simulate disease transmission based on an SEIR (susceptible-exposed-infectious-recovered) compartmental epidemiological model. Heat map representation has been used to depict disease dynamics.

The sections that follow describe the design and implementation of the simulator.

3.1. Overview

The simulator models an epidemic outbreak and displays a visualization of the disease dynamics. It draws relevant demographic data from census information of a county and models the disease spread based on initial infected population, infection parameters, demographics and spatial distribution of the population in that region. The state of the epidemic as it propagates is then depicted on a geographic map of the region to give a visual representation of the extent, pattern and intensity of the disease spread.

The user-input for the simulator include the disease parameters for the epidemic, demographic constraints based on which the epidemic has to be modeled and the region for which disease spread has to be simulated. Entities representing census blocks in the region are created and populated using census information classified on user-selected demographic constraints. These census block entities are represented on the regions geographic map with the help of GeoTools [5], a code library that provides standards compliant methods for manipulation and representation of geospatial data. The user may then add infectious population to the desired census blocks to initiate the epidemic simulation. Infectious contacts are generated randomly for every population group in each census block, on the basis of disease and demographic parameters. As the epidemic propagates through the region with time, infection state of each census block at that time-step is depicted on the geographic map using a color code representing the infection state. This provides users a visual representation of the state of infection in the region at each time-step. The simulation ends when there are no more people in the infected or latent compartments of the population. At the end of the simulation, disease dynamics, duration and extent of the epidemic, pattern of spread, census blocks prone to the infection are some of the aspects that can be studied and interpreted.

3.1.1. Input and Output

The simulator provides a platform for the user to model disease spread under different settings and helps analyze the disease dynamics through visualization. The user may experiment with different demographic and disease parameters, or initial conditions to analyze disease dynamics in various scenarios. The results produced from the simulation may be

analyzed with the help of visualization through heat map representation or by the means of SEIR plots.

The following are the inputs to be provided by the user before the start of simulation:

- Disease parameters: Infectivity, infectious and exposed periods of the disease.
- Demographic parameters: These are parameters specific to a demographic group. They include average contact rate for each demographic group, social behavioral parameters - affinity, mobility and reach.
- Census data: Census data pertaining to demographics and spatial population distribution for the geographic region of interest on a census block basis.
- Geometry: Geometry of the geographic region to enable rendering the regions map using GeoTools.
- Initial infected population: The initial infection scenario i.e. numbers of infected and exposed individuals in infected census blocks before the start of the simulation.

The graphical user interface (GUI) for the simulator facilitates users to input disease parameters for the epidemic, social behavioral parameters for desired demographic groups. Infectious or exposed/latent population can be added either by selecting target census blocks from a list or by choosing census blocks by clicking on the geographic map of the region. Disease dynamics and epidemic propagation can be studied by iterating through each time step of the simulation. At the end of each time step, the simulator outputs statistics of susceptible, exposed, infectious and recovered individuals for every sub-population as wells as cumulative numbers for the region. These per-time-step statistics can be used to develop plots to study the outbreak. Additionally, for every time step, a heat map representation of intensity of infection in each census block is rendered on a geographic map of the region. With the aid of a color scheme to represent different infection intensities, this representation helps in analyzing the dynamics of the outbreak.

3.2. Simulator Design

The simulator retrieves population distribution and demographics from census data based on user-specified demographic constraints. Epidemic propagation is simulated by means of contacts between population sub-groups. Visualization of the simulation is facilitated with the help of a heat-map representation of percentage of population that is infected.

The simulator is implemented using an object-oriented design paradigm. Population statistics pertaining to each census block in the region are stored in objects representing the census blocks. The contact model is derived from the global stochastic field simulation modeling paradigm. SEIR model is used to represent the disease compartments. Methods from GeoTools [5] are used to produce the heat-map representation.

This section describes the design of the simulator. The architecture, design choices and motivation behind those choices, along with the technologies used are detailed in this section.

3.2.1. Architecture

The simulators architecture is designed to retrieve relevant demographic data and spatial distribution of host population from data repositories; model epidemic outbreaks based on contact and disease transmission models; and provide a visualization of the disease dynamics. The core components of the architecture, as shown in Figure 3.1 include:

- Databases that provide census, geographic data
- GUI for user input
- Contact model and transmission model for epidemic simulation
- Visualization methods

3.2.1.1. *Functional Blocks.* The processes performed by the simulator to produce and visualize the epidemic as described above can be classified broadly into the categories below. Figure 3.2 shows the core functional blocks of the simulator.

- Initialization: A platform is set up for the infection spread to be simulated. User inputs infection parameters, demographic parameters and social behavioral constraints

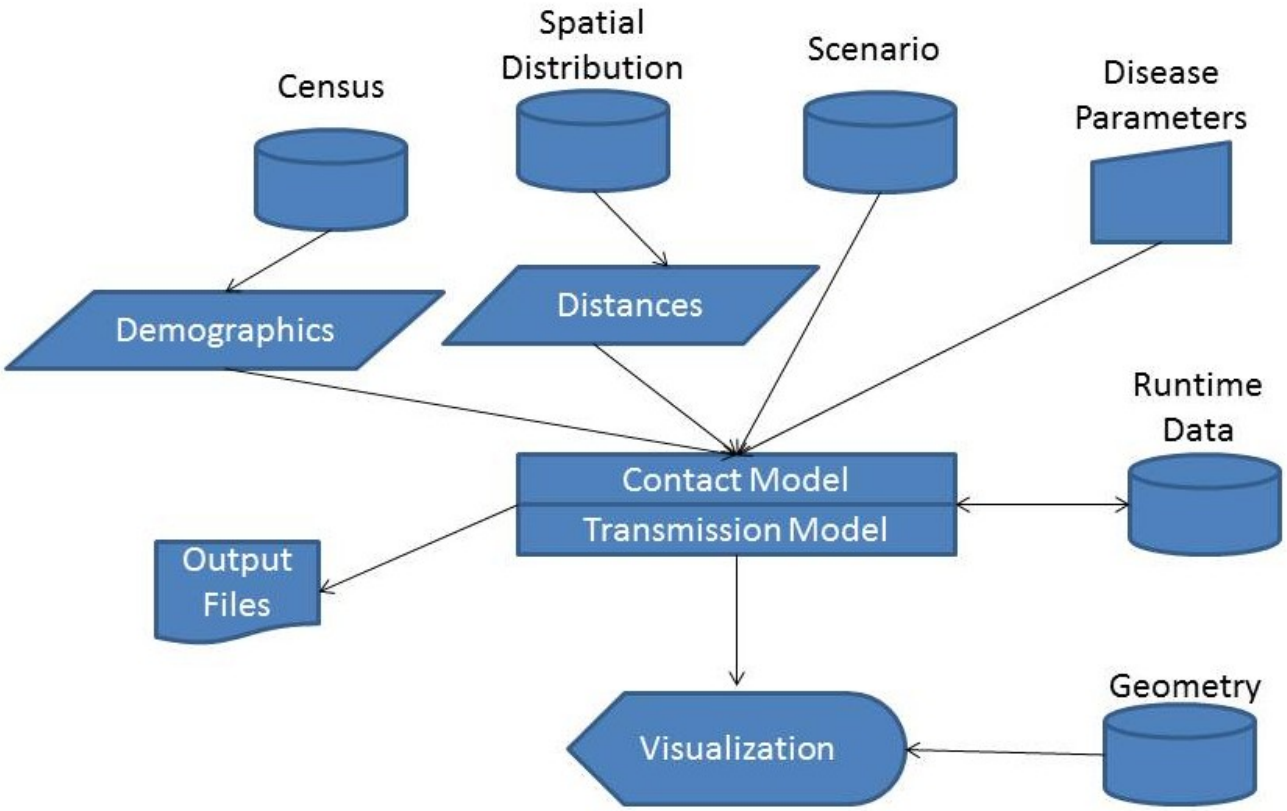

Figure 3.1. Architecture of the simulator

for demographic groups via a graphical user interface. Database connections are set up and demographic information is retrieved from the census database on the basis of user selected demographic classification. Census block objects are created to represent each census block of the region, and are populated using the census information. Necessary modifications are made to the database to hold the infection state of each census block.

- Contact model: Contacts are generated between different demographic groups within a census block and with other blocks to simulate the infection spread. Geographic distances between census blocks and population densities are made use of to choose blocks to be contacted. The contact model is executed for every time unit and the numbers of people in each disease compartment are updated at the

end of each time unit.

- Visualization: A heat map is generated to represent the intensity of infection in each census block. Methods from the GeoTools code library are made use of to represent infection intensity of each block on the geographic map of the region, at the end of each time unit. This provides a visual depiction of the disease spread in the region, and an intuitive method to grasp the disease dynamics.

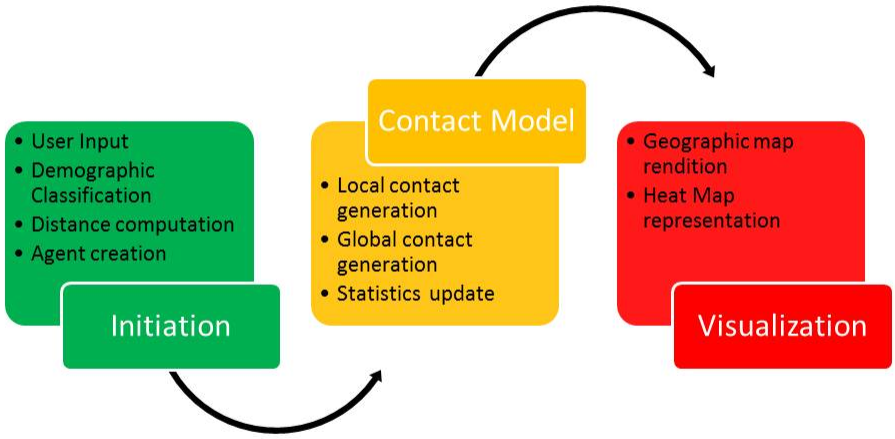

Figure 3.2. Core functional blocks of the simulator

3.2.2. Design Choices

Described in the sections that follow is the motivation behind the different design choices made in developing the simulator:

3.2.2.1. *Compartmental Model.* The SEIR compartmental model of epidemiology has been used to represent infection state. People are grouped into either of the S, E, I or R disease compartments based on their exposure to the epidemic. Within each census block object of the simulator, numbers of individuals in each compartment are maintained for every demographic group in that block. As the epidemic progresses and people move from one

infection state to another, they are moved in between the disease compartments. Changes to the disease compartments due to births, deaths or migration are not considered in this model, as these changes are insignificant in the shorter time frame of the epidemic.

This simulator focuses on diseases where the host gains immunity to the disease after recovering from the infection, i.e. diseases that adhere to the SEIR epidemiological model. Outbreak simulations for contagious diseases that spread by the means of direct or indirect interactions between infected and susceptible individuals, such as physical interactions or interactions via a medium like air, can be developed using this simulator.

3.2.2.2. *Geographic Granularity & Population Distribution.* This simulator is designed to simulate infectious disease spread for a geographic region such as a county. As census data made available by the US Census Bureau is used by the simulator, the representation of geographic entities as provided by the US Census Bureau is adopted. The hierarchical representation of geographic entities provided by the Census Bureau, starting bottom-up from a census block to a county can be seen in Figure 3.3 :

A census block, being the smallest geographic unit for which the Census Bureau tabulates cumulative data [2], has been chosen as the geographical unit at the lowest level of granularity in this thesis. A county usually consists of thousands of census blocks, with each block providing cumulative social and demographic information. Each census block is identified by its block identification number obtained from the census data. Furthermore, every census block is associated with unique latitude and longitude coordinates. Spread of the disease between different census blocks is estimated based on population counts of participating census blocks and geographic distance between them. In the simulator, a county is represented as an array of census block objects, with each object having its own data fields to hold information about the census block it represents.

Most geographic regions with significant populations have a non-homogenous population distribution. Census blocks vary in population densities ranging from vast unpopulated blocks to smaller blocks with large populations. Figure 3.4 shows the population distribution of Denton county. It can be seen that the population distribution is non-uniform, with certain

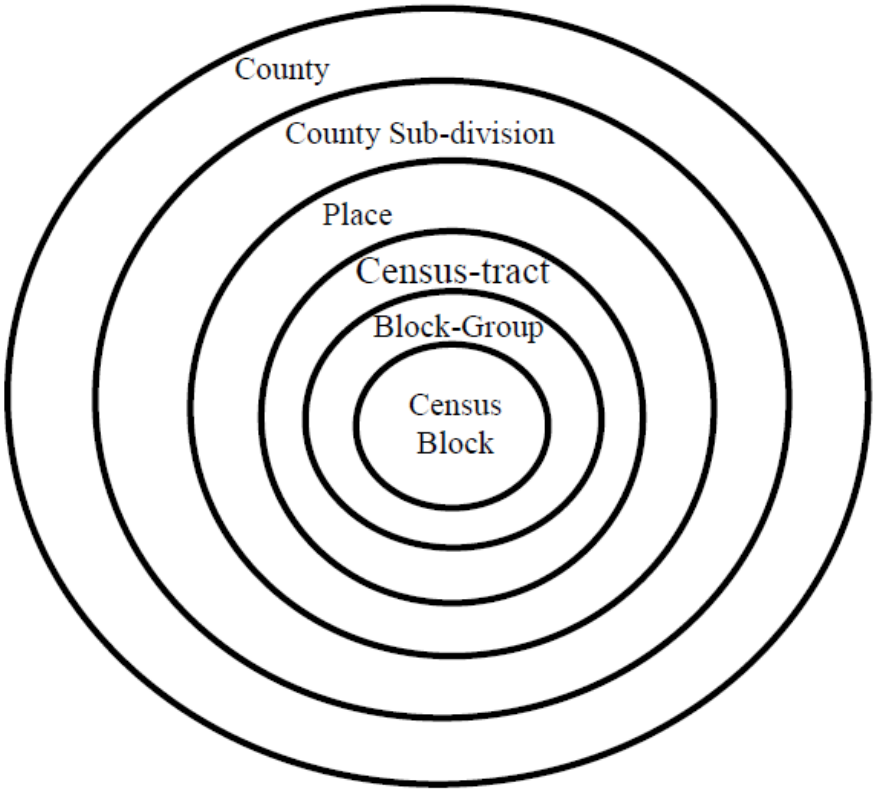

Figure 3.3. Hirerarchy of geographic entities

Hierarchy of geographic entities as represented by US Census Bureau in Census 2000.

pockets having higher population densities while others are barely populated. The spatial distribution of population in a region plays a critical role in determining the pattern of disease spread and disease dynamics. Hence, it is important to consider the non-homogeneity and spatial spread of the population while modeling an epidemic outbreak for a region.

In order to represent the actual spatial distribution and demographics of the region in the simulation, real data has to be made use of. The 2000 US Census data, made publicly available by the US Census Bureau through the American FactFinder web resource is used to obtain the demographics and spatial distribution of population up to a census block granularity for the region of interest. Within a census block, population is classified into sub-groups, referred to as demographic groups, based on disease-specific demographic constraints such as age or gender.

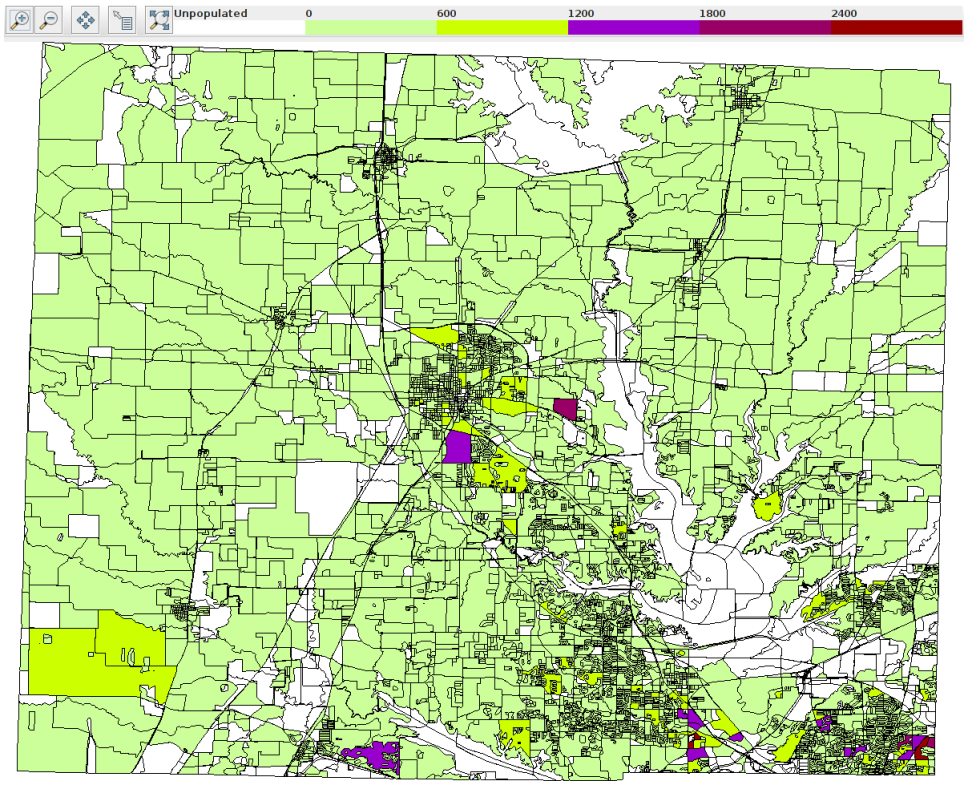

Figure 3.4. Population distribution of Denton county, Texas

Population distribution of Denton county, Texas. At the top is the color scale; darker colors represent larger population

3.2.2.3. *Modeling Social Behavior.* As mentioned above, the population is classified into demographic groups, on the basis of disease-specific demographic constraints such as age or gender. The simulator must recognize the fact that individuals belonging to different demographic groups behave differently, not only with respect to contracting the infection, but also differ in their social behavior. Social behavior of a demographic group affects the spatial and temporal nature of disease spread. For example, a disease may have different effects on people of different age groups. Children could be more prone to the infection than middle-aged or old-aged people. The way children interact with middle-aged individuals may differ from the way they interact with the old-aged ones. Old-aged individuals may tend to stay within the vicinity of their homes, while middle-aged may tend to travel farther [28]. These differences affect the way interactions between sub populations are modeled

while simulating the epidemic outbreak. Hence, the differences in social behaviors of different demographic groups must be accounted for, and accommodated in the disease spread model.

The parameters affinity, mobility and reach have been used to incorporate the differences in social behaviors into the simulator. Each demographic group is associated with user assigned values of affinities, mobility and reach. Affinity defines the likelihood of one demographic group making a contact with another. The greater the affinity between two participating demographic groups, the greater is the likelihood of the first group making a contact with the second. Mobility influences the likelihood of a demographic group making a contact within its home census block. The Smaller the value of mobility, the greater is the likelihood of a contact being made within the same census block. Reach sets a limit on the farthest census block from the home census block that can be reached by a demographic group to make a contact. Each demographic group has its user assigned contact rate. It should be noted that these parameters are defined for a demographic group, not an individual. They tend to reflect the effective average behavior of all individuals of the demographic group.

3.2.2.4. *Contact Model.* An epidemic propagates in a population by means of infectious contacts. A contact is an interaction between two individuals in the population that is conducive to the spread of the disease. An infectious contact is a contact that actually results in the disease passing on from the individual making the contact to the individual with whom the contact is made. It should be noted that an infectious contact occurs only when there is an interaction between an infected individual and a susceptible individual. In order to simulate the disease spread, infectious contacts are generated within the population.

The people whom a person contacts on any given day are not pre-determined. Based on the nature of disease transmission, the people who are being contacted may not even know that they have been exposed to the infection. For instance, in the case of certain air borne diseases, disease causing pathogens can remain suspended in air or travel distances on air currents by means of dust particles or respiratory droplets. These pathogens can be contracted by unrelated individuals. In order to model this unpredictability that exists in the

transmission of a disease in the real world, stochasticity needs to be incorporated into the contact model.

There have been different approaches to generating infectious contacts. These range from simple deterministic models like the SIR model to stochastic agent based models where contacts are generated probabilistically for every individual of the population. Deterministic models tend to assume a homogeneous population distribution, ignoring the difference in the ways different demographic groups behave during an epidemic. Agent based models, which simulate the epidemic on a per-individual basis recognize the heterogeneity in a population distribution, but model the behavior of every individual in the population resulting in a higher computational overhead.

In order to take advantage of similarities in the day-to-day routines of people belonging to the same demographic groups, at the same time recognizing the non-homogeneity in the population, a hybrid approach has been chosen. In this thesis, a stochastic approach where contacts are generated locally and globally on a per-demographic-group basis is used in an attempt towards modeling the randomness that exists in the spread of an infection in the real world. Generating contacts for each demographic-group rather than for every individual, taking advantage of similarities in behaviors of individuals belonging to the same demographic group, reduces the total number of contact generation operations. This results in a lower computational overhead and shorter turn-around time, without making a compromise on modeling heterogeneity in the population.

3.2.2.5. *Visualization.* Epidemiologists and public health personnel need to plan ahead and be prepared to mitigate the outbreak of an epidemic. Having a prior idea of the sub populations and regions likely to be effected helps in planning activities like stock piling vaccinations, creating awareness etc. For this purpose, an estimation of the duration of epidemic, numbers of infected people in different risk groups, and regions likely to be affected would be immensely helpful. A simulation of the disease spread provides the above details concerning the epidemic to aid the process of planning.

When the disease spread is simulated in a region as large as a county with thousands

of census blocks, and hundreds of thousands of people, a large amount of data is produced at each time step. It would be difficult to study the statistics corresponding to each census block every time step. This brings in the need for an elegant way to study both the individual-block-level and cumulative effect of the epidemic. In order to interpret all the data generated and to understand the disease dynamics in an effective and intuitive way, it is important to have a method to visualize the simulation. So as to serve the above purpose, this visualization mechanism must depict the following:

- Localities affected by the epidemic
- Intensity of infection within a census block and
- Spatial and temporal propagation of the infection

In this simulator, visualization is provided by means of depicting disease dynamics on a geographic map of the region where the epidemic is simulated. At each time step, intensity of infection in each census block is shown using a heat map representation where infection intensity is visualized by means of a color code. GeoTools, an open source Java code library that provides methods to render maps using geospatial data is made use of to generate and display heat maps of the region at each time step. Methods from GeoTools are used to provide an interface to add infectious population, read shape files of the region from the census database, and render a map along with intensity of infection for each block.

3.2.2.6. *Choice of Technologies.* The simulator has been implemented in Java programming language on the NetBeans Development Environment. JDBC is used to connect to a POSTGRESQL database that holds census data and the required data is retrieved from the database using POSTGRESQL queries. Methods from GEOTools code library are used to help visualize the disease dynamics on a geographic map of the county selected by the user. The program is organized into different Java classes, each with appropriate methods and variables, to perform sub tasks involved in producing the simulation.

3.3. Simulator Implementation

Various approaches have been adopted by infectious disease spread simulators to model an epidemic outbreak. In this thesis, a hybrid between the Global Stochastic Field Simulation model and an agent-based model is proposed. This simulator uses an approach where the disease spread in a region is modeled by utilizing the spatial distribution of population and social behavioral constraints of demographic groups involved. Population is distributed into different compartments on the basis of infection states and demographics. A global stochastic contact model where infection spreads by the means of local and global contacts between demographic groups in census blocks is applied. The geographic entity at the lowest level of granularity that has been used in this model is the census block, while, a group of individuals belonging to the same demographic group within a census block is the population unit at the lowest level of granularity.

Figure 3.5 shows the process flow for the simulator. The simulator is initialized by creating census block objects using demographic data retrieved from the census database. Census data and user-inputs including disease parameters, social behavioral parameters are inputted to the contact model to generate infectious contacts so as to simulate an outbreak. Contacts are iteratively generated for all population sub-groups and statistics of infection compartments are updated at the end of every time step. The intensity of infection in each census block is displayed on a heat map generated using geometry of the region. The following sections describe the implementation of these processes.

3.3.1. Compartmental Model

At any given time, different people have different levels of exposure to an epidemic. As the epidemic progresses, people move from one state of infection to another. In order to capture these differences in exposure to the epidemic, the SEIR compartmental model is used in this thesis. Within a population sub group, people are classified into one of the susceptible(S), exposed(E), infectious(I), or recovered(R) compartments based on the infection state they are in.

Individuals who are in the S compartment are susceptible to contracting the infection,

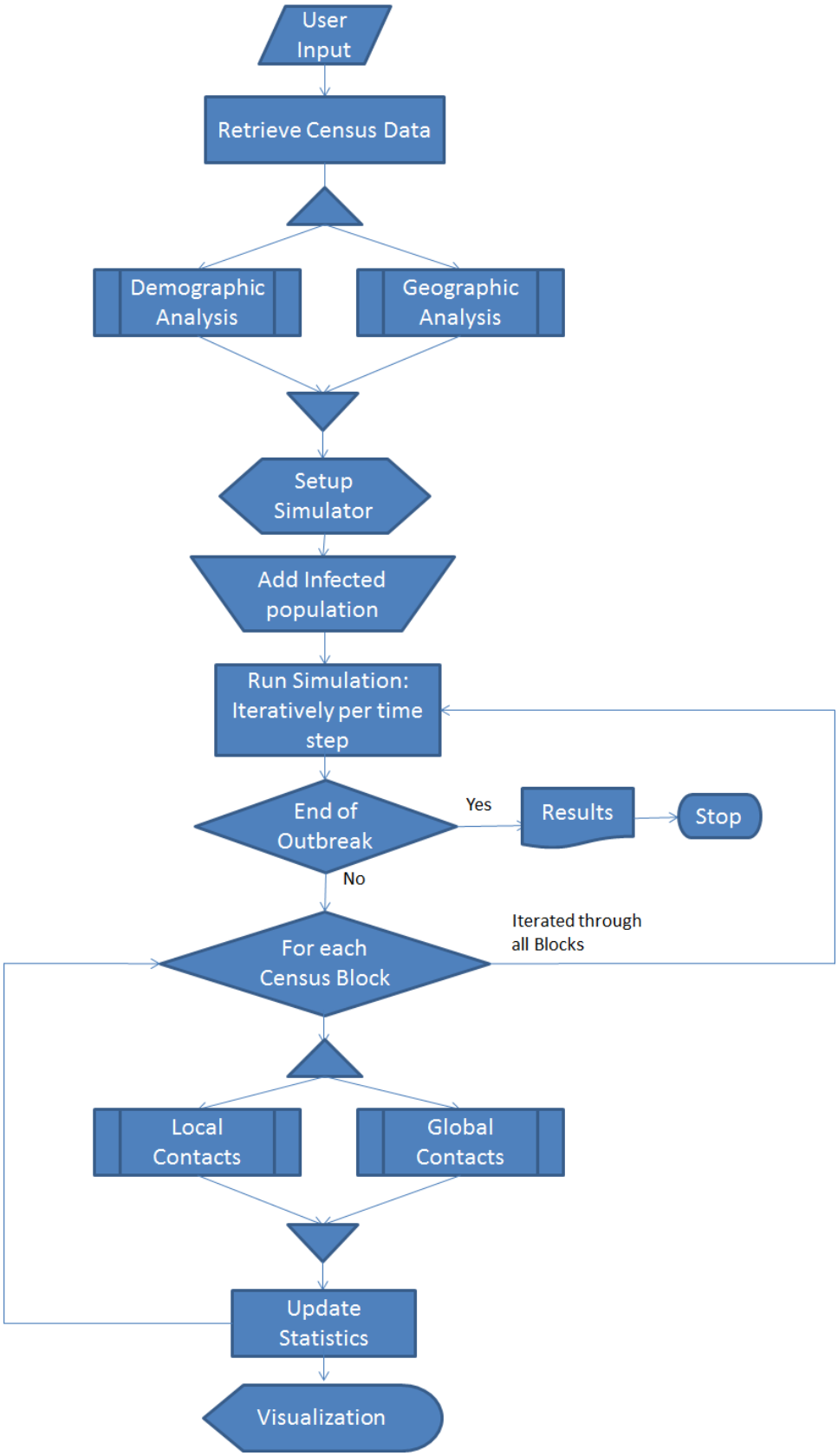

Figure 3.5. Process flow diagram for the simulator

while those who belong to the R compartment have recovered from the infection and acquired immunity. E represents the ones who are infected, but are in the latent stage of the infection, thus, show symptoms but do not start transmitting the infection. The compartment I constitutes of those individuals who are infected and actively transmit the infection to others.

The epidemic propagates by means of contacts made by individuals in infected (I) compartment with the ones who are in susceptible(S) compartment. When a susceptible person contracts the infection, he is moved to the exposed (E) compartment, beginning the exposed or latent period of infection. At the end of the exposed (or latent) period, he is moved to the infected (I) compartment, starting the infectious period when he contributes to infection transmission. The person recovers at the end of infectious period and is then moved to the recovered(R) compartment. All individuals who are in the recovered(R) compartment acquire immunity to the infection. Figure 3.6 describes the transition of people between infection compartments as they move from one infection state to another.

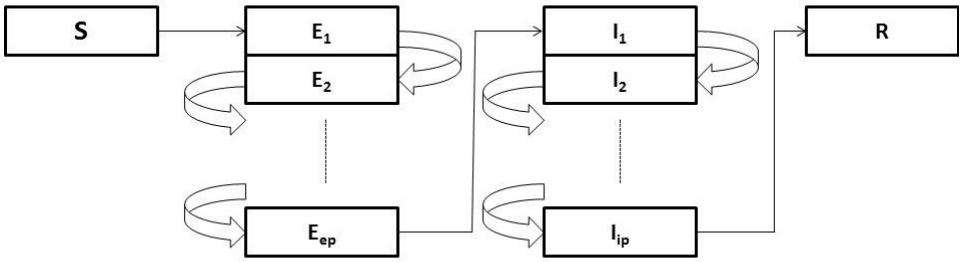

Figure 3.6. Representation of the SEIR disease transmission model

A representation of the SEIR disease transmission model used. This schematic shows the progression of a host between disease compartments. Here, ep represents the duration of exposed period and ip represents the duration of infectious period.

3.3.2. Geographic Layout

The simulator models disease spread for a geographic region such as a county. A census block, being the smallest geographic unit for which complete cumulative census data

is available, has been chosen as the geographic unit based upon which population distribution is modeled . Hence, the region is represented as a conglomerate of individual census blocks, which are unique in themselves.

3.3.2.1. *Census Block.* Every census block is identified by a unique identifier, referred to as block-id that is deduced from the census block IDs in the census data. The block-id is an integer between 0 and N-1 (both included), where N is the number of census blocks in the region. Each census block is assigned a unique integer in $[0, N-1]$ as its block-ID. Every census block is associated with unique latitude and longitude coordinates that represent its geographic location. A census block with block-ID i, where $i \in [0, N-1]$ is denoted as C_i.

As mentioned previously, the simulator employs an object-oriented design. Each of the census blocks that constitute the region of interest is instantiated as a census block object. Population statistics of the census block are stored in its respective census block object. At any time step t, a census block object C_i maintains S_{ij}, E_{ij}, I_{ij} and R_{ij} where S_{ij} represents the number of susceptible individuals, E_{ij} - exposed, I_{ij} - infectious and R_{ij} - recovered individuals in C_i who belong to the demographic group D_j.

3.3.2.2. *Region.* The geographic region for which the epidemic outbreak is modeled is a collection of individual census blocks. It is represented as a collection of all the census block objects that correspond to individual blocks. The geographical region R is defined as follows:

$$R = \{C_i\} \text{ where} i \in [0, N-1]$$

Additionally,

$$\forall C_i, C_j \in R, C_i \cap C_j = \Phi$$

3.3.2.3. *Interaction Coefficient.* Disease spread between census blocks is modeled by means of interactions between demographic groups in participating census blocks. It is evident that a greater number of interactions tend to occur with populous blocks rather than sparsely populated ones. Additionally, there is a smaller likelihood of interactions taking place between blocks that are farther apart. Interaction coefficient, proposed by Mikler et al. in

[24] models the likelihood of an interaction taking place between any two census blocks. It is computed based on their population counts and the distance between them.

Interaction coefficient between two blocks is the product of the populations of the blocks divided by the distance between them. For any two census blocks C_p, C_q, the interaction coefficient is computed as shown in (4)

$$IC(p,q) = \frac{P_p * P_q}{dist(C_p, C_q)} \tag{4}$$

The greater is the value of interaction coefficient; greater is the likelihood of interactions between two census blocks.

Normalized Interaction Coefficient (IC_i). The values of populations and distances between census blocks can vary significantly for different pairs of census blocks. This may lead to large variations in the values of interaction coefficients between different combinations of census blocks, which necessitate normalization of interaction coefficients. The values of interaction coefficients are normalized to a scale of 0 to 1, so as to minimize the variations. The normalized interaction coefficients are stored in two-dimensional matrix $\dot{N}$ with each row holding the normalized interaction coefficients for the block represented by it with every other block. The normalized interaction coefficients for a census block, i.e. a row in the two-dimensional matrix are computed as follows:

The j^{th} column in the i^{th} row of the matrix, representing the normalized interaction coefficient between census blocks C_i and C_j respectively is given by

$$\dot{N}_i[j] = \frac{IC(i,j)}{\sum_{k=0,k\neq i}^{N-1} IC(i,k)} \tag{5}$$

The normalized interaction coefficient matrix $\dot{N}$ is computed and stored at the beginning of the simulation. The values of normalized interaction coefficients from $\dot{N}$ are used to calculate global contacts made by a census block with external blocks.

3.3.2.4. *Calculating Distances.* The contact model which will be described in the subsequent sections, takes into account geographic distances between participating census blocks when infectious contacts are generated. Latitude and longitude coordinates of census blocks

obtained from the census database are used to calculate distance between two census blocks. The method used to calculate distances between census blocks is detailed in this section.

As a consequence of the earths surface being almost spherical, euclidian distance which computes the straight line distance between two points in euclidian space, does not give an accurate measure of the distance between any two points on the earth. In order to calculate the shortest distance between two points over the earths surface, neglecting the differences in elevation, great-circle distance between them is computed. Great-circle distance, which is the shortest distance between any two points on the surface of a sphere, gives a more accurate measure of the distance between any two points on the earths surface. The Haversine formula, derived from the law of haversines in spherical trigonometry, gives sufficiently accurate results for great-circle distances [28, 1]. Given, the latitude and longitude coordinates of two points on the earths surface, the Haversine formula computes shortest distance over the earths surface between those points.

The US Census Bureau uses a census blocks internal point to represent its geographic location. Census data provides latitude and longitude coordinates of the internal points of all census blocks. An internal point is a single point within a geographic entity - in this case a census block - that represents its approximate geographic center. If the entities shape causes the internal point to be located outside its boundary or in a water body, it is relocated to a land area within the entity [2]. It represents the approximate geographic center of the block.

To calculate the distance between two census blocks, the Haversine formula is made use of. Given the latitude and longitude coordinates of the internal points of any two census blocks, the Havesrine formula is used to computes the shortest distance between them over the earths surface. Shown below in (6) is the method used to compute distance between census blocks P and Q using Haversine formula [28, 1]:

The terms latP and longP represent latitude and longitude coordinates respectively of census block P, while latQ and longQ represent latitude and longitude coordinates respectively of census block Q. Here, c is the angular distance in radians, and a is the square of half the

chord length between census blocks P and Q.

$$dist\,(P,Q) = R * c \tag{6}$$

Where,

$$c = 2 * atan2(\sqrt{a}, \sqrt{1-a})$$

$$a = \sin^2(\delta lat/2) + \cos(latP) \cdot \cos(latQ) \cdot \sin^2(\delta long/2)$$

$$\delta lat = latQ - latP$$

$$\delta long = longQ - longP$$

Here, R is the earth's radius (mean radius = 6,317km)

3.3.3. Population Distribution

Population distribution of a geographic region with significant population is usually non-homogeneous with respect to space and demographics. In order to model the non-homogeneity and spatial distribution of the population, census data is made use of. Population statistics and geographic coordinates of each census block are retrieved from the census database for this purpose. Census blocks in this model differ from each other in terms of population statistics, demographics and geographic location. So as to depict the spatial distribution of population and differences in constitution of individual census blocks with respect to population statistics, population distribution is modeled along the following lines, as can be seen from Figure 3.7 :

- The population of the region is distributed into respective census blocks as per the census data.
- Within a census block the population is classified into demographic groups based on user-specified demographics.
- Additionally, within a demographic group, population is distributed into different disease compartments corresponding to infection state.

Let P be the set representing the human population in the region R, and P_i be the population in a census block C_i. The population belonging to a demographic group D_j in block C_i is represented as P_{ij}. Within the set P_{ij}; $S_{ij}, E_{ij}, I_{ij}, R_{ij}$ represent the people in susceptible, exposed, infectious and recovered states of disease respectively. The population distribution at any time t is modeled as per the following expressions:

$$P(t) = \sum_{i=0}^{N-1} \sum_{j=0}^{D-1} [S_{ij}(t) + E_{ij}(t) + I_{ij}(t) + R_{ij}(t)] \tag{7a}$$

$$P = P_1 \cup P_2 \cup \ldots \cup P_N \tag{7b}$$

$$P_i \cap P_j = \Phi \text{for any (i,j)} \mid C_i, C_j \in R \tag{7c}$$

$$S_{ij} \cap E_{ij} \cap I_{ij} \cap R_{ij} = \Phi \tag{7d}$$

$$S_{ij} \cup E_{ij} \cup I_{ij} \cup R_{ij} = P_{ij} \tag{7e}$$

As a census block is the geographical entity at the lowest level of granularity, the spatial distribution of population within a census block is not taken into account. It is assumed that all the population of a census block is located at the approximate geographic center of the block, as indicated by its latitude and coordinates in the census data. Furthermore, it is assumed that there is no migration between census blocks, or migration from or migration to external regions, since changes to population caused by migration are insignificant in the time frame of an outbreak. At the end of any time step, the population or demographics of a census block remain the same as that at the beginning. The changes within a census block are limited to the epidemiological compartments.

3.3.3.1. *Demographic Groups.* The population within each census block is classified into demographic groups based on demographics relevant to the epidemic being modeled. As mentioned previously, these groups are chosen corresponding to demographics of the risk groups for the epidemic. Each demographic group has differing social behaviors with respect to the disease and when members of the group interact with other groups. These differences in social behavior are captured using the parameters affinity, mobility and reach.

For a demographic group $D_j \in D$, the set of all demographic groups within a census

block, and $j \in [0, d-1]$ where d is the number of demographic groups:

- Mobility of a demographic group D_j is denoted as $m_j, m_j \in (0, 1)$. It is a measure of the fraction of contacts made by the j^{th} demographic group that are directed towards groups in external blocks. A mobility value of 0 means that all contacts made by the demographic group stay within the same census block, while a value of 1 means that all contacts are global, i.e. they happen with external census blocks.
- Reach denoted as $reach_j$ is a fractional number between 0 and 1. It is a measure of the farthest census block that can be reached by the demographic group D_j. If $reach_j$ equals 0, D_j can make contacts with blocks no farther than its home block, while a value of 1 allows D_j to make contacts with any census block in the region.
- Affinity is a measure of the likelihood of a demographic group interacting with another demographic group. It is a fractional value that measures the proportion of total contacts made by a demographic group that are directed towards the other group. Each demographic group has a value of affinity for every other group including itself.

The values of affinities between different demographic groups are stored in the affinity matrix. Affinity matrix A is defined as follows:

$$A = [a_{ij}]_{i=0,1,\ldots d-1; j=0,1,\ldots d-1} \tag{8}$$

where a_{ij} is the affinity of demographic group D_i with respect to D_j and,

$$\sum_{j=0}^{d-1} a_{ij} = 1$$

3.3.4. Contact Model

The disease spreads by the means of contacts between infected and susceptible populations. In this disease spread simulation, contacts are modeled between demographic groups. People belonging to a demographic group in a census block can make contacts with people belonging to the same group or another group either within the same census block or in an external block. Not all contacts lead to spread of the disease. Only the contacts between an

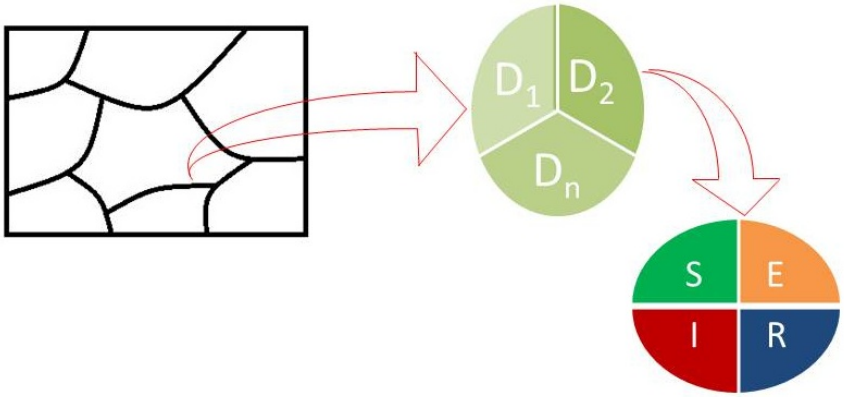

Figure 3.7. Population distribution model

Population distribution model. The geographic region on which the epidemic is modeled consists of census blocks. Within a census block, the population is classified into demographic groups. The population of a demographic group in turn is placed into disease compartments.

infected person and a susceptible person lead to the spread of the disease. Hence, contacts are generated only for the infected population. Infection parameters and factors like contact rate, mobility, affinity and reach of the demographic groups involved influence the number and extent of the infectious contacts generated.

During each time step in the course of the epidemic simulation, infectious contacts are generated iteratively for every demographic group in each of the census blocks. These contacts are distributed among susceptible populations based on the contact model described below. The population statistics for a census block are updated at the end of every time step to reflect the changes occurring in the numbers of people in different infection compartments due to propagation of the infection.

Contacts between different demographic groups are generated on the basis of a contact model. The contact model is implemented in accordance with the following ideology:

- Disease spread happens only by means of contacts originating from infected population and directed towards susceptible population.

- Different demographic groups differ in their day to day routines, and social interactions with other groups.
- Contacts are more likely to happen with census blocks that are closer and have a larger population rather than blocks that are farther away and have a smaller population.
- There is an element of randomness that exists in the spread of an epidemic in the real world, which brings in the need to incorporate stochasticity into the contact model.

Algorithm 1 shows the pseudocode for implementation of contact model.

Input: Infection and social parameters, demographics and coordinates of census blocks

Output: Numbers of infection transmissions to demographic groups in each census block

foreach $C_i \epsilon R$ **do**
  foreach *population group* C_{ij} *in* C_i **do**
    $L_{ij} = generateLocalContacts(C_{ij})$;
    $G_{ij} = genrateGlobalContacts(C_{ij})$;
    foreach *population group* C_{ik} *in* C_i **do**
      $\dot{L}_{ij}(k) = \psi * LInf_{ij} * A[j,k] * \frac{S_{ik}}{P_{ik}}$;
      $assignLocalContcats(\dot{L}_{ij}(k))$;
    end
    $disributeGlobalContacts(G_{ij})$;
  end
end

Algorithm 1: Algorithm for contact model

3.3.4.1. *Contact Rate.* Contact rate is the average number of contacts an individual makes in a particular time step, as defined earlier. In this model, each demographic group is assigned a contact rate. All individuals in a census block belonging to the same demographic group have the same contact rate. Average contact rate for a demographic group is user

defined. However, in an effort to mimic real world variations, an element of randomness is incorporated to contact rates. The average contact rate for a demographic group in a census block is temporally distributed. At the end of every time step, the contact rate for the next time step is drawn from a normal distribution.

3.3.4.2. *Infectious Contacts.* As mentioned above, infectious contacts are generated between demographics groups belonging to either the same or different census blocks. Within a census block infectious contacts are generated iteratively for the infected population of each demographic group. They are directed towards susceptible populations belonging to different demographic groups in different census blocks. These infectious contacts can either be local, i.e. within the same block or global, i.e. distributed among blocks other than the originating block. The fraction of local and global infectious contacts depends on the parameter - mobility of the demographic group that initiates the contacts. It should be noted that not all infectious contact are effective, i.e. not all infectious contacts lead to the actual transmission of the disease, due to factors such as immunity. The number of effective infectious contacts is limited by the availability of susceptible population and infectivity of the epidemic.

At every time step, infectious contacts are generated iteratively for each demographic group in a census block. The methodology for infectious contact generation for C_{ij} i.e. the population represented by a demographic group D_jD in a census block $C_i \in R$ is described below.

For any demographic group D_j in a census block C_i with a contact rate of $_{ij}$, the total number of contacts generated at time t is

$$P_{ij} * \beta_{ij}(t) \tag{9}$$

where P_{ij} is the number of people belonging to demographic group D_j in census block C_i. However, we are interested in contacts originating from people in the infectious state of the disease. Hence number of infectious contacts would be a subset of (7),

$$Inf_{ij} = I_{ij} * CR_{ij}(t) \tag{10}$$

The infectious contacts Inf_{ij} can either be local or global, i.e. intra-block or inter-block

respectively. Let f_{ij} (13) be the fraction of local infectious contacts, or in other words the infectious contacts that are initiated by the demographic group D_j of census block C_i and terminate within the same block. The number of local infectious contacts in this case is given by $LInf_{ij}$ (11). Rest of the infectious contacts is global infectious contacts, i.e. infectious contacts made with blocks other than C_i. The number of global infectious contacts is given by $Ginf_{ij}$ (12).

$$LInf_{ij} = f_{ij} * Inf_{ij} \tag{11}$$

$$GInf_{ij} = (1 - f_{ij}) * Inf_{ij} \tag{12}$$

The fraction f_{ij} is computed as below:

$$f_{ij} = (1 - m_j) \tag{13}$$

where, m_j is the value of mobility for demographic group D_j.

Not every non-infectious individual who is part of an infectious contact ends up being infected, due to factors like virulence of the disease and immunity of the host. Infectivity of the disease ψ, is a parameter that measures the number of individuals that are actually infected on exposure to the infection. Taking into account the effect of infectivity, only a fraction of the infectious contacts generated are effective in actual transmission of the disease. Furthermore, it should to be noted that only the infectious contacts directed towards susceptible population contribute towards infection propagation. Hence, the number of susceptible individuals in the target group also influences the number of effective infectious contacts.

3.3.4.3. *Local Contacts.* The portion of local contacts $Linf_{ij}$ generated by C_{ij} that result in infecting susceptible population is denoted by $\dot{L}_{ij}$ (14). These contacts are distributed among all the demographic groups present in the block C_i. The distribution of these contacts is decided based on the affinity values between demographic groups obtained from the affinity matrix. The number of susceptible individuals in the population group C_{ik} who are infected

as a result of the local contacts $\dot{L}_{ij}$ is given by:

$$\dot{L}_{ij}(k) = \psi * LInf_{ij} * A[j,k] * \frac{S_{ik}}{P_{ik}} \tag{14}$$

where ψ denotes the infectivity of the disease, and $A[j,k]$ gives the value of affinity between demographic groups D_j and D_k from the affinity matrix A. Hence, the number of people belonging to the sub-group C_{ik}, who contract the infection as a result of local contacts initiated by the infectious population belonging to the sub-group C_{ij} is $\dot{L}_{ij}(k)$. Those susceptible individuals who contracted the infection now move into the latent state of infection. Numbers of people in appropriate disease compartments are updated to reflect the changes due to disease transmission as described in the section .

3.3.4.4. *Global Contacts.* Global infectious contacts initiated by any demographic group in a census block, are targeted towards susceptible populations of various demographic groups in external census blocks. In order to complete these inter-block contacts, target blocks are chosen randomly and infectious contacts are distributed among those randomly selected blocks on the basis of interaction coefficient between participating blocks and reach of the demographic group initiating the contact. Within the contacted census block, infectious contacts are distributed among constituent demographic groups proportional to the value of affinity between the demographic groups involved. The number of actual infection transmissions is weighed down by the availability of susceptible population and infectivity of the disease. Below is a brief description of the methodology used to model global infectious contacts:

- A target block is randomly chosen by drawing a random fractional number and comparing it successively with cumulative normalized interaction coefficients until a cumulative value that is greater than the random fractional number is found.
- A contact is assigned to the block represented by the cumulative interaction coefficient obtained above, if it is within the reach of the demographic group initiating the contact.
- The above process is repeated until all infectious global contacts initiated by the

source group are distributed.

- Within each contacted census block, infectious contacts are distributed among constituent demographic groups proportional to affinities between source and target demographic groups.

  For the population group C_{ij}, belonging to demographic group D_j in census block C_i, the number of global infectious contacts generated is given by $GInf_{ij}$ as computed in (12). Let $\dot{G}_{ij}(C_x)$ be the number of contacts made by C_{ij} with a census block C_x, randomly chosen confirming to Algorithm 2 such that,

$$C_x \in [0, N-1], x \neq i \text{and} dist(C_x, C_i) < maxDist * \Upsilon_j$$

where Υ_j represents the value of reach associated with D_j.

The infectious contacts $\dot{G}_{ij}(C_x)$ are distributed among the constituent demographic groups of the contacted block C_x, weighted by their affinities. The number of effective infectious contacts, i.e. the contacts that result in susceptible individuals getting infected, made with a demographic group D_k in block C_x is given by (15).

$$\dot{G}_{ij}(C_{xk}) = \dot{G}_{ij}(C_x) * \psi * A[j,k] * \frac{S_{xk}}{P_{xk}} \tag{15}$$

Hence, $\dot{G}_{ij}(C_{xk})$ is the number of people in the group C_{xk} who contract the infection as a consequence of global infectious contacts made by the group C_{ij}. They are moved to the corresponding exposed compartment and population statistics are updated as described in section . Global infectious contacts $GInf_{ij}$ are distributed among all contacted external census blocks C_x in accordance with Algorithm 2 so that,

$$GInf_{ij} = \sum_{x=0, a\neq i}^{N-1} \dot{G}_{ij}(C_x) \tag{16}$$

3.3.4.5. *Update of Population Statistics.* During each time step infectious contacts are generated iteratively for every demographic group in each of the census blocks, as described above. The disease transmissions that take place as a consequence of these infectious contacts result in newly exposed population in the respective groups. Numbers of individuals in susceptible and exposed disease compartments are updated to reflect this, as individuals who

Input: G_{ij} - Number of contacts to be made globally by the group C_{ij}

Output: Array $\dot{N}_i$ of normalized interaction coefficients for block C_i

Array $\dot{G}$ of contacts made with each census block

Initialize $\dot{G}$;

while $G_{ij} > 0$ **do**

  Choose a fractional number $x : 0 < x < 1$ randomly ;

  $ic_cumulative = \dot{N}_i[0]$ $block = 0$;

  while $ic_cumulative < xanddist(C_{block}, C_i) < maxDist * \Upsilon_j$ **do**

    $ic_cumulative = ic_cumulative + \dot{N}_i[block]$;

    $block = block + 1$;

  end

  $\dot{G}[block] = \dot{G}[block] + 1$;

  $G_{ij} = G_{ij}1$;

end

Algorithm 2: Algorithm for generating global contacts

contracted the infection move from susceptible to exposed state of the disease. At the end of a time step, the population statistics of each group within a census block are updated as and when infected people recover, or people in the exposed state become infectious. Specifically, the following updates are made to the population statistics of a demographic group within a census block to reflect the changes occurring due to disease transmission.

- Updates are made to susceptible and exposed populations of each group in a census block to reflect disease transmissions that are a result of local and global contacts made with it.
- People who recover from the disease, i.e. those who have been in the infectious disease compartment for the duration of infectious period are moved to recovered compartment.
- People who become infectious i.e. those who have been in the exposed compartment for the duration of exposed period are moved to the infectious compartment.

Susceptible, Exposed, Infectious, and Recovered populations S_{ij}, E_{ij}, I_{ij} and R_{ij} respectively of the group C_{ij} are updated as below:

$$S_{ij}(t+1) = S_{ij}(t) - \sum_{k=0}^{D-1} \dot{L}_{ik}(j) - \sum_{a=0}^{N-1} \sum_{b=0}^{D-1} \dot{G}_{ab}(C_{ij}) \tag{17a}$$

$$E_{ij}(t+1) = E_{ij}(t) + \sum_{k=0}^{D-1} \dot{L}_{ik}(j) + \sum_{a=0}^{N-1} \sum_{b=0}^{D-1} \dot{G}_{ab}(C_{ij}) - \hat{e}_{ij}(t) \tag{17b}$$

$$I_{ij}(t+1) = I_{ij}(t) + \hat{e}_{ij}(t) - \hat{i}_{ij}(t) \tag{17c}$$

$$R_{ij}(t+1) = R_{ij}(t) + \hat{i}_{ij}(t) \tag{17d}$$

Where $\hat{e}_{ij}(t)$ and $\hat{i}_{ij}(t)$ represent the number of individuals in the last day of their exposed and infectious periods respectively in C_{ij}, D is the number of demographic groups and N is the number of census blocks.

In a real-life system the number of contacts made by an individual does not remain constant over time. In order to more realistically model the contact rate of each demographic group, an element of randomness is introduced into the corresponding contact rates. Contact rate for each demographic group in a census block is considered to be temporally distributed. At the end of a time-step contact rate is updated using an adjustment factor δ, drawn from a normal distribution.

At a time-step t, contact rate of individuals in a census-block i belonging to the demographic group j at time $t+1$ is computed as follows:

$$CR_{ij}(t+1) = CR_{ij}(t) * (1 + \delta * \rho) \tag{18}$$

Where,

$$-1 < \delta < 1 \text{ and } 0 < \rho < 1$$

Here δ is a pseudo-random number drawn from a normal distribution with mean 0.0 and standard deviation 1.0, and ρ is the parameter which sets an upper limit on the fraction by which $CR_{ij}(t+1)$ can vary from $CR_{ij}(t)$.

3.3.5. Stochasticity

In order to model the inherent unpredictability that exists in disease transmission, stochasticity is incorporated into the disease spread model. A factor of randomness is used in the simulator in the following ways to incorporate stochasticity:

- Each time a new census block has to be contacted while global contacts are being generated; it is chosen randomly with the help of a random number generator. Feasibility of making contacts, and strength of contacts made with this block is then decided based on factors spatial and social parameters.
- At the end of each time step, contact rates for all sub groups are updated by drawing the new value number from a Gaussian distribution.

3.3.6. Visualization

A region such as a county usually has thousands of census blocks with hundreds of thousands of people in it. When an epidemic outbreak simulation is executed for such a region over a period of time, a large amount of data is produced, analysis of which is difficult. This necessitates methods to enable study and analysis of the data produced through visualization. The spread of the disease is displayed on an interactive geographic map of the region. A heat map representation is used to depict the percentage of each census blocks population that is infected, to provide a visual perception of the disease dynamics. Methods from GeoTools code library are used to render maps using geometry of the region available in the census database.

The visual interface that displays geographic map of the region in the context of the disease spread has been designed to serve two main purposes:

- Displaying disease dynamics and intensity of infection in a census block on a geographic map as the epidemic progresses.
- Enabling inputting the initial infectious population by selecting target census blocks on the geographic map.

3.3.6.1. *Heat Map.* Intensity of infection in a census block is measured on the basis of proportion of infected people in that block at a given time. It is represented as an integer on a scale of 1 to 10, 1 being the least intensity and 10 being the highest. Intensity of infection Inti in a census block C_i at a time t is given by:

$$Int_i = \lceil \frac{I_i}{P_i} * 10 \rceil \tag{19}$$

Where I_i gives the number of infected people and P_i gives the total number of people in census block C_i.

During each time step of the simulation, intensity of infection for every census block is displayed on the geographic map using a color scheme. A color is associated with every value of intensity of infection. Darker colors are used to depict larger values of intensity of infection to provide a visual perception of disease dynamics.

GeoTools allows colors to be rendered as a layer over the geographic map. Colors are associated with attributes of features of the geometry such as block-id or latitude and longitude coordinates using certain rules. To render the heat map, rules that bind colors to values of intensity of infection are applied. The color that is associated with the value of intensity of infection in each census block is painted as a layer over the census block on the map. As the simulation progresses, values of intensity of infection in each census block are updated, and the map is re-rendered for every time step.

3.3.6.2. *Block Selection.* An initial infected population is needed to start off the epidemic outbreak simulation. In order to experiment with different scenarios, the set of infected people before the start of the simulation can be varied. Locations of the initial infectious people and their numbers are some of the initial parameters that can be varied. The initial infectious population can be inputted by selecting their locations on the map through the visual interface as it is intuitive to choose census blocks by clicking on a map.

Once a user clicks on the map to select a census block so as to add infected population, the block's block-id and geographic coordinates are needed to update its statistics. A mouse click event returns the screen position of the click event. This screen position has to

be transformed to geographic coordinates to get hold of the census block that the user tries to select. However, as mentioned previously, a census block's geographic location is represented by the coordinates of its approximate geographic center. The transformed geographic coordinates may not be matching with coordinates of the census blocks. In this case, the block with geographic coordinates closest to the transformed ones are chosen. The following approach is used to select the census block whose coordinates have the closest match to the ones returned by the mouse click event:

- A 2x2 (in pixels) bounding box is constructed around the screen position of the mouse-click.
- This bounding box is transformed to a bounding box with geographic map coordinates.
- Features of the geometry (census blocks in this case) that overlap with this bounding box are computed, and their attributes (ID, latitude and longitude) are retrieved.
- If there is more than one block overlapping with the bounding box, the block that has the least distance from the selected point is chosen as the selected block.

CHAPTER 4

EXPERIMENTS AND RESULTS

In this section, epidemic outbreak experiments performed using the simulator presented above, and results from those experiments are presented. Disease spread simulations under various conditions have been studied to analyze the behavior of the simulator to changes in underlying parameters. The focus has been on studying the effect of disease parameters, social behavioral parameters and demographic stratification on the propagation of an epidemic. Experiments were designed so as to study the effect of these parameters on disease dynamics. Simulations have been performed by varying the values of infectivity, number of demographic groups and social behavioral parameters of those groups. The different experiments performed and results obtained are presented in the sections that follow.

4.1. Experiments

All the experiments were conducted on Census 2000 data obtained for Denton county in Texas. The census data made available by the US Census Bureau, provide the population and demographic distribution used in these experiments. Denton county has a population of 432976,and is divided into 7355 census blocks, as per Census 2000. The population distribution of Denton county is shown on a geographic map in Figure 3.7.

As mentioned previously, the experiments are designed so as to examine the effect of disease parameters, social behavioral parameters and demographic stratification on the propagation of an epidemic. Comparative studies were performed by varying one of the parameters and using fixed values for all other parameters so as to examine the effect of individual parameters on the spread of disease. The experiments below study the effects of the parameters - infectivity, contact rate, mobility, affinity and reach on the spread of the disease. Different scenarios where the population is stratified using different demographic constraints are presented below.

4.1.1. Disease Parameters

In this section, experiments designed to measure the sensitivity of the simulator to infectivity of the disease are presented.

4.1.1.1. *Experiment: Infectivity.* This experiment used Census 2000 data for Denton county. In this experiment, no demographic stratification was made use of. Within a census block all the population is considered as one group. Simulations were performed using different values of infectivity. The spread of disease when the infectivity is 0.05 is compared with the case where infectivity is 0.01. The parameters used in this experiment are shown in Table 4.1. All the parameters here are set to model the ideal case, where the social behavioral parameters do not limit the behavior of the simulator.

Table 4.1. Parameters used in experiment 1

Parameter	Value
Number of Demographic Groups	1
Latent Period	2
Infectious Period	5
Affinity	1.0
Mobility	0.5
Reach	1.0
Contact Rate	25

To start off the simulation, 100 individuals belonging to the census block 2850 which a population of 877, were infected.

4.1.2. Social Behavioral Parameters

The experiments designed to study the effect of social behavioral parameters on the spread of disease are presented here. One of the parameters - contact rate, mobility, reach, and affinity is varied, keeping all other parameters constant to compare differences in disease spread patterns in different cases.

4.1.2.1. *Contact Rate.* In this experiment, the effect of contact rate on the spread of infection is analyzed. Simulations were performed using contact rates of 10 and 40. No demographic stratification was used in this experiment. The values of parameters used in this experiment are shown in Table 4.2.

Table 4.2. Parameters used in contact rate experiment

Parameter	Value
Number of Demographic Groups	1
Latent Period	2
Infectious Period	5
Affinity	1.0
Mobility	0.5
Reach	1.0

To start off the simulation, 100 individuals belonging to the census block 2850 which has a population of 877, were infected.

4.1.2.2. *Mobility.* In this experiment, the effect of the parameter - mobility - on the spread of infection is analyzed. Simulations were performed using mobility values of 0.05 and 0.9. No demographic stratification was used in this experiment. The values of parameters used in this experiment are shown in Table 4.3.

Table 4.3. Parameters used in mobility experiment

Parameter	Value
Number of Demographic Groups	1
Latent Period	2
Infectious Period	5
Contact Rate	25
Affinity	1.0
Reach	1.0

To start off the simulation, 100 individuals belonging to the census block 2850 which has a population of 877, were infected.

4.1.2.3. *Reach.* In this experiment, the effect of the parameter - reach - on the spread of infection is analyzed. Simulations were performed using reach values of 0.1 and 1.0. No demographic stratification was used in this experiment. The values of parameters used in this experiment are shown in Table 4.4.

Table 4.4. Parameters used in reach experiment

Parameter	Value
Number of Demographic Groups	1
Latent Period	2
Infectious Period	5
Contact Rate	25
Affinity	1.0
Mobility	0.5

To start off the simulation, 100 individuals belonging to the census block 2850 which has a population of 877, were infected.

4.1.3. Demographic Stratification

In the following experiments, spread of the disease when demographic stratification is used is studied. Population is stratified on the basis of gender or age, and disease spread patterns in different demographic groups are analyzed, when different social behavioral parameters are used for different groups.

4.1.3.1. *Demographic: Gender.* In this experiment, gender was used as the demographic. Denton county has 215368 male individuals and 217608 female individuals in all as per Census 2010. Within each census block of Denton, the population was classified into male and female demographic groups. The value of infectivity used here was 0.05. Parameters used in this experiment are shown in Table 4.5. The male demographic group has a contact rate of 20

while the female demographic group has a contact rate of 15. Affinity matrix used in this experiment is shown in Table 4.6.

The purpose of this experiment is to demonstrate the applicability of the framework to perform simulations using multiple demographic groups. For this purpose, similar values have been used for the social behavioral parameters for both the groups, expecting similarities in disease dynamics.

Table 4.5. Parameters used in gender experiment

Parameters	Male	Female
Contact Rate	20	15
Infectious Period	5	5
Latent Period	2	2
Mobility	0.7	0.5
Reach	0.75	0.5

Table 4.6. Affinity matrix for gender experiment

Affinity matrix	Male	Female
Male	0.6	0.4
Female	0.4	0.6

To start off the simulation, 85 male and 15 female individuals belonging to the census block 2850 which has a population of 877, were infected.

4.1.3.2. *Demographic: Age.* This experiment used age as the demographic constraint. The population was classified into 3 age groups: young-aged (0-17), middle-aged (18-54), old aged (55+). Denton county has 131156 young-aged, 253186 middle-aged and 48634 old-aged individuals respectively, as per Census 2000. An Infectivity of 0.05 was used in this experiment. Table 4.7 shows the parameters used in this experiment. The contact rates for each demographic group and values of affinities between groups are based on the study by Eubank et al. on social interactions between individuals of different age groups that lead to

airborne diseases [16]. Affinity matrix, shown in Table 4.8, is constructed noting from [16] that individuals of each age group tend to make most of the interactions with individuals of the same group. In this experiment, mobility and reach are set based on the assumption that middle-aged individuals travel further away from home, and make more interactions with individuals from other census blocks, while old-aged individuals tend to stay closer to home.

Table 4.7. Parameters used in age experiment

Parameters	Young(0-17)	Middle(18-54)	Old(55+)
Contact Rate	18	15	9
Infectious Period	5	5	5
Latent Period	2	2	2
Mobility	0.5	0.75	0.4
Reach	0.5	0.9	0.6

Table 4.8. Affinity matrix for experiment 3

Affinity matrix	Young	Middle	Old
Young	0.6	0.3	0.1
Middle	0.3	0.5	0.2
Old	0.1	0.3	0.6

4.1.3.3. *Effects of Demographic Stratification.* In this experiment, the effects of demographic stratification on the epidemic outbreak are studied. The outbreak obtained for age-stratification is compared with the case where there is no demographic stratification. The parameters used in the experiment, as can be seen from Table 4.9 are approximately average values of those used in the experiment with 3 age-groups.

The second part of this experiment is a comparative study between gender-stratification and age-stratification.

To start off the simulation, 11 young-aged, 86 middle-aged and 3 old-aged individuals belonging to the census block 2850 which has a population of 877, were infected.

Table 4.9. Parameters used in experiment with no demographic stratification

Parameter	Value
Number of Demographic Groups	1
Latent Period	2
Infectious Period	5
Affinity	1.0
Mobility	0.67
Reach	0.7
Contact Rate	15

4.2. Results

The results obtained after performing the experiments described in the previous section are presented in this section. The results are analyzed and interpreted by means of plots and heat maps to study the disease dynamics under various scenarios.

4.2.1. Disease Parameters

Experiments were performed to study the effect of diseases parameters on the spread of disease. In particular, the effect of infectivity on disease dynamics is analyzed below.

4.2.1.1. *Infectivity.* In this experiment, all the social behavioral parameters were set to model the ideal case, so as to study the disease spread when it is not influenced by any of the social behavioral parameters. Two simulations were performed, the first one with an infectivity of 0.05 and the other with an infectivity of 0.01, as described in the Experiments section. The epidemic touched its peak on 29^{th} day, when a little more than 4% of the population was infectious, in the former case, while it reached the peak on 66^{th} day when about 1% of the population was infectious in case of the latter. It can be clearly seen from Figure 4.2, which shows a comparison between the percentage of infectious individuals in both the cases, that a higher value of infectivity leads to a shorter epidemic that infects a larger number of individuals. The epidemic takes longer to reach its peak and die down when infectivity is 0.01. Figure 4.3 shows a comparison of the intensity of infection when epidemic

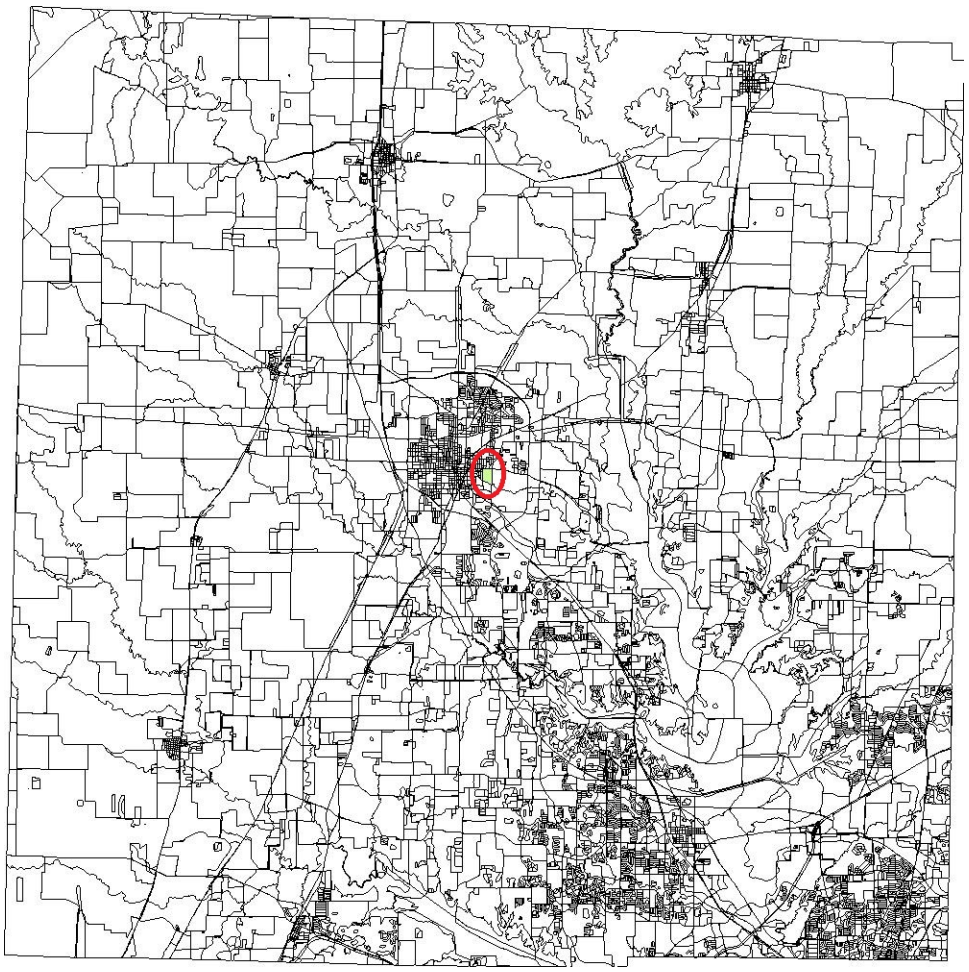

Figure 4.1. Infected population at start of simulation

The colored block at the bottom right is the initially infected block.

reaches its peak in both the cases.

4.2.2. Social Behavioral Parameters

The results obtained after performing experiments designed to study the effect of social behavioral parameters on the disease dynamics are analyzed below.

4.2.2.1. *Contact Rate.* In this experiment, all the parameters were set to values shown in Table 4.2. Experiments were performed using differing values of contact rate. Two simulations were performed, the first one with a contact rate of 40 and the other with a contact rate of 10, as described in the Experiments section. The epidemic touched its peak on 32^{nd} day, when about 3.5% of the population was infectious, in the former case, while

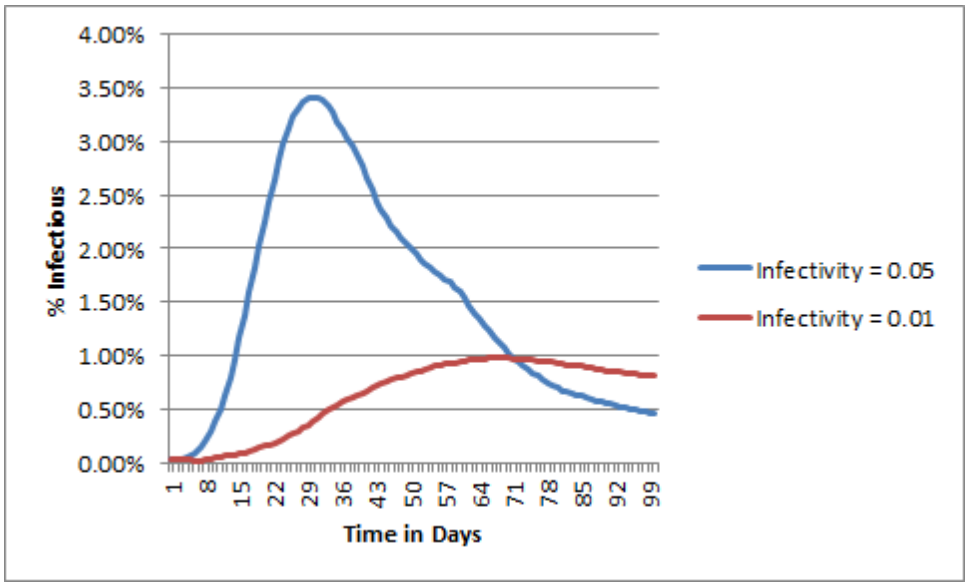

Figure 4.2. Infectivity experiment: Comparison of % infectious infectivity = 0.05 vs infectivity = 0.01

it reached the peak on 77^{th} day when about 1% of the population was infectious in case of the latter. It can be clearly seen from Figure 4.4, which shows a comparison between the percentage of infectious individuals in both the cases, that a higher value of contact rate leads to an epidemic that last for a shorter duration and infects a larger number of individuals. The epidemic takes longer to reach its peak and die down when contact rate is 10. Figure 4.5 shows a comparison of the intensity of infection when epidemic reaches its peak in both the cases.

4.2.2.2. *Mobility.* In this experiment, all the parameters were set to values shown in Table 4.3. Experiments were performed using differing values of mobility. Two simulations were performed, the first one with a mobility of 0.05 and the other with a mobility of 0.9, as described in the Experiments section. The epidemic touched its peak on 72^{nd} day, when a little over 1% of the population was infectious, in the former case, while it reached the peak on 25^{th} day when about 6% of the population was infectious in case of the latter. It can be clearly seen from Figure 4.6, which shows a comparison between the percentage of infectious individuals in both the cases, that a higher value of mobility leads to an epidemic that last

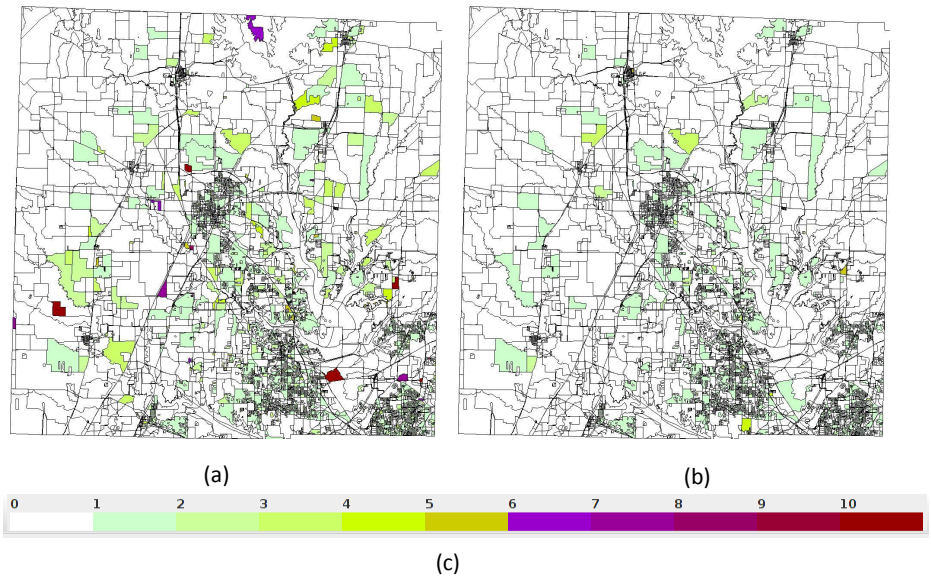

Figure 4.3. Infectivity experiment: Intensity of infection
(a) Day 27 - Peak of the epidemic when infectivity = 0.05. (b) Day 66 - Peak of the epidemic when infectivity = 0.01. (c) Color scale.

for a shorter duration and infects a larger number of individuals. The epidemic takes longer to reach its peak and die down when mobility is 10. Figure 4.7 shows a comparison of the intensity of infection when epidemic reaches its peak in both the cases. A larger value of mobility leads to a larger number of global contacts being generated, resulting in a quicker, more intense spread of the disease.

4.2.2.3. *Reach.* In this experiment, all the parameters were set to values shown in Table 4.4. Experiments were performed using differing values of reach. Two simulations were performed, the first one with a reach of 0.1 and the other with a mobility of 1.0, as described in the Experiments section. The epidemic touched its peak on 23^{rd} day, when close to 1% of the population was infectious, in the former case, while it reached the peak on 25^{th} day when about 4% of the population was infectious in case of the latter. It can be clearly seen from Figure 4.8, which shows a comparison between the percentage of infectious individuals

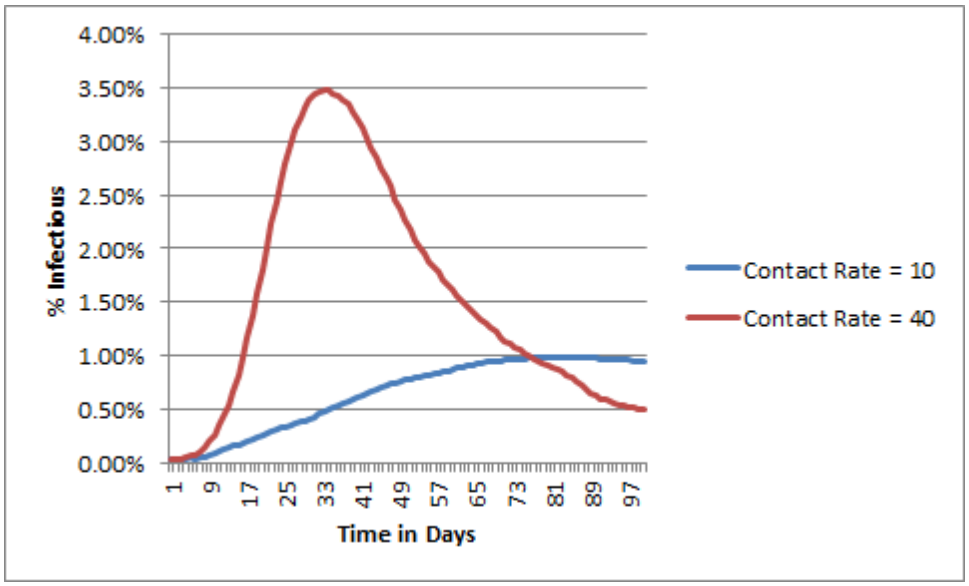

Figure 4.4. Contact rate experiment: Comparison of % infectious contact rate = 40 vs contact rate = 10

in both the cases, that a higher value of reach leads to an epidemic that last for a shorter duration and infects a larger number of individuals. The epidemic takes longer to reach its peak and die down when reach is 0.1. Figure 4.9 shows a comparison of the intensity of infection when epidemic reaches its peak in both the cases. A larger value of reach leads to a more extensive spread of the disease, while, a smaller value leads to a more localized spread.

4.2.3. Demographic Stratification

4.2.3.1. *Demographic: Gender.* In this experiment, population was classified based on gender, as described earlier. The value of infectivity used was 0.05. Figure 4.10 shows the Exposed-Infectious plot for this simulation. The epidemic reached its peak level on 21^{st} day, with about 45000 individuals being infectious. Figure 4.11 shows a comparison between the percentages of infectious individuals in male and female population groups. The male group reached its peak on the 20^{th} day, while the female group reached its peak on 21^{st} day of the epidemic, both with about 10% of their respective populations being infectious.

Figure 4.12 shows the population distribution of male and female demographic groups in Denton county at a census block level. It can be seen that there is similarity in the pop-

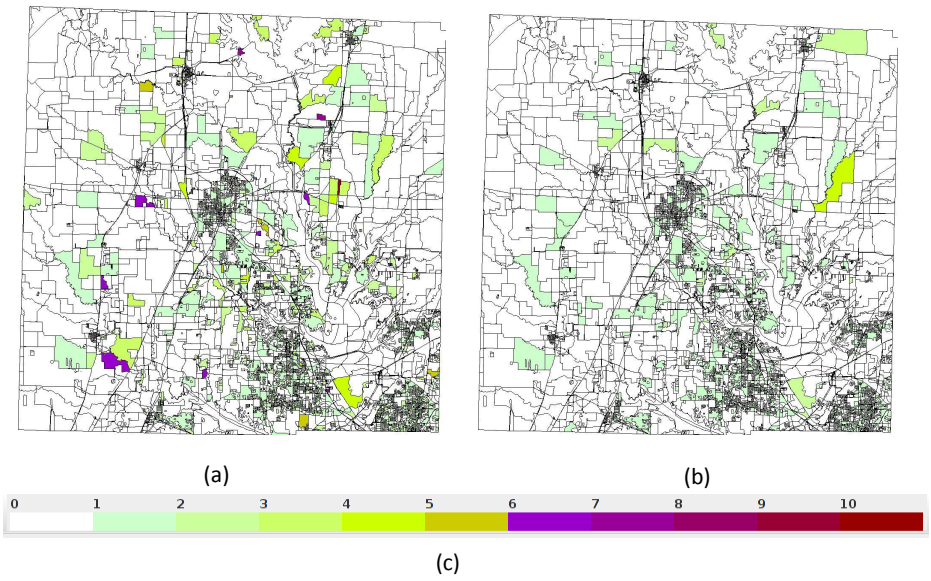

Figure 4.5. Contact rate experiment: Intensity of infection
(a) Day 32 - Peak of the epidemic when contact rate = 40. (b) Day 77 - Peak of the epidemic when contact rate = 10. (c) Color scale.

ulation distribution of both the demographic groups, with most blocks having a good mix of male and female populations. As mentioned previously, this experiment was designed to demonstrate the applicability of the framework to perform simulations on multiple demographic groups. Similar social behavioral parameter values were used for both the groups. As a result, we can observe similarities in the disease dynamics in both the groups as can be seen from Figure 4.13.

4.2.3.2. *Demographic: Age.* The demographic constraint used in this experiment was age, with the population being classified into 3 age-groups as described earlier. The value of infectivity used in this experiment was 0.05. The exposed-infectious plot for this simulation is shown in Figure 4.14. The epidemic reached its peak on the 22^{nd} day with about 42000 individuals being infectious on that day. Figure 4.15 shows a comparison between the percentages of infectious individuals in the three age-groups. It can be seen that the epidemic reaches

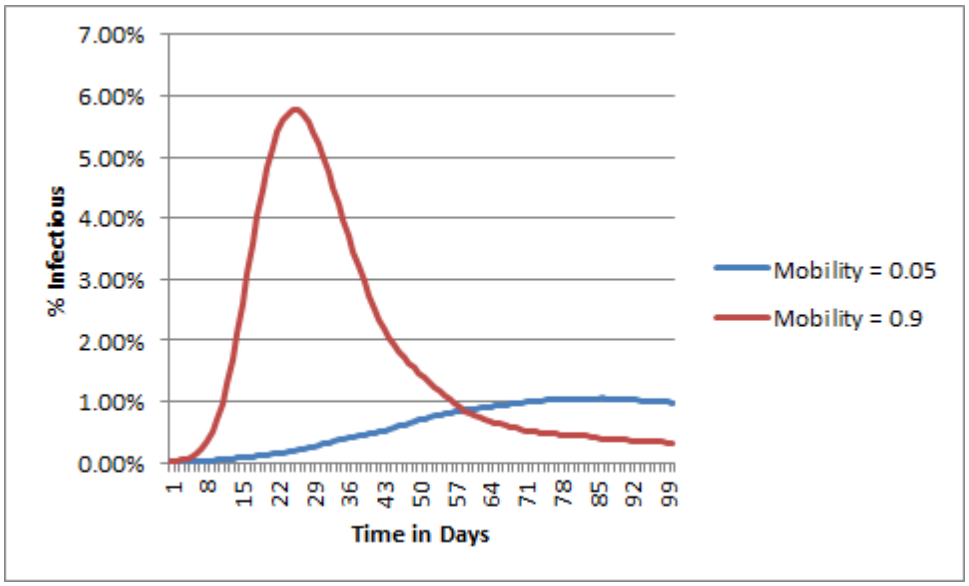

Figure 4.6. Mobility experiment: Comparison of % infectious mobility = 0.05 vs mobility = 0.9

the peak quicker in young and old age-groups compared to middle age-group. Additionally, the percentage of infectious at peak is more for young and old age-groups in contrast to middle-aged. This can be attributed to higher values of affinities of these groups for contacts among themselves, and smaller populations of these groups. It can also be observed that larger populations lead to longer epidemics with shorter peaks.

Figure 4.16 shows the population distribution for the 3 age-groups. While young and middle-aged are somewhat uniformly distributed, old-aged are more likely to be in certain pockets. Figure 4.17 shows the intensity of infection in different census blocks of Denton county, on the days when the epidemic touched its peak in each demographic group. At its peak, infection prevalence in young-aged individuals shifts from blocks with a higher percentage of young-aged people to closer blocks with a moderate percentage of them, due to a higher affinity for contact within the same group and moderate values of reach and mobility. Owing to larger values of mobility and reach, and some-what uniform affinities with all groups, the infection prevalence in middle-aged population is widely spread out. In the case of the old-aged demographic group, the infection spread is mostly confined to blocks around

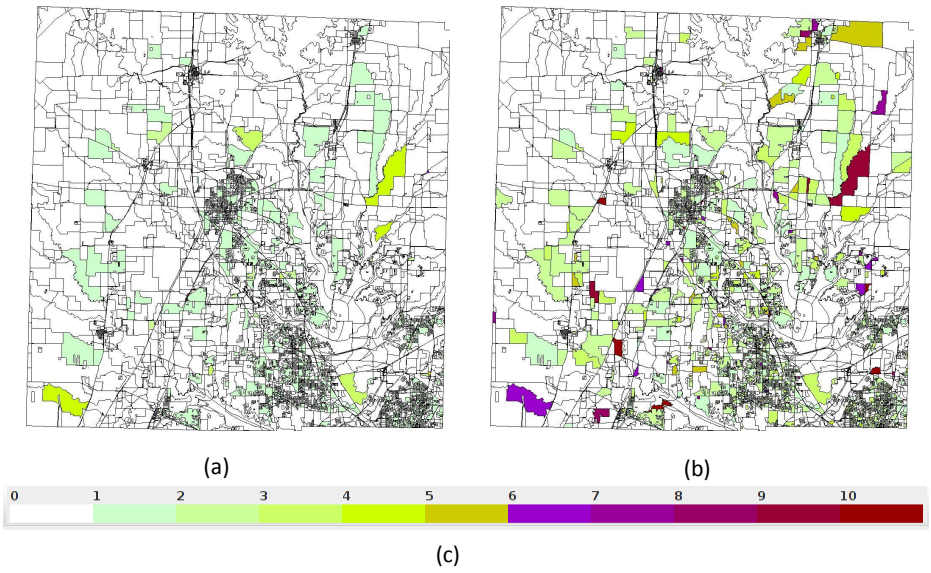

Figure 4.7. Mobility experiment: Intensity of infection
(a) Day 72 - Peak of the epidemic when mobility = 0.05. (b) Day 25 - Peak of the epidemic when mobility = 0.9. (c) Color scale.

the pockets where there is a larger density of old-aged individuals.

4.2.3.3. *Effects of Demographic Stratification.* Figure 4.18 shows a comparison between the cases of no demographic stratification and age-stratification into 3 groups, with respect to the number of infectious individuals on each day. It can be observed that the epidemic peaks earlier and the number of infectious individuals when the epidemic peaks is larger in case of 3 demographic groups. This behavior can be attributed to demographic stratification. As infectious contacts are directed towards particular demographic groups, disease spread in these groups is quicker. The epidemic recedes sooner due to fewer susceptible individuals available, thus producing a shorter epidemic where larger numbers of people are infected.

Figure 4.19 shows a comparison between the numbers of infectious individuals in the population in cases of age-stratification and gender-stratification. As the average values of social behavioral parameters are similar in both the cases, we can see similarities in disease

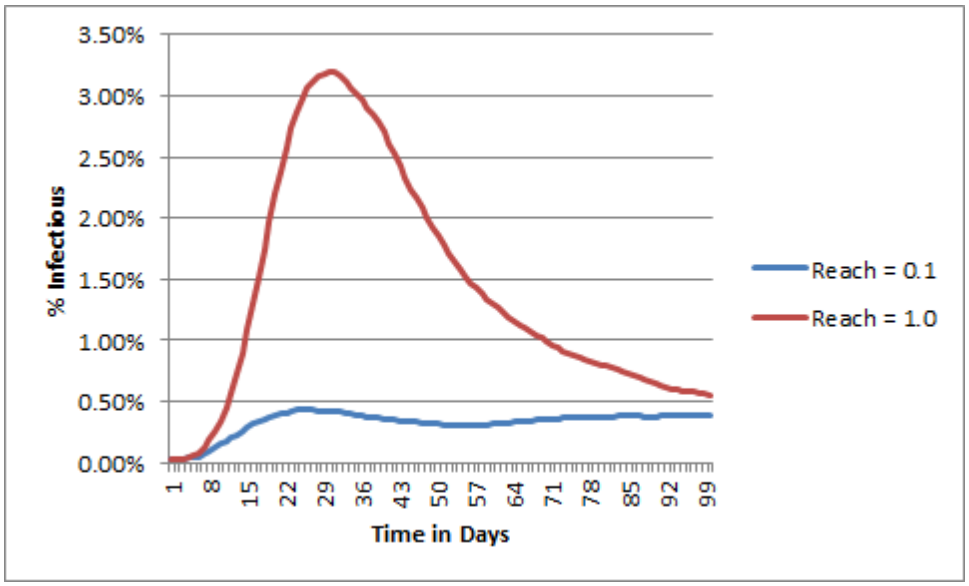

Figure 4.8. Reach experiment: Comparison of % infectious

reach = 0.1 vs reach = 1.0

dynamics.

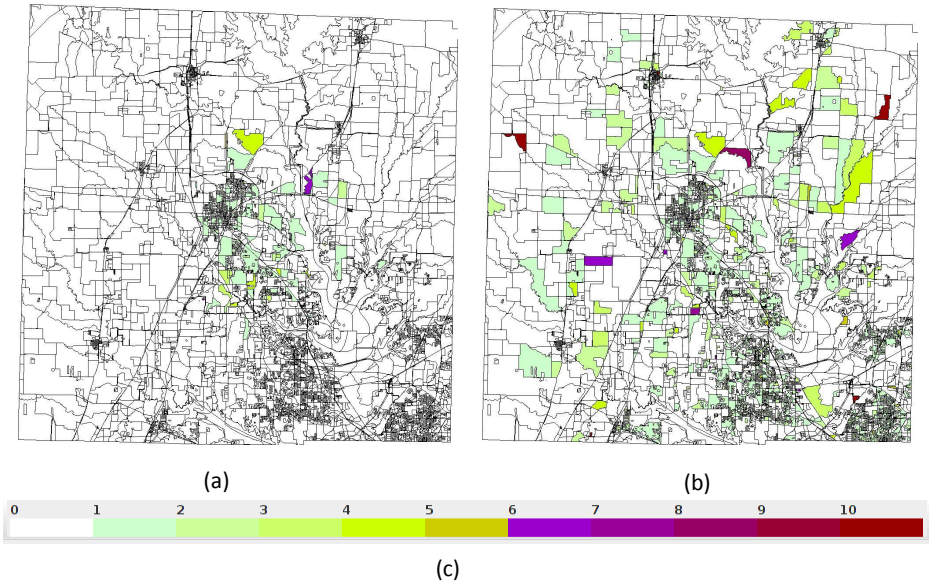

Figure 4.9. Reach experiment: Intensity of infection

(a) Day 23 - Peak of the epidemic when reach = 0.1. (b) Day 25 - Peak of the epidemic when reach = 1.0. (c) Color scale.

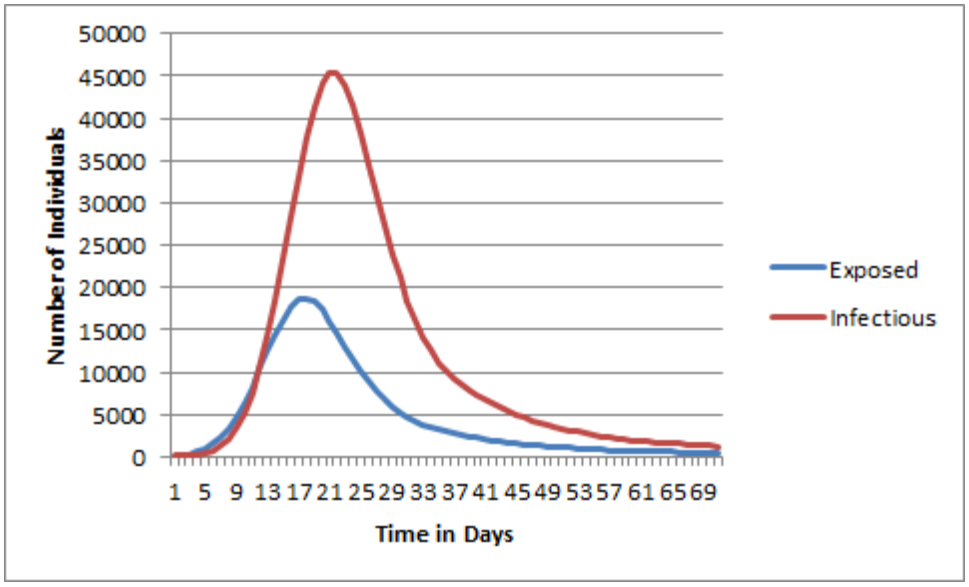

Figure 4.10. Exposed-infectious plot for gender experiment

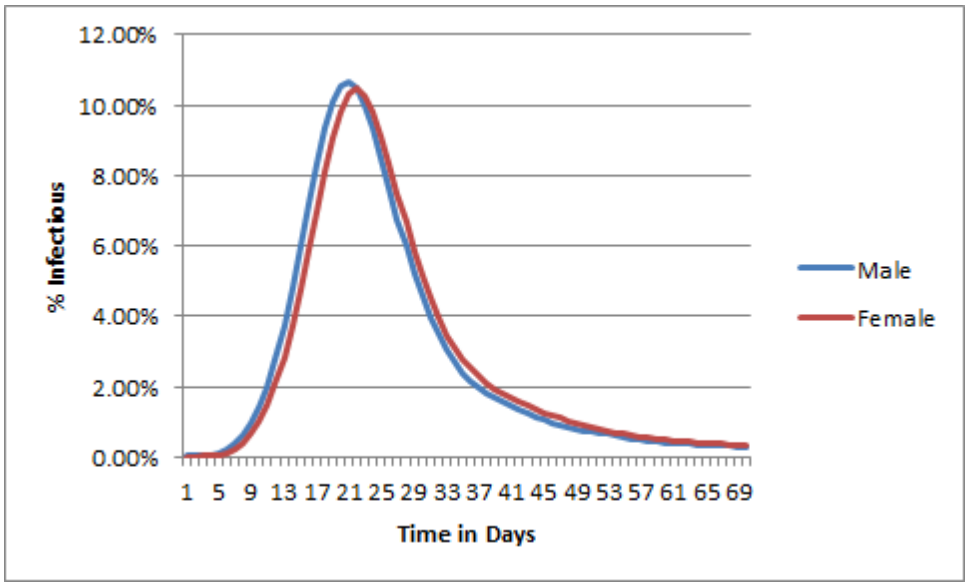

Figure 4.11. Gender experiment: % infectious: Male vs female

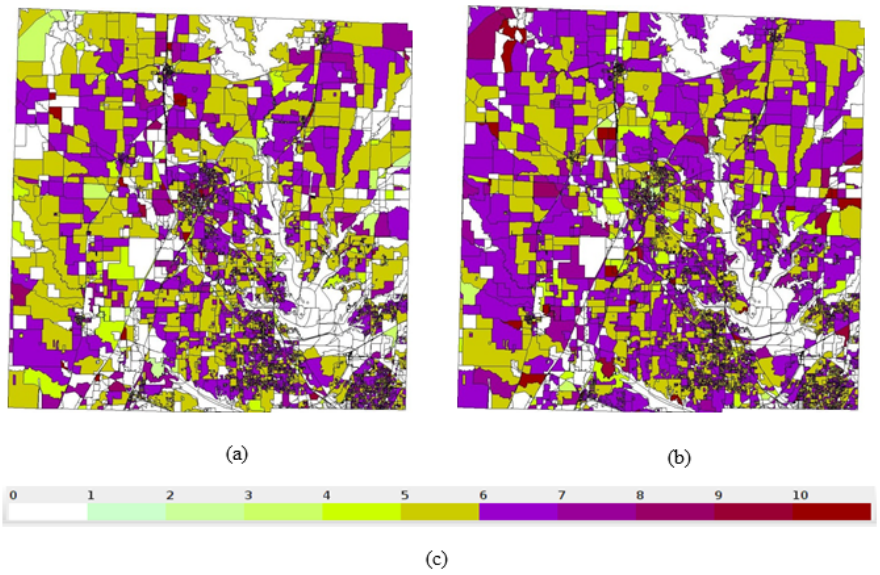

Figure 4.12. Gender experiment: Population distribution: male vs female (a) Population distribution of male demographic group in Denton. (b) Population distribution of female demographic group in Denton. (c) Color scale

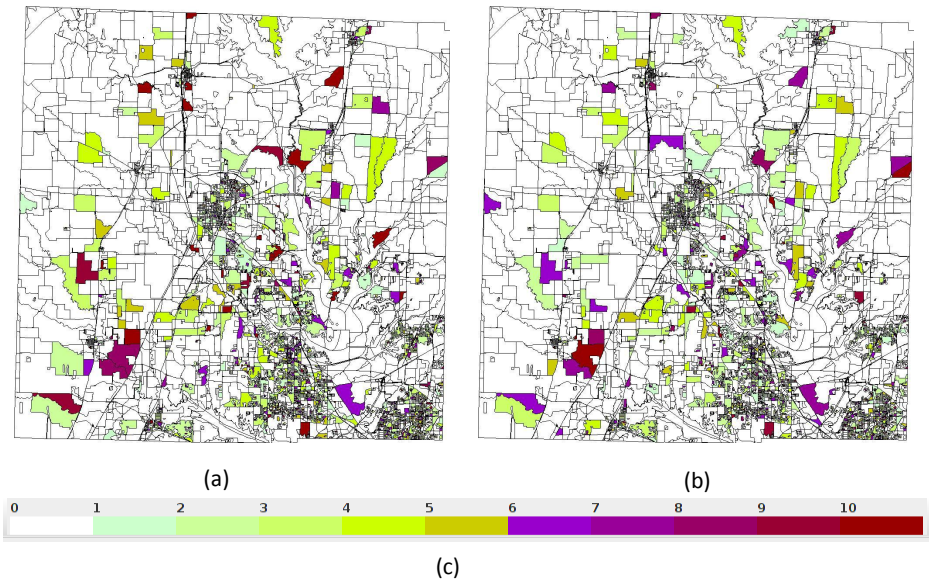

Figure 4.13. Gender experiment: Intensity of infection

(a) Intensity of infection on day 20 among males. (b) Intensity of infection on day 21 among females. (c) Color scale

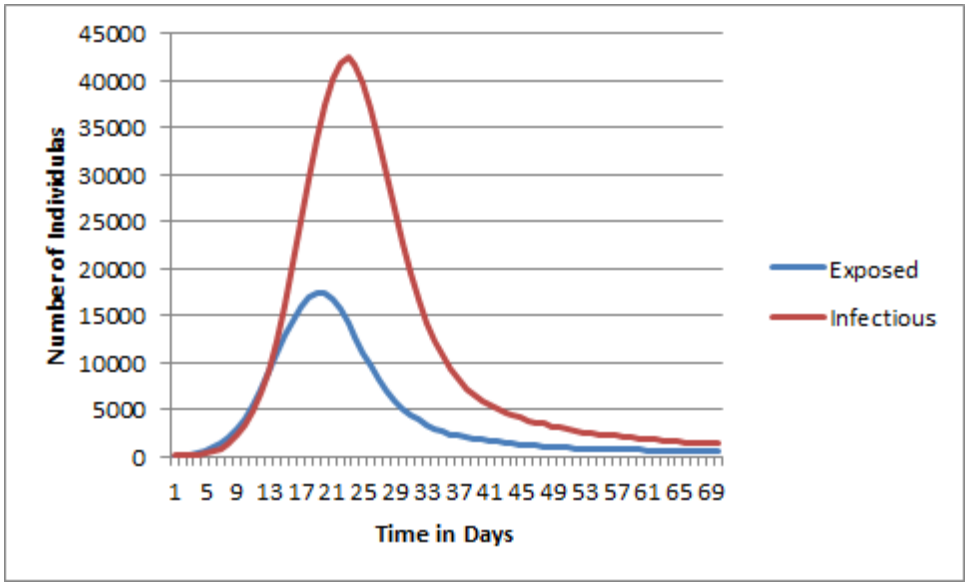

Figure 4.14. Exposed-infectious plot for experiment 3

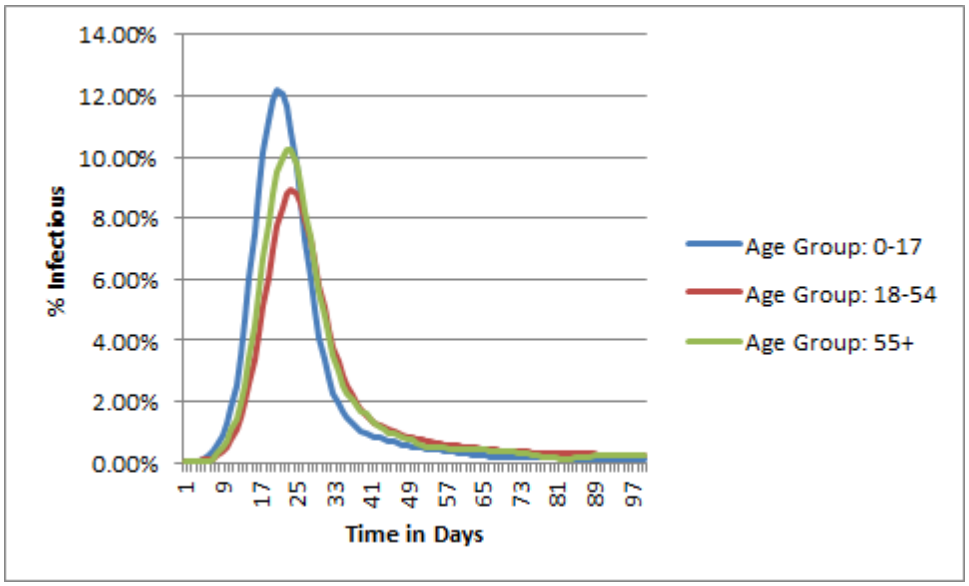

Figure 4.15. Age experiment- % infectious: age-group 1 vs age-group 2 vs age-group 3

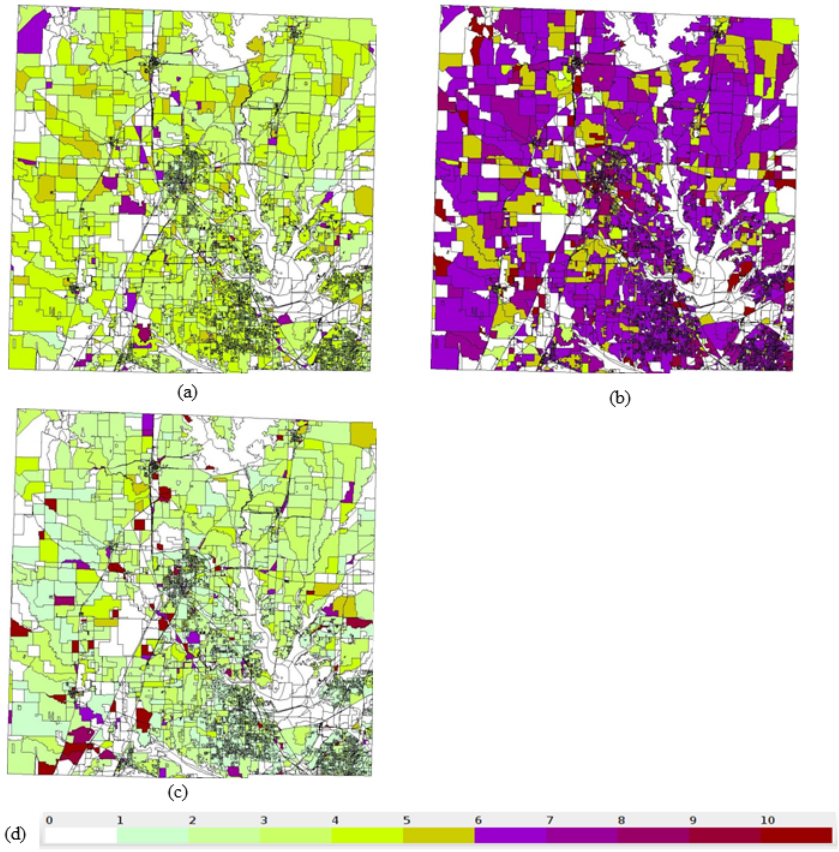

Figure 4.16. Age experiment: Population distribution

(a) Age-group 1. (b) Age-group 2. (c) Age-group 3. (d) Color scale.

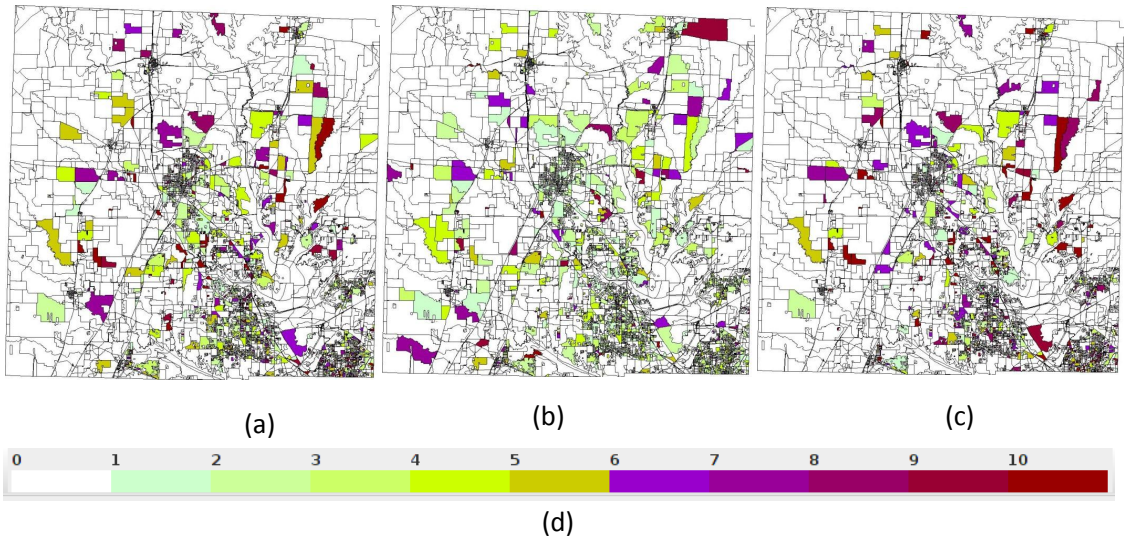

Figure 4.17. Age experiment: Intensity of infection

(a) Day 19 - age-group 1. (b) Day 27 - age-group 2. (c) Day 22 - age-group 3. (d) Color scale

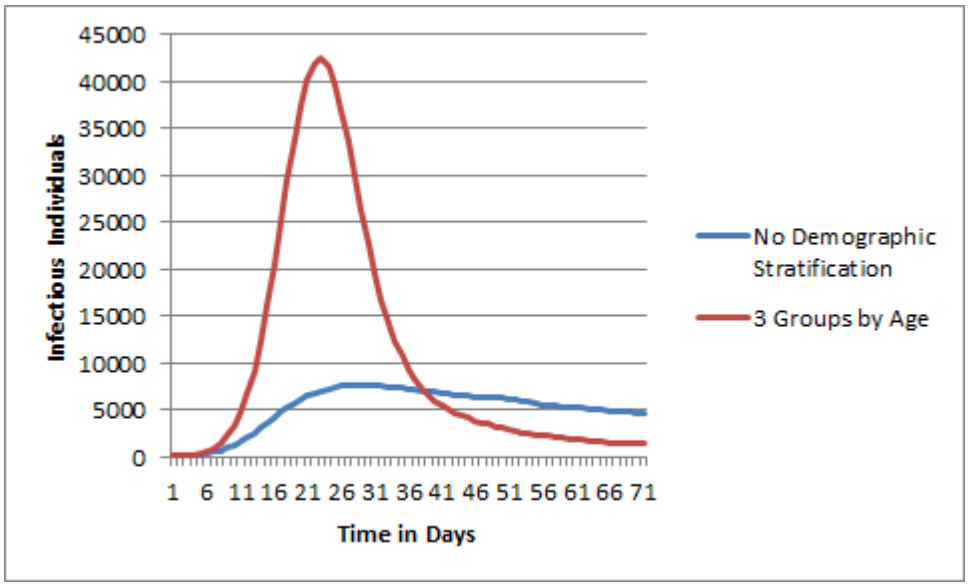

Figure 4.18. Number of infectious individuals: No demographic classification vs 3 age-groups

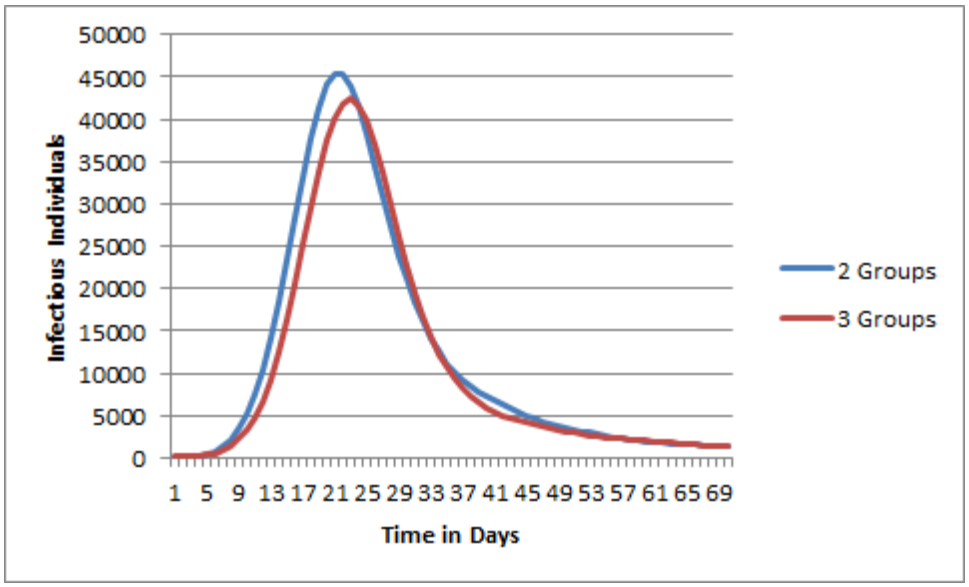

Figure 4.19. Number of infectious indiviuals: Gender vs age

CHAPTER 5

SUMMARY AND CONCLUSION

5.1. Summary

Epidemics have been recorded through-out the history and have caused significant human and monetary losses. Localized epidemics, which have the potential to grow into pandemics in an increasingly well-connected world, continue to pose a major threat. Prior planning to prevent or control the occurrence of epidemics is crucial, and public health officials and epidemiologists work towards developing response plans to mitigate an impending epidemic outbreak. However, the lack of reliable historic data due to under-reporting, emerging and re-emerging strains of diseases, and changing infrastructure in communities pose a major challenge for public health professionals in developing response plans. This necessitates simulation of disease spread to facilitate what-if analyses that help the process of preparing for an epidemic outbreak.

As part of this thesis, a framework to simulate and visualize infectious disease spread has been developed. The simulator uses census data to obtain spatial distribution of population. The population is stratified into demographic groups based on disease-specific demographic constraints. Disease spread is simulated by means of contacts between infectious and susceptible populations. The contact model is based on global stochastic field simulation paradigm, where contacts are generated locally and globally for every demographic group in each census block, randomly on the basis of population counts, geographic distance, disease and demographic parameters. The disease dynamics are visualized by means of a heat-map representation where a color scheme maps a color to the percentage of infected population in a census block.

Experiments were performed to study the spread of disease under various scenarios. Disease dynamics were studied by performing simulations using different disease and demographic parameters. Infectivity of the disease and contact rate, affinity, mobility and reach of different demographic groups were varied in different experiments. Experiments were

performed using no demographic classification, gender as the demographic, and age as the demographic for stratifying population. The census data for Denton county, Texas from US Census 2000 was used in the experiments.

Varying infectivity affected the extent and duration of the epidemic. Greater values of infectivity lead to outbreaks which peaked earlier, and infected larger number of people. Spatial distribution of population played a significant role in the spread of the disease. Demographic groups with differing values of affinity, mobility and reach had differences in disease dynamics, highlighting the role played by these parameters. Spatial distribution of population in a region and population mix with respect to disease-specific demographics, play an important role in the disease dynamics and pattern of disease spread in the region.

5.2. Future Work

While this framework provides a tool to study the spread of infectious diseases in a geographic region based on spatial distribution of population and social behavioral parameters of demographic groups, there is a lot of scope to enhance and extend it to provide more intuitive and beneficial results.

The simulator now uses a single demographic such as age or gender as the demographic constraint to stratify the population. This can be extended to include multiple demographics on the basis of which, the population is stratified. For instance, a combination of race and income levels can be used to stratify the population based on socio-economic factors. Furthermore, the idea of local contacts, which are limited to the same census block where the contacts originate, can be extended to include a group of adjoining census blocks, to model places like universities, schools or large work-places.

Additional visualization methods to visualize the pattern of disease spread would be very helpful in understanding the disease spread patterns. The representation of the region as a graph with census blocks as vertices and infectious contacts as edges, with number of infectious contacts being the weight of each edge would produce interesting results that would be worth studying.

GLOSSARY

Affinity: Affinity is the likelihood of a demographic group making a contact with another demographic group, including itself. There is a value of affinity for every pair of demographic groups. It is measured as the proportion of total contacts made by a group that are directed towards the other group.

Census block: A census block is the smallest geographic entity for which the census Bureau tabulates 100-percent data. Many census blocks correspond to individual city blocks bounded by streets, but blocks especially in rural areas may include many square miles and may have some boundaries that are not streets.

Contact: A contact is an interaction between any two individuals that is conducive to the transmission of a disease, or disease causing pathogens. A contact may or may not involve direct physical contact, depending on the nature of the disease involved.

Contact rate: Contact Rate is defined as the total number of contacts an individual makes with all other individuals in the population per time unit.

Demographic group: A demographic group is a sub-group of the population defined by demographic criteria like age, sex, ethnicity etc. The population is classified into demographic groups based on a demographic that is relevant to the epidemic being modeled.

Exposed/Latent period: The time period between a host's exposure to an infection and becoming infectious.

Geometry: Geometry of a feature such as a census block is the shape of the region available to GeoTools as represented in the census data.

Heat map: A heat map is a graphical representation of data where the values taken by a variable in a two-dimensional map are represented as colors.

Infectious period: The duration for which an infected individual is capable of transmitting the pathogen to a susceptible individual.

Infectivity: The proportion of people exposed to a pathogen that become infected.

Interaction coefficient: Interaction coefficient between any two census blocks is a measure of the strength of interaction between them. It is directly proportion to the population densities of the participating blocks and inversely proportional to the distance between them.

Internal point: An internal point is a single point within a geographic entity that represents its approximate geographic center. It is used by the Census Bureau to denote the geographic coordinates of an entity such as a census block which it represents.

Mobility: Mobility of a demographic group is the likelihood of a contact made by an individual of the group being with a group in an external block.

Reach: Reach indicates the farthest census block that can be reached by a demographic group in a particular census block to make a contact. It is measured as a fraction of the distance between census blocks that are farthest apart in the region.

Risk group: Risk group for a disease is a group of population identified by a demographic, which is at an elevated risk of contracting the disease than others. //check for accuracy

Time step: A time step is the smallest unit of time used in the simulation. The epidemic simulation is updated for every time step.

Transmission probability: Transmission probability is the probability of the successful transfer of a pathogen from one host to another when a contact is made.

BIBLIOGRAPHY

1. *Geographic information systems faq*, 1997.
2. *Census 2000 geographic definitions"*, 2000.
3. *Census 2000 geographic definitions*, 2007.
4. *Ecdc daily update - pandemic (h1n1) 2009*, 2010.
5. *Geotools - the open source java gis toolkit*, 2011.
6. *Cellular automaton*, 2012.
7. K Abbas, Mikler AR, Ramezani AR, and Menezes S, *Computational epidemiology: Bayesian disease surveillance*, (2004).
8. H Abbey, *An examination of the reed-frost theory of epidemics*, (1952).
9. Linda J. S. Allen, *An introduction to stochastic epidemic models part 1*.
10. N. Bailey, *The mathematical theory of epidemics*, 1957.
11. Fred Brauer, *Compartmental models in epidemiology*, (2008).
12. John S. Brownstein, Cecily J. Wolfe, and Kenneth D. Mandl, *Empirical evidence for the effect of airline travel on inter-regional influenza spread in the united states*, (2006).
13. K.M. Carley, D.B. Fridsma, and E. et. al. Casman, *Biowar: scalable agent-based model of bioattacks*, (2006).
14. CD Corley and Mikler AR, *A computational framework to study public health epidemiology*, (2009).
15. Angus Deaton, *Health in an age of globalization*, Tech. report, NATIONAL BUREAU OF ECONOMIC RESEARCH, .
16. S. Eubank, *Scalable, efficient epidemiological simulation*, (2002).
17. Justus Friedrich and Carl Hecker, *The epidemics of middle ages*, , 1835.
18. Manfred S. Green and Tiberio Swartz et al., *When is an epidemic an epidemic*, (2002).
19. Peter Haggett, *The geographical structure of epidemics*, , .
20. Valerie Isham, *Stochastic models for epidemics*, (2004).

21. Merrill and Timmreck, *Introduction to epidemiology*, , 2006.
22. AR Mikler, Jacob R, Gunupudi V, and Patolla P, *Agent-based simulation tools in computational epidemiology*, (2004).
23. AR Mikler and Venkatachalam S., *Modeling infectious diseases using global stochastic automata*, (2005).
24. AR Mikler, Venkatachalam S, and Ramisetty-Mikler S, *Decisions under uncertainty a computational framework for quantification of policies to address infectious disease epidemics*, (2007).
25. KD Patterson and Pyle GF, *The geography and mortality of the 1918 influenza pandemic*, (1991).
26. Fu S and Milne G, *Epidemic modelling using cellular automata*, (2003).
27. Wolfram S, *Statistical mechanics of cellular automata*, (1983).
28. R.W. Sinnott, *Virtues of the haversine*, (1984).