COMPUTATIONAL MODELLING FOR SOCIAL AND BEHAVIOURAL DYNAMICS AND IMPACT ASSESSMENT

GROUP G13

JAYASOORIYA K.P.	E/18/153
THENNAKOON E.M.T.G.P	E/18/359
THIKSIGA R.	E/18/362

Supervisors

Prof. G.M.R.I. Godaliyadda (University of Peradeniya)

Prof. J.B. Ekanayake (University of Peradeniya)

Prof. M.P.B. Ekanayake (University of Peradeniya)

Prof. V.R. Herath (University of Peradeniya)

Department of Electrical and Electronic Engineering

Faculty of Engineering

University of Peradeniya

Sri Lanka

September 2024

ABSTRACT

Modelling is one of the main methods we can use to study and research our world. An agent-based Model (ABM) is a software model that contains agents who are capable of autonomously interacting with each other and with the environment. The flexibility of AMBs gives the ability to use them in many novel applications. They are ideal as virtual test beds, especially in social sciences where direct experimentation is unethical and infeasible. ABMs are widely used in modelling disease propagation, combat modelling, supply chains and in social sciences. Many of the existing ABMs rely heavily on mathematical equations to model human behavior. There are some that attempt to model human behavior with some amount of real-world data obtained from people through surveys, attendance reports and call recordings. But due to lack of raw data, this approach cannot predict human behavior accurately enough. In our study of ABMs, we propose an ABM that uses GPS data to model human behavior. This is a novel approach to study and model human motion dynamics. Using advanced machine learning algorithms to effectively study human behavior patterns and using that knowledge in building an ABM is the main goal of this project. This ABM will have the ability to predict human behavior with a much higher accuracy compared to current ABMs.

ACKNOWLEDGMENTS

We would like to take this opportunity to express our heartfelt gratitude to all those who have contributed to the successful completion of this project.

First and foremost, we are deeply indebted to our supervisors, Prof. Roshan Godaliyadda, Prof. Janaka Ekanayake, Prof. Parakrama Ekanayake and Prof. Vijitha Herath for their guidance, support, and valuable insights throughout this journey. Their expertise and encouragement have been instrumental in shaping this project and pushing us to strive for excellence.

We owe our deep gratitude to Dr. R.D.B. Ranaweera, the course coordinator for the encouragement, and cooperation extended and especially for structuring the course in an effective manner. Last but not least, we extend our appreciation to our families and friends, for their unwavering support, throughout the entire duration of this project.

In conclusion, this project would not have come to fruition without the collective effort and support of all the individuals mentioned above. We are truly grateful for their involvement, and We acknowledge that their contributions have played a crucial role in shaping the outcome of this work.

Thank you all.

Jayasooriya K.P. (E/18/153)
Thennakoon E.M.T.G.P. (E/18/359)
Thiksiga Ragulakaran (E/18/362)

Department of Electrical and Electronic
Engineering, Faculty of Engineering,
University of Peradeniya.
Sri Lanka.

CONTENTS

ABSTRACT i
ACKNOWLEDGMENTS ii
LIST OF TABLES vii
LIST OF ABBREVIATIONS viii
CHAPTER 01 1
INTRODUCTION 1
1.1 Agent-Based Model 1
1.1.1 The Agents 2
1.1.2 The Environment 3
1.2 Literature review 3
1.3 Applications 4
1.4 Scope of this project 4
CHAPTER 02 5
METHODOLOGY 5
2.1 Obtaining Data 5
2.2 Cleaning Data 6
2.3 Clustering Data 7
2.3.1 Utilizing K-means 7
2.3.2 Utilizing DBSCAN 8
2.4 Labeling Data 9
2.5 Domain Conversion 11
CHAPTER 03 13
Behavioral Modeling 13
3.1 Literature review 13
3.2 Mobility patterns 13
3.3 Hybrid model 14
3.4 Leveraging Hybrid Models in Our Study: A Strategic Approach 15
CHAPTER 04 17
Spectral Clustering 17
4.1 What is spectral Clustering? 17
4.2 Why spectral Clustering? 17
4.3 Data encoding 17
4.4 Literature review and algorithm for spectral clustering. 21
4.5 Spectral clustering results. 24
4.5.1 Use of different Laplacians. 24

4.5.2 Spectral clustering on different occupations. 27
4.6 Conclusion 31
CHAPTER 5 **32**
Introduction to the Agent-Based Modeling Simulator 32
5.1 Brief introduction 32
5.2 Components of the Simulation 32
5.2.1 Environment 32
5.2.2 Agents 34
CHAPTER 6 **36**
Rules that define human motion 36
6.1 Human Motion Model 36
6.1.1 Location Visit Probability Matrix 36
6.1.2 Stay Duration Matrix 36
6.2 Generating Random Variables 36
6.2.1 Theory 36
6.2.2 Implementation in Python 37
CHAPTER 7 **39**
Transportation 39
7.1 Introduction 39
7.2 Public transport 39
7.3 Three wheels 39
7.4 Distance calculation 40
CHAPTER 8 **41**
Disease Modeling 41
8.1 Literature Review 41
8.2 Markov Model 41
8.3 Contact Modeling 42
8.4 Disease propagation 43
CHAPTER 9 **44**
Results and Discussion 44
9.1 Disease Spread patterns 44
9.2 Effect of the Transport network on the Disease Spread patterns. 44
9.3 Discussion 45
CHAPTER 10 **46**
Conclusion 46
REFERENCES **47**

LIST OF FIGURES

Figure 2.1	Sample Data Format	5
Figure 2.2	K-means clustering method applied on GPS data	7
Figure 2.3	DBSCAN clustering method applied on GPS data	8
Figure 2.4	Representation of an Outlier Cluster	11
Figure 2.5	Numbered cluster with the catalog of the cluster	11
Figure 2.6	Trajectory of each person in the Time and Location domain	12
Figure 3.1	Mobility patterns	13
Figure 3.2	Hybrid model according to PDSIM authors	15
Figure 3.3	Expected format of Location occupancy probability matrix	15
Figure 3.4	Expected format of Location visit probability matrix	16
Figure 4.1	Gaussian Kernel function in three dimensions	22
Figure 4.2	Sigma sweep performed on nurse attendants for self-similarity (ε) = 0.1	23
Figure 4.3	For teachers (UL)	24
Figure 4.4	For Bank workers (UL)	24
Figure 4.5	For teachers (SNL)	25
Figure 4.6	For bank workers (SNL)	25
Figure 4.7	For teachers (RWL)	26
Figure 4.8	For bank workers (RWL)	26
Figure 4.9	Prominent mode is mode 34 at $\sigma = 7$	27
Figure 4.10	The geographical cluster formed by the outlier teacher at his workplace	27
Figure 4.11	The string list of the outlier teacher	28
Figure 4.12	Sigma sweep plot after removing the outlier teacher	28
Figure 5.1	Real-world example for geographical hierarchy	33

Figure 5.2	Map illustrating the Kandy District with annotated cities and functional zones.	33
Figure 5.3	Hierarchical tree structure of the environment.	34
Figure 6.1	Converting the PDF to CDF.	37
Figure 6.2	Getting the desired output from the probability matrix,given a random number.	38
Figure 8.1	The Markov Model used in the Simulation	42
Figure 9.1	Number of agents in each state over days.	44
Figure 9.2	Variation of exposed and Infected number of agents with and without transport network	45

LIST OF TABLES

Table 2.1	Data collection details	6
Table 2.2	Location Names	9
Table 4.1	Location names and Assigned number	18
Table 4.2	Location names with modified New Assigned Number	19
Table 4.3	Spectral Clustering Results	30

LIST OF ABBREVIATIONS

ABM	Agent-Based Model
BTS	Base Transceiver Stations
CSV	Comma-separated value
eps	Epsilon Value
GPS	Global Positioning System
minPTS	Minimum Points
PDSIM	Pandemic Simulator
RWL	Random Walk Laplacian
SNL	Symmetric Normalized Laplacian
UL	Un-normalized Laplacian
SIR	Susceptible, Infected, Recovered
SEIR	Susceptible, Exposed, Infected, Recovered

CHAPTER 01
INTRODUCTION

1.1 Agent-Based Model

Over the last forty years, Agent-based modeling (ABM) has gained significant popularity in social sciences[1]. This modeling technique has proved to be instrumental in building explanations of social processes based on the emergence of complex behavior from simple activities [2]. ABM allows us to study emergent orders resulting from local interactions among many independent components. This, in turn, helps us understand how such emergent orders can influence or constrain individual actions of those components. Self-organization is a crucial process that facilitates the emergence of orders in social systems. The said process is characterized by the concepts of "bottom-up" and "downward causation." This modeling technique plays a vital role in comprehending the operations of social systems in a more nuanced manner. The use of ABM to simulate social processes provides researchers with an opportunity to make predictions, test hypotheses, and gain insights into complex social phenomena. Therefore, it is imperative to utilize ABM in social research to gain a better understanding of social systems.

ABM is gaining interest in social sciences as it enables researchers to build computational models with direct representation of individual entities, cognition, and interactions. ABM is a type of computational model that simulates the actions and interactions of individual agents and their interactions with their environment. This helps in gaining a better understanding of how the system functions dynamically [3]. Thus, ABMS is particularly suitable for the analysis of complex adaptive systems and emergent phenomena in social sciences, traffic, biology, and others.

To create an agent-based model, the following three elements have to be explicitly dealt with. First, the set of agents is the most characteristic element. These agents are autonomous with respect to the other entities within the simulated environment. Next comes the specification of the interactions of the agents among themselves and with their shared environment. Since these interactions are responsible to produce the overall outcome, the design of all involved aspects is central. Interactions need not be explicitly represented within, for example, organizational structures. Rather, they may occur implicitly, as is the case with stigmergic interactions. However, in the implemented ABS, although organizational structures may not be obvious as such, it is important that they be explicitly considered. The third element, the simulated environment, contains all other elements. These may be resources, other objects without active behavior, as well as global properties.

When creating an agent-based model, three key elements must be addressed. Firstly, the set of agents, which are independent of the other entities in the simulated environment. Secondly, the interactions between the agents and the shared environment, which are crucial to producing the overall outcome. These interactions may be represented implicitly, as in the case of stigmergic interactions, or explicitly, as with organizational structures. In the implemented ABM, it is important to consider organizational structures explicitly. Finally, the simulated environment encompasses all other elements, including resources, non-active objects, and global properties.

1.1.1 **The Agents**

Agent-based modeling involves the representation of social actors, including individual persons, animals, organizations, and nation-states, as computational entities. These agents possess the ability to interact with one another, exchanging informational messages and acting upon the information they receive. Each agent within the model functions as an autonomous entity, thereby enabling the inclusion of heterogeneous agents in the artificial population. Such heterogeneity proves useful when building models of various phenomena, with agents possessing diverse capabilities, roles, perspectives, or knowledge stocks. In ABM, agents are characterized by four critical traits, allowing for their proper description and utilization within the model[4].

1.1.2 The Environment

ABM, or Agent-Based Modeling, simulates the behavior of individual agents within a given environment to observe how their interactions affect the overall system. The environment can be a simple, neutral space or a complex geographical space that affects the agents' decision-making and interactions. ABM offers a more detailed and realistic view of how systems function, identifying patterns and trends not apparent through other methods.

1.2 Literature review

It has been observed that compartmental models fail to capture the complexity of human behavior, especially mobility patterns and social networks. Meta population models attempt to overcome these limitations but still suffer from behavioral generalizations within the meta populations. In ABMs, the behavior of each unique agent is captured, making agent-based epidemiological simulations more powerful than meta population models[5].

In a study conducted by Barret et al.[6] , an agent-based simulator was developed using a synthetic population model built from the United States Census. The model was created by feeding 163 different variables into the agent, which were then mapped to geographical locations. Daily activities were modeled using data sets obtained from school attendance, transportation surveys, and other sources.

An agent-based system that uses call detail records to model virus spreading through social interactions and mobility patterns[7]. To capture realistic human mobility patterns and social dynamics, the ubiquitous infrastructure provided by a cell phone network is used. Cell phone networks are built using a set of cell towers, called Base Transceiver Stations (BTS), that connect the cell phones to the network. Each BTS has a specific geolocation represented by its latitude and longitude. The BTSs provide cellular coverage to specific areas, referred to as sectors. A sector of each BTS is assumed to be a two-dimensional, non-overlapping polygon. To define the coverage area of each sector, a Voronoi tessellation is used. This gives an indication of the geographical position of a user at a given moment in time. The Voronoi tessellation was used for the ABM.

1.3 Applications

In public health, agent-based modeling has historically been used almost exclusively to model infectious disease transmission and control in populations. ABMs are a good fit to analyze disease transmission since it can be used to find the interactions between individuals and the environments, often giving rise to population patterns of infectious disease incidence and persistence[8].

Research has been conducted in the area of sea transportation regarding several ABM approaches. These approaches include exploring various wind-assisted ship propulsion technologies to improve maritime sustainability and efficiency, simulating nautical services within ports to better understand their complex dynamics, modeling the disruptions of maritime trade chokepoints and global liquefied natural gas, modeling maritime traffic in piracy-affected waters, and exploring the effects of policy instruments on transitioning the maritime fuel system away from heavy fuel oil[9].

1.4 Scope of this project

There has been limited research on Agent-Based Models (ABMs) that use real data. In most cases, pure mathematical modeling is used. Even when real data is available, there are issues like low accuracy or insufficient data. To improve the accuracy of agents, it is crucial to use real data collected from people in the real world. This project aims to develop an ABM using practical statistical data (GPS) obtained from 100 individuals of varying occupations, collected for 21 days, and applying existing theories. The model proposed in this report aims to achieve a more realistic representation of human behavior.

We later discuss how the developed ABM can be utilized to analyse airborne disease spread in a given environment (Kandy District, Sri Lanka).

CHAPTER 02

METHODOLOGY

2.1 Obtaining Data

During the data collection phase, we obtained geospatial time series data documenting the daily mobility trajectories of 82 individuals across 19 distinct professional classes over a continuous period of 21 days with the resolution of 5 minutes. This comprehensive dataset comprises latitude and longitude coordinates paired with timestamps, facilitating the tracking of individuals as they move between various locations throughout the day. A part of the collected data is shown in figure 2.1.

	A	B	C	D	E	F
1	Latitude	Longitude	Altitude	Date	Time	
2	7.2286471	80.5888317	0	8/17/2023	19:51:00	
3	7.2286471	80.5888317	0	8/17/2023	19:54:00	
4	7.2286462	80.5888315	0	8/17/2023	19:59:00	
5	7.2286469	80.5888299	0	8/17/2023	20:04:00	
6	7.2286469	80.5888299	0	8/17/2023	20:09:00	
7	7.2287717	80.58895	467.8	8/17/2023	20:14:00	
8	7.2287289	80.5889549	467.7	8/17/2023	20:19:00	
9	7.2284415	80.589319	467.6	8/17/2023	20:24:00	
10	7.2282686	80.5892551	467.7	8/17/2023	20:29:00	
11	7.2286471	80.5888317	467.7	8/17/2023	20:34:00	
12	7.2285988	80.5888317	476	8/17/2023	20:39:00	
13	7.2286466	80.5889744	476	8/17/2023	20:44:00	
14	7.2286469	80.5888299	476	8/17/2023	20:49:00	
15	7.2286469	80.5888299	476	8/17/2023	20:54:00	
16	7.2286677	80.5888657	478.4	8/17/2023	20:59:00	
17	7.2286469	80.5888299	478.4	8/17/2023	21:04:00	
18	7.2286469	80.5888299	478.1	8/17/2023	21:11:00	
19	7.228825	80.5887662	538.1	8/17/2023	21:14:00	
20	7.2286909	80.5888922	538.1	8/17/2023	21:19:00	
21	7.2286935	80.5890828	537.4	8/17/2023	21:24:00	
22	7.2286812	80.5888894	538.4	8/17/2023	21:29:00	
23	7.229275	80.5893683	538.3	8/17/2023	21:34:00	
24	7.2288311	80.5891871	538.3	8/17/2023	21:39:00	
25	7.2286517	80.5889639	538.3	8/17/2023	21:44:00	
26	7.2287145	80.5889757	538.3	8/17/2023	21:49:00	
27	7.2288703	80.5890896	538.2	8/17/2023	21:54:00	

01an@tiec.lk

Figure 2.1: Sample Data Format

Table 2.1 documents the profession classes and the data collection process.

Table 2.1: Data collection details

Category		Code	Minimum Requirement	Ideal Requirement	Ongoing	Completed	# Needed for ideal req	# needed for min req
Transport	Bus Drivers	BD	5	8	0	2	6	3
	Other drivers	OD	2	2	0	2	0	0
	Tuk Drivers	TD	5	7	0	6	1	0
Educational	Teachers	TC	7	10	0	9	1	0
	Students	ST	7	10	1	7	2	0
Medical	Doctors	DC	3	4	0	4	0	0
	Nurses/Attendants	NA	5	7	0	7	0	0
	Midwife/PHI	MP	3	4	0	4	0	0
Garment	Garment Admins	GA	3	4	0	3	1	0
	Garment Workers	GW	3	4	1	3	0	0
Commercial	Retail shop workers	RW	3	4	0	5	0	0
	Super market workers	SW	3	4	0	4	0	0
Financial	Bank workers	BA	3	4	0	4	0	0
Agents	Sales Reps	SR	2	3	0	3	0	0
Agriculture	Farmers	FM	3	4	0	2	2	1
	Estate Workers	EW	3	4	0	0	4	3
	Livestock Farmers	AN	2	3	0	3	0	0
Administrative	Office workers	OW	7	10	0	10	0	0
	Field Officers	FO	3	4	0	4	0	0
			72	**100**	**2**	**82**	**17**	**7**

Legend	
	Ideal requirement met
	Minimum requirement met
	Minimum requirement NOT yet met
	Ideal requirement exceeded

2.2 Cleaning Data

It could be observed that, data missing and the Un-evenness in the data spread. Therefore, we had to go for cleaning process.

Initially, the data we got from the participants of the data collection had some data chunks missing here and there due to the signal issues, power drop and etc.…, which lead to the inconsistency of data. But we can allow some small data loss up to a time limit for each day. From the analysis of dataset, it was concluded that 4.8 hours data loss in a day was bearable. Since we sample GPS data at each 5 minute interval, we have total of 288 data points per day. Thus, the days which consist of less than 80% of the data points were removed. That means, days consist of data points above 231 were considered as useful data. Rest of the days were was omitted from our study.

2.3 Clustering Data

For our research, which aims to analyze location patterns irrespective of latitude and longitude coordinates, it is crucial to extract location information solely based on temporal data. Thus, we concentrated on identifying locations without considering geographic coordinates. This approach allows us to focus exclusively on the temporal aspect of the data, aligning with the objectives of our study. In this regard, each individual's daily data was processed through clustering algorithms to identify clusters representing locations, with both k-means clustering and DBSCAN explored for this purpose.

2.3.1 Utilizing K-means

K-means clustering is a centroid-based algorithm that partitions the data into a predetermined number of clusters, aiming to minimize the distance between data points and the centroid of their assigned cluster. However, when applied to our data, K-means yielded large clusters that included transportation points, making it challenging to accurately identify locations and stay points. As seen in figure 2.2, these transportation points caused the clusters to span across multiple distinct locations, hindering the algorithm's ability to precisely capture individual locations. Furthermore, K-means needs to set a predetermined cluster number, which we cannot know beforehand.

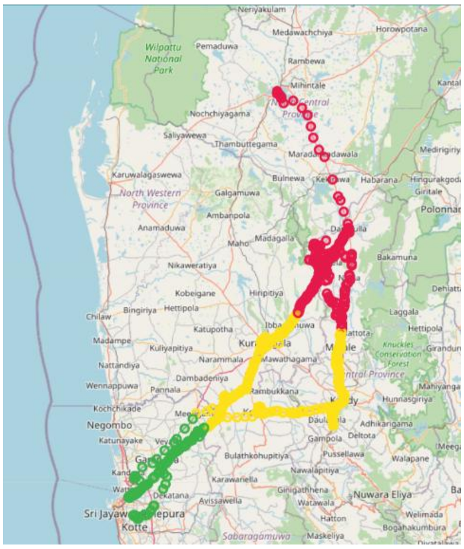


Figure 2.2: K-means clustering method applied on GPS data

2.3.2 Utilizing DBSCAN

On the other hand, DBSCAN is a density-based algorithm that identifies clusters based on regions of high density separated by regions of low density. It does not require the user to specify the number of clusters in advance, and it can find clusters of arbitrary shapes and sizes. DBSCAN is particularly effective for datasets with noise and outliers, as it can classify these points as noise rather than assigning them to a cluster.

Considering above reasons, DBSCAN was implemented as an alternative to address the challenge posed by transportation points, which act as outliers to our desired location clusters. This approach resulted in the formation of convenient clusters that effectively identified stay points. In DBSCAN, transportation points are identified as outliers due to their low density, while locations with high density are designated as stay points. In figure 3.3 we can see that the Leveraging the density-based nature of DBSCAN allowed for the accurate differentiation between stay points and transportation points, facilitating the identification of meaningful location clusters. Outlier points are marked in black dots.

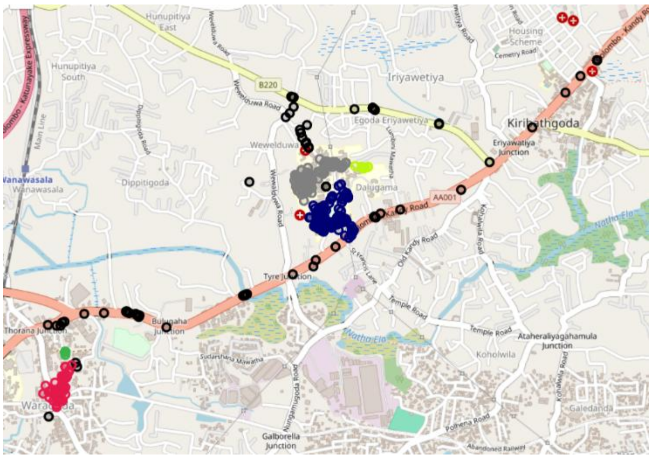


Figure 2.3: DBSCAN clustering method applied on GPS data

Within DBSCAN, we can define the clustering parameters, specifically the maximum distance between two samples to be considered part of the same neighborhood (Epsilon value) and the minimum number of samples required for a cluster to be formed (min points). By adjusting these values, we were able to obtain location clusters and more refined sub clusters. This flexibility allowed us to tailor the clustering process to our specific dataset, ultimately yielding meaningful and detailed clusters that accurately captured the underlying structure of the data.

Example:-

eps=1, minPTS=5 :- Location inside the university like canteen, playground, labs, as individual clusters

eps=5, minPTS=10 :- Entire university as one cluster.

2.4 Labeling Data

After carrying out a clustering process and eliminating transportation routes to identify specific locations that individuals frequented in their daily routines, our primary objective was to convert the longitudinal and latitudinal time domain data to a location time domain to enhance the generalization of the case. After obtaining the clustered locations, we needed to assign them appropriate names. To do so, we utilized the data from Table 2.2 to assign suitable names to the identified places. This process allowed us to gain valuable insights into the daily routines of the individuals under study.

Table 2.2: Location Names

Zone	Place
Residential Zone	Retail Shops
	Residential Park
	Gathering Place
	Home
Transport	Bus Station
	TukTuk Station
Administrative Zone	Administrative Zone
	Admin Office
	Admin Work Area
UrbanZone	Commercial Building
	Commercial Canteen
	Commercial and Financial Zone
	Commercial Work Area

	Gathering Place
	Fast Food Joint
	Restaurants
	Hospital
	Medical Zone
	Bank
	Shopping Mall
	Super Markets
	Retail Shops
Agricultural Zone	Agricultural Zone
	Estate
	Livestock Cultivate Area
	Plant Cultivate Area
Educational Zone	Classroom
	Education Zone
	School
	School Canteen
Industrial and Manufacturing Zone	Garment Building
	Garment Canteen
	Garment Office
	Garment Work Area

During the analysis of the dataset, the process of identifying clusters proved to be a challenging task. While some locations were successfully identified and assigned names, others were not possible to be considered as a location due to various reasons such as insufficient data or errors in the data collection process. These unidentifiable clusters were labeled as outlier clusters, as depicted in Figure 2.4.

The identification of these outlier clusters allowed for a more accurate and comprehensive understanding of the dataset, as they represented data points that did not fit the expected patterns of the identified clusters.

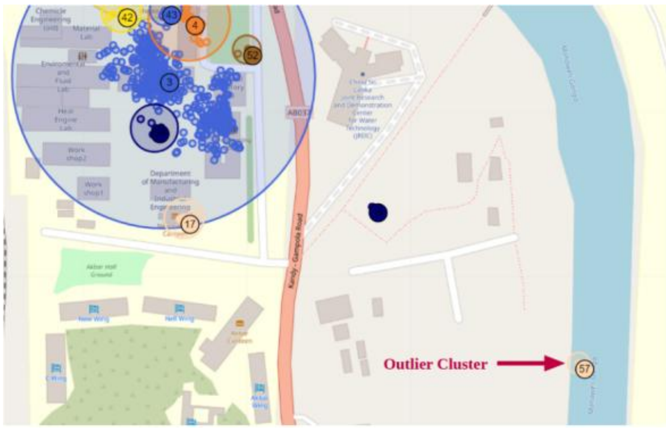
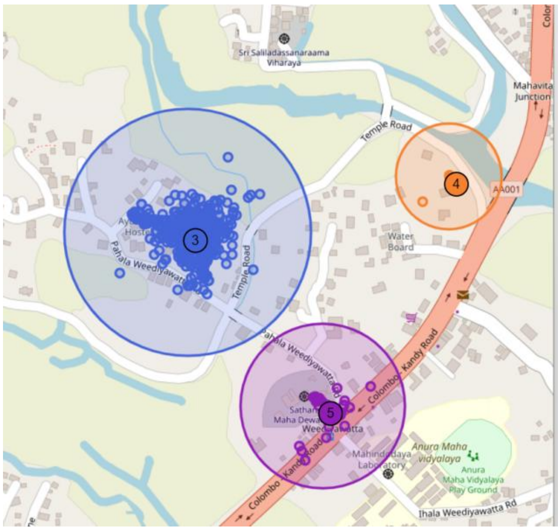


Figure 2.4 Representation of an Outlier Cluster.

In order to maintain consistency in the catalog, the assigned location name for these unidentifiable clusters was designated as "-1," which is shown in Figure 2.5. This allowed for easy identification and differentiation of the outlier clusters from the other identified clusters. The creation of catalogs for each individual was the final step in the analysis process, which provided a more detailed and personalized understanding of the data. Overall, this analysis was successful in identifying the clusters and outlier clusters in the dataset.



No	Location Name
0	Bank
1	Bank
2	Restaurants
3	Home
4	ResidentialZone
5	GatheringPlace
6	ResidentialZone
7	EducationZone
8	Restaurants
9	ResidentialZone
10	_w_home
11	RetailShops
12	Bank
13	Bank
14	Hospital
15	Hospital
16	ResidentialZone
17	Bank
18	_w_home
19	RetailShops
20	RetailShops
21	RetailShops
22	_home
23	ResidentialZone
24	Bank
25	RetailShops
26	_w_home
27	-1
28	-1

Figure 2.5 Numbered cluster with the catalog of the cluster.

2.5 **Domain Conversion**

Utilizing the above mentioned catalogs, each cluster number was mapped to corresponding string labels, effectively transforming the trajectory data into a format independent of geospatial coordinates.

Due to the non-uniform distribution of collected data, it was necessary to smooth it by recovering missing data without compromising reliability. A perceptual study was conducted on our dataset to address this issue, yielding the following findings:

- The first 5 hours of the day were typically spent at corresponding resting places (ie: Home, Weekend home, etc..).
- The last 4 hours of the day were typically spent at corresponding resting places.
- Movement between locations was observed during the middle of the day, making it difficult to identify a strict pattern through perception.

Based on the above observations, we concluded that interpolation is the most suitable solution for rectifying the data meaningfully. Having omitted harmful data losses during the cleaning process beforehand, the interpolation process is unlikely to significantly affect the accuracy of the trajectory.

Consequently, interpolation was performed differently for different time segments as follows:

- Instances where missing data occurred at the beginning or end of the day were interpreted as corresponding resting places, while missing data during daytime hours were interpolated with the preceding value.
- Subsequently, the data were converted to a 1-minute resolution, with 5-minute gaps filled using the preceding value. This process yielded a meaningful trajectory, embedded with logical assumptions.

Finally, the output comprised location strings with corresponding timestamps of 1 minute interval (Figure 2.6).

	2023-09-22_01tc	2023-07-13_03tc	2023-08-03_03tc	2023-08-05_03tc	2023-08-08_03tc	2023-08-21_03tc	2023-08-28_03tc	2023-09-02_03tc	2023-07-03_05tc
00:00:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:01:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:02:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:03:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:04:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:05:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:06:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:07:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:08:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:09:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:10:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:11:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:12:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:13:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:14:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:15:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:16:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:17:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:18:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:19:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:20:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:21:00	Home	Home	Home	Home	Home	Home	Home	Home	Home
00:22:00	Home	Home	Home	Home	Home	Home	Home	Home	Home

Figure 2.6: Trajectory of each person in Time and Location domain

CHAPTER 03

Behavioral Modeling

The primary objective of this research centers around the development of a sophisticated artificial intelligence (AI) model that can accurately model human mobility patterns. This model is engineered to analyze the spatial movements of individuals across diverse locations, meticulously tracking the temporal duration of their engagements at each locale.

3.1 **Literature review**

To model human behavior some researches have taken the approach of pure mathematical modeling using differential equations and such [10]. But in our research, we aim to model human behavior using practical data.

3.2 **Mobility patterns**

Typically, movement is conceptualized as random variations in position within a 2D space. However as explained in [11] a, real-world human mobility often exhibits a combined discrete and random pattern as shown in figure 3.1. Random movement occurs when individuals transition between different locations, while discreteness arises when individuals remain in a particular location for a period of time. This combined discrete and random fashion of movement reflects the complex dynamics of human mobility and is essential to consider when modeling and analyzing mobility patterns.

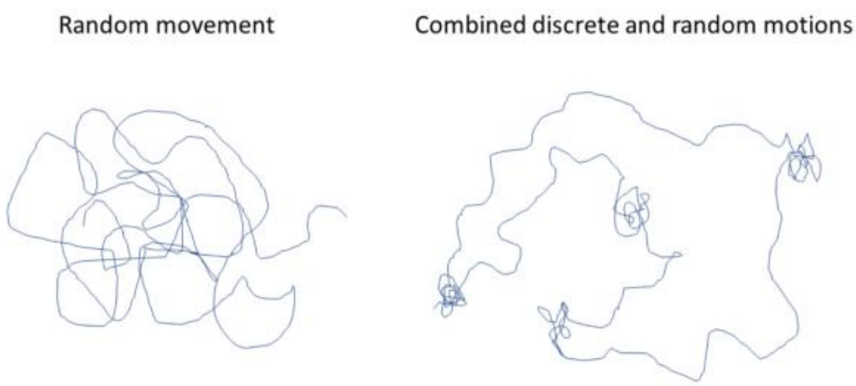


Figure 3.1: Mobility patterns[11]

According to the authors of [11], modeling human choices from a uniform random distributions is making a very serious claim about human behavior.

This implies that all choices are equally likely. But it is very clear that this is not true. People have preferences and their choices to where they will be, will depend on the person, occupation, their past memory, and the time of the day and the day itself. This is why modeling human behavior needs analyzing large amounts of data to gather insights about human motion dynamics.

3.3 Hybrid model

In the paper, PDSIM [11], An agent based framework, authors have proposed about a hybrid model that can be used to model human behavior. To model human mobility effectively, they have used a split functionality called "*Functional Bimodality".* This method approach two distinct functionalities tailored to estimate mobility dynamics:

- Stay Duration Estimation: First, we start by trying to figure out how long people stay in certain places. We look at past information to see if there are any trends in how long people tend to stay. The stay duration details come from analyzing the data, so they match up with how things actually happen in the real world.
- Location Visit Probability Estimation: After understanding how long people stay in different places, we then figure out how likely they are to move from one place to another once they're done. By looking at our data set, we can figure out these probabilities, which helps us understand the complex patterns of how people move around.

Combining these functionalities yields a hybrid model that encapsulates the multifaceted nature of human mobility (Figure 3.2). At each time step, individuals determine their stay duration based on the estimated distribution, subsequently utilizing the location visit probability distribution to dictate their next destination. This iterative process is supposed to accurately simulate human movement over time according to the authors of PDSIM paper[11].

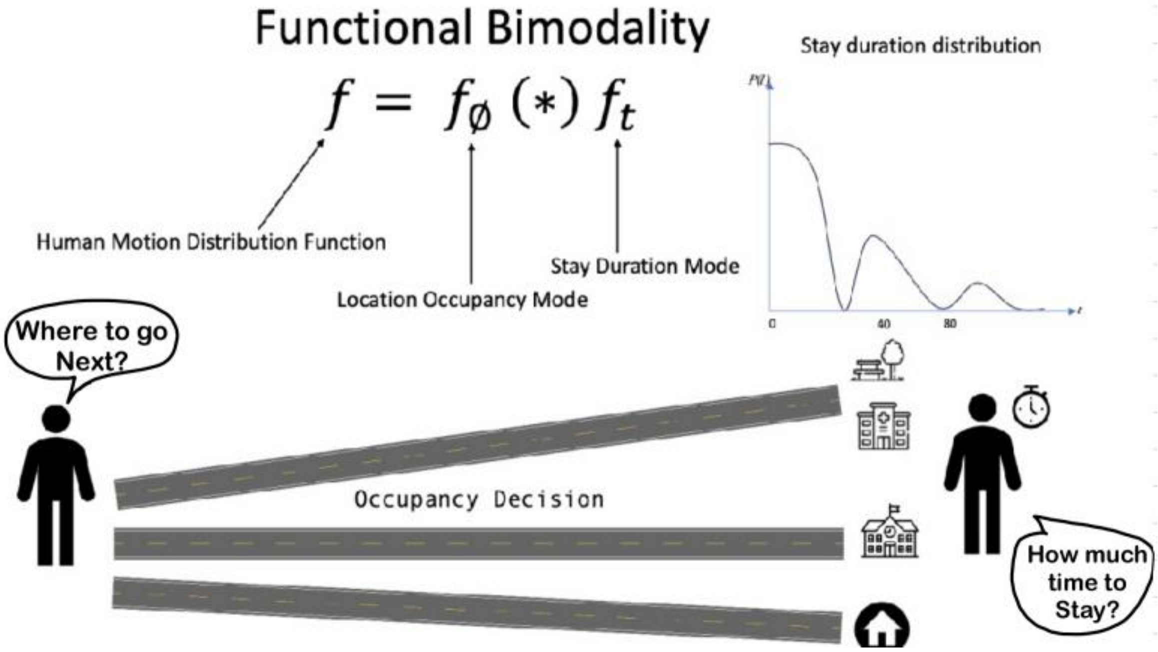


Figure 3.2: Hybrid model according to PDSIM authors [11].

3.4 Leveraging Hybrid Models in Our Study: A Strategic Approach

In our study, we thoroughly examine how people move around, paying close attention to how this differs among different job occupation classes. We use artificial intelligence techniques to dig deep into our data and find hidden patterns. Our main goal is to use these patterns to predict what might happen in the future.

Our ABM relies heavily on the complex hybrid model that combines two important matrices: the Location occupancy probability matrix (shown in Figure 3.3) and the Location visit probability matrix (illustrated in Figure 3.4). These matrices are proposed in the PDSIM paper[11]. These matrices are crucial for understanding the detailed dynamics of how people move around.

Stay Duration (min)

		001	002	003	004	005	………..
Location	**_home**	0.6429	0.6429	…….			
	_w_home	0.2143	0.2143	…….			
	_work	0.0000	0.0000	…….			
	AdministrativeZone	0.0000	0.0000	…….			
	AdminOffice	0.0000	0.0000	…….			
	AdminWorkArea	0.0000	0.0000	…….			
	AgriculturalZone	…….	…….	…….			
	AvgProvince	…….	…….	…….			

Figure 3.3: Expected format of Location occupancy probability matrix

Timestamp

		00:00:01	00:00:02	00:00:03	00:00:04	00:00:05	....
Location	_home	0.6429	0.6429	…….			
	_w_home	0.2143	0.2143	…….			
	_work	0.0000	0.0000	…….			
	AdministrativeZone	0.0000	0.0000	…….			
	AdminOffice	0.0000	0.0000	…….			
	AdminWorkArea	0.0000	0.0000	…….			
	AgriculturalZone	…….	…….	…….			
	AvgProvince	…….	…….	…….			
	Bank						
	BusStation						

Figure 3.4: Expected format of Location visit probability matrix

In the explained hybrid model (Figure 3.2), an individual's path follows a series of time-based stays and then moves to different locations. The Stay Duration Matrix controls how long someone stays at each place, determining when they decide to move to the next location. Once a stay expires, the Location Visit Probability Matrix helps the individual choose where to go next, representing the probabilistic way people make decisions about where to go.

This research doesn't just look at how people of individual occupation types move around. We also study the dynamics within each of these profession classes and human dynamics depending on the type of the day and many other factors. By figuring out and understanding these human dynamics, we can learn more about what causes people to move from one place to another. This will much accurately model human behavior.

To find such patters in humans, we need to use advanced machine learning clustering techniques on our data set.

CHAPTER 04

Spectral Clustering

4.1 What is spectral Clustering?

Spectral clustering is a technique used in machine learning for partitioning a data set into clusters based on their underlying structure. This method is very useful for data that is not linearly separable in original data space. In spectral clustering, a graph Laplacian matrix is created and its eigen values and eigen vectors are obtained. This eigen space will give us information about connections in our data. Overall, spectral clustering is a very powerful clustering algorithm.

4.2 Why spectral Clustering?

As explained in the previous chapter, we need a method to find human motion patterns within classes. Considering the complexity of the data set, it is not efficient to use simple clustering algorithms like K Means clustering. But the most important point is that we are mainly interested in the number of clusters there will be in our data. K Means and many other algorithms require us to initialize them with the number of clusters. But in context of our clustering problem, this is simply not the case. This is the reason, spectral clustering algorithm was selected above other clustering algorithms.

4.3 Data encoding

However, in order to use any algorithm for location-based clustering, we need to represent the location strings using numerical values. In our study, we assigned a number between 0 and 1 placed in equal intervals for each location, as shown in Table 4.1.

After assigning these numbers, we fed them into the spectral clustering algorithm. However, the results were not as promising as we had hoped. This was due to the fact that some frequently used locations had adjacent numbers, causing the algorithm to identify them as almost the same location. As a result, the clusters generated by the algorithm were not accurate.

Table 4.1 Location names and Assigned number

Location	Assigned no
_home	0.0227
_w_home	0.0455
_work	0.0682
AdministrativeZone	0.0909
AdminOffice	0.1136
AdminWorkArea	0.1364
AgriculturalZone	0.1591
AvgProvince	0.1818
Bank	0.2045
BusStation	0.2273
Classroom	0.25
CommercialBuilding	0.2727
CommercialCanteen	0.2955
CommercialFinancialZone	0.3182
CommercialWorkArea	0.3409
COVIDQuarantineZone	0.3636
DenseDistrict	0.3864
EducationZone	0.4091
Estate	0.4318
FastFoodJoint	0.4545
GarmentBuilding	0.4773
GarmentCanteen	0.5
GarmentOffice	0.5227
GarmentWorkArea	0.5455
GatheringPlace	0.5682
Home	0.5909
Hospital	0.6136
IndustrialManufactureZone	0.6364
LivestockCultivateArea	0.6591
MedicalZone	0.6818
PlantCultivateArea	0.7045
ResidentialPark	0.7273
ResidentialZone	0.75
Restaurants	0.7727
RetailShops	0.7955
RuralBlock	0.8182
School	0.8409
SchoolCanteen	0.8636
ShoppingMall	0.8864
SparseDistrict	0.9091

SuperMarkets	0.9318
TestCenter	0.9545
TukTukStation	0.9773
UrbanBlock	1

After encountering an issue with the representation of locations, a new approach was used. To address this problem, alternative representations of the locations were explored. First, all the locations were divided into main zones. After analyzing the behavioral pattern of the GPS data, the zones were sorted to give the maximum distance between the most used zones. This allowed us to reduce the complexity of the data and focus on the most relevant information.

To give more distance between the zones, three bits were assigned to each zone, and four bits were for each zone's location, as shown in Table 4.2. Then, the zone and location numbers were concatenated, creating a new location binary number. This binary number was then converted into a decimal number. Then, all the numbers were mapped between zero and one.

This approach allowed a more compact and accurate representation of the locations, enabling us to analyze the data more efficiently. By reducing the complexity of data, we could focus on the most relevant information, making our analysis more effective and informative. Overall, this method proved valuable since, after being fed into the spectral clustering algorithm, it got acceptable results.

Table 4.2 Location names with modified New Assigned Number

Zone	Zone Bin no	Location	Location Bin no	Final bin no	Dec no	New assigned no
ResidentialZone	000	_home	0000	0000000	0	0
		_w_home	0001	0000001	1	0.0086
		ResidentialZone	0010	0000010	2	0.0172

		ResidentialPark	0011	0000011	3	0.0259
		RuralBlock	0100	0000100	4	0.0345
		Home	0101	0000101	5	0.0431
Other	001	AvgProvince	0000	0010000	16	0.1379
		COVIDQuarantineZone	0001	0010001	17	0.1466
		DenseDistrict	0010	0010010	18	0.1552
		SparseDistrict	0011	0010011	19	0.1638
		TestCenter	0100	0010100	20	0.1724
Transport	010	BusStation	0000	0100000	32	0.2759
		TukTukStation	0001	0100001	33	0.2845
AdministrativeZone	011	AdministrativeZone	0000	0110000	48	0.4138
		AdminOffice	0001	0110001	49	0.4224
		AdminWorkArea	0010	0110010	50	0.431
UrbanZone	100	_work	0000	1000000	64	0.5517
		CommercialBuilding	0001	1000001	65	0.5603
		CommercialCanteen	0010	1000010	66	0.569
		CommercialFinancialZone	0011	1000011	67	0.5776
		CommercialWorkArea	0100	1000100	68	0.5862
		GatheringPlace	0101	1000101	69	0.5948
		UrbanBlock	0110	1000110	70	0.6034
		FastFoodJoint	0111	1000111	71	0.6121
		Restaurants	1000	1001000	72	0.6207
		Hospital	1001	1001001	73	0.6293
		MedicalZone	1010	1001010	74	0.6379
		Bank	1011	1001011	75	0.6466

		ShoppingMall	1100	1001100	76	0.6552
		SuperMarkets	1101	1001101	77	0.6638
		RetailShops	1110	1001110	78	0.6724
AgriculturalZone	101	AgriculturalZone	0000	1010000	80	0.6897
		Estate	0001	1010001	81	0.6983
		LivestockCultivateArea	0010	1010010	82	0.7069
		PlantCultivateArea	0011	1010011	83	0.7155
EducatioinZone	110	Classroom	0000	1100000	96	0.8276
		EducationZone	0001	1100001	97	0.8362
		School	0010	1100010	98	0.8448
		SchoolCanteen	0011	1100011	99	0.8534
IndustrialManufactureZone	111	IndustrialManufactureZone	0000	1110000	112	0.9655
		GarmentBuilding	0001	1110001	113	0.9741
		GarmentCanteen	0010	1110010	114	0.9828
		GarmentOffice	0011	1110011	115	0.9914
		GarmentWorkArea	0100	1110100	116	1

4.4 Literature review and algorithm for spectral clustering.

There are many literature written about spectral clustering approach for data science. The paper by Andrew Y. Ng[12], and the reference, [13] proved to be very useful in our study.

Let us consider our data points as $x_1, x_2, x_3, \ldots x_n$ and notion of similarity as $S_{ij} > 0$ between x_i, x_j.

To define this similarity in mathematical terms, we can use a Gaussian kernel.

$$S_{ij} = \begin{cases} exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) & i \neq j \\ \mu & i = j \end{cases}$$

Where sigma is the scaling parameter. And μ is the predefined self similarity; which is a positive value in the range of [0,1].

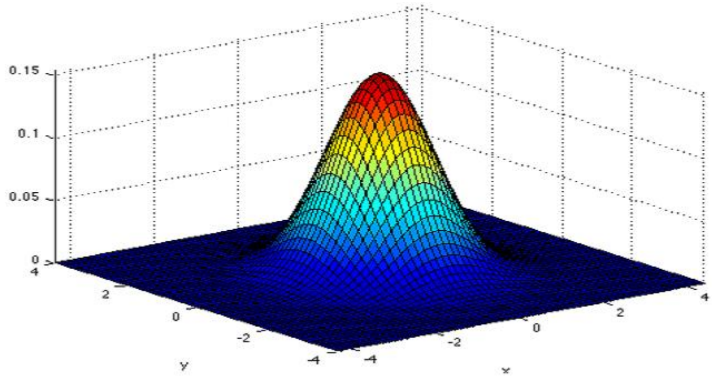


Figure 4.1: Gaussian Kernel function in three dimensions.

The Gaussian kernel will give a high similarity to close by data points. This will encourage them to be in the same cluster.
First, a similarity matrix called Adjacency matrix is created as shown below.

$$W_{[i,j]} = \begin{bmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{n1} & \cdots & S_{nn} \end{bmatrix}$$

Then we can create the degree matrix ans,

$$D_{[i,j]} = \begin{cases} \sum_j W_{[i,j]} & i = j \\ \mu & otherwise \end{cases}$$

Now there are several methods to obtain the Graph Laplacian,

Un-normalized Laplacian (UL),

$$L = D - W$$

Symmetric normalized Laplacian (SNL),

$$L = I - D^{-1/2} W D^{-1/2}$$

Random walk Laplacian (RWL),

$$L = I - D^{-1} W$$

Then the eigen values and eigen vectors of the Laplacian were calculated.

Sort the eigen values from smallest to largest.

Perform the sigma sweep (figure 4.2) by plotting eigen gaps between the sorted eigen values against sigma. Note that mode xy corresponds to the eigen gap between x^{th} and y^{th} eigen values after sorting.

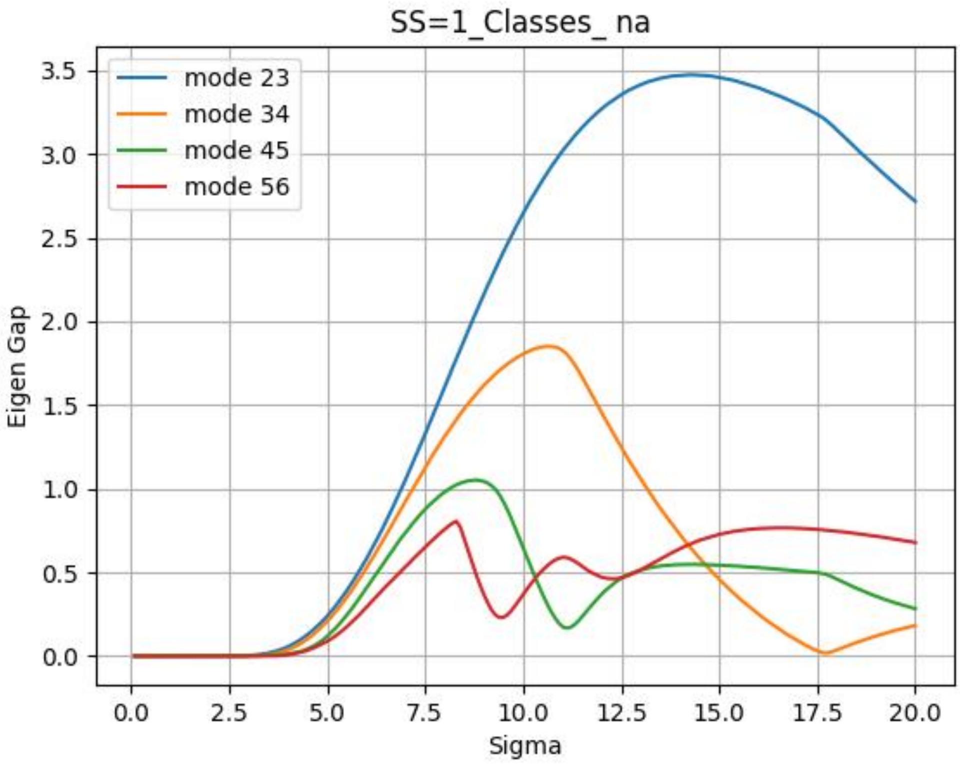


Figure 4.2: Sigma sweep performed on nurse attendants for self similarity(ε) = 0.1

From this sigma sweep plot, the number of optimal clusters can be concluded by using below selection algorithm.

- Largest gap value: the eigen gap with the highest gap value was considered as the prominent mode.

Once the prominent mode is selected, the σ value for which the highest gap value was recorded for the prominent mode was selected as the dominant σ under the selection algorithm[13].

Once we have these two parameters (prominent mode and dominant σ) clustering can be done in below steps as described in [13].

Form the matrix, $X = [v_1 v_2 ... v_n] \in \mathbb{R}^{n \times n}$ by stacking eigen vectors in columns. Treating each row of X as a point in $\mathbb{R}^n$, cluster them into the desired number of clusters using K means clustering.

4.5 Spectral clustering results.

4.5.1 Use of different Laplacians.

1. Unnormalized laplacian- It can be observed that, sigma sweep lines are not smooth when we use the UL. Sweep plot is not clear for all the traces and they tend to lose some important details (Observe figure 4.3 and figure 4.5).

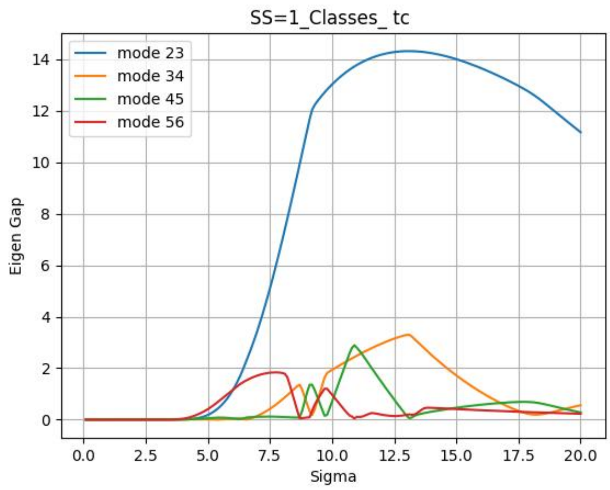


Figure 4.3: For Teachers (UL)

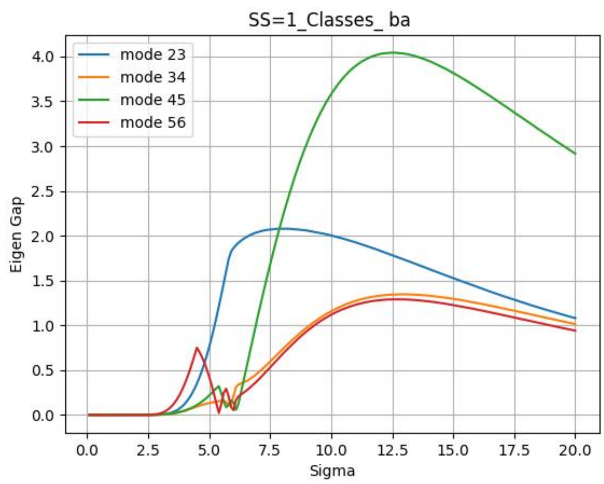


Figure 4.4: For Bank workers (UL)

2. Symmetric normalized laplacian- Gives much better and smoother curves compared to unnormalized laplacian (Figure 4.5 and Figure 4.6).

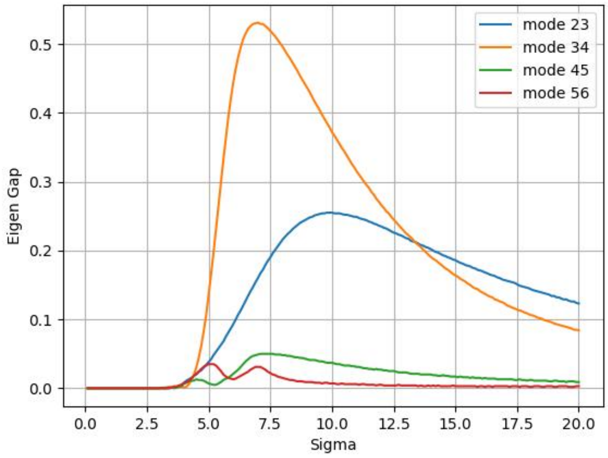


Figure 4.5: For Teachers (SNL)

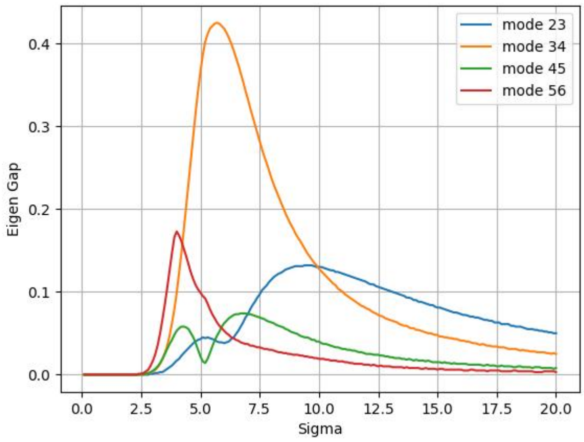


Figure 4.6: For bank workers (SNL)

3. Random walk laplacian - Results are identical as for symmetric normalized laplacian (Figure 4.7 and 4.8). This is also confirmed in [14].

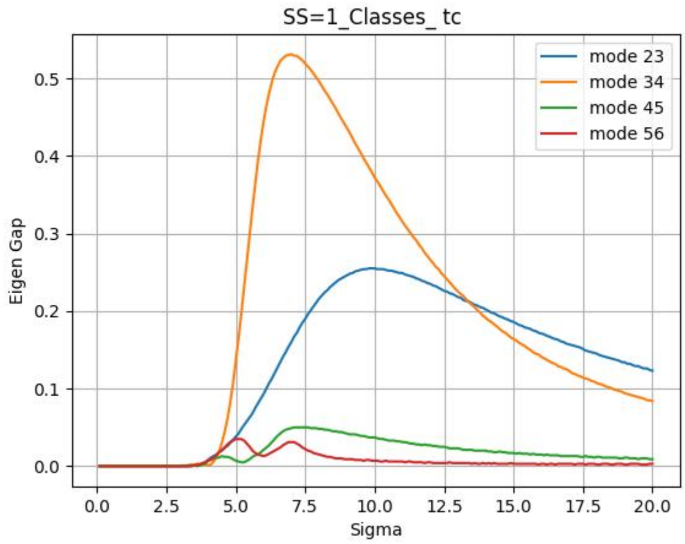


Figure 4.7: For Teachers (RWL)

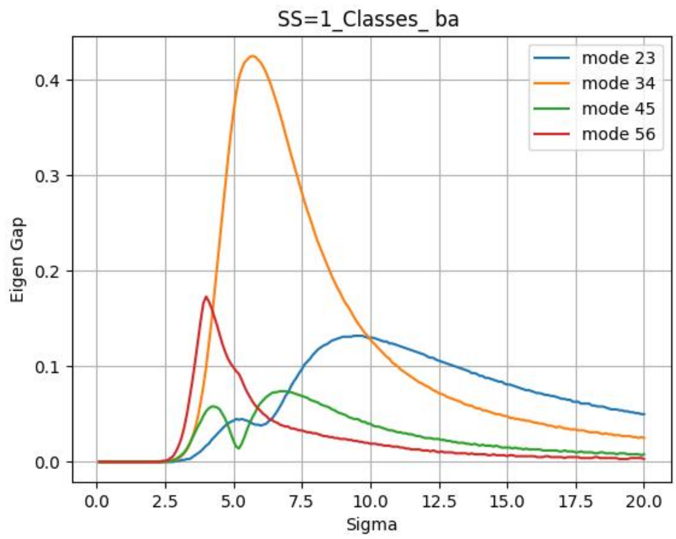


Figure 4.8: For Bank workers (RWL)

Considering these results, it can be seen that symmetric normalized laplacian is much more reliable at all cases. Therefore, it was selected to carry out further research on human behavior.

4.5.2 Spectral clustering on different occupations.

4.5.2.1 Teacher Class

Figure 4.9 shows the sigma sweep plot for Teacher class.

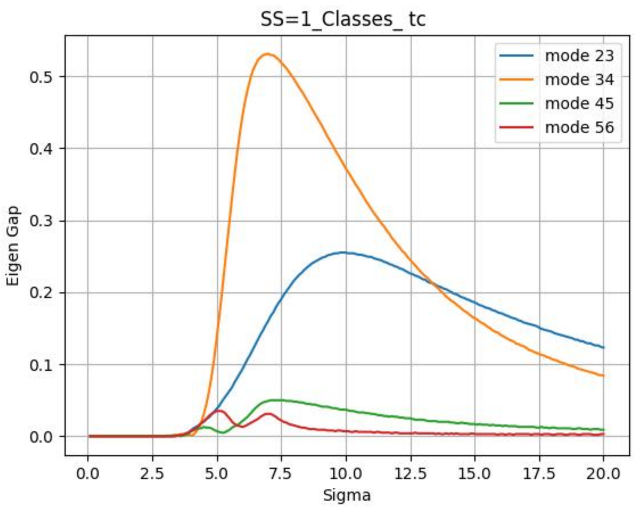


Figure 4.9: Prominent mode is mode 34 at σ = 7.

<u>For the case of two clusters,</u>

I. A teacher makes a separate cluster.

II. Normal working days.

III. Holidays.

This is an interesting behavior. Upon analyzing this teacher, it was found that he stayed at the school quarters without going home. This unusual behavior caused him to be in a separate cluster. This can be clearly seen in figure 4.10 and 4.11.

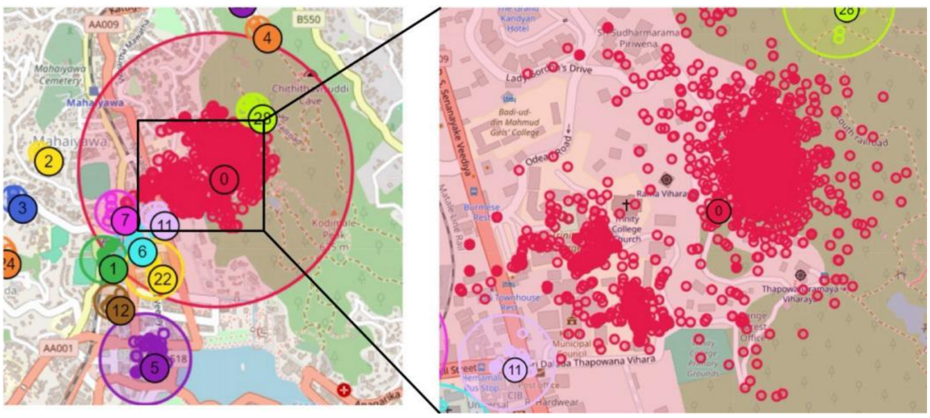


Figure 4.10: The geographical cluster formed by the outlier teacher at his workplace.

A	AT	AU	AV	AW	AX	AY
TimeStamp	2023-07-30_06tc	2023-08-01_06tc	2023-07-07_07tc	2023-07-08_07tc	2023-07-09_07tc	2023-07-10_07tc
00:00:00	_home	_home	School	School	School	School
00:01:00	_home	_home	School	School	School	School
00:02:00	_home	_home	School	School	School	School
00:03:00	_home	_home	School	School	School	School
00:04:00	_home	_home	School	School	School	School
00:05:00	_home	_home	School	School	School	School
00:06:00	_home	_home	School	School	School	School
00:07:00	_home	_home	School	School	School	School
00:08:00	_home	_home	School	School	School	School
00:09:00	_home	_home	School	School	School	School
00:10:00	_home	_home	School	School	School	School
00:11:00	_home	_home	School	School	School	School
00:12:00	_home	_home	School	School	School	School
00:13:00	_home	_home	School	School	School	School
00:14:00	_home	_home	School	School	School	School
00:15:00	_home	_home	School	School	School	School
00:16:00	_home	_home	School	School	School	School
00:17:00	_home	_home	School	School	School	School
00:18:00	_home	_home	School	School	School	School
00:19:00	_home	_home	School	School	School	School
00:20:00	_home	_home	School	School	School	School
00:21:00	_home	_home	School	School	School	School
00:22:00	_home	_home	School	School	School	School
00:23:00	_home	_home	School	School	School	School
00:24:00	_home	_home	School	School	School	School
00:25:00	_home	_home	School	School	School	School
00:26:00	_home	_home	School	School	School	School
00:27:00	home	home	School	School	School	School

Figure 4.11: The string list of the outlier teacher.

It was decided remove this teacher from the data set and run the clustering algorithm again since this was an anomaly in our data set. The obtained sigma sweep plot is shown below. We notice that there are two prominent clusters.

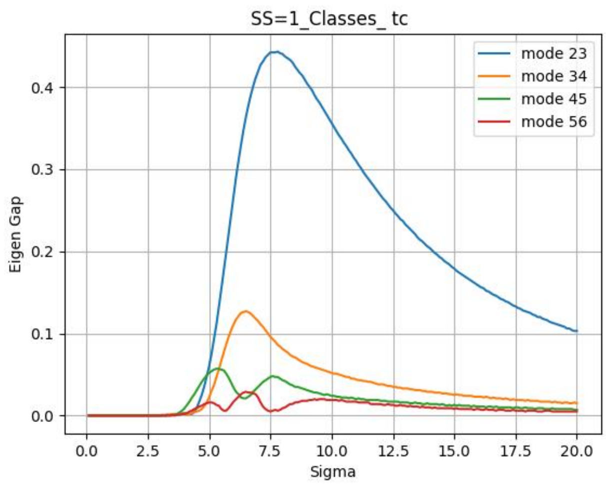


Figure 4.12: Sigma sweep plot after removing the outlier teacher.

After performing K means on eigen vectors, we obtain the following results for two clusters. Figure

I. First cluster contains weekend or holiday behavior.

II. Second cluster contains normal workdays.

Therefore, it was concluded that for teacher class, their behavior split into week day and week end (Holiday) behavior.

4.5.2.2 Doctors

For doctors, below clusters were observed.

I. The regular work routine of a Doctor. (Morning to evening regular shift).

II. The work routine of a Doctor with a late-night to next-day evening shift.

III. The work routine of a Doctor with a night to morning shift

IV. Doctor behavior during weekends and holidays

V. The work routine of a Doctor who manages a night shift alongside duties at a school.

We conclude that behavior of doctors is based on their work shift.

4.5.2.3 Bank workers

For bank workers, below clusters were observed.

I. Holidays/weekends

II. Normal days.

III. An outlier who spent 3 consecutive days in the bank.

The behavior of this third cluster in abnormal. This is likely due to the fact that this particular person has forgot their GPS device at bank or other reason. We can conclude that the behaviour of bank workers depends on the type of the day.

4.5.2.4 Results for other classes

For other classes this process was carried out and such similar results were obtained. The obtained results are summarized in the table 4.3.

Table 4.3 : Spectral Clustering Results

Class	Identified Clusters
Animal Farmer	Work area is located close to residence
	Work area is located far away from residence
	Lives and works in a designated Agricultural Zone
Bank Worker	Typical work day behaviour
	Weekend/Holiday behavior
	Outlier behaviour (seems to have spent all day at the bank for multiple consecutive days)
Doctor	Normal work day behaviour
	Night to evening shift
	Night to morning shift
	Weekend/Holiday behavior
	Night shift at hospital and works at a school medical center during the day
Farmer	Work area is located close to residence
	Works at an Agricultural Zone
Field Officer	Work related behavior
	Non-work related behaviour
Garment Admin	Weekend/Holiday behaviour
	Normal work day behaviour
Garment Worker	Work area is located far away from residence
	Lives in/close to the work area
Midwife/PHI	Work is done on the move (visiting residences)
	Works at a Medical Zone
Sales Rep	Normal workday behaviour
	Holiday behavior
Student	Normal behavior of student travelling from home
	Weekend/Holiday behavior with irregular attendance to school
	Hostel student
Super Market Worker	Morning to night shift behaviour
	Afternoon to night shift
	Morning to evening shift
	Works morning to afternoon & evening to night with a break in between. Also includes holiday behaviour
Teacher	Teacher lives in or close by to school premises
	Normal teacher behaviour (travelling from home)
	Weekend/holiday behavior

4.6 Conclusion

After examining the results, it can be observed that the spectral clustering algorithm excels at producing clusters with distinct behavioral patterns, ranging from regular work routines to holiday behaviors, in alignment with real-world intuition. Furthermore, the algorithm effectively filters out anomalous patterns, as demonstrated in some of the instances above. This approach, therefore, provides a strong framework for understanding and modeling complex social behaviors. This human motion trajectory dataset together with the above observed results can be utilized in Agent-based Models to improve their accuracy.

CHAPTER 5
Introduction to the Agent-Based Modeling Simulator

5.1 Brief introduction

Since we sorted out all the occupation classes with their hidden' sub class and the required statistics (Location Visit Probability Matrix and the Stay Duration Probability Matrix), now we can move on to create the ABM model for our data set.

ABMs can simulate the actions of each agent in the simulation and agents behave on a given set of rules. This reflects how the real world behaves. Earlier we discussed how the ABMs can be improved with the real-time GPS data of humans. Using the developed ABM, now going to simulate how an airborne disease going to spread and how it is going to affect the agents in various occupation classes.

5.2 Components of the Simulation

5.2.1 Environment

The environment is a crucial component in the ABM to get accurate simulation of transportation and disease propagation. The environment is used to eject complexities of the real world into the simulation. To show the geographical hierarchy, a tree-like hierarchy is used for the environment as shown in Figure 5.1. This helps the agents to travel between places which is similar to real-world travel tragedies. When an agent moves from one location to another, that agent must pass through some locations/zones. To identify these locations and zones, the tree structure hierarchy was used. This is very useful to understand and predict the spread of diseases.

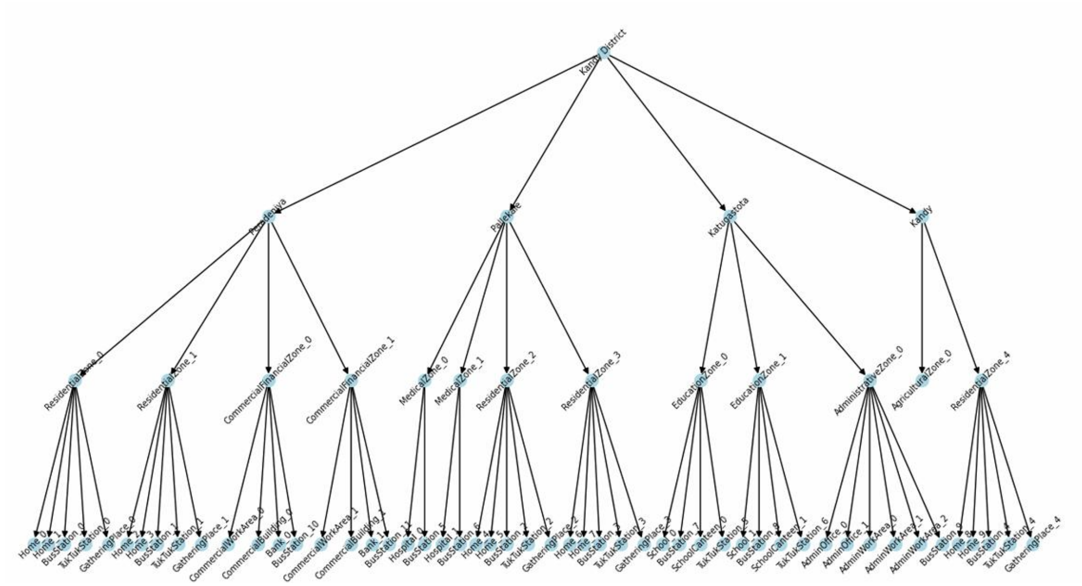


Figure 5.1 Real-world example for geographical hierarchy

To do the simulation as the case study area we selected the Kandy district in Sri Lanka. For this, Pallekele, Kandy, and Gampola main cities were selected. In our ABM model, we can select the boundaries of a real-world map as the environment and can transfer it into the ABM environment which gives more realistic data to the system.

First, above mentioned cities were carefully annotated in the google maps using polygon shapes. Then within the sub-cities, zones were identified and marked, such as educational, medical, and residential zones. The names of these zones were chosen based on the dataset to accurately represent the spatial distribution and functional characteristics of various locations within the district. Figure 5.2 illustrates the annotated cities and zones.

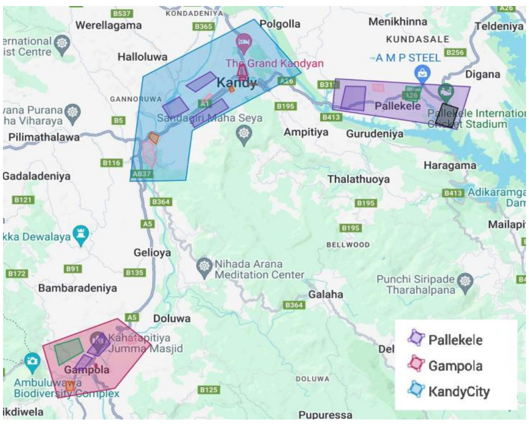


Figure 5.2 Map illustrating the Kandy District with annotated cities and functional zones.

The coordinates of the polygons were saved in a CSV file by exporting the annotations. These coordinates were used to create the environment in the ABM. Using the 'NetworkX' library, a hierarchical structure was built including a detailed and organized representation of the environment. Primary locations such as Schools, Hospitals, Homes, supermarkets are at the base level of the hierarchy. Then These locations were randomly placed within designated zones, such as educational, residential, and medical zones. Since specific placement of location within the zones doesn't impact much on the disease propagation, this method minimizes the need for manual annotation. The dataset ensures consistent location and zone names with real-world data, and Figure 5.3 displays the hierarchical tree structure of the environment which is a visual representation of the tree structure, with a single zone expanded to highlight its individual locations for better clarity.

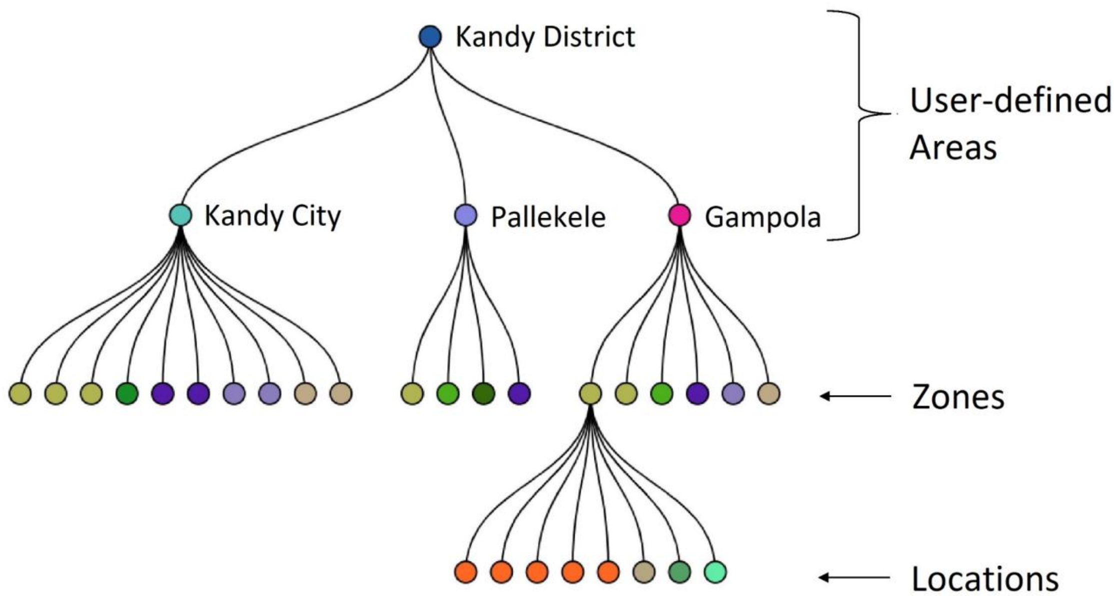


Figure 5.3 Hierarchical tree structure of the environment.

5.2.2 Agents

To simulate real human behavior in the environment, agents play a crucial role in the ABM. Agent interactions are used to model a wide range of phenomena, including disease propagation, smart city planning, transportation systems, and social dynamics.

In our model to represent the agents, class, and object structure were selected. This agent object contains various information such as home and work locations, the age range, and information related to disease propagation risks.

First, we defined the total number of agents that going to be present in the environment. Then the respective number of agents were created according to the given percentages in the occupation classes. Then, they are assigned to the sub-classes according to the given sub-class percentages. Each Node object within the 'Home' location class was initially assigned a total agent count based on simulation parameters. For a given home adult to child ratio was defined as a simulation parameter. Then adults and child agents are randomly assigned to the homes. When assigning the child agents to the homes, there must be assigned at least one adult into the home assuming that at least one adult is required for the proper functioning of a home. Then all the agents were assigned to a work location based on the workplace class.

CHAPTER 6
Rules that define human motion

6.1 Human Motion Model

For each agent, a personalized schedule was created at the beginning of each day. This schedule includes the location that the agent going to visit and the time duration that the agent going to spend in the given location as we mentioned in Chapter 3. To create the schedule two cumulative probability matrix were used which were the Location Visit Probability Matrix and the Stay Duration Probability Matrix.

6.1.1 Location Visit Probability Matrix

This matrix gives the probability of being at each location for a given time. Since now each occupation class has sub-classes under it, new Location Visit Matrices were created for each sub-class. These matrices were created from the real GPS data trajectories.

6.1.2 Stay Duration Matrix

This matrix gives the probability of stay duration for a given location. Since now each occupation class has sub-classes under it, new Stay Duration Matrices were created for each sub-class. These matrices were created from the real GPS data trajectories.

6.2 Generating Random Variables

When starting the day, each and every agent get the time table for them. This personalized time table shows what are the locations to visit and how much time needs to spend on that location. For this calculations Stay Duration Matrix and Location Visit Probability Matrix were used.

6.2.1 Theory

Let's assume,

- $f_x\ (x)$ is uniformly distributed in [0,1] and,
- Y = $g_{(x)}$ to be a monotonically increasing function.

Then,

$$f_y(y) = \frac{f_x(x)\ |_{x=g^{-1}(y)}}{\frac{dy}{dx}\ |_{x=g^{-1}(y)}} = \frac{1}{\frac{dy}{dx}}$$

Also

$$f_y(y) = \frac{dF_y(y)}{dy}$$

Therefore,

$$\frac{dF_y(y)}{dy} = \frac{dx}{dy} \quad \Rightarrow \quad F_y(y) = x$$

Finally; can be obtain that

$$y = F_y^{-1}(x)$$

6.2.2 Implementation in Python

First, we select the Location Visit Probability Matrix according to the occupation class and its sub class. Then we get the cumulative function of the given probability for a specific time step as shown in the figure 6.1.

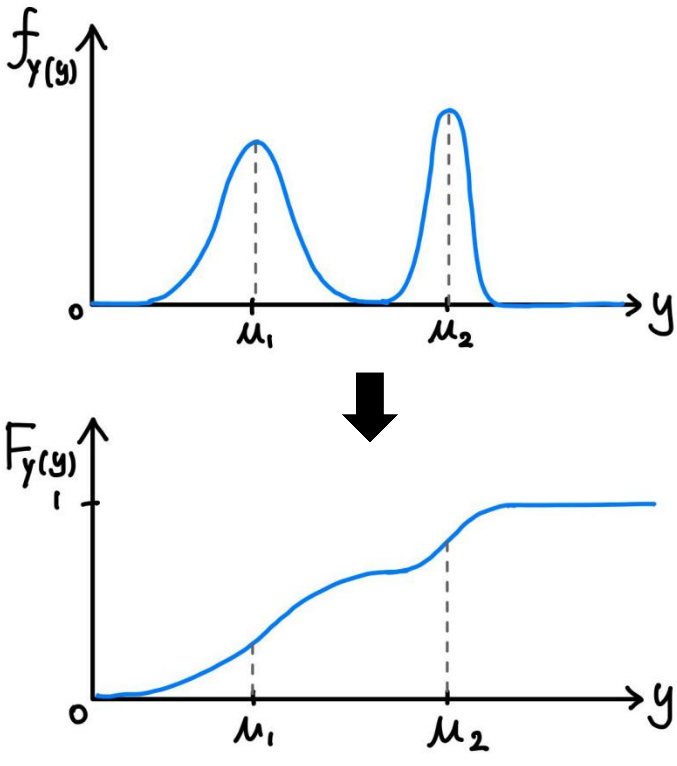


Figure 6.1 converting the PDF to CDF.

Then a numerical value between 0 and 1 was randomly generated. Then it was mapped using the inverse cumulative function of location visit probability matrix as shown in the figure 6.2 to generate new location random variables.

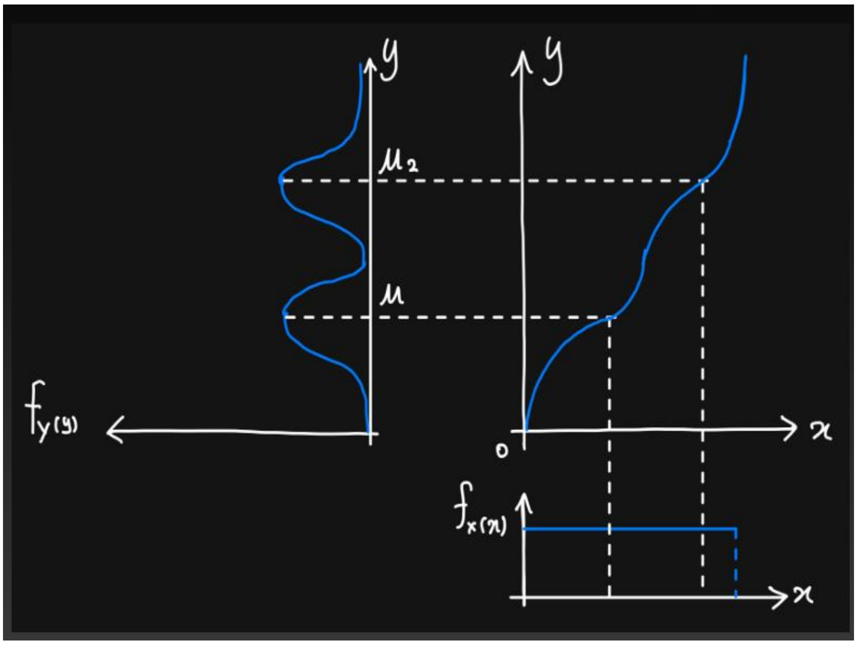


Figure 6.2 Getting the desired output from the probability matrix, given a random number.

Same method is carried out for the Stay Duration Matrix to identify how much time is going to spend at a given location. This method ensures the realistic behavior of agents while also preserving an acceptable level of randomness. The simulation allows agents to move in each minute to the locations specified in their schedules, with suitable transport modes assigned. Then each agent stays at the given location till given stay duration is expired. Within that period, agent go in a random motion withing the given location boundary. Finally, the agent moves on to the next scheduled location as per their personal schedule.

CHAPTER 7
Transportation

7.1 Introduction

Most of the existing ABMs do not consider transport elements [15], [16]. In some applications like disease modeling, it is crucial to have a transport layer in the stimulation due to the frequent interactions between agents while in transport medium. In our simulation, we modeled a realistic transport system embedded to the Kandy environment. We implemented a virtual transport system that closely mirrored practical conditions, incorporating buses and three-wheelers as the primary public transport modes. Each zone had a bus and a three-wheeler station, allowing agents to travel between zones.

7.2 Public transport

We have simulated buses and three-wheelers as public transport medium. When agents need to move to a new location, if the location is very close by, they will walk to that location. Other-wise they will reach the nearest bus stop to catch a public transport medium. We have initialized two bus routes, Gampola-Kandy, and Pallekele-Kandy, to cover all the intermediate zones in the three cities. These bus routes can be changed according to the real-world bus schedule data. Then bus objects were created on each route to start their journeys at defined intervals. This interval was set to ensure that most agents could find a bus without waiting more than 15 minutes at a bus station.

Each bus station, referred to as a Bus Node, maintained an updated dictionary of available buses and their next destinations. Each Bus object waited for 5 minutes at each Bus Node. When a bus arrived, agents who had been waiting for transportation for no more than 15 minutes were assigned to that bus, provided the bus's next destination matched the agent's next location in the bus's dictionary of destinations.

7.3 Three wheels

If the 15-minute waiting time were exceeded, each agent would be assigned to a Three-Wheeler object to travel to the next destination. The number of Three-Wheeler objects assigned to each zone was decided based on the population at that zone. This ensures every agent who missed the bus can reach their target destination using a three-wheeler. When a three-wheeler arrived at a destination zone, it stayed there until the next trip was assigned.

7.4 Distance calculation

The travel time between zone to zone for the buses and the three-wheelers was calculated by dividing a defined speed for the transport class by the straight-line distance between zones. The distances were taken from the modeled environment which consist of real longitude and latitude information. This speed was decided considering the practical time for travel using those transportation mediums.

If *(lon1, lat1)* are the coordinates of the starting location and *(lon2, lat2)* are the coordinates of the destination zone, the distance between them is given by,

$$d = 2 * R\,\mathrm{atan}\,2(\sqrt{a}, \sqrt{(1-a)})$$

Where;

$$a = \sin^2\left(\frac{dlat}{2}\right) + \cos(lat1) * \cos(lat2) * \sin^2\left(\frac{dlon}{2}\right),$$

$$dlon = lon2 - lon2, dlat = lat2 - lat1$$

Here, R was the radius of the Earth (Mean value of the radius 6371 Km was taken)

CHAPTER 8
Disease Modeling

8.1 Literature Review

Among the many applications of Agent Based Models, disease modeling is of no less importance. Pandemics and epidemics pose significant threats due to their rapid global spread, overwhelming healthcare systems, and causing high mortality and morbidity rates. The most recent COVID-19 pandemic has underscored the urgent need for advanced tools to predict and manage infectious disease transmission in complex environments.

Earliest disease modeling methods were purely analytical. They did not consider an environment, transportation or object oriented agents. These models are called the SIR models. Although largely positive results have been obtained in a handful of studies employing these traditional analytical approaches [17], [18], [19] ,they fail to capture micro level insights about the simulation.

Also there have been numerous attempts to model epidemics using ABMs. But many existing models oversimplify environments and often fail to consider disease transmission through public transportation[15], [20]. Because the ABM proposed in above chapters consists of a realistic environment and a transportation network, it is capable of predicting disease spread patterns very accurately for a given location.

8.2 Markov Model

In the studies[20], [11] they have proposed a Markov decision chains to model the disease propagation. In our study we will use a similar method together with some improvements over them. Here, agents in the total population were categorized into states as in the SEIR model. In the simulation, initially, a small portion from the total number of agents were randomly set to the infected state while others were in the susceptible state. Then the disease propagation can be studied. The Markov state model used in this study is depicted in figure 8.1.

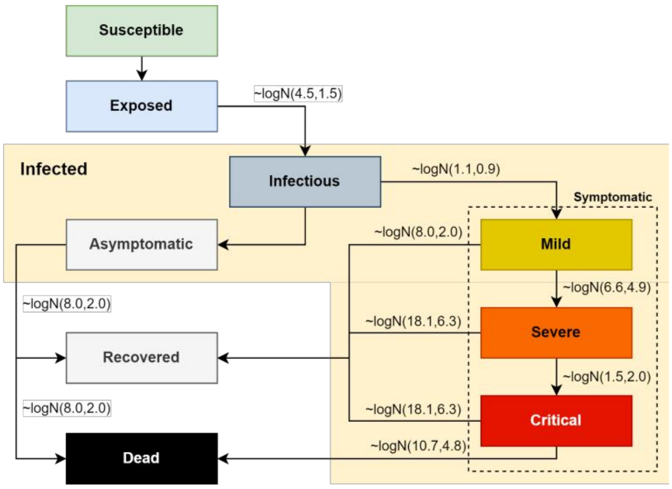


Figure 8.1 : The Markov Model used in the Simulation

The transition times were taken from log normal distributions. It was discussed in [21] that transition times follow a log normal distribution, and the transition times were taken from [11].

8.2 Contact Modeling

Each minute, the simulation allows agents to move to the locations specified in their schedules, agents will use appropriate transportation methods to reach their destination. Once an agent has reached their destination, they will stay there for some time until the stay duration expires. During their stay, agents engage in random motion within the location's boundaries.

This will allow agents to randomly meet and interact with other agents. This way, it is possible to get contacts within agents. If an infected agent comes into contact with a non-infected agent, the non-infected agent can get infected with some probability. This probability will depend on the contact time and the disease state of the disease state of the infected agent.

While in transport, agents can also get infected if there are other infected agents in the public transportation method. The disease state of each agent is updated in regular intervals. If there are more infected agents in a given transport object, susceptible agents have a high probability to get the disease. Since we only assign one agent at a time to a three-wheel object, disease propagation was only considered for buses.

8.4 Disease propagation

When an agent gets infected, they will enter the Markov state model and, they will pass through the states shown in figure 8.1. When that agent travels through the environment, it will meet other agents and spread the disease accordingly.

CHAPTER 9
Results and Discussion

9.1 Disease Spread patterns

The simulation started with a population of 170 agents. Out of these, 5% were randomly selected and injected into the simulation as infected, while the remaining agents were designated as susceptible. The transportation model enables agents to move within the environment according to their respective schedules. Figure 9.1 shows the number of agents in different states (susceptible, exposed, infected, recovered, and others) across the simulation over each of the 50 days.

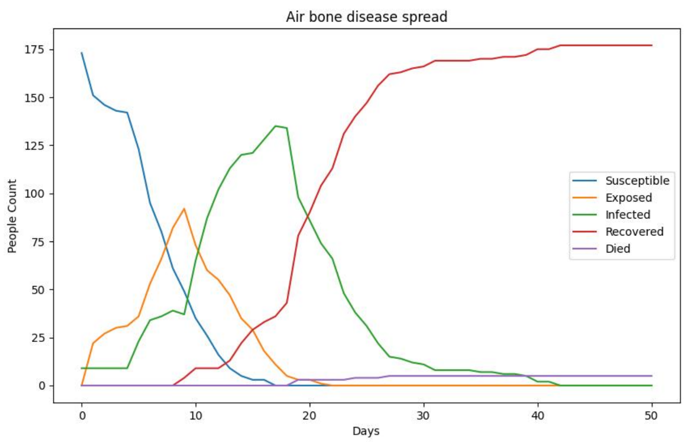


Figure 9.1 : Number of agents in each state over days.

The results obtained using the ABM closely match with the theatrical results obtained by analytical methods using differential equations [22]. This demonstrates the accuracy of the designed ABM when simulating disease propagation.

9.2 Effect of the Transport network on the Disease Spread patterns.

It was decided to study the disease spread with and without transport network. Figure 9.2 shows the number of exposed and infected agents with and without including public transport network in disease propagation.

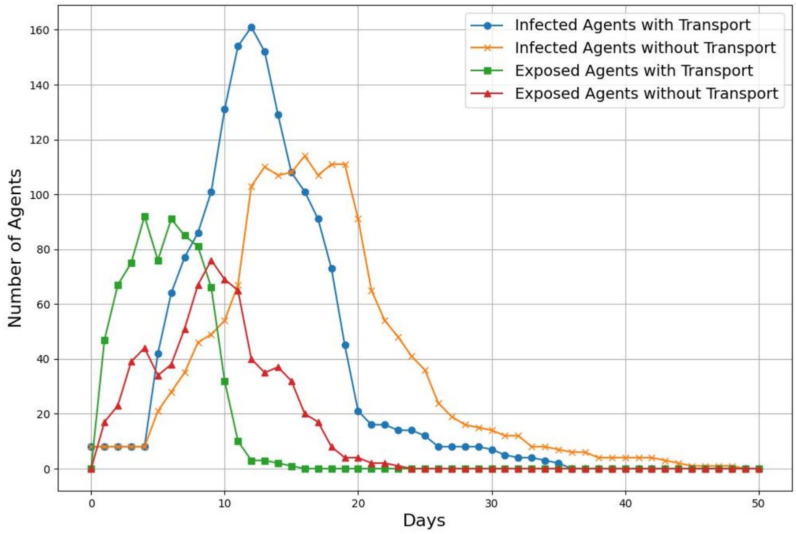


Figure 9.2: Variation of exposed and Infected number of agents with and without transport network

The figure 9.2 indicates that disease exposure and infection occur much earlier when transportation is included. This highlights the significant impact of public transportation on the rate of disease exposure and transmission.

9.3 Discussion

From above results, it is evident that model performs well and the results align with real world intuition. The disease model is flexible, which can be allowed to model wide range of diseases in any given environment. The information that can be gathered from this simulation can be very useful in making policies by government. For an example, the zones what are needed to be locked down can be easily identified beforehand.

CHAPTER 10
Conclusion

ABMs are a very powerful tools that can model complex environments which cannot be analyzed by traditional analytical methods. They are used over a wide variety of applications. In this project, we have proposed an advanced ABM framework that integrates realistic environmental structures and public transportation models. This approach enhances the model's accuracy in simulating airborne disease transmission by reflecting real-world complexities more effectively.

Our ABM model is flexible so that it can simulate different environments, transportation systems. It is also equipped with different agent classes which significantly improves the accuracy. This can not only be used in modeling disease propagation, but also in traffic control, economics and sociological studies. Over all this is a very powerful tool which can be very desirable for policy makers.

REFERENCES

[1] T. Teo, Ed., *Handbook of Quantitative Methods for Educational Research*. Rotterdam: SensePublishers, 2013. doi: 10.1007/978-94-6209-404-8.

[2] de V. Benny, 'Using Simulation To Study School Effectiveness.', Sep. 1995.

[3] R. L. Axtell and J. L. Epstein, 'AGENT-BASED MODELING: UNDERSTANDING OUR CREATIONS'.

[4] M. Abdou, L. Hamill, and N. Gilbert, 'Designing and Building an Agent-Based Model', 2012, pp. 141–165. doi: 10.1007/978-90-481-8927-4_8.

[5] A. Apolloni, V. S. A. Kumar, M. V. Marathe, and S. Swarup, 'Computational Epidemiology in a Connected World', *Computer*, vol. 42, no. 12, pp. 83–86, Dec. 2009, doi: 10.1109/MC.2009.386.

[6] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. V. Marathe, 'EpiSimdemics: An efficient algorithm for simulating the spread of infectious disease over large realistic social networks', in *SC '08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, Nov. 2008, pp. 1–12. doi: 10.1109/SC.2008.5214892.

[7] E. Frias-Martinez, G. Williamson, and V. Frias-Martinez, 'An Agent-Based Model of Epidemic Spread Using Human Mobility and Social Network Information', in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, Oct. 2011, pp. 57–64. doi: 10.1109/PASSAT/SocialCom.2011.142.

[8] M. Tracy, M. Cerdá, and K. M. Keyes, 'Agent-Based Modeling in Public Health: Current Applications and Future Directions', *Annu Rev Public Health*, vol. 39, pp. 77–94, Apr. 2018, doi: 10.1146/annurev-publhealth-040617-014317.

[9] '(PDF) Exploring the Applications of Agent-Based Modeling in Transportation'. Accessed: Mar. 04, 2024. [Online]. Available: https://www.researchgate.net/publication/373533262_Exploring_the_Applications_of_Agent-Based_Modeling_in_Transportation

[10] C. W. Weimer, J. O. Miller, and R. R. Hill, 'Agent-based modeling: An introduction and primer', in *2016 Winter Simulation Conference (WSC)*, Dec. 2016, pp. 65–79. doi: 10.1109/WSC.2016.7822080.

[11] H. Weligampola *et al.*, 'Pandemic Simulator: An Agent-Based Framework with Human Behavior Modeling for Pandemic-Impact Assessment to Build Sustainable Communities', *Sustainability*, vol. 15, no. 14, Art. no. 14, Jan. 2023, doi: 10.3390/su151411120.

[12] A. Ng, M. Jordan, and Y. Weiss, 'On Spectral Clustering: Analysis and an algorithm', in *Advances in Neural Information Processing Systems*, MIT Press, 2001. Accessed: Mar. 04, 2024. [Online]. Available: https://papers.nips.cc/paper_files/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html

[13] D. Y. L. Ranasinghe *et al.*, 'Transmittance Multispectral Imaging for Reheated Coconut Oil Differentiation', *IEEE Access*, vol. 10, pp. 12530–12547, 2022, doi: 10.1109/ACCESS.2022.3144841.

[14] J. F. Lutzeyer and A. T. Walden, 'Comparing Graph Spectra of Adjacency and Laplacian Matrices', vol. 882, 2020, pp. 191–202. doi: 10.1007/978-3-030-36683-4_16.

[15] R. Hinch *et al.*, 'OpenABM-Covid19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing', *PLoS Comput Biol*, vol. 17, no. 7, p. e1009146, Jul. 2021, doi: 10.1371/journal.pcbi.1009146.

[16] A. Trivedi, N. Sreenivas, and S. Rao, 'Modeling the Spread and Control of COVID-19', *Systems*, vol. 9, p. 53, Jul. 2021, doi: 10.3390/systems9030053.

[17] S. Side, U. Mulbar, S. Sidjara, and W. Sanusi, 'A SEIR model for transmission of tuberculosis', presented at the THE 4TH INTERNATIONAL CONFERENCE ON MATHEMATICAL SCIENCES: Mathematical Sciences: Championing the Way in a

Problem Based and Data Driven Society, Putrajaya, Malaysia, 2017, p. 020004. doi: 10.1063/1.4980867.
[18] M. H. A. Biswas, L. T. Paiva, M. De Pinho, and 1. Faculdade de Engenharia da Universidade do Porto, DEEC and ISR-Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, 'A SEIR model for control of infectious diseases with constraints', *Mathematical Biosciences and Engineering*, vol. 11, no. 4, pp. 761–784, 2014, doi: 10.3934/mbe.2014.11.761.
[19] A. A. Ogholbake and H. Khamfroush, 'A Modified SEIR Model for the Spread of COVID-19 Considering Different Vaccine Types', 2021, *arXiv*. doi: 10.48550/ARXIV.2106.08820.
[20] C. C. Kerr *et al.*, 'Covasim: An agent-based model of COVID-19 dynamics and interventions', *PLoS Comput Biol*, vol. 17, no. 7, p. e1009149, Jul. 2021, doi: 10.1371/journal.pcbi.1009149.
[21] J. Qin, C. You, Q. Lin, T. Hu, S. Yu, and X.-H. Zhou, 'Estimation of incubation period distribution of COVID-19 using disease onset forward time: a novel cross-sectional and forward follow-up study', Mar. 10, 2020. doi: 10.1101/2020.03.06.20032417.
[22] J. M. Carcione, J. E. Santos, C. Bagaini, and J. Ba, 'A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model', *Front. Public Health*, vol. 8, p. 230, May 2020, doi: 10.3389/fpubh.2020.00230.