Spring 2014

A Framework for Synthesizing Agent-Based Heterogeneous Population Model for Epidemic Simulation

Madih Sahar
Purdue University

Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Madiha Sahar

Entitled

A Frameworks for Synthesizing Agent-Based Heterogeneous Population Model for Epidemic Simulation

For the degree of Master of Science in Electrical and Computer Engineering

Is approved by the final examining committee:

ARIF GHAFOOR
Chair

CHARLIE HU

WALID G. AREF

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): ARIF GHAFOOR

Approved by: M. R. Melloch 12-17-2013
Head of the Graduate Program Date

A FRAMEWORK FOR SYNTHESIZING AGENT-BASED HETEROGENEOUS POPULATION MODEL FOR EPIDEMIC SIMULATION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Madiha Sahar

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Electrical and Computer Engineering

May 2014

Purdue University

West Lafayette, Indiana

I dedicate this thesis to my beloved parents Amjad Javed and Rukhsana Amjad, my brothers Arslan Amjad and Haseeb Amjad, and my mamoo jaan without their support, encouragement and love this would not have been possible.

ACKNOWLEDGMENTS

I would like to convey my special thanks to my advisor, Professor Arif Ghafoor for his kind, warm, and unwavering support and advice, and most of all for his patience throughout my graduate studies. I will always be thankful to him for his affectionate guidance throughout my graduate studies. I would also like to express my thanks to my committee members Dr. Y. Charlie Hu and Dr. Walid Aref for their valuable inputs.

The research presented in this thesis is supported by grants from the Defense Threat Reduction Agency (DTRA)(Grant Number: HDTRA-1-10-1-0083), the National Science Foundation (Grant Number IIS-0964639) and the Cyber Center at Purdue University.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Sahar, Madiha M.S.E.C.E., Purdue University, May 2014. A Framework for Synthesizing Agent-Based Heterogeneous Population Model for Epidemic Simulation. Major Professor: Arif Ghafoor.

Social interactions play an important role in spread of a disease. In this thesis we propose a probabilistic approach to synthesize an agent-based heterogeneous population interaction model to study the spatio-temporal dynamics of an air-born epidemic, such as influenza, in a metropolitan area. The proposed methodology is generic in nature and can generate a baseline population for the cities for which detailed population summary tables are not available. The joint probabilities of population demographics are estimated using the International Public Use Microsimulation Data (IPUMS) sample data set. Based on the population density and the socio-economic status, the population is divided into three types of residential areas. Agents, representing individuals, are assigned various activities based on their education, age, and gender. Since transportation can also influence the spread of a disease, this activity, with a finite time span, is also assigned to individuals. The proposed approach is used for the city of Lahore, Pakistan. The agent-based model for Lahore is synthesized and a rule based disease spread model of influenza is simulated for the city population. The simulation results are visualized to analyze the spatio-temporal dynamics of an influenza epidemic for Lahore.

1. INTRODUCTION

Infectious disease in humans occurs when an infecting agent impacts the immune system of an individual. The compromised individual shows symptoms of infection such as fever. Infections may occur occasionally to any individual but when an infectious disease impact individuals on a large scale in any area frequently, it is called endemic. When the number of persons being affected is higher than the expected number of cases at a time of year in any particular area then it becomes epidemic. Epidemics are contained in small areas such as parts of a country or a city. If there are cases of an infection spread out in different parts of the world then the infectious disease is known to be a pandemic. The most frequent and notorious epidemics and pandemics reported in history are of influenza and respiratory infections. For instance in the 1918 influenza pandemic, healthy adults between 20 and 40 had the highest morbidity and mortality rate [1]. The twentieth century has witnessed three influenza pandemics and scientists agree that it is a question of when, not if, the next pandemic will occur that can cause colossal human and economic losses. Estimates of deaths from the 1918 Spanish influenza pandemic range from 21 million to 100 million. Similarly, the 1968 Hong Kong pandemic killed millions of people (estimates run between 1 million to 4 million) [2]. With the recent spread of the H5N1 influenza virus, commonly called avian flu, the World Health Organization described the world's status as in Phase 3 (Pandemic Alert Period) of six pandemic phases. Thousands of human cases were confirmed with H5N1 and hundreds of deaths [3]. It is believed that all three of the 20th century influenza pandemics have avian origin [4].

1.1 Surveillance of Infectious Disease Epidemics

To protect both human lives and national security a nation in collaboration with international partners must be prepared for and able to respond quickly to localized and global events, as well as the next pandemic. Coordinated global surveillance is necessary for early detection and action. Increased global connectivity enables migration of microbes, leading to drastic changes in the pattern of global health and disease. Sub-Saharan Africa has a big dilemma in combating HIV/AIDS. It is reported that more than 25 million people are infected with HIV/AIDS. The problem is global in nature and sub-Saharan African countries are not the only ones that face this problem; countries in Latin America and the Caribbean have a similar situation and their strong global ties can lead to a global pandemic. In essence, the health of Americans and the health of people around the world are more closely linked than ever before. Epidemics of novel re-emerging infectious diseases (IDs) can quickly spread globally through various avenues, the leading one being air travel. An outbreak in a distant country can threaten the health of public at home. Greater movement of people and of products including food, drugs, and medical devices can increase exposure to potential health risks originating outside the United States. As highlighted in a recent report by the White House : The threat of contagious disease transcends political boundaries, and the ability to prevent, quickly detect and contain outbreaks with pandemic potential has never been so important. An epidemic that begins in a single community can quickly evolve into a multinational health crisis that causes millions to suffer, as well as spark major disruptions to travel and trade. Addressing these transnational risks requires advance preparation, extensive collaboration with the global community, and the development of a resilient population at home. To protect both human lives and national security our nation in collaboration with international partners must be prepared for and able to respond quickly to localized and global events, as well as the next pandemic. In the specific case of a pandemic, it is estimated that 30% of the US population could become

ill, including first responders, health care professionals, and government policy and decision makers. It is unclear what portion of the population will be stricken most severely [5].

1.2 Our Contribution

Global epidemiological monitoring is a daunting challenge involving the collection and analysis of massive, time-evolving data of varying accuracy and reliability. Health incidents often have an incidence (signal) profile that is less than the statistical noise as experienced through the two most widely used health care monitoring systems [6], [7]. This is due to several reasons including the data quality and the sensitivity of the detection method. While conventional epidemiology has achieved significant successes in managing diseases and epidemics, the approach is inadequate in dealing with the high noise to signal ratio in case of bio attacks where the focus is on early detection [8]. Part of the reason for this is that conventional epidemiology has inherent limitations as it does not account for spatial, geographical, and social dimensions in modeling of epidemics. This challenge exacerbates with the ever increasing global connectivity of people. The intersection of global epidemic surveillance and the national security of a country needs to understood in the context of bio security, bio surveillance and medical countermeasures. The goal of this research is to propose a methodology and develop a tool that can be used as generic standardized platform that enables modeling of epidemic for arbitrary large metropolitan cities in various part of the world. A social networking based technique to study disease spread model for prediction and control of epidemic and pandemic spreads is proposed. A methodology for agent model synthetic population is generated using a disaggregated sample data. This data provides information about household and demographic attributes of every person living in household. Every individual in the generated population is called an agent. Agents are then assigned various activities such as staying home,

engaging in work places, attending educational institutes and commuting to different locations.

The main contributions of this thesis are summarized as follows.

- We propose a methodology to synthesize baseline population by computing demographic attributes of persons/agents for socio-economic groups in a city. The city is classified into various socrio-economic groups based on domain knowledge of the demographic attributes. The joint distributions are calculated from the Public Use Microsimulation Data, International (IPUMS). In our methodology we synthesize a baseline population by calculating the joint distributions for each socio-economic region.

- We describe the development of spatio-temporal agent model based on the set of assumptions using a demographic domain knowledge about the interactions and activities performed by the general population in a city. The domain knowledge includes percentage distribution of work types, work and education activity locations, and the rules of assignment of various activities across the city. The work type percentage distributions vary across socio-economic groups, as mentioned in the first contribution. The percentage distribution of work types and activity assignment rules allow us to closely capture the work and education activity patterns of the city.

- The synthesized spatio-temporal activity based agent model can be abstracted as a temporal agent graph. This graph is used to analyze the spatio-temporal spread of epidemic through a rule based simulation. The rules govern the spread of disease among humans. The simulation results are visualized to observe the spatio-temporal spread of the disease for a given initial triggering point of the epidemic. The visualization provides an insight of the epidemic and assists in making effective decision measures to prevent an epidemic from spreading by applying various countermeasures [9].

1.3 Thesis Organization

The organization of this thesis is as follows. Chapter 2 describes the background of epidemics, various disease spread models, the approaches used to synthesize population and preventive methodologies. In Chapter 3 we present a methodology to synthesis baseline population using the joint distributions related to various socio-economic groups. We synthesize population for the city of Lahore to elaborate the synthesis process. In Chapter 4 we describe the process of generating an activity based agent model system. Chapter 5 shows visualization results based on simulation of an epidemic. In Chapter 6 we propose possible future research tasks and the conclusions of the research presented in this thesis.

2. LITERATURE REVIEW

2.1 Characterization of Infectious Agents

Infectious agents are classified into various groups such as virus, bacteria, and protozoa etc. The classification of the infectious agents is based on their presence in atmosphere such as in air, soil, humans, water etc. and the process of transmission from one person to another. The diseases are categorized in air borne, water borne, vector borne and direct contact if the mode of transmission is air, water, vector(carrier of disease such as mosquito) and direct, indirect or prenatal contact respectively. The infectious agents when present is air or water infect people through respiratory tract, and oral or fecal ways [10]. Disease transmission requiring contact may transfer through blood or saliva or physical touch e.g. HIV/AIDS. Perinatal diseases are similar to disease transmitted by contact but are transmitted in the womb of the mother before the child is born e.g. hepatitis B [11].

2.2 Infectious Disease Dynamics

There are three types of roles persons play in disease dynamics. The individuals can be classified as susceptible, exposed, infected, recovered etc. Susceptible persons are those who can get infected due to some infectious agent. A susceptible individual may get infected by interacting with pathogens present in air, water, or soil etc. The infected individual is known as the host. The host interacts with susceptible individuals and transmit the infectious agent through air, water etc. There is a duration of infection which is specific to each infectious agent and the person's age. If the recovered person can never get infected again, it is called Susceptible, Infected, Recovered (SIR) model [12].

Fig. 2.1.: State Diagram of SIR Model

The variations of SIR model includes Susceptible, Infected, Susceptible (SIS) model and Susceptible, Exposed, Infected, Recovered (SEIR) model. Figure 2.2 shows the states of SEIR model [13].

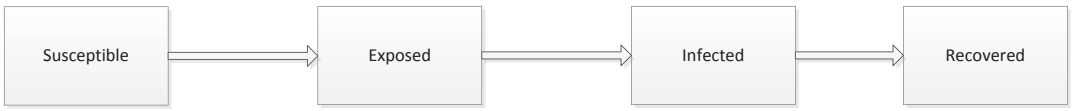

Fig. 2.2.: State Diagram of SEIR Model

The infection time of an individual is divided into two parts: latent period and infectious period. In latent period of infection an infectious agent is transmitted to an individual and the infected person cannot transmit infection to other persons. The infectious time is further divided into two time periods i.e. at first an individual is transmitting infection to other humans without having any clinical symptoms called incubation period and then the person starts to have clinical symptoms as well called infected period. The duration of the incubation period depends on the type of infectious disease.

The quantity of disease or force of infection is estimated to understand how long the disease may take to spread in various areas. Estimating prevalence of an epidemic is one way to measure the force of infection by calculating the proportion of infected persons at a given area at a given time. This can also be defined as incidence of

disease times the days in an infectious period. Measuring the incidence of disease is another way to compute the force of infection by computing how many individuals can get infected by one infected individual. Reproduction number (Ro), the average number of infected individuals caused by a single host, can also be calculated for a similar purpose. If Ro is greater than one, it indicates the possibility of an outbreak of an epidemic [5], [11].

2.3 Disease Transmission Process

Direct or indirect human contact plays important role in epidemic outbreak and its duration. Direct contact requires an infected and a susceptible person to be in close proximity such as in a house or work place etc. Indirect contact is when an infected and susceptible are not interacting directly and the environment acts as a medium to transmit disease such as at market places, hospitals etc. The intensity and spread depends on duration and frequency of effective contact of persons.

2.3.1 Spatial Transmission Models

There are three commonly known disease spread models: deterministic models, stochastic models and agent based models. In a simple deterministic model a homogeneous, uniform mixing, closed population is assumed and every person at any time a person is either susceptible, infected or recovered [14]. In stochastic model infection is transmitted if a susceptible person is in close proximity of an infected person. The spatial dimension of population is not considered in these models.

To consider spatial and temporal information of population for analyzing epidemic spread, various approaches have been proposed. One of them is the patch model [15]. In this model a city is spatially divided in several patches and population in each patch is assumed to be classified as three or more compartments such as susceptible, infected, recovered etc. Each patch has homogeneous, uniformly mixed and closed population. Every infected person can transmit infection to susceptible persons on

same rate. The infection transmission can also be computed as a function of distance between infected and susceptible person. In patch models population is divided into groups based on SIR states.

Multigroup disease spread model assumes that the population is divided into small groups depending on their spatial locations such as household, work locations etc. Every person in a group is connected to each other. One infected person can infect everyone in the same group.

Network transmission is another approach to study epidemic spread. Like multi-group models it considers population of groups and in order for disease to transmit an infected person must be in the same group as those of susceptible. The transmission of infection is zero intra groups. In network model all individuals in one group are not connected to every person of the group [15].

2.3.2 Agent Based Models

Agent based models allow to incorporate heterogeneity aspect to the traditional epidemic models. The agent based models requires data of every person which is difficult to collect. In order to deal with this problem, synthetic agent population is synthesized. The population synthesis process is highly data dependent and often have the shortcoming of unavailability of data requiring several assumptions to be made. Synthesizing an agent population to study disease spread network requires a connection network that is critical for disease spread. This contact network plays a key role in the transmission of disease among the population. The location information of agents in space and time is required to build the transmission network for disease. The network is built by understanding the types of activities performed by the individuals, the places for activities, the time required to perform an activity and estimation of the number of persons engaged in an activity. The person level contact is initiated by the activities individuals perform and is either direct or indirect. The direct contact in social networks is caused by social or personal level interactions, for instance, a

person interacting with his family members in a household. The indirect contact is when a infected person can transmit disease to a susceptible person without having a close social connection. For instance a person traveling in a bus can get infected because of an infected person traveling in the same bus. The indirect contact means that there is a disease transmission network between infected and susceptible persons without having them interacting on a one-to-one basis. This indirect contact is not a personal choice to make [16], [17].

2.3.3 Micordata Synthesis

Microsimulation is a process of simulating individual behavior to study and predict the effect of various complex and dynamic processes. In order to run a microsimulation a micro dataset is required. This data is created such that it represents the desirable characteristics of agents. Once agent model is synthesized, the behavior of the agents is simulated using some rule based or mathematical models. In activity and transportation based microsimulation models, agents are usually households i.e. persons living in a household performing a set of daily activities of interest, and using transportation in the area of interest. To date most of the microsimulation models synthesized are application specific and are based on a conventional approach proposed in [18]. This approach utilizes public use microsimulation aggregated data and disaggregated statistics of population such as public use microsimulation data (PUMA), summary files of individuals (SF) and small area statistics (SAS) [16], [5], [19], [20], [21].

Agent networks allow users to analyze the behavior of persons that is not static and it may change over time. An agent is an entity having attributes assigned to it. For instance, an agent has age, gender, and education assigned to it. The list of activities that it may perform, the locations of potential activities and their duration is also attached to every agent in the population. The decision made by agents to

perform activities may or may not be dynamic depending upon the requirements of the system [22], [16].

2.3.4 Baseline Population Synthesis Approach

The challenge in synthesizing a microsimulation dataset is to generate a baseline population which is used to construct the agent model. This process requires sample data from real social networks, the demographic details, the education and work and information about other activities to create the baseline agent population [23], [16], [5]. Iterative Proportional Fitting (IPF) is a well established and commonly used approach to synthesize agent population data [18]. This approach is used to synthesize population in travel demand systems [19], [24], [25], [22], and [8]. The data sets required by IPF includes:

- U.S Census Bureau Summary Tape File 3A (STF-3A) data

  STF-3A files are summary tables of demographic attributes of population from 1990 census data. These tables are available for every block group and census tract.

- Public Use Microdata Area (PUMA)

  If more than one block group or census tract are combined, they are called PUMA. Each PUMA has a Public use Microdata Sample (PUMS) which is 5% representative of a sample population. This sample data contains the complete structure of a household. Family members living in a sample household, annual income etc are tabulated in PUMS data set.

- TIGER/Line

  Master Area Block Level Equivalency or Geographic Correspondence Engine (MABLE / Geocorr) MABLE / Geocorr is a web portal that generates files containing relationships of geography of U.S. according to the 1990 census. It generates a correlation list known as "equivalency files", "crosswalks", and " geo-

graphic corresponding files". Census block is the smallest unit of MABLE/Geocorr files. This file contains land use geographic layout of census data of census block groups.

- Forecast Marginal File

  The population synthesizer requires a file containing forecast marginal distribution of attributes selected in STF-3A and PUMS as a function of the census block group. These forecasts are provided by the transportation agencies.

- Network Data

  The network data contains location coordinates where activities are performed.

  IPF requires generation of multidimensional contingency tables to compute the household distributions. These tables are computed using PUMS dataset that corresponds to the selected STF-3A table [22]. The dimensions of both the tables must correspond to each other for the synthesis of population using IPF [26]. A weight is assigned to every household type which represents the possibility of having a household. To calculate the proportion of household in every block group, IPF is run as a two stage process. This methodology works only if marginal totals of all the attributes are available for the area. The detailed algorithm is explained in [18]. Figure 2.3 shows the overview of population synthesis using IPF.

  IPF is a mechanism to adjust the values in the data table to get the desired marginal totals. Each row cell value is proportionally adjusted by the desired marginal value. This is done by multiplying each cell value by marginal row sum and dividing it by actual row sum. This step is called row adjustment. Similarly, each cell value is proportionally adjusted column wise by the desired marginal value. This step is called column adjustment. These two steps complete the first iteration of the IPF algorithm. The end result of this algorithm is probability distributions of household types as tabulated in STF-3A and PUMA dataset [27].

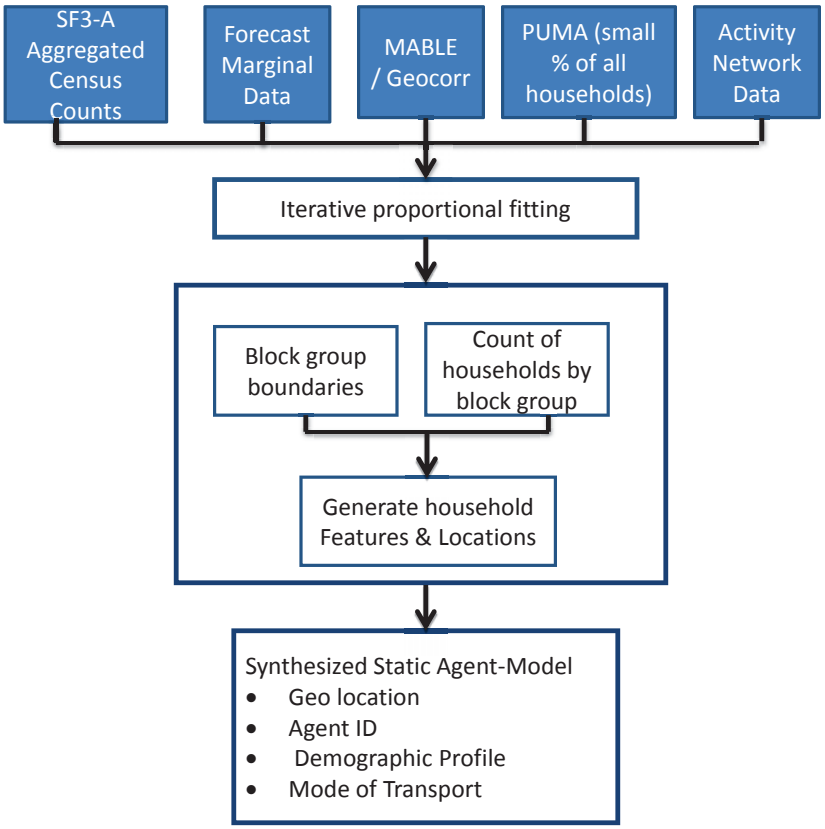

Fig. 2.3.: Overview of Population Synthesis Using IPF

Challenges Faced to Use IPF

There are some limitations to Iterative Proportional Fitting methodology when it is used to synthesize the baseline population. These limitations are widely discussed in various research such as [21], [26], and [28]. For example the IPF can only compute joint distributions of households level for any input data. It cannot compute the probabilities of having certain demographic attributes. Its second drawback is that if the input STF-3A data is not consistent with PUMA dataset, we face the zero cell problem where the IPF assigns a zero probability to every household for which the the input datasets are not consistent. Therefore, this approach cannot be used to synthesize population when incomplete marginal data is present. Several modified

algorithms and optimizing techniques are proposed to deal with these issues. The approaches are discussed in [21] and [26]

For example the methodology using Markov Chain Monte Carlo method to calculate joint distributions is presented in [29]. Other technique which optimize the IPF include classical optimizing and heuristic optimizing techniques. The simulated annealing and genetic algorithms are explained in [30] and [28]. These algorithms are computationally intense. Population synthesis of small population, or selecting less demographic attributes can be synthesized with these techniques [31].

2.4 Preventive Measures

Individuals with different socio-economic backgrounds have different responses when there is an epidemic outbreak. To contain an epidemic, the preventive measures are needed both at individual and public level. Individuals living in high income area are more likely to respond to vaccination and other antiviral kits while population from the low income areas may rely on their immunity to fight the disease [9].

Public interventions include quarantining an area by closing schools or work places, isolating an area by shutting down all the transportation to and from that area, and increasing social awareness about various techniques to fight the epidemic [32], [33].

3. POPULATION SYNTHESIS FOR SPATIO-TEMPORAL AGENT MODEL

3.1 Introduction

In this chapter, by using the city of Lahore as a case study, we propose a methodology for synthesizing a generic agent based model for epidemiology analysis and predictions about disease spread in large metropolitan cities. The model incorporates human interaction and is generic in terms of its applicability to any metropolitan city. The purpose of this chapter is to describe the process of generating the *baseline population* of agents with different demographic attributes including education. We define baseline population in terms of the percentage of individuals in each socio-economic class. The overall methodology of agent model synthesis is divided in five components. Figure 3.1 shows these processes.

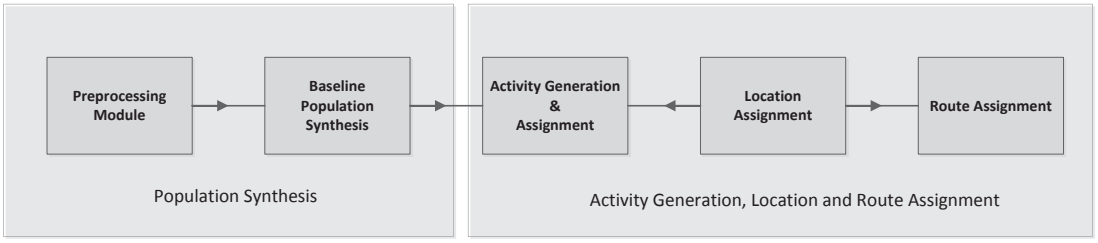

Fig. 3.1.: Synthesis of Spatio-Temporal Agent Model

In this chapter we elaborate the first two components of the agent model synthesis process. Our approach of synthesizing baseline population is based on computing joint distributions of various socio-economic groups using the Integrated Public Use Micro-data Series, International (IPUMS) sample data containing various demographic attributes of a representative population. A city can be divided into small

geographic units such as union councils to create a geo-centered synthesis of heterogeneous population.

The first challenge is to identify demographic attributes which are important to analyze disease spread in the city. The census data, and the IPUMS sample data are collected according to the selected demographic attributes, also known as the control variables. The socio-economic grouping of union councils is required to incorporate the difference that may exist in the distributions of demographic attributes, education, and other activities. Using the above mentioned datasets, the baseline population is synthesized and is used to generate temporal activities among individuals as explained in Chapter 4. The process of computing the joint distributions for various demographic groups illustrated in next section.

3.2 Population Synthesis System and Demographic Categorization

Constructing an individual based population to study epidemic spread is a highly data dependent process. IPUMS sample data and demographic domain knowledge is used to synthesize a population's demographic attributes, social interactions and activity behaviors which represent the actual population of a city. All the individuals or agents for the synthesized population must be consistent with the percentage of sample data available for the city under consideration, which in our case is Lahore. The geographic data (Geo-referenced shapefile) is available for Lahore which provides continuous space for the agents perform various activities including transportation. Using this file, the spatial patterns of an agent's activities are observed. Figure 3.2 shows the percentages of a person's demographic attributes including education. The population of the city is divided into three groups. One group of population lives in areas which can be highly dense and the average education level attained by the residents of these areas is low. The other type of areas can be characterized as having a low population density, but with habitants who are highly educated. These two population groups represent, the extreme cases and must be analyzed separately

since the spread of disease in these areas can have different spatio-temporal patterns. In particular, the people living in a dense area can have a different interaction pattern in terms of number of individuals with whom direct or indirect interaction takes place. People living in other areas can have a less frequency of interaction which depends on the type of work they perform and their life style. The major part of the city can have mixed population which can be represented by the attribute values falling in the middle of the two extreme groups. The probabilities of the first two groups can be extracted from the sample data and the probabilities of the third group are assumed to be the average of the above mentioned two extreme groups.

3.2.1 Input Datasets

Below is the description of data sets needed to synthesize the agent model including the Integrated Public Use Microdata Series, International (IPUMS), and the Census data. For the proposed methodology, we select age, gender, the type of area of residence, and the education level from the available demographic attributes in IPUMS.

Integrated Public Use Microdata Series, International (IPUMS)

The first data set used in the synthesis of the agent model is the Integrated Public Use Micro Data Series, International (IPUMS) which is disseminate census microdata. Generally IPUMS is available for a small percentage of households. The methodology proposed in [34] is used to use this data for generating micro data records for all the households upto 100%. IPUMS for the city of Lahore is available by the University of Minnesota [35]. This data represents a 2% sample of the census data collected in 1973 for the whole city and contains about 100,000 individuals divided into urban and non urban categories. Every individual has several attributes such as household size, demographic attributes (age, gender, Marital Status), and the maximum education level attained.

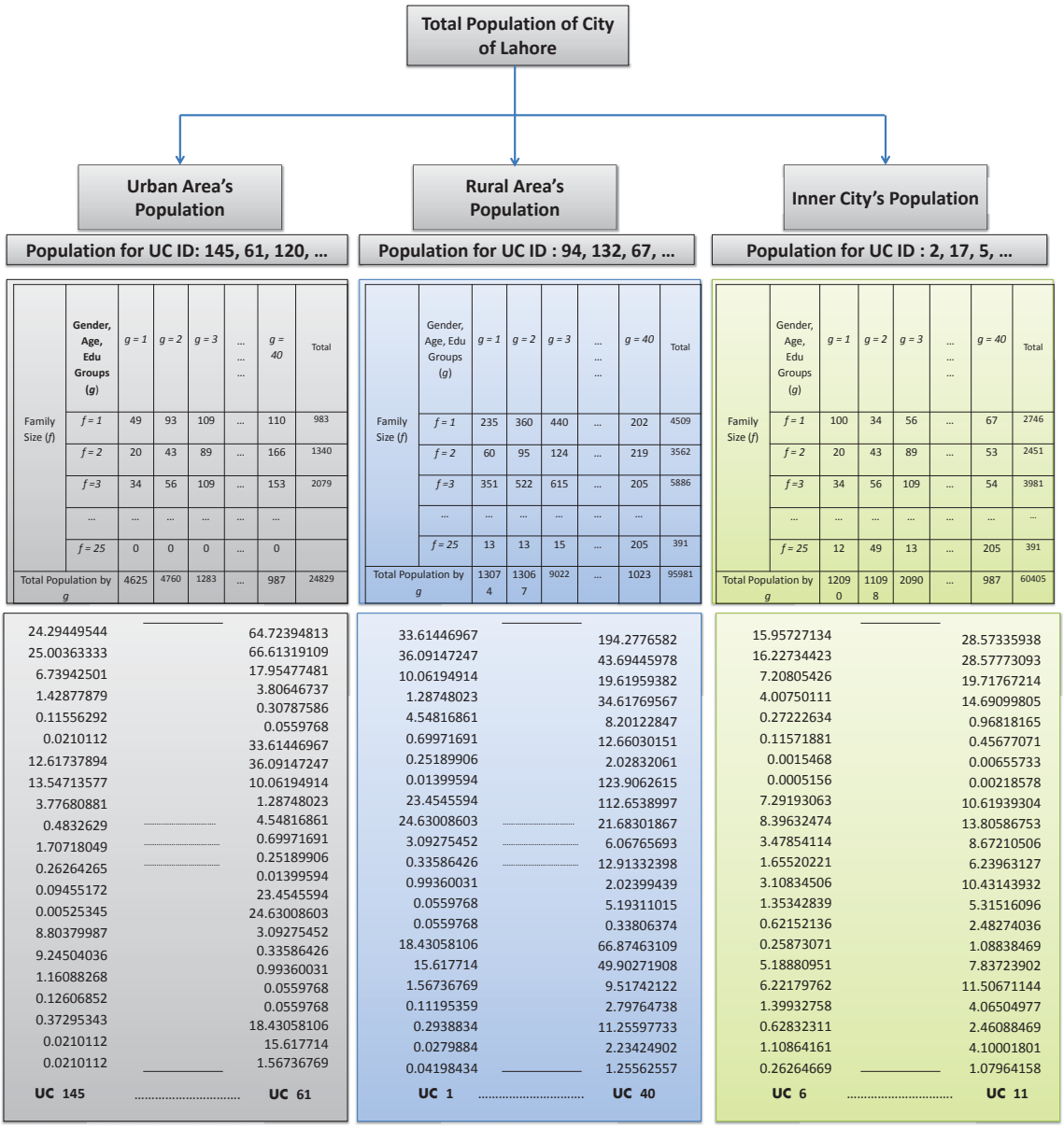

Fig. 3.2.: Example of Baseline Population for Case Study of Lahore

The smallest unit of geographical region in IPUMS is a district. The smallest unit of individual grouping is the household as discussed in Section 3.4 . There are two types of household. In the sample data, age of an individual can have a value between 0 and 99. For this research, we divide the age attribute into five groups. Each

Table 3.1: Age Group of Population

Age (years)	Age Group
0-15	1
16-30	2
31-45	3
46-60	4
60 +	5

of the first four groups corresponds to a duration of 15 years. The fifth age group represents all the individuals who are of age 60 years or above. Table 3.1 summarizes the distributions of the groups.

There are 10 levels of education reported in IPUMS data. Education and age of individuals are combined to provide a higher level of grouping. This aggregation considerably reduces the complexity to compute the joint distributions without losing important detail. We consider four levels of education which include: (i) Basic level Education, (ii) Intermediate level education, (iii) Higher level education and (iv)No education. These levels are used to assign activities to agents as discussed in next chapter. Table 3.1 represents the grouping of education with age.

We now elaborate the first two modules of Figure 3.1.

3.3 Preprocessing Module

As mentioned above, for our methodology we choose age, gender, type of area of residence, and education level from available demographic attributes in IPUMS dataset. Table 3.2 represents this data. The first column (Persons) represents the family size of each individual considered in the sample. The second column (Urban) in table represents the residential area of the city. In this data all the urban areas are given a label (urban) with value 1, and all the rural areas have code 2. The third

Table 3.2: IPUMS Sample Data

Persons	Urban	Pernum	Age	Edupk
7	2	1	19	320
7	2	2	18	310
7	2	3	15	310
7	2	4	13	230
7	2	5	10	210
7	2	6	9	210
7	2	7	6	210
3	2	1	27	320
3	2	2	25	220
3	2	3	3	000
2	2	1	23	310
2	2	2	17	230

column (Pernum) is identifier attached to individuals living in one family. Note, all individuals with Persons value = 7, i.e. having (Pernum) from 1 to 7 represents one household. Next two columns, (Age) and (Education) level, are grouped as mentioned in Section 3.2.1.

One of the limitations of the IPUMS sample data is that it only identifies each person either living in urban area or rural area. In other words the data does not provide any information about the geographical locations of these two areas. Another limitation of this data is that it contains sample population of only two types of residential areas which are urban and rural. It is possible that a city may have some areas which are not classified as rural or urban. According to the census data, "inner" city areas are also identified, which are not classified in IPUMS data. To address this discrepancy, we assume that the population in inner city has the population distribution which is mixed of both urban and rural areas. For the inner city areas,

the joint distributions are computed by taking an average of both urban and rural populations.

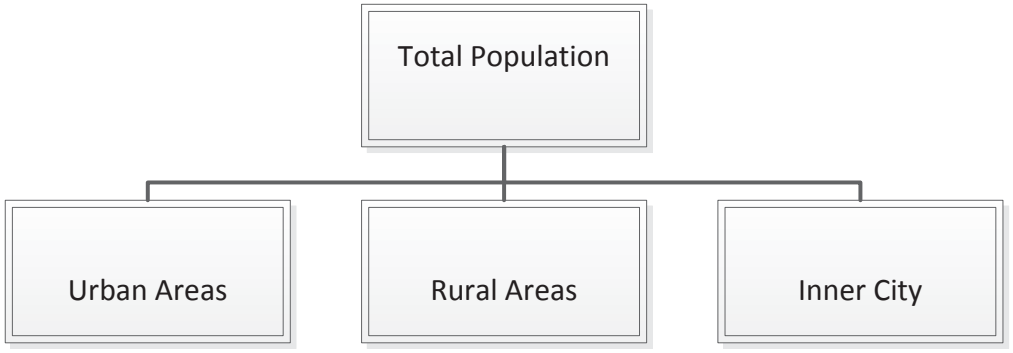

Fig. 3.3.: Population Distribution

Accordingly, Figure 3.3 shows the overall classification of the city population in three residential areas. *We use our knowledge of the city* to associate union councils to each of the three groups. Table 3.4 provides an association this table is pair of the *demographic domain knowledge.*

Table 3.3: Union Councils in Three Groups

Urban	Rural	Rural	Inner City	Inner City	Inner City	Inner City
145	67	78	47	2	17	130
61	69	1	82	73	33	71
120	94	95	101	139	113	12
63	132	14	18	24	5	79
147	44	127	29	75	54	87
62	96	115	81	107	86	135
144	99	52	20	10	138	106
123	42	40	27	19	28	85
64	97	16	134	76	38	11
53	32	39	25	114	30	23
66	133	141	102	131	100	110
148	84	116	88	46	68	6
150	60	143	21	9	74	7
152	35	112	34	58	55	80
65	109	51	3	45	59	93
50	126	146	15	36	129	83
122	41	118	70	140	22	72
119	31	49	92	89	111	108
	117	37	26	105	103	43
	121	142	56	48	128	77
			98	90	8	13
			137	57	91	136

3.4 Baseline Population Synthesis Module Description

Joint distributions for demographic attributes are calculated in this module and are used to synthesize the baseline population. This module takes the IPUMS sample data, the grouping of union councils and information about the percentage population for each union council as input datasets. The output generated by this module is the percentage distribution of the overall population over the whole city. In particular, every person is assigned an age, gender, an education level and the family size. Figure 3.4 shows the interconnection of different modules of population synthesis process in term of inputs and outputs

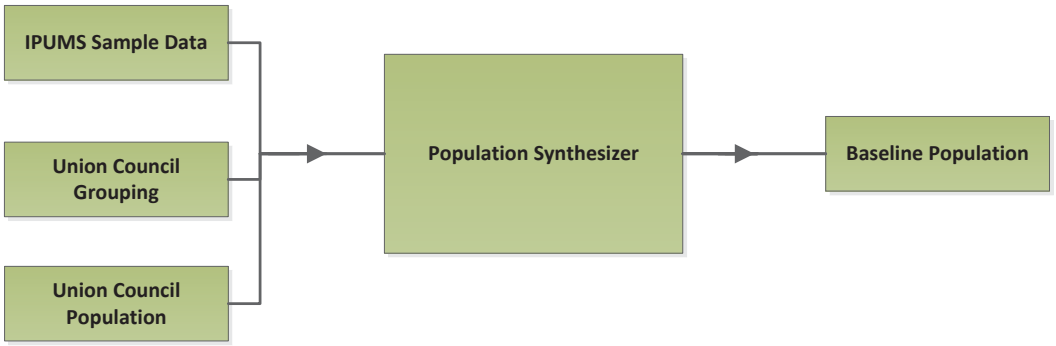

Fig. 3.4.: Context Diagram of Baseline Population

Every household in the baseline population is assumed to be either a measure of a household or does not have any family household. Households with family size less than 15 are assumed to be family households. A household with size above fifteen is assumed to be a non family household or a *group quarter* [22]. Every household is assumed to have at least one person living in it. All households with more than 25 persons living in it are grouped as one early. Household family demographics are extracted from the IPUMS sample data.

The data flow diagram to synthesize the baseline population is displayed in Figure 3.5. It can be noticed that the synthesis of baseline population of agents for any

metropolitan city requires small datasets which include IPUMS and census data. The first module, called Grouping and Joint Distribution Calculation Module performs a two step process. It groups the population based on their age and education level as explained in Section 3.2.1. Subsequently, the probability distribution for each group of age, education, gender, and family size is computed. We consider population in groups of age, gender, and education and compute the joint probabilities in terms of size of every group. The next module uses the union council population data and union council group data to compute the probability of households of every size in every socioeconomic group of population. Using the household size probabilities, the probability of houses with each family size is every union council is computed. Each household is then populated with persons using the joint distributions of demographic attributes extracted in Grouping and Joint Distribution Calculation Module. The output of this module is a file with number of persons in every group of population. The Person Identifier Assignment Module generates agent model and each agent is assigned a unique identifier along with demographic attributes. Household Identifier Assignment Module assigns household ID to the population to complete the process of synthesis of baseline population.

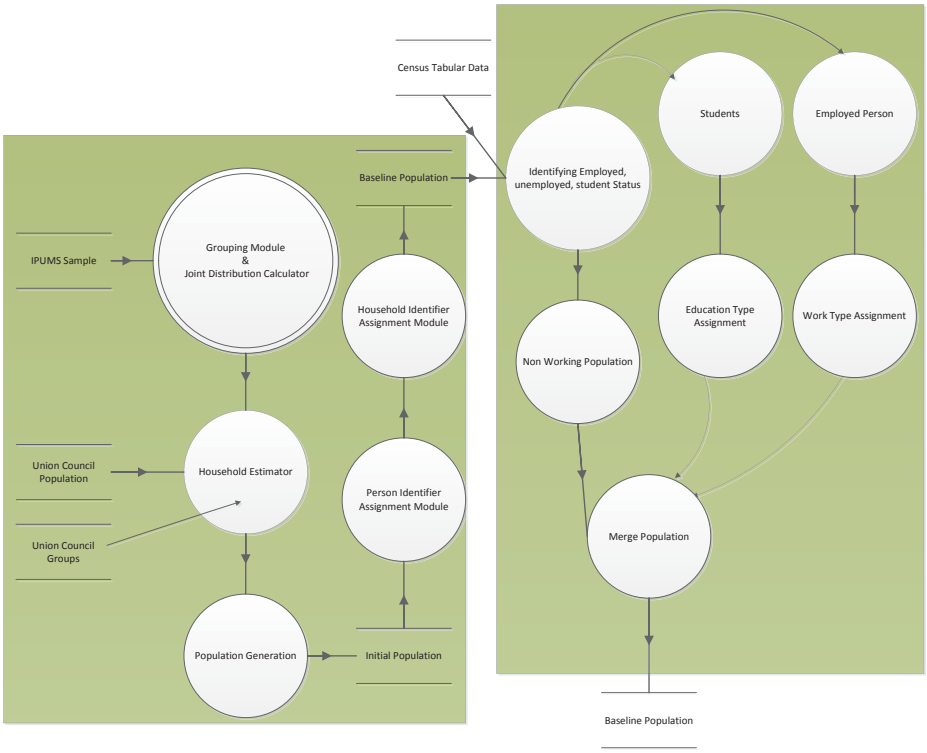

Fig. 3.5.: Process of Baseline Population Synthesis

3.5 Case Study: Baseline Population Synthesis for City of Lahore

We now illustrate the process of population synthesis for the city of Lahore. For this purpose we assume four control variables which are age *(A)*, gender *(G)*, education *(E)*, and family size *(F)*. These variables are random variables and can assume various values as described below. We compute number of households in every socio-economic group as shown in Figure 3.2 the process of computation is an existing approach extending the approach proposed in [34].

For our case, the possible values for A, G, $E$and F are $\{1, 2, 3, 4, 5\}$, $\{1, 2\}$, $\{E0, E1, E2, E3\}$, and $\{1, 2, 3 \ldots 25\}$ respectively.

Let U be a random variable representing ID of a union council for the city of Lahore. The possible values for U are $\{1, 2, 3 \ldots, 151\}$.

Let S be the random variable that represents the socio-economic status of the union councils as grouped in Table 3.4. The possible values for S are $\{1, 2, 3\}$.

Let n_f = Number of persons with family size f

Note: n_f is a multiple of family size.

N = Total population

n_s = Number of persons in socio-economic group s.

The probability of a person being in group s is computed as follows.

$Prob(S = s) = n_s/N$

The probability of having a household of family size $F = f$ is calculated for group $S = s$ as follows.

$Prob(House\ with\ F{=}f \mid S{=}s) = n_f/N$

For a given population size, we can identify the number of persons in every UC with the family size f. We define matrix M representing the joint percentage distribution of the overall population with values represented in A, G and E matrices. The total number of elements in M are equal to product of all possible values in $|A| \times |E| \times |G|$ which in our case is 40. For the IPUMS sample data, the conditional probability distribution of each group represented as element of M is calculated as follows.

Note, each member of matrix M, m_{ijk} is a triplet where

$i \in [1,2,3,4,5]$

$j \in [1,2,3,4]$

$k \in [1,2]$

e.g. *[Age 1, Education 1, Male]* represents one such tuple.

n_{ijk} represents the total population with attributes i, j, k.

n_{fs} represents the total population in socioeconomic group s with family size f.

$Prob(\ tuple\ takes\ value\ i,\ j,\ k \mid F{=}f,\ S{=}s) = n_{ijk} \ / \sum f \ \sum s \ n_{fs}$

This joint probability distribution indicates the number of persons for each of the socio-economic groups. Population of each UC is computed using the joint distributions of demographic attributes for the socio-economic groups. The joint probability distribution gives us a number of persons in every UC that falls under these groups

and is displayed. Figure 3.2 provides partial information of socio-economic groups and there computed probabilities for the case study of Lahore.

The computed number of individuals in every group may not be a whole number. Note, the rounding of these numbers can give error by adding extra persons or having less persons. To fix this issue, extra persons are deleted or added to the population. Every person in the group is then assigned a unique identifier and represents and agent. In addition, each person is assigned his/her own demographic information. One person is written in each row of the output data file, known as population file. Each person is assigned its own demographic attributes and a household identification number. Every household is assigned a unique identifier. All persons living in same household have the same household ID.

3.6 Conclusion

In this chapter we describe a process to synthesize baseline agent population for any metropolitan city. We demonstrated the process by synthesizing baseline population of city of Lahore using IPUMS and census percentage data. The joint distributions of demographic attributes of agents are calculated for every socio-economic area of the city. This synthesizes a heterogeneous population of the city in terms of demographic attributes.

4. ACTIVITY GENERATION AND LOCATION ASSIGNMENT FOR SPATIO-TEMPORAL AGENT MODEL

4.1 Introduction

Based on the baseline population data, we propose a methodology for synthesizing an agent model in this chapter. The process entails assigning temporal activities to agents as well as the locations where these activities are performed. We assume 4 activity types including work, education, household, and transportation. The process of assigning activities to the agents is implemented by last three components shown in Figure 3.1. In this chapter we discuss three components of the system which consult the demographic domain knowledge as discussed in Section 4.2. These modules are labeled as Activity Generation, Location Assignment, and Route Assignment. Note, household, work, and education activities require assignment of geo-location on the map of city while the route does not have such requirements. In the following section we describe the dataset used to synthesize the agent model synthesis require.

4.2 Input Datases

The overall process of activity assignment requires following datasets:

- **Baseline Population**: This is the population file synthesized as discussed in Chapter 3.
- **GIS Dataset**: This dataset includes the shapefile of city of Lahore which provides the geo-coordinates of the city. This file is used to synthesize partial domain knowledge datasets, discussed below.

- **Demographic Domain Knowledge Dataset**: This dataset provides the various statistics, assumptions, and rules about the activities of population based on a person's knowledge. The demographic domain knowledge is explained in following section.

4.2.1 Demographic Domain Knowledge Dataset

The domain knowledge datasets can be categorized as:

- *Percentage Distributions of Work Types*
- *Activity (Work, Education, Transportation, Household) Centered Geographic Location Data*
- *Rules of Assignment of Work, Education, and Route Activity*

4.2.2 Percentage Distributions of Work Types

Census data provides aggregated summary tables containing the information about the percentage of employed and the unemployed persons in a city, the number of working persons per household and the average number of employed persons per household. The census data containing ten categories of work as well as the percentage of population in each category for both male and female population. Table 4.1 shows the work types available in the census data.

Percentage Distributions of Employment Status

Census data provides with percentage distributions of the employed, unemployed persons and the percentages of people working in each work type. The challenge here is that this percentage is for the total population of the city. The percentages of the employed and unemployed and work types can be different in the three areas of the city we have identified in the previous section.

Using the knowledge about the city of Lahore, we use an estimate about the employment percentages and work percentages for each group of population in the city.

Table 4.1: Work Types from Census Data

1	Legislators, senior Officials, and Managers
2	Professionals
3	Technicians and Associate Professionals
4	Clerks
5	Service, shop, market sales workers
6	Skilled agriculture and fishery workers
7	Crafts and related trade workers
8	Machine operators and Assemblers
9	Elementary occupations
10	Armed forces

Table 4.2: Employment and Unemployment Percentages

SocioEconocmic Status	Urban	Urban	Rural	Rural	Inner City	Inner City
Gender	Male	Female	Male	Female	Male	Female
Employed	80	40	60	10	80	25
UnEmployed	20	60	40	90	20	75

Table 4.3: Work Type Percentage Distributions for Both Genders

	Urban	Urban	Rural	Rural	Inner City	Inner City
W1	.23	.13	.12	.1	.12	.16
W2	.12	.29	.10	.12	.32	.29
...	...	...	...	...	...	...
W10	.09	.01	.1	0	.2	.01
Total	1	1	1	1	1	1

Table 4.2 shows the employment percentage distribution of the city assumed for the model.

Percentage Distribution of Work Type Activity

Again, *using the knowledge about the city of Lahore*we assume that the high income households are more likely to be populated in urban residential areas, whereas the low income households reside mostly in rural areas. The reason for this division is the difference in the levels of education attained by the individuals in each residential area. Table 4.3 shows percentages of various types of work assumed for persons of every residential areas. Accordingly, Table 4.3 shows the work distribution assumed in synthesizing the agent model.

Percentage Distributions of Workers per Household

Census data also provide information about average number of working individuals in every household, along with the percentages of households with the average

Table 4.4: Census Data

Number of Employed Persons	Percentages
Male	80.91
Female	19.09
Percentage of Houses with Persons Working	Percentage
One Person	46.92
Two Persons	29.02
Three Persons	13.07
Four Persons	6.54
Five and More Persons	4.47

number of persons working per household. These percentages are assumed for all the three residential areas. Table 4.4 shows the percentages of individuals working per household and the average number of working individuals per household.

4.2.3 Activity Centered GIS Location Data

Geo-referenced GIS Layers or shapefiles are the third required input dataset. The shapefile for the city of Lahore contains geographic coordinates. The proposed methodology requires the location coordinates for assigning locations to activities performed by the agents. From the shapefile of Lahore, activity locations pertaining to residential areas, work areas, educational institutes, and transportation routes are identified.

In essence, the synthetic population is distributed over various areas of the city. The fundamental unit of region in this file is a Union Council (UC). The associated database file, with this shapefile, provides information about the population per union council and the size of its area. This data is used in grouping union councils to generate socio-economic groups and baseline population synthesis. Lahore is capital

of province Punjab and is second largest city of country. There are 9 towns in the city and each town is further divided in union councils. There are 151 union councils in the city. The city is culturally rich and one of easily accessible cities of Pakistan where public transportation service is available around the clock. The shapefile for Lahore provides us with boundaries of union councils in every town of the city. For our case study, the union council is considered as the unit of geography for Lahore. The associated database file of the shapefile contains other important information such as population in every union council, and the size of the area of every union council. Figure 4.1 shows the number of union councils on the map of city. The color shows the classification of union councils based on their population density. The pink color represents the UCs with low population density, yellow shows the UCs with high population density and white shows the population with the average population density.

Using this shapefile, we identify residential areas, work locations, educational institutes, and transportation routes as discussed below.

Residential Area Geo-Location GIS Layer

Residential areas are the designated places where people reside in houses, hostels, and apartments. In synthesizing our model we assume that the residential areas do not overlap with any other activity location such as work, educational institutes etc. Other non residential areas such as rivers, and parks are also excluded from the residential areas. To ensure a uniform distribution of residential places in the city of Lahore, we divide the area in small grids. Each grid represents a block and population is assumed to be uniformly distributed in every grid. Figure 4.2 shows the grid map for residential area of the city of Lahore.

Identification of Work Locations

Coordinates for work locations are extracted from the shapefile of Lahore. Work types are known from the census data as shown in Table 4.1. All the different activities performed at one place are assigned the same activity location type. Activity locations are identified on map of Lahore *using the existing knowledge* of city to ex-

Fig. 4.1.: Shapefile of Lahore

tract coordinates of locations. Accordingly, six different types of work locations are identified. Unlike residential areas, each coordinate represents a work location and multiple individuals are assigned to one location. For the city of Lahore, there are more than 50,000 work locations are identified.

Identification of Educational Institutes

Likewise we select educational institutes (schools, colleges, universities) from the shapefile of Lahore. We assume that there are five colleges and five schools in every union council. There are ten universities in whole city.

Identification of Transportation

Individuals require a mode of transportation to perform their daily activities. The proposed model assumes that there is only one mode of transportation available which

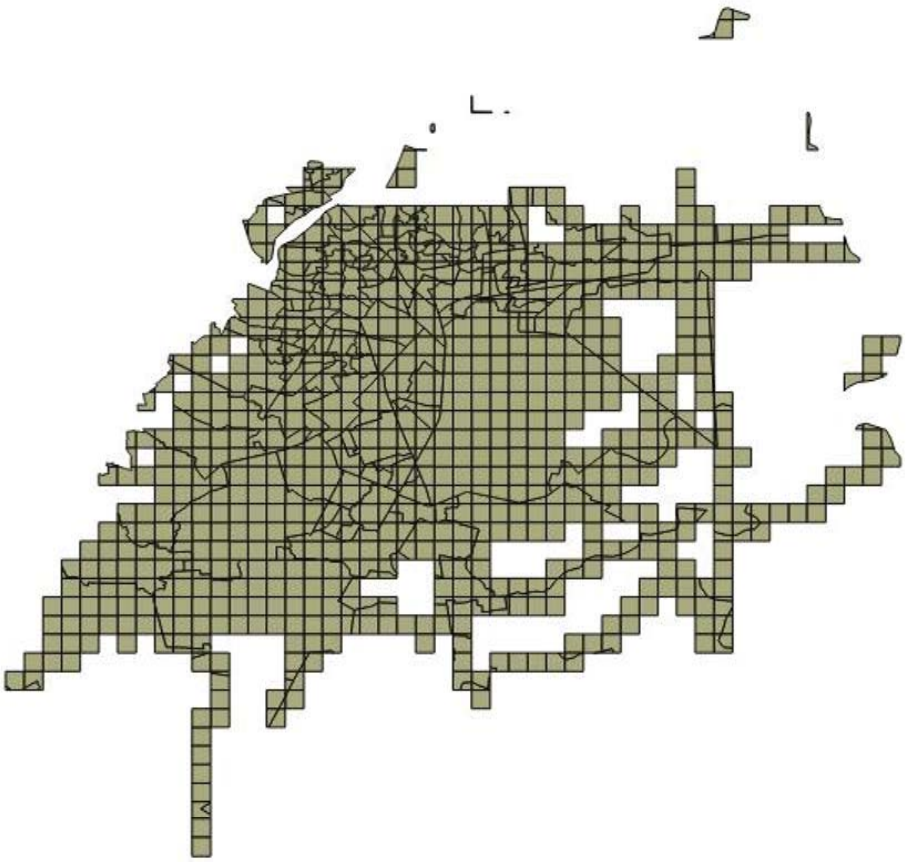

Fig. 4.2.: Residential Area of Lahore

is public transportation. By reviewing the public transportation of city of Lahore, we identify that there are 9 main bus routes in the city. These routes cover the whole city and are widely used by people to commute within the city. One of the challenges is that road network shapefile for bus routes are not available. To deal with this challenge, we identify the routes of buses on the shapefile according to general information available about route. Each bus route has multiple stop points which are used to in the route assignment component of the proposed system.

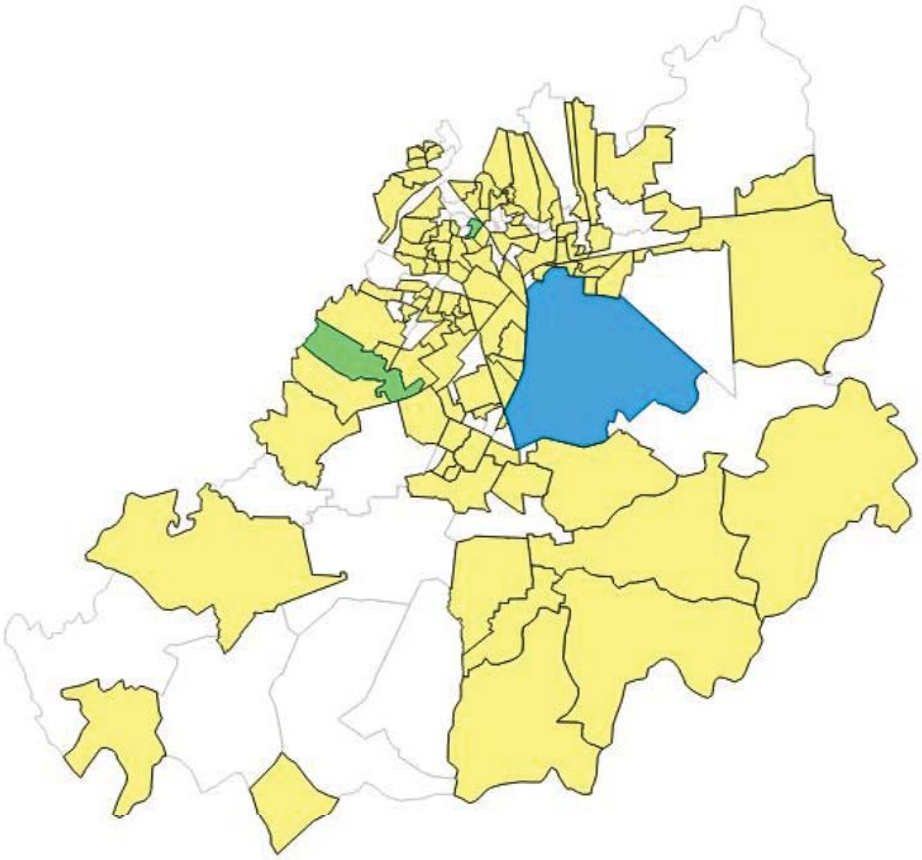

Fig. 4.3.: Blue: high work density Areas, Yellow: average work density areas, Green: low work density areas

4.2.4 Activity Assignment Rules Based on Demographic Attributes

In this section we describe three types of knowledge based rules for assigning work, education, and route assignment to agents. The work and education rules are used for assignment of work and education. This assignment is performed by Activity Generation module of Figure 3.1. Similarly, the route assignment rules are used by Route Assignment module to allocate routes to agents.

Work and Education Rules

The existing knowledge about the population of Lahore indicates that the type of work performed by an individual depends mainly on age, gender, and education.

A person must fulfill certain requirements of age and education before he/she is considered eligible for the type of work. We assume experience is required for certain work types. The requirement of experience establishes a relation between the work type and age of the worker. For instance, a managerial position usually require individual to have a degree in management sciences and expected experience is of several years. It can be inferred that a person employed for such a position has an age which is above 35.

Using the domain knowledge of the city we can stipulate the percentage distribution of work for different socio-economic areas. People living in higher income areas are more likely to be more educated than the ones living in low income areas. Also, the people in high income areas are less likely to be unemployed as compared to other areas of the city. This knowledge helps us making the assumption that people living in low income areas are more likely to be unemployed, less educated, and working in elementary occupations. Also, the percentage of women working in elementary (low income) occupations is higher than any other area of the city. Using this knowledge we make assumptions about the percentage distribution of various work types across the population of city along with requirements a person must meet in order to be assigned a specific work type. Work activity types can be represented as a vector of ten elements.

W = [W1, W2, W3 ... W10]

Similarly, education activity type can be represented as the following vector.

E = [E0, E1, E2, E3]

By using the domain knowledge of the city we know that literacy rate in Pakistan is very low. There are many people working as sales persons or doing other mediocre jobs which do not require any formal education. Under these assumptions, individuals with *E0* (no education) are considered eligible to perform certain types of work.

Rules of assigning education levels are shown in Table 4.6.

Rules for Route Assignment

Table 4.5: Rules for Work Assignment Based on Domain Knowledge of the City

Work Activity Type	**Education**	**Age Group**
W1	E3	3, 4
W2	E3	2,3,4
W3	E2	2,3,4
W4	E1, E2	2,3,4
W5	E0, E1,E2	2,3,4
W6	E0, E1	2,3,4
W7	E0, E1	2,3,4
W8	E1	2,3,4
W9	E0	2,3,4
W10	E1, E2	2,3,4

Table 4.6: Education Type Activity Assignment Rules

Education Type Activity	Working	Age	Education in Baseline Population
E0	NA	Any	E0
E1	No	1	E1
E2	No	1,2,3	E2
E3	No	2,3	E3

Table 4.7: Route Assignment Rules

Activity Performed	Closest to Work	Closest to House	Closest to Educational Institute
Work	Yes	No	No
Education	No	No	Yes
Stay Home	No	Yes	No

Route assignment module assumes that the only mode of transportation available to the general population is the public transport (bus). There are 9 different bus routes in the city. We estimate the map of bus routes using the shapefile of the city. Transportation is an activity performed by every individual. During weekdays only working population of the city is assumed to use transportation for commuting back and forth between residence and the work place. We further assume each individual use a single bus route for commuting that is closest to the location of the household. For some locations, multiple closest bus routes may exist. In this case, we randomly select a bus route for the person.

4.3 Synthesis of Agent Model

Figure 4.4 illustrates process flow for activity generation and assignment of activity locations and transportation routes. The first component is called Activity Generation Module which requires the baseline population (discussed in Chapter 3), Employment Percentage data, Work Distribution Percentage data, Work Type Assignment Rule data, and Education Level Assignment Rule data. It groups the baseline population into working and non working population prior to assigning the work types to agents. This module then uses the information about the percentage of work activity type in the percentage distribution file, randomly selects a person from

the working population, checks the rules of assignment of the work type, analyzes if the requirements are met to decide weather or not an activity should be assigned to the selected person. After completing the work assignment process, it combines the working and non working agents to assign education levels. The education level assignment module randomly selects a person who is not working, checks the education level assignment rules to assign education type.

After work and education activity types are assigned to persons, location choice module selects a location according to the activity type assigned to a person. If the activity type location is empty, the location is assigned to the person. For household, one household location is assigned to only one household.

Routes are assigned to the population depending upon their work, education, and household activity location. Working and person attending an educational institute is assigned to that bus route that is closer to his/her work location.

In the following sections we explain the agent synthesis process in detail.

4.3.1 Activity Generation

Our assumption for the activity generation module is that every individual lives in a house and there are no homeless people in the city. Table 4.2 shows the employment percentages of working and non working population in every socio-economic group. The employed persons are those who are eligible to work. On the contrary, a person is not eligible to work who has retired from his work and is above 60 years old or is less than 15 years of age. The employed persons are selected following the percentages from Table 4.2 such that no individual under age 15 years is eligible to work and no individual above age 60 years is eligible to work. All persons with the age that fall in age range 16 years to 60 years are eligible to be grouped as employed or unemployed. These employment percentages for both genders in each socio-economic group provides us with information of how many people are working in every UC. Assignment of work type to each working individual depends on their age, gender,

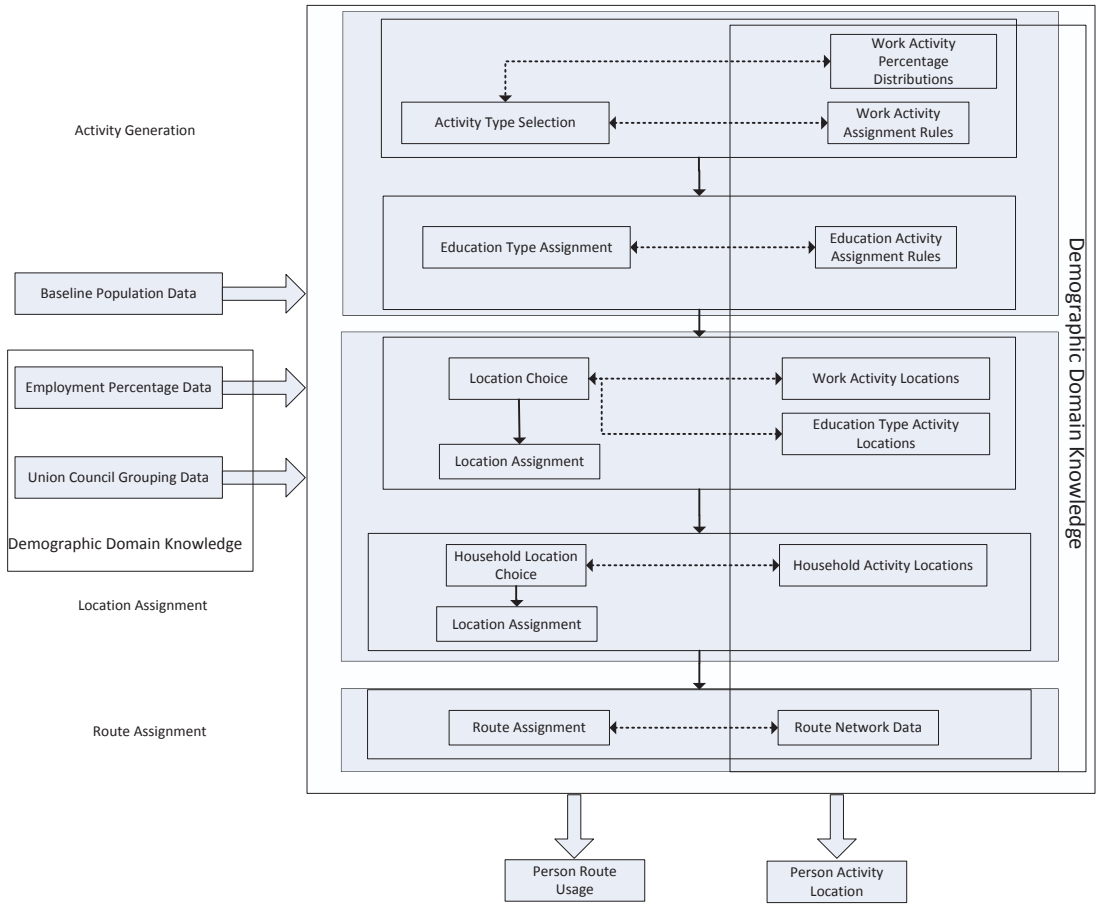

Fig. 4.4.: Conceptual Diagram of Activity Assignment, Location Assignment, Route Assignment Modules

and education. We calculate the number of persons working in every work type by using the percentages from Table 4.3 for each socio-economic group. Every person is equally likely to be selected as employed and unemployed by this module. Also, all the individuals satisfying the requirement for every work type are selected randomly with uniform distribution.

Figure 4.5 shows the data flow of work activity assignment module. It takes employment percentage data, baseline population from previous module, and union council grouping data as input. Population is divided in two groups: one that is

eligible to work and other that is not eligible to work. The work eligible population is then further identified as employed and unemployed persons. The employed persons are then assigned work types considering the above mentioned conditions. The final population is synthesized after merging all the population groups together. This file is used as input in education assignment module.

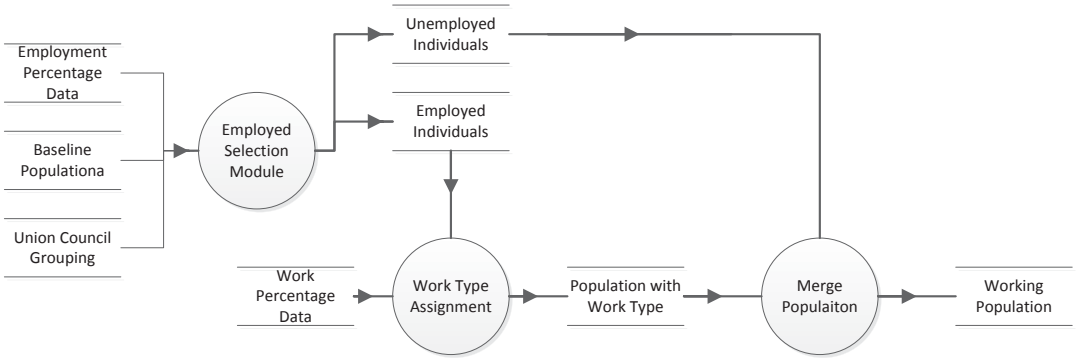

Fig. 4.5.: Work Assignment Module

The other type of non household activity is education. Using the baseline population synthesized in previous section provides us with the information of highest education level attained by every individual. Table 3.2 displays the grouping of education based on age of individuals in years. There are three levels of education, basic, medium, and higher. The 'Activity Generation' component following assumptions:

Table 4.6 shows the following rules of education assignment.

1: This module assumes if the person is assigned an education level higher than E0 and age of the person match with the one displayed in Table 4.6, if not working, are assigned a respective educational institute. Individuals who are assigned E1 (basic) or E2 (middle) or E3 (higher) if declared students will go to schools, colleges, and universities respectively. Baseline synthesized population has education levels assigned to every individuals.

2: The second assumption is that there are no students in the city with age 30 years or older. The age range between 16 years and 30 years (Age group 2) is overlapping age group between employed and students.

3: The third assumption is that every person in the city is either working or is a student. Every individual otherwise between age 0 years to 15 years is a student if assigned some education level.

Figure 4.6 shows flow of the education assignment module in detail. Education assignment module process the population that has not assigned any work type by activity generation module. The same module also assigns education type activity to individuals following rules mentioned in Table 4.6.

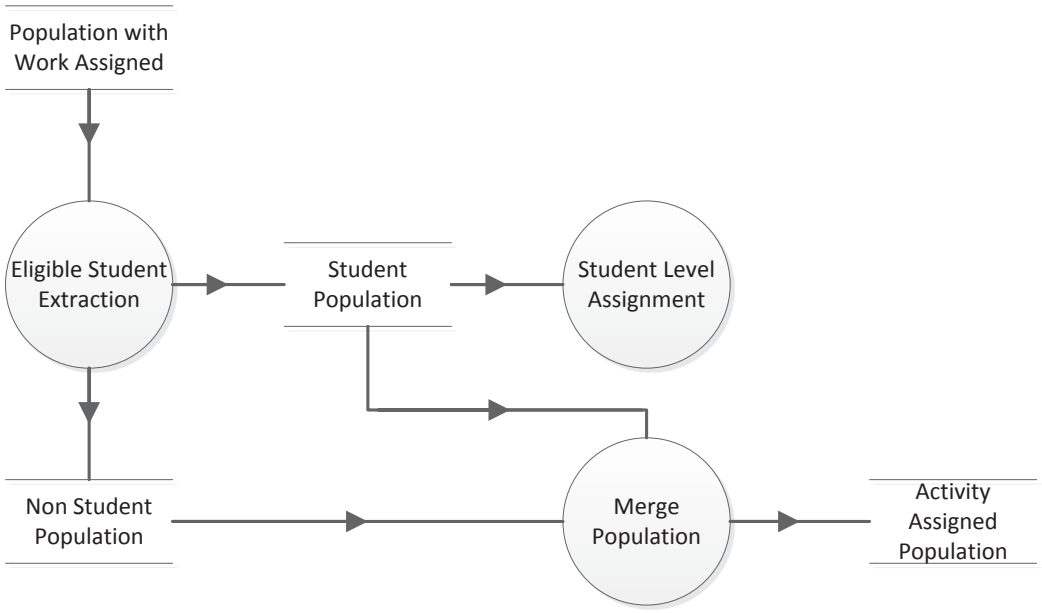

Fig. 4.6.: Education Assignment Data Flow Diagram

4.3.2 Location Assignment

Location assignment module performs two tasks. First task is to assign work locations and educational institutes to the individuals. Second task is the assignment of household locations. The rules of these two assignment tasks are different and are explained below. The proposed module assumes that more than one type of work may be performed at one location. In case of one type of work is performed at a location, the maximum number of persons at the location is assumed to be 100. If more than one type of work is being performed at the location, the maximum capacity of location is multiple of number of types of work performed at a location times 100. A location is assumed to have no less than 25 persons. Figure 4.7 shows the overall process of activity assignment.

Figure 4.8 shows the process of location assignment in terms of inputs and outputs.

In this process, we first spatially spread households over whole city. The data flow diagram for the process of coordinate assignment to different households in every union council is shown in Figure 4.9. The process takes two input files: the baseline population and a point-coordinate file which contains coordinates for every union council. This file is extracted from the baseline geo-referenced GIS layer of city of Lahore. The points are randomly selected from every union council. Every point is assumed to represent only one household.

Household assignment is similar to the work activity location assignment. The only difference is that each household is assigned just one point of x-y coordinate. Depending on the size of population, approximate number of coordinate points can be extracted from GIS file of city of Lahore if needed.

4.3.3 Route Assignment

Routes are assigned to agents following the rules explained in Section 4.2.4. During weekdays, the transportation activity is work dependent. Associated with the activity is the time to perform it a persons use the transportation. We use Euclidean

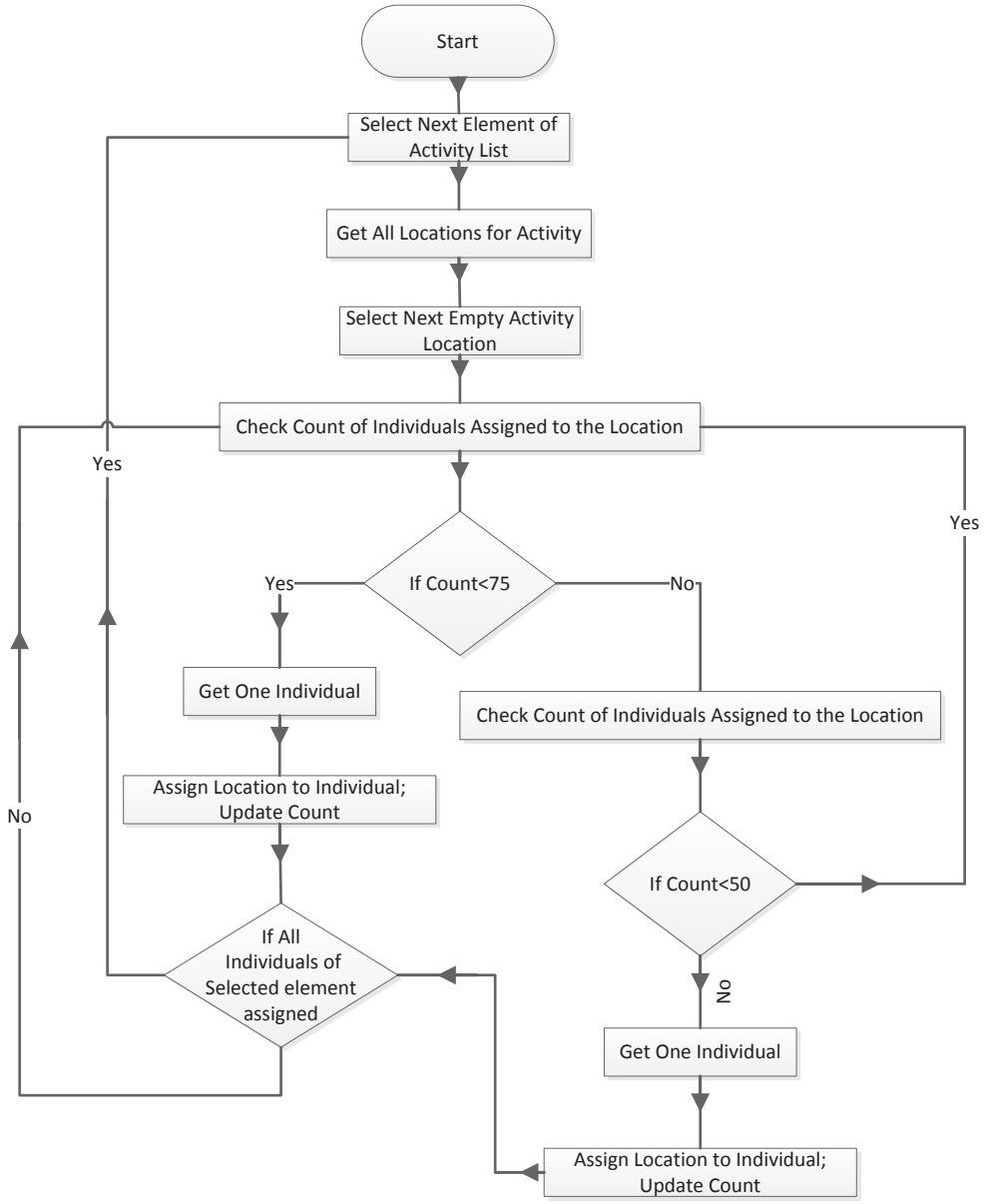

Fig. 4.7.: Activity Location Assignment FlowChart

distance between the individual's house location and coordinates of the route of bus which is closest to the house. Multiple buses are assigned on each route. Individu-

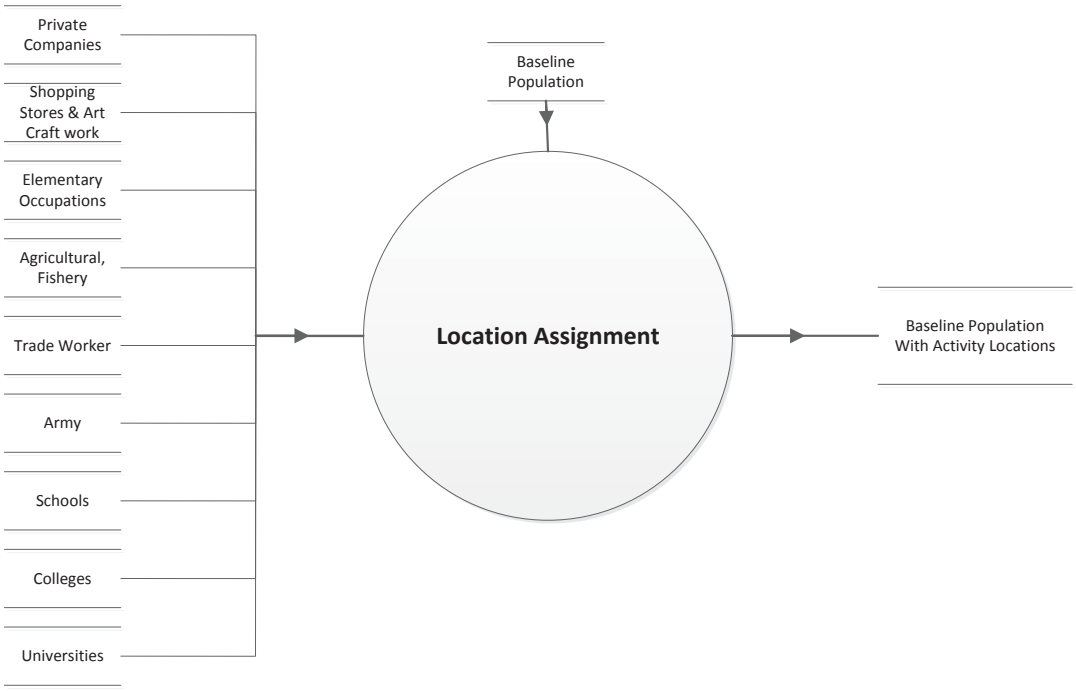

Fig. 4.8.: Location Assignment Data Flow Diagram

als are assigned a bus number which is randomly selected. This is a basic model of transportation used to understand the impact of transportation on disease spread.

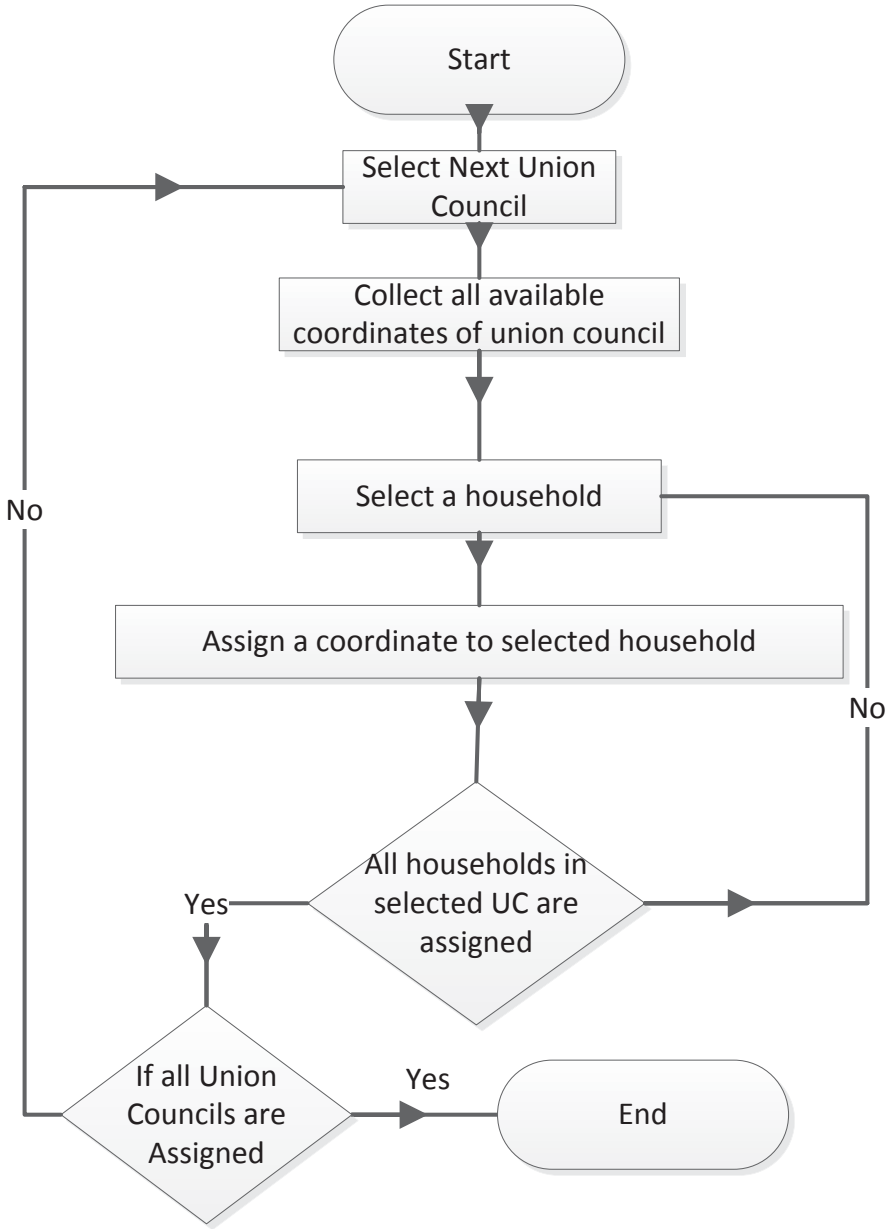

Fig. 4.9.: Data Flow Diagram for Household Coordinate Assignment

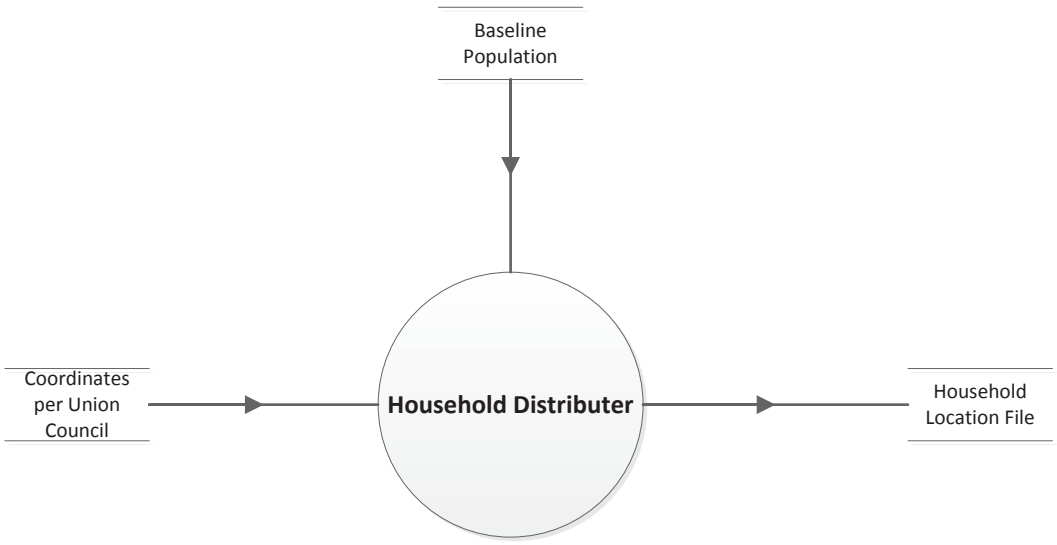

Fig. 4.10.: Data Flow Diagram for Household Coordinate Assignment

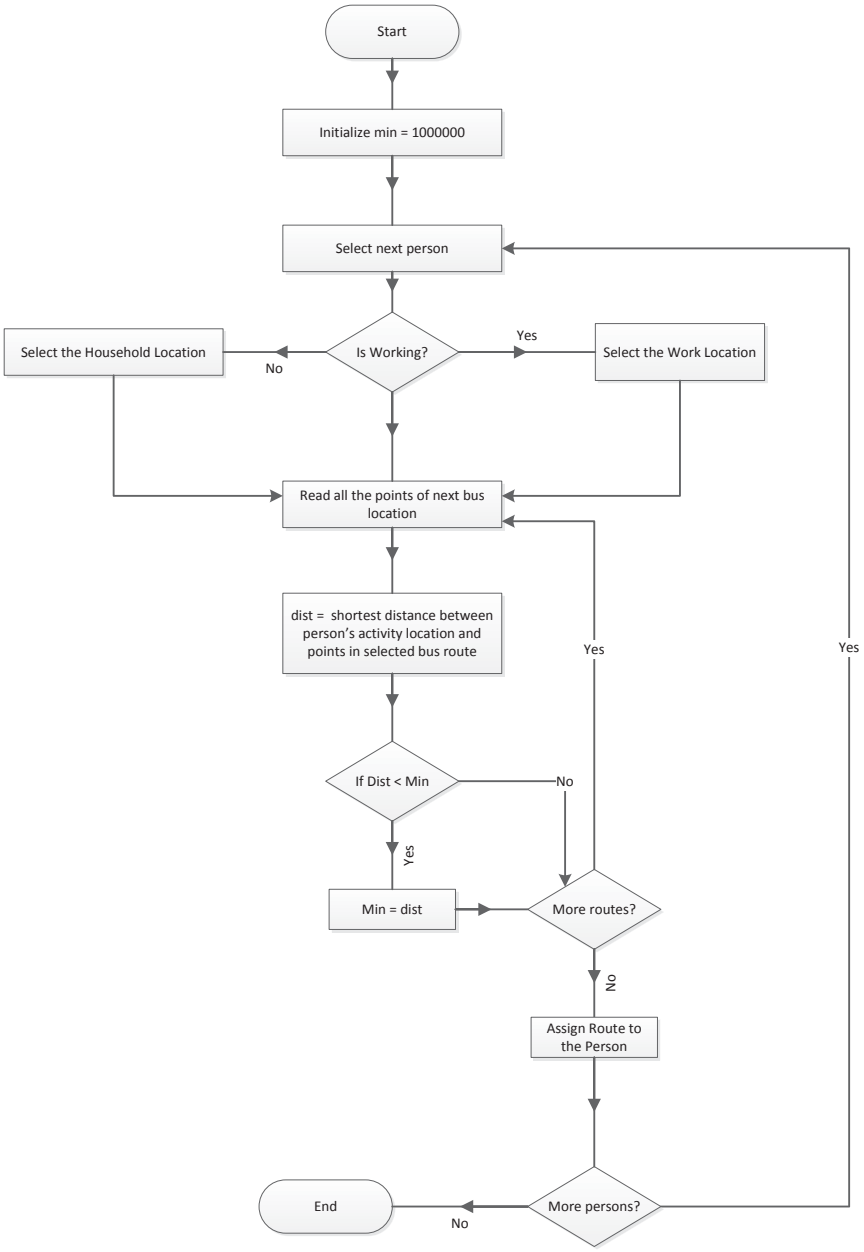

Fig. 4.11.: Data Flow Diagram for Route Assignment

4.4 Agent Model and Implementation

Based on the activities and routes assigned to the baseline population as explained above to the agent population synthesized in Chapter 3, agent activity model is synthesized. In this model, an agent can be represented as an entity with its demographic attributes such as age, gender, etc. Similarly, work, education, and route activities can be represented as entities. Figure 4.12 shows the Entity Relationship Diagram of the agent model.

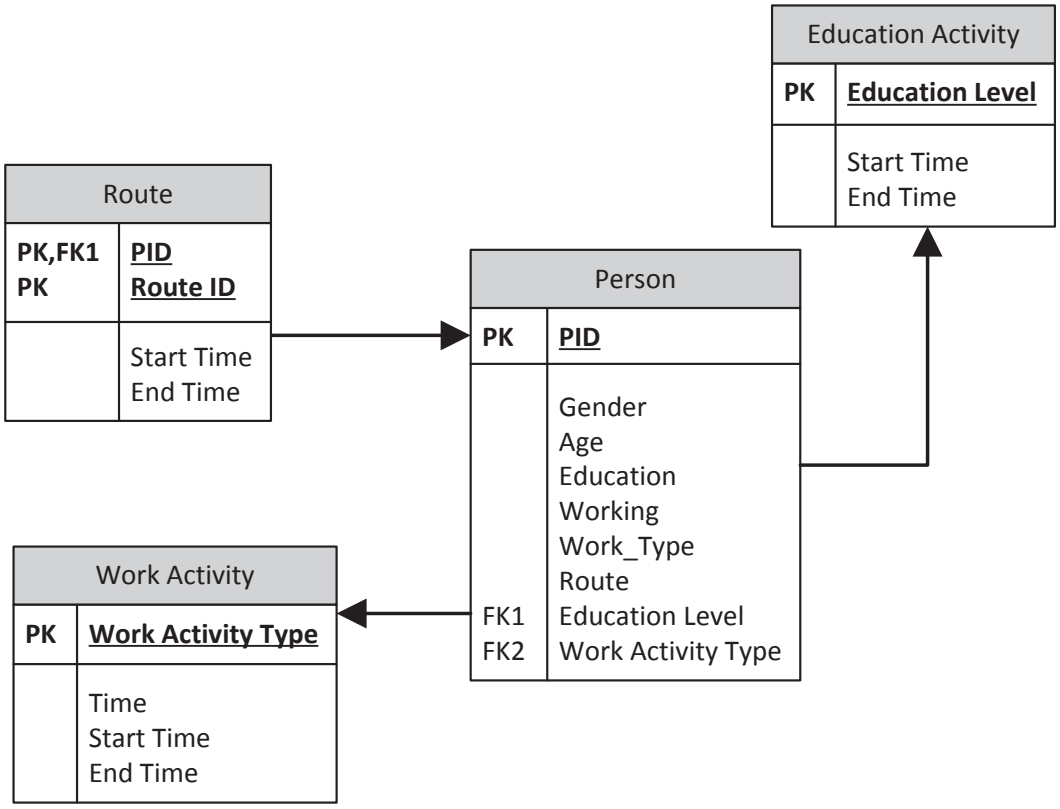

Fig. 4.12.: Entity Relationship Diagram of Agent Graph

The model can be represented as an abstract temporal graph. In this abstraction, agents constitute vertices and activities are represented as edges with temporal attributes between agents. Edges can be static or dynamic depending upon the type

of activity. The static edges correspond to activity associated with people living in the same household and working or studying at the same location. Agents riding a bus have dynamic connection since they interact with certain number of randomly selected agents within the same bus. The graph represents the population of the city which plays its role in disease transmission. Figure 4.13 shows the abstract representation of an agent graph. The interaction among people of city is used for simulation of epidemics described in Chapter 5.

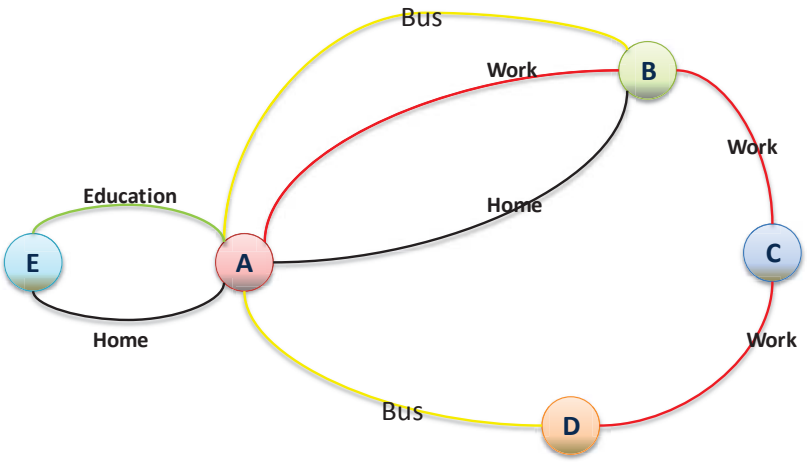

Fig. 4.13.: An Agent Graph Model

4.5 Conclusion

Using demographic attributes of population of a city, a process to synthesize a spatio-temporal activity based agent model is presented in this chapter. The process requires demographic domain knowledge of the city about work distributions, education and employment distributions. Agents are assigned activity types based on the demographic domain knowledge of the city. The synthesized agent graph model closely represents the interactions of actual city population. Agents are then assigned locations of activities following the rules of activity assignment. Public transportation is the only mode of transportation is assumed in this model. Agents are assigned a bus route to commute within the city to perform other activities.

5. GEOGRAPHICAL CONTEXT DRIVEN VISUALIZATION OF EPIDEMIC

5.1 Introduction

The spatio-temporal agent model synthesized in chapter 4 is subsequently used to simulate epidemic visualization. We elaborate how various activities and their locations play a critical role in disease spread. We can also compute the density of infectious agents both at the activity location of transfer of infection and there residential areas. This visualization assists us in preplanning exercises to predict a possible epidemic and to understand the effect of various preventive measures that might be effective in containing the epidemic in small area. The simulation allows us to select initially infected persons on the map of a city to analyze difference in disease dynamics based on demographic attributes.

The medium for infection to transfer is the activities performed by the agents. As mentioned in previous chapters, these activities are represented as edges of the network and agents represent vertices. The main contribution in this visualization of disease spread is mapping of locations of agents on city map as discussed in previous chapters. Such mapping can be used to visualize simulation of spread to facilitate decision making and visualization analytic. The spatio-temporal nature of model allows visualization of the spread of infectious disease assuming some initial triggering points of disease. We identify activity locations as markers on the map of city Lahore which help us in identifying the main cause of infection transmission in terms of activity type locations. Further enhancement of visualization by coloring union councils where color intensity is based on number of infectious people in the area.

5.2 Epidemic Domain Knowledge

We run a rule based disease spread simulation on the agent network synthesized in previous chapters. The simulation uses two sets of rules pertaining to epidemic spread. The rules are listed below.

Rule Set 1

Throughout this chapter, we refer to rule set 1 as Case 1.

Effective Contact

We assume four or more persons which are already infected can infect a susceptible person when they meet simultaneously with the susceptible persons.

Contact Duration

The duration of contact between susceptible and infected persons lasts more than or equal to 30 minutes.

Recovery Rate

Every infected person recovers after exactly 5 days.

Rule Set 2

Rule set 2 is referred to as Case 2. This rule is similar to Case 1 except the effective contact is reduced to 2 persons. The complete set of rules is explained as follows.

Effective Contact

We assume two or more persons which are already infected can infect a susceptible person when they meet simultaneously with the susceptible persons.

Contact Duration

The duration of contact between susceptible and infected persons lasts more than or equal to 30 minutes.

Recovery Rate

Every infected person recovers after exactly 5 days.

5.3 Visualization of Disease Spread - Day To Day

In this section, we discuss the results of epidemic spread simulation. We run the simulation for Case 1 and Case 2 with same initial infected individuals. The simulation follows Susceptible, Exposed, Infected, Recover (SEIR) model of disease spread [5]. This model assumes that every agent is susceptible when simulation starts. Some of the susceptible agents, when exposed to the infection (initial infected) triggers an epidemic. Exposed individual stays in this state for a day and it cannot transmit infection and do not display any clinical symptoms. The symptoms of infection appear after one day when agents are in Infected state. They can now transmit infection to other susceptible agents. The duration of infected state depends on the type of pathogen, and age of individuals. In our simulation, we assume duration of infection to be 5 days. After infected period, individuals are in Recovered state. A recovered individual cannot get infected again. In this simulation, we assume dynamic contact between individuals in using transportation services at same time and route. Every agent will interact with different agents every time the transportation activity is performed. The selection of agents interacting with each other during transportation activity is random. However, static contact is assumed at work activity locations.

Day 1

On day 1, epidemic is triggered by introducing initial infected agents in population of the city. For this example visualization, we select 147 agents to be initially infected agents working in inner city at 12:00 PM. The agents are selected from various work locations to trigger an epidemic. We run two different simulations on the selected initial infected agents. Figure 5.1 displays the location of initial infected agents.

As explained in Section 5.3, no new infected agents are seen on day 2 of simulation.

Day 3 For case 1 simulation, on day 3 there are 425 cases of new infections observed in the city. All of the infected individuals got infected at work locations. For case 2 simulation, new infected individuals count 352 and points of transmission

Fig. 5.1.: Day 1 Initial Infected Residential Location

of infection are work locations. Infected agents for both cases are from all three types of areas of the city. Figure 5.2 shows the areas of infection incidence on the city map.

Fig. 5.2.: Day 3: Case 1 (Left), Case 2 (Right)

Day 4

On day 4, new infections observed in case 1 and case 2 simulations are 90 and 16 respectively. All the incident infections of both simulations occur in households. The

infected agents are from inner city and rural areas of city in both simulations of case 1 and case 2. Figure 5.3 shows the areas of infection incidence on the city map.

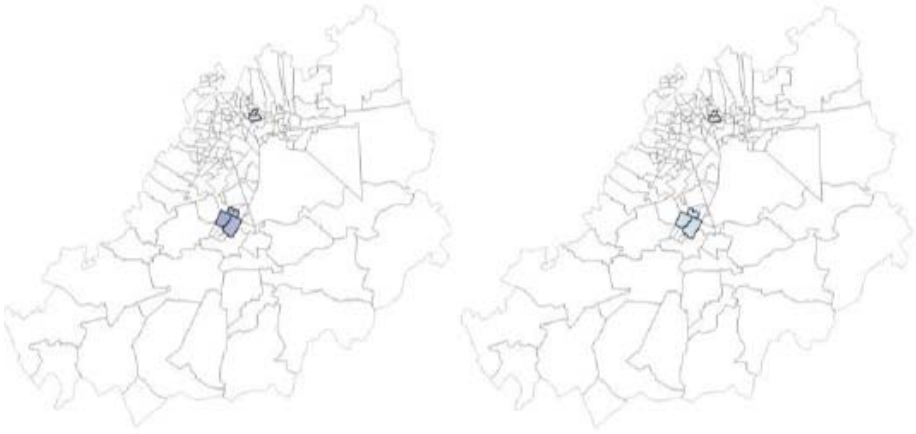

Fig. 5.3.: Day 4: Case 1 (Left), Case 2 (Right)

Day 5

The total new infection on Day 5 is 390 and 47 for case 1 and case 2 respectively. The transmission of infection in case 2 simulation is slower because threshold of infection transmission is higher than case 1. We observe that for case 1, incidence of infection occurs at both household locations and transportation location whereas infection incidence of case 2 occurs only at household locations. The infection transmission occur in inner city and rural areas of the city. The detail of disease dynamics on day five if provided in table 5.1. Figure 5.4 shows the areas of infection incidence on the city map.

Table 5.1: Table: Day 5 Disease Dynamics

Activity	Case 1 Infected	Case 2 Infected
Transportation	193	0
Household	196	47
Work	0	0

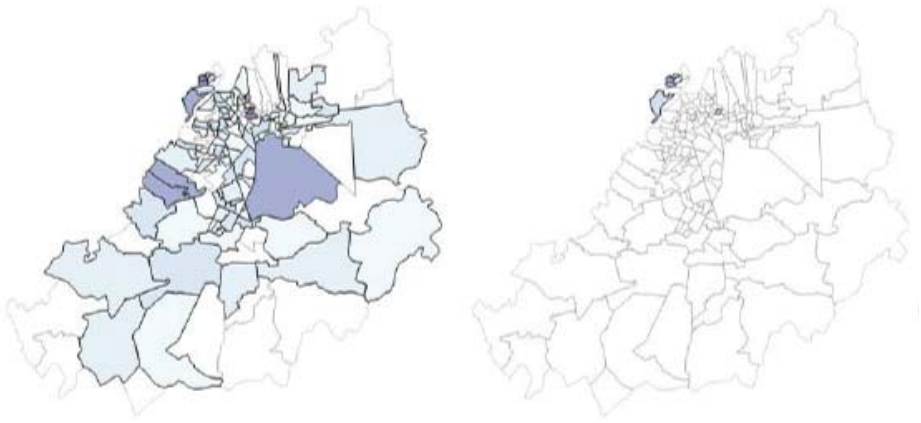

Fig. 5.4.: Day 5: Case 1 (Left), Case 2 (Right)

Day 6 Onward

The transmission of infection during transportation activity is responsible for spread of infection across the population of a city. We observe most of the infection transmission occur during transportation activity on day 5. On day 5 a few infections are observed in urban areas of the city. Most of the infections are observed in inner city and rural areas. On day 6 we observe 108 new cases of infection where the points of infection transmissions are work activity locations.Majority of new infected agents belong to rural areas of the city. On day 7, 1074 new infections are observed and the infection incidence occurred on household, work, and transportation activity locations. Cases of infection are observed in urban areas of the city as well as rural and inner city areas of city. There is only one more case is observed on day 8 of case 1 simulation. No new infection is observed after day 5 of case 2 simulation. Disease dynamics of incident infection of day 6, 7, and 8 are displayed in Table 5.2.

Table 5.2: Day 6, Day 7, Day 8 Disease Dynamics

Activity	Day	Case 1 Infected
Household	6	17
Transportation	6	56
Work	6	33
Household	7	110
Transportation	7	158
Work	7	806
Household	8	1
Transportation	8	0
Work	8	0

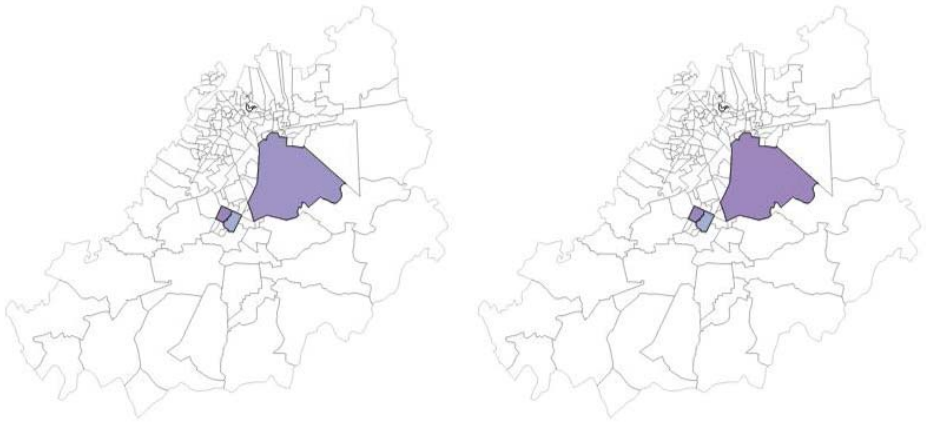

Fig. 5.5.: Day 6 (Left), Day 7 (Right)

6. CONCLUSIONS AND FUTURE WORK

In this thesis we have presented a technique to synthesize an activity based agent model. We have elaborated the technique by synthesizing an agent model for the city of Lahore as testbed in Chapter 3. The proposed approach is generic and can be used to synthesize a population for any metropolitan city. We assume that a city can be classified into three groups based on its demographics including socio-economic status of the persons. These socio-economic groups can further be distributed over smaller geographic areas. For the city of Lahore, each socio-economic group is divided into smaller units of geography called Union Council. There are various union councils in each socio-economic group. The process of synthesizing the activity agent model is a two step process. The first step is to synthesize a baseline population of agents using joint distributions of demographic attributes tabulated in the IPUMS sample data. These joint probability distributions are computed for each socio-economic group. The population of each union council is then estimated using the computed joint distributions.

The second step of activity agent model is to assign activity types to the agents as discussed in Chapter 4. We assume four activities for this model. These include household, educations, work, and transportation. We use three types of demographic domain knowledge datasets of city of Lahore. These include percentage work distributions and employment percentages, activity locations data, and rule of work, education, and transportation assignment. Using this demographic domain knowledge and baseline population, agents are assigned activity type and an activity location. The transportation is then assigned to the agents keeping their work locations in consideration. This gives us a temporal abstract agent network where agents represent the vertices and activities represent edges of the agent graph.

A rule based simulation of a disease spread is run on the synthesized activity based agent graph. The results of simulation are visualized in Chapter 5. The epidemic domain knowledge governs rules of epidemic spread. The initial infected agents are selected considering the socio-economic status. This visualization is a preplanning exercise to predict an upcoming epidemic and for decision making and other visual analytic.

In the work presented in this thesis, there are various directions to extend the research. Below are some of the possible research task that can be done.

- Application of various techniques to control disease spread in order to improve decision making and pre-planning exercises. These control measures can be pharmaceutical and non pharmaceutical such as vaccination, social awareness and quarantining an area. Also, the response of people from different socio-economic background may be different to an epidemic [32].

- Development of advanced activity agent model using fine grain assignment rules of various activity types such as work, education and transportation. In the research presented in this thesis, the agents are assigned activity type locations only. These agents can further be assigned activity sub-locations. The only means of transportation assumed in the Route Assignment module of this thesis is public transportation. The personal mode of transportation as well as advanced rules for public transportation may be applied.

- Incorporating various context parameters such as weather in the synthesis of agent model as well as other geographic features which can critically impact the spread of a disease.

LIST OF REFERENCES

LIST OF REFERENCES

[1] Y. Zhang, X. Lin, F. Zhang, J. Wu, W. Tan, S. Bi, J. Zhou, Y. Shu, and Y. Wang, "Hemagglutinin and neuraminidase matching patterns of two influenza a virus strains related to the 1918 and 2009 global pandemics," *Biochemical and biophysical research communications*, vol. 387, no. 2, pp. 405–408, 2009.

[2] C. Viboud, R. F. Grais, B. A. Lafont, M. A. Miller, and L. Simonsen, "Multinational impact of the 1968 hong kong influenza pandemic: evidence for a smoldering pandemic," *Journal of Infectious Diseases*, vol. 192, no. 2, pp. 233–248, 2005.

[3] E. C. Claas, A. D. Osterhaus, R. van Beek, J. C. De Jong, G. F. Rimmelzwaan, D. A. Senne, S. Krauss, K. F. Shortridge, and R. G. Webster, "Human influenza a h5n1 virus related to a highly pathogenic avian influenza virus," *The Lancet*, vol. 351, no. 9101, pp. 472–477, 1998.

[4] G. Neumann, T. Noda, and Y. Kawaoka, "Emergence and pandemic potential of swine-origin h1n1 influenza virus," *Nature*, vol. 459, no. 7249, pp. 931–939, 2009.

[5] A. J. Heppenstall, A. T. Crooks, and L. M. See, *Agent-based models of geographical systems*. Springer, 2012.

[6] F.-C. Tsui, J. U. Espino, V. M. Dato, P. H. Gesteland, J. Hutman, and M. M. Wagner, "Technical description of rods: a real-time public health surveillance system," *Journal of the American Medical Informatics Association*, vol. 10, no. 5, pp. 399–408, 2003.

[7] P. E. Plsek and T. Greenhalgh, "Complexity science: The challenge of complexity in health care," *BMJ: British Medical Journal*, vol. 323, no. 7313, p. 625, 2001.

[8] K. M. Carley, D. B. Fridsma, E. Casman, A. Yahja, N. Altman, L.-C. Chen, B. Kaminsky, and D. Nave, "Biowar: scalable agent-based model of bioattacks," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 36, no. 2, pp. 252–265, 2006.

[9] C. L. Barrett, K. R. Bisset, J. Leidig, A. Marathe, and M. V. Marathe, "Estimating the impact of public and private strategies for controlling an epidemic: A multi-agent approach.," in *IAAI*, 2009.

[10] N. Kenrad and C. Masters, "Infectious disease epidemiology: Theory and practice," 2006.

[11] E. Vynnycky and R. White, *An introduction to infectious disease modelling*. Oxford University Press, 2010.

[12] J. Satsuma, R. Willox, A. Ramani, B. Grammaticos, and A. Carstea, "Extending the sir epidemic model," *Physica A: Statistical Mechanics and its Applications*, vol. 336, no. 3, pp. 369–375, 2004.

[13] M. E. Newman, "Spread of epidemic disease on networks," *Physical review E*, vol. 66, no. 1, p. 016128, 2002.

[14] A. R. McLean, "Infectious disease modeling," in *Infectious Diseases*, pp. 99–115, Springer, 2013.

[15] Y. Yang, P. Atkinson, and D. Ettema, "Individual space–time activity-based modelling of infectious disease transmission within a city," *Journal of The Royal Society Interface*, vol. 5, no. 24, pp. 759–772, 2008.

[16] D. Ballas, G. Clarke, and I. Turton, "Exploring microsimulation methodologies for the estimation of household attributes," in *4th International Conference on GeoComputation, Mary Washington College, Virginia, USA*, 1999.

[17] S. Riley, "Large-scale spatial-transmission models of infectious disease," *Science*, vol. 316, no. 5829, pp. 1298–1301, 2007.

[18] R. J. Beckman, K. A. Baggerly, and M. D. McKay, "Creating synthetic baseline populations," *Transportation Research Part A: Policy and Practice*, vol. 30, no. 6, pp. 415–429, 1996.

[19] M. G. McNally, "An activity-based microsimulation model for travel demand forecasting," 1996.

[20] M. W. Macy and R. Willer, "From factors to actors: Computational sociology and agent-based modeling," *Annual review of sociology*, pp. 143–166, 2002.

[21] J. Y. Guo and C. R. Bhat, "Population synthesis for microsimulating travel behavior," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2014, no. 1, pp. 92–101, 2007.

[22] J. L. Bowman, "A comparison of population synthesizers used in microsimulation models of activity and travel demand," *Unpublished working paper. http://jbowman. net/papers/2004. Bowman. Comparison_of_PopSyns. pdf*, 2004.

[23] C. Bauch, A. dOnofrio, and P. Manfredi, "Behavioral epidemiology of infectious diseases: an overview," in *Modeling the Interplay Between Human Behavior and the Spread of Infectious Diseases*, pp. 1–19, Springer, 2013.

[24] N. Jonnalagadda, J. Freedman, W. A. Davidson, and J. D. Hunt, "Development of microsimulation activity-based model for san francisco: destination and mode choice models," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1777, no. 1, pp. 25–35, 2001.

[25] A. R. Pinjari, N. Eluru, R. B. Copperman, I. N. Sener, J. Y. Guo, S. Srinivasan, and C. R. Bhat, "Activity-based travel-demand analysis for metropolitan areas in texas: Cemdap models, framework, software architecture and application results," tech. rep., 2006.

[26] T. Arentze, H. Timmermans, and F. Hofman, "Creating synthetic household populations: problems and approach," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2014, no. 1, pp. 85–91, 2007.

[27] P. Norman, "Putting iterative proportional fitting on the researcher's desk," 1999.

[28] P. Williamson, M. Birkin, P. H. Rees, *et al.*, "The estimation of population microdata by using data from small area statistics and samples of anonymised records," *Environment and Planning A*, vol. 30, no. 5, pp. 785–816, 1998.

[29] K. Müller, K. W. Axhausen, K. W. Axhausen, and K. W. Axhausen, *Population synthesis for microsimulation: State of the art*. ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT), 2010.

[30] P. Williamson, M. Birkin, P. H. Rees, *et al.*, "The estimation of population microdata by using data from small area statistics and samples of anonymised records," *Environment and Planning A*, vol. 30, no. 5, pp. 785–816, 1998.

[31] A. B. Slavkovic and S. E. Fienberg, "Bounds for cell entries in two-way tables given conditional relative frequencies," in *Privacy in Statistical Databases*, pp. 30–43, Springer, 2004.

[32] M. Ventresca and D. Aleman, "Evaluation of strategies to mitigate contagion spread using social network characteristics," *Social Networks*, 2013.

[33] I. M. Longini, A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. Cummings, and M. E. Halloran, "Containing pandemic influenza at the source," *Science*, vol. 309, no. 5737, pp. 1083–1087, 2005.

[34] N. Wongchavalidkul and M. Piantanakulchai, "Estimating synthetic baseline population distribution when only partial marginal information is available," in *Proceedings of the Eastern Asia Society for Transportation Studies*, vol. 7, 2009.

[35] S. Ruggles, T. Alexander, K. Genadek, R. Goeken, M. Schroeder, and M. Sobek, "Integrated public use microdata series (ipums): Version 5.0 [machine-readable database]," *University of Minnesota, Minneapolis, available at http://usa. ipums. org/usa*, 2010.

VITA

VITA

Madiha Sahar was born in Lahore, Pakistan. She received her B.S. degree in Computer Engineering from University of Engineering and Technology, Lahore, Pakistan, in 2009. She received her M.S. degree in Electrical and Computer Engineering in 2014 from the School of Electrical and Computer Engineering at Purdue University.