TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Azer Ramazanli 194229IASM

Disease Spread Modelling

Master's thesis

Supervisor: Aleksei Tepljakov
PhD

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Azer Ramazanli 194229IASM

Haiguse leviku modelleerimine

Magistritöö

Juhendaja: Aleksei Tepljakov
PhD

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Azer Ramazanli

02.08.2021

Abstract

As the world population figures rise and interstate mobility becomes increasingly ubiquitous, infectious disease outbreaks become more and more dangerous, having high transmission rates and causing more impact on our lives. The ongoing COVID-19 pandemic which according to Worldometers.info, has reached more than 196 million confirmed cases and caused more than 4 million deaths as of 29 July 2021 [1], shows how vital it is to understand the spread mechanisms of the viral pathogens and to be able to detect probable reasons and sources triggering higher disease transmission rates.

This thesis work is conducted as part of the project entitled “COVSG22: Monte-Carlo analysis of the spreading rate of a virus as a function of human mobility and social distancing” run jointly by the Department of Cybernetics and the Department of Computer Systems at Tallinn University of Technology. The project aims to simulate a country-level infectious disease outbreak scenario based on the example of the Republic of Estonia during the COVID-19 pandemic. One of the principal goals of the project is to predict the number of infected cases for various spatial scales, the largest one being the country itself, based on a complex set of parameters with the help of simulation and modelling. The successfully developed model is expected to contribute to determination of the optimal prevention methods against the transmission of infectious diseases and the strategies in the control of an outbreak.

The main contribution of this thesis work is the synthetic population generator software which besides building the geospatial environment of the Republic of Estonia, also generates a synthetic population reflecting the overall characteristics of the Estonian society in itself. The biggest advantage of the software is that, it achieves the mentioned features utilizing only open-access aggregated data, without requiring the use of any personal information. Another benefit of the generator which can be noted is that, even though, the outcome of the synthetic population generator is originally intended to be used in the country-level scale-free network-based simulations of different disease spread scenarios, the capabilities of the software provide possibility to employ it also for other epidemic models like agent-based and Monte Carlo simulations, or even in a larger context, for simulations dealing with other types of spread like information spread in the future.

This thesis is written in English and is 57 pages long, including 6 chapters, 16 figures and 10 tables.

List of abbreviations and terms

ABM	Agent Based Modelling
ADS	Address Data System
ARIMA-WBF Model	Autoregressive Integrated Moving Average – Wavelet-based Forecasting Model
BMA	Bayesian Model Averaging
CFR	Case Fatality Rate
COVID-19	Coronavirus Disease of 2019
CSV	Comma-separated Values
EHAK	Estonian Administrative and Settlement Division
GDP	Gross Domestic Product
GIS	Geographic Information System
GPS	Global Positioning System
IPF	Iterative Proportional Fitting
LAMBERT-EST	Lambert Conformal Conic Projection – Estonia
LEST-97	Lambert Estonia Orthogonal Coordinate System of 1997
LSTM	Long Short-term Memory
MC	Monte Carlo
ML	Machine Learning
OSM	OpenStreetMap
OSMnx	OpenStreetMap - NetworkX
SI Model	Susceptible – Infected Model
SIPHERD Model	Susceptible – Infected – Purely asymptomatic – Hospitalized – Recovered – Deceased Model
SIR Model	Susceptible – Infected – Removed Model
SIRSi	Susceptible – Infected – Removed – Sick Model
SIS Model	Susceptible – Infected – Susceptible Model
WGS	World Geodetic System

Table of contents

List of figures

List of tables

1 Introduction

Even though the interest in disease spread modelling has exponentially increased starting from the late 2019, after the novel coronavirus caused a global pandemic by quickly spreading to all over the world, it has actually always been a very important research topic for the last few centuries of human history. The concept of using modelling to explore disease spread mechanisms dates back to 18th century, when famous Swiss mathematician and physicist Daniel Bernoulli formulated and solved a mathematical model in order to assess the effect of variolation against smallpox disease [2]. Starting from the early 20th century, modelling has become an established tool in epidemiology with the emergence of several deterministic compartmental models [2].

Due to having a complex and chaotic structure, modelling of an infectious disease spread is a very challenging task including a composite set of parameters, indicators, and variables. Therefore, nowadays a diverse number of methodologies are adapted with this purpose, including traditional compartmental models, agent-based models, scale-free network-based models, ML-based models etc. Each of the mentioned methodologies has its own advantages and disadvantages which are discussed comprehensively in Section 2, but, the COVSG22 project, which this thesis work stems from, utilizes scale-free networks in order to conduct Monte-Carlo simulations for disease outbreak. Scale-free networks are the group of networks whose degree distribution follows a power law. This choice can be justified with the fact that, many real-world networks, especially the ones which are intended to be modeled in the context of epidemiology are large-scale complex networks and studies shows that, in most cases the degrees of the nodes in large-scale networks are distributed with power-law [3].

The study reviewing the fundamentals of epidemiological theory and network theory [4], emphasizes the superiority of networks constructed on real-world data over the artificially generated networks when the application of the networks to the epidemic modelling are considered. However, the authors disregard the possibility of modelling entire population network and the social interactions among them with high accuracy, based on real-world data, stating that producing this kind of data is “impractically time-consuming task” and also determining the contacts among the nodes of the network requires gathering personal information from all of the network. Instead, three alternative methods, namely, infection tracing, contact tracing and diary-based study are proposed to be used for creating “real” networks. The proposed methodologies still do require personal information gathering, but from a subgroup of the network, instead of the entire population.

The implementation of the synthetic population generator software, which is the principal focus point of this thesis work, aims to prove that it is possible to model the population in such a way that, most of the major repetitive spatial and social interaction patterns emerging in the society are covered in the model, and more importantly it can be built based on aggregated statistical data without using any personal information. The scale-free network-based model currently used in the COVSG22 project has only one edge type in the network defining all kind of connections among the individuals, and lacks the usage of the real-world data at the moment. The dataset produced as the outcome of the software will be used in converting the current network to a weighted network and diversifying the edge types, where the weight for each connection type will be determined as the result of Monte Carlo simulations. The reason for the new network to be a weighted network is that, even though the interaction patterns reflected in the output of the generator such as being member of the same household, attending the same school, working in the same enterprise, or being neighbors living in the same building all can be classified as a social connection, each of them has different level of influence on the transmission rate of a disease.

Another strong argument justifying the usage of the population generator is provided in [5], where the importance of having community structure in the real networks used for modelling a disease spread is underlined. Communities are the phenomenon emerging in the network when a group of nodes have dense connections among each other, while the connections with other communities are sparser. The result dataset of the synthetic population generator implemented in this thesis work can help to exhibit community structure in the network based on real-world communities existing in the society. For example, the individuals living in the same household compose a community, while a larger scale community structure can be observed based on which settlement the individuals reside.

The ultimate goal of this thesis work is to supply the COVSG22 project with a dataset reflecting all the aforementioned features in itself and this objective is achieved by performing the following tasks:

1. Forming the set of parameters, affecting the behaviour of the infectious disease outbreak and infection transmission rates, and classifying them by the level of importance.
2. Getting acquainted with the epidemic modelling theory, terms, disease spread models built with different methodologies, datasets used in those models and how the datasets are constructed.
3. Determining the list of input data needed for implementing the software which will produce the dataset.
4. Exploring the open-access data sources like geospatial databases and electronic portals of the administrative registers, acquiring the input data and pre-processing it.
5. Implementing the synthetic population generator software producing the dataset by following the steps given below.

    - Constructing the spatial hierarchy of the country.
    - Integrating the vector data of some geospatial entities to the system, whose importance in terms of disease spread modelling is observed based on the literature review.
    - Generating the individuals and distributing them to the households based on the aggregated statistical data.
    - Assigning the activities containing repetitive behaviour like residence, work, and education to the generated individuals and allocating location coordinates for those activities.

The thesis is organized as follows. Section 2 starts by determining the key factors affecting the behaviour of viral pathogen spreads based on the literature review, then presents the existing classification methods for disease spread models, and examines models built by different methodologies under two main categories, namely, temporal and spatiotemporal models, from the perspective of the characteristics of the dataset used in the implementation process. Finally, the synthetic population generation is discussed in Section 2 in terms of generic implementation methodologies, common problems, and solution methods offered in the literature. Section 3 provides the steps followed for building the generator software and the methodologies utilized in each phase. In Section 4, after the reader is introduced to the list of data sources and software dependencies, the source code of the software executing the steps discussed in the previous section is explained by using pseudo-codes, selected source code parts and flowchart diagrams. Section 5 besides presenting the final results of the software, also demonstrates the outcomes of the intermediate stages. Finally, in Section 6, the conclusion remarks are given and further improvement possibilities and adaptation perspectives are discussed.

2 Literature review

In this section the reader is provided with a comprehensive literature review conducted in epidemic modelling and synthetic population generation spheres. The literature review is organized in the following manner. Section 2.1 provides the list of the foremost factors and parameters in epidemic modelling under three main categories: disease transmission rate related parameters, spatial factors, and socio-demographic factors. In Section 2.2, firstly, the reader is introduced to the common concepts, the terminology, and the established classification methods in epidemic modelling, and then, different models existing in literature are analysed in two subsections, namely temporal (Section 2.2.1) and spatiotemporal (Section 2.2.2) models, depending on whether the spatial aspects of the disease spread is taken into consideration or not. Section 2.3 presents the theory about conventional implementation steps followed in synthetic population generation field, common problems faced, and the solution methods offered in different studies.

2.1 Key factors and parameters in disease spread modelling

Infectious disease spread has a very compound group of factors affecting its behaviour, and determining the principal factors correctly is arguably the most crucial task to consider while modelling the disease dynamics.

The first group of parameters which has to be examined is disease transmission rate related parameters whose values are disease-specific and depend on the type of the infectious disease intended to be modelled. Basic reproduction number – R_0, which represents the average number of secondary infections caused by an infectious individual, is determined as the essential disease transmission parameter in several studies [6]–[8]. The basic reproduction number is often perceived as a threshold variable deciding whether the spread will proceed to a serious state or not. In most deterministic models, an infectious disease survives enough to advance to a pandemic state if and only if the value of the basic reproduction number exceeds one, and otherwise if the value is smaller than one, then the number of infected cases gradually decreases and eventually becomes extinct [6]. Serial interval (time between symptom onsets in an infector-infectee pair), generation interval (time between infection events in an infector-infectee pair) and incubation period (time between moment of infection and symptom onset) are also often listed between key parameters [7], [8].

Another aspect to inspect, while modelling viral pathogen dynamics is the spatial dimension of the spread. A research exploring the spatial factors of COVID-19 in New York City through ordinary least squares regression and geographically weighted regression, indicates correlation between population density, medical density, green space density, mean distance travelled, gender distribution, commuting (walking, carpooling, and public transit) and the rate of COVID-19 positive cases [9]. The protective effect (negative correlation with the outcome) of the distance from the capital in municipalities is also noted alongside the correlation between the demographic density and COVID-19 positive cases in a study investigating spatial and demographic factors affecting vulnerability to COVID-19 in the State of São Paulo, Brazil [10]. Closed environment social interaction hotspots also cannot be disregarded while talking about key spatial factors affecting disease spread rates. A study conducted in Japan, shows that, closed environment facilities contribute to secondary transmission of COVID-19 and promote superspreading events [11].

Socio-demographic factors also have a weighty influence on the characteristics of a disease outbreak such as the spread and fatality rates. Several researches indicate the excessive proportion of the aging population in society as an important factor for high CFRs [12]–[14], while some others analyse the gender imbalance in CFRs and reveal higher case fatality among men [15], [16]. Underdeveloped healthcare system is also among the socio-demographic reasons which can lead to a higher spread and case fatality [17], [18]. Another study exploring the socio-economic determinants of the coronavirus pandemic, divides them into four disjoint groups based on the results of the BMA estimation [19]. As a result, three determinants are found to have a strong evidence indicating significant impact on the number of the coronavirus cases, which are the overweight prevalence in the country, the population density and the number of international tourist arrivals. The fraction of elderly population in the country is classified as the determinant with medium evidence, whereas only weak evidence is found for the average household size and the fraction of young population in the country. The authors conclude that all other potential determinants considered to possibly give meaningful explanation for difference between the number of coronavirus cases in different countries are found to have negligible evidence. Higher socio-economic status and per capita GDP are observed to be positively correlated with the number of reported incident cases in the early phase of the epidemic [20]–[22]. This phenomenon is explained with numerous reasons including more social interactions as a result of more economic activity [20], widespread testing, greater transparency with reporting and better national surveillance systems in the countries with a higher per capita GDP [22]. However, with the adaptation of social distancing and public health measures, the correlation is observed to either become insignificant [20], or invert over time resulting in lower growth rate in countries with better SES values at the end of the observation period [21].

2.2 Classification and analysis of disease spread models

Numerous infectious disease epidemiology models representing spread dynamics have been designed since the beginning of the 20th century due to its crucial importance and different classification methods have emerged for categorizing them. For example, one of the prevalent views is to categorize those models into two groups, namely deterministic and probabilistic models. Deterministic models are those which don't contain any randomness in the implementation. In the other words, whenever a deterministic model is executed with the same data and initial conditions it will produce the same outcomes. A probabilistic model, on the other hand, benefits from the usage of randomness elements and dynamic parameters and consequently, can offer different possible results in each run [23].

Epidemic models can also be categorized under six main groups by the methodology used in the implementation, based on the literature review done:

- Conventional compartmental models.
- Models using Monte Carlo methods.
- Scale-free network-based models.
- Agent-based models.
- ML-based models.
- Hybrid models.

Compartmental models are general technique used in epidemiology for the mathematical modelling of disease transmission, where the population is divided into different compartments with the possibility of the transition between the compartments. Generally, the model is formulated as a system of differential equations, by expressing the transfer rates between the subgroups as derivatives of the sizes of the compartments with respect to the independent variable of time t [24].

Monte Carlo methods-based models maintain the idea of classifying the individuals based on the health status in most cases. But, unlike the compartmental models, the population is not divided into disjoint groups; instead, addressed as a set of interacting particles, each one representing an individual and having a state characterizing its current health status. Thus, the system dynamics are modelled with the help of various MC algorithms which rely on repeated random sampling instead of the differential equations. In this context, the epidemic models based on the MC methods can be evaluated as the stochastic approach using randomness to enhance the conventional compartmental models which are deterministic in principle.

In a similar manner, scale-free network-based models also preserve the concept of attaching a state associated with the health status to each individual separately. The difference with other methods is that, this group of models handles the population dynamics in the context of networks. The networks constructed in epidemic models for representing the relations between individuals in a population are composed of numerous nodes – individuals, and edges – the connections between the individuals. The set of nodes an individual is connected to is called its neighbourhood and the size of this neighbourhood is the individual's degree in the network [4]. A study investigating the behaviour of complex networks [3], reports that the degree distribution of large-scale networks follows a power law. In other words, the probability *P(k)* that a vertex in the network interacts with *k* other vertices decays in the form of

$$\mathrm{P}(k) \sim k^{-\gamma} \tag{1}$$

where γ is the scale parameter whose value is determined by the given network. The networks demonstrating this behaviour are called "scale-free" networks. Since, the networks used for modelling the disease spread are large-scale complex networks, they are also classified as scale-free networks.

Another established methodology in epidemic modelling is the use of agent-based simulations. In agent-based modelling, the population is modelled with the help of autonomous agents representing the individuals, which demonstrate repetitive interactions among themselves. ABM is especially useful for modelling real-world systems, while the goal is to gain valuable information about the dynamics of the system it emulates. It can be explained with the fact that, real-world systems are composed of behavioural entities and ABM is the most natural method for this kind of tasks, since it is capable of capturing complex phenomena resulting from the interaction of individual entities [25]. In [26], the authors state that ABM has been used in modelling of a wide variety of diseases including COVID-19, malaria, smallpox, tuberculosis, avian influenza, etc., and its effectiveness has been proven.

Unlike the aforementioned methodologies, ML-based disease spread models are very diverse in terms of the purpose of ML adaptation, the used algorithm, and the expected outcome, which is why, they cannot be described with one generic implementation methodology. Hence, the comparative analysis of these models is not possible and they must be investigated separately.

Lastly, there exist hybrid models utilizing more than one of the mentioned methodologies together. By doing so, these models aim to exploit the strong sides of each approach that collectively compose the hybrid model.

Despite having some other more common categorization systems as the ones mentioned above, the following subchapters analyse disease spread models under two main categories which are temporal models and spatiotemporal models (Section 2.2.1 and 2.2.2, respectively) from the perspective of the main focus point of the thesis work. Temporal models handle spread dynamics in the time domain only by ignoring the spatial data correlation, while in the spatiotemporal models, data obtained both on time and space domains are utilized for constructing the model, which has at least one temporal and one spatial attribute. In this respect, the spatiotemporal models provide additional advantages for discovering and observing pattern persistence in the data over all space-time domain. On the other hand, the main challenge characteristic to all spatiotemporal models is that considering both the temporal and spatial data while designing the model adds additional complexity to the data analysis process [27].

Among all the model groups classified by the used methodology, the compartmental models are the only class of models, which are homogeneously temporal. In other words, none of the conventional compartmental models benefits from the usage of the spatiotemporal datasets, and therefore, the example representatives of this group are analysed under Section 2.2.1 only. Opposingly, considering the fact that one of the two core components of an agent-based model is the environment, which cannot be constructed without handling the spatial aspects, all agent-based epidemic models are spatiotemporal in principle. Thus, all of the agent-based models selected for analysis are discussed under Section 2.2.2, solely. All other model groups classified by the used methodology have both temporal and spatiotemporal examples examined under Sections 2.2.1 and 2.2.2, respectively.

2.2.1 Temporal models

The common feature of all models discussed under this subchapter is that, the data used for constructing the model or the outcome of the model is time series data, which don’t include any spatial characteristic. Compartmental models are the most straightforward temporal approach to model a disease spread based just on a very simple dataset composed of few disease transmission rate-related parameters. Despite the compartmental models not being the most potent methodology for describing the internal dynamics of the epidemy, their analysis is especially important due to fact that, most of the more established models are built on the compartmental models, adapting the idea of health status categorization. Thus, the well understanding of the idea behind this classification and being familiar with the recent categorization methodologies associated with the current outbreaks is a must in the context of disease spread modelling. Three of the well-known examples of the compartmental models, including the SIR model, creating the basis for more comprehensive representatives of this class of models are described below.

One of the first and simplest models in disease spread modelling is the Ross Epidemic (SI) model, developed in 1911. In this model, the entire population is divided into two mutually exclusive subgroups labelled as S and I, where S stands for susceptible and I stands for infected, and mixed homogeneously making sure that each individual has an equal chance of infection. Transition rate from S subgroup to I subgroup is then calculated as proportional to β – the average subgroup contact constant and to the number of individuals in S subgroup at any given instance of time [28].

SIS (susceptible – infected – susceptible) epidemic model is another mathematical model built on SI model, with only difference being that, in this method, the number of the infected population can also decrease as some infected individuals change status to the susceptible. Therefore, another average subgroup contact constant – σ is also integrated to the model for the transitions from I compartment to S compartment [28].

SIR (susceptible – infected – removed) model developed by Kermack & McKendrik in 1927 plays a role of the premise for nearly all compartmental models developed until modern times. The SIR model extends the SI model with the possibility that some members of the infected subgroup at any given instance of time are moved to R subgroup by recovery or by death and are no longer considered as the source of further infection [29]. In the simplest form, the dynamics of the system developed by using the SIR model can be described by following system of ordinary differential equations:

$$\frac{dS}{dt} = -\frac{\beta IS}{N}, \qquad S(0) = S_0 \geq 0, \tag{2}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I, \qquad I(0) = I_0 \geq 0, \tag{3}$$

$$\frac{dR}{dt} = \gamma I, \qquad R(0) = R_0 \geq 0, \tag{4}$$

where $S(t) + I(t) + R(t) = N$ and β and γ stands for the infection and the removal rate constants, respectively [30].

As the amount of research on this specific topic have increased due to COVID-19 pandemic, some new compartmental models considering extensions to the simple SIR model have emerged in recent times which can be referred to as the state-of-art models examining temporal dimension of the infectious disease outbreak. Some of those models are analysed below.

SIRSi (susceptible – infected – removed – sick) model proposed in [31] integrates three major additions to the SIR model. First of all, considering the possibility of the unreported or asymptomatic cases, the infected subgroup in traditional SIR model is divided into infected people who shows the symptoms and those who do not. The subgroup I in this model, represents the infected population in the incubation stage before the onset of the symptoms and those individuals who are tested positive are moved to the Si compartment. Another aspect taken into account is that the acquired immunity among the individuals who have recovered from the infection is not always permanent, and at least a group of the individuals in R compartment can be prone to getting infected again after some time has passed. In order to address this issue, the model contains a feedback loop from the R compartment to the S compartment. Additionally, the birth and non-disease-related deaths of the individuals in the population are also represented in the model for better reflection of the real-life scenarios. Figure 1 describes the overall scheme of the SIRSi model.

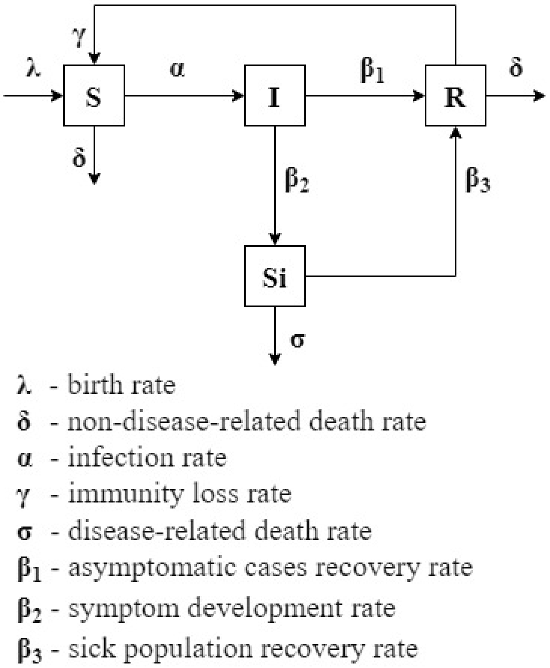

γ
λ
S
α
I
β1
R
δ
δ
β2
β3
Si
σ
λ - birth rate
δ - non-disease-related death rate
α - infection rate
γ - immunity loss rate
σ - disease-related death rate
β1 - asymptomatic cases recovery rate
β2 - symptom development rate
β3 - sick population recovery rate

Figure 1. The SIRSi model scheme.

Another compartmental model implemented recently, is called the SIPHERD model which proposes to divide the population into seven mutually exclusive subgroups, namely, S – susceptible, E – exposed, I – symptomatic, P – purely asymptomatic, H – hospitalized or quarantined, R – recovered, D – deceased. Transitions of the individuals between these compartments are formulated with the following system of differential equations [32]:

$$\frac{dS}{dt} = -S(\alpha E + \beta I + \gamma P + \delta H) \quad (5)$$

$$\frac{dE}{dt} = S(\alpha E + \beta I + \gamma P + \delta H) - (\mu + \xi_I + \xi_P)E \quad (6)$$

$$\frac{dH}{dt} = \mu(E+P) + \nu I - \sigma H(t - t_R) - \tau H(t - t_D) \quad (7)$$

$$\frac{dI}{dt} = \xi_I E - (\nu + \omega)I \quad (8)$$

$$\frac{dP}{dt} = \xi_P E - (\mu + \eta)P \quad (9)$$

$$\frac{dR}{dt} = \omega I + \eta P + \sigma H(t - t_R) \quad (10)$$

$$\frac{dD}{dt} = \tau H(t - t_D) \quad (11)$$

where the detection of the asymptomatic and symptomatic cases is calculated dependent on the number of tests done per day (T_{PD}) with the formulas expressed in Equation (12) and Equation (13).

$$\nu = \nu_0 + \nu_1 T_{PD} \quad (12)$$

$$\mu = \mu_0 + \mu_1 T_{PD} \quad (13)$$

Definitions for the parameters and factors used in the system equations of the SIPHERD model are provided in Table 1.

Table 1. Definitions for the parameters used in the SIPHERD model.

Parameter	Definition	Parameter	Definition
α	Transfer rate, E to S	μ_0	Detection prob. of E
β	Transfer rate, I to S	μ_1	Detection prob. coefficient
γ	Transfer rate, P to S	ν_0	Detection prob. of I
δ	Transfer rate, H to S	ν_1	Detection prob. coefficient
ξ_I	Conversion rate, E to I	η	Home recovery rate of asym. cases
ξ_P	Conversion rate, E to P	ω	Home recovery rate of sym. cases
σ	Recovery rate	t_R	Recovery delay
τ	Death rate	t_D	Mortality delay

The SIPHERD model is especially notable due to fact that, dividing I subgroup in traditional compartmental models into E – undetected exposed, I – symptomatic and P – purely asymptomatic categories is not only very worthy in terms of maximizing the model accuracy, considering the latest research developments emphasizing the proportion of asymptomatic and undetected cases in the coronavirus outbreak [33], but also can be interpreted as novelty in the area. The authors also analyse some social distancing and lockdown scenarios generated for testing and validating the model in the same paper [32], however, since the used method is not a computational simulation, but a mathematical model, the effects of the scenarios are not obtained as a result of some learning process by an algorithm or derived depending on other parameters, but rather given predefined as 5% decrease in the transmission rates due to the lockdown.

A similar approach is followed by [34], where several important characteristics related to the spread dynamics of the COVID-19 disease, such as the possibility of the undetected infectious individuals and the effect of the control measures are also considered in the implementation of the model named θ-SEIHRD. Additionally, the model introduces a novel perspective which takes the fraction θ of detected cases over the total infected cases into consideration and contributes to further exploration of the importance of this ratio on the spread of the disease.

Although the models mentioned as the state-of-art compartmental models which were developed recently, have shown quite significant improvements on the mathematical modelling of the infection spread, still there are some limitations common for nearly all proposed mathematical methodologies. These constraints are explained in [35] with an example of a disease spreading by person-to-person contact in the context of a large country with huge number of discrete geographical regions. Assuming that the time needed for movements from one geographical region to another one is smaller than the serial interval of the disease, it can be claimed that the propagation of the disease takes place only at the destination location during these movements. This scenario is equivalent to a directed graph, where the nodes represent discrete geographical regions and the edges represent the links between these regions. Using differential equations for describing the transmission rates in models intended for this kind of scenarios would not be the most appropriate method due to having very high dimensionality. In particular, a model for n regions and p different compartments of individuals can have up to pn^2 equations. Thus, the compartmental models which are not effective when the spatial dimension is taken into consideration, most of the time completely disregard it and focus on temporal parameters only.

Another major obstacle of the compartmental models is directly associated with one of its core concepts. The problem is that, dividing the population into disjoint groups and formulating the transmission of the disease with a differential equation is equivalent to assuming that the probability of any individual to be infected by any other individual is equal for the entire population, which is obviously not a realistic approach considering the complex structure of human societies. Several temporal models [36]–[38] try to overcome this problem by representing the society with scale-free networks, in which an infected node can only propagate the disease to an uninfected node, if and only if, they have a direct connection in the network. The difference between these models is in the approach selected for formulating the probability of disease propagation at each time step between a neighbour infector – infectee pair. In [36], an infected node can propagate the infection along each of its connections independently with the predefined probability *p* at each time step, while the model presented in [37] formulates this probability based on the number of days passed after the infector node got infected. The transmission of the disease between a connected infector – infectee pair can happen starting from 2 days after the node got infected till 14 days after the infection took place. The probability of the transmission reaches its maximum value between 4th and 6th days. Unlike the aforementioned two models, the propagation probability is handled from the perspective of the infectee, rather than the infector in [38]. This model evaluates the probability of a node to get infected based on the number of its already infected neighbours and uses the variable $\beta_S(k)$ to represent the transmission rate for a susceptible node with degree greater than or equal to *k*, if it is connected to *k* already infected nodes.

A Monte Carlo simulation model similar to a stochastic point process model, proposed in [39] can also be classified as a temporal model, due to fact that the spread dynamics of COVID-19 epidemic are simulated based essentially on just two disease-specific parameters: average reproduction rate as a function of time and average serial interval. The actual reproduction number is calculated by using Poisson distribution with mean R_t and in order to balance the indeterministic characteristics of MCS, the simulation is run thousand times and the median output values are accepted as the most realistic values. The authors claim that the simulation outcomes obtained by testing the model with the real COVID-19 data collected in Australia and United Kingdom, proved to be consistent with the real-life results.

Another temporal model developed in recent times with the aim of forecasting the number of coronavirus cases, the mortality rate and the recovery rate over time uses a type of the recurrent neural networks, more precisely LSTM networks for this purpose. Even though, the model scores a very high accuracy rate of 93.4% for short term predictions and 92.67% for long term predictions, the authors mention that spatial movements between provinces create problems for modelling the pandemic based on time-series dataset of confirmed case numbers [40].

Regression techniques are also among the most extensively used class of ML methods in epidemic modelling, alongside deep learning algorithms. They are especially useful when applied to time-series forecasting problems. For example, hybrid ARIMA-WBF model developed in [41] is used for the real-time forecasting of daily COVID-19 cases, meanwhile, the regression tree algorithm is also adopted for the risk assessment with CFR dataset in the same study.

2.2.2 Spatiotemporal models

Reflecting spatiotemporal characteristics of almost any system in the model built for simulating its internal dynamics is quite a challenging task. Especially when the system intended to be simulated is a large-scale real-world human society, which is the case for epidemic models, it requires more time and energy to build a spatiotemporal model than a temporal model. The most important reason is that, a human society is a very complex and dynamic system. Therefore, a spatiotemporal model must include spatial and socio-demographic factors discussed in Section 2.1 and at the same time consider spatial dynamics of social interactions between the individuals as well as human mobility patterns between discrete geographical entities in order to be able to correctly reflect the characteristics of a large-scale human society during the epidemic. The reason why spatial dynamics of social interactions are emphasized in the previous sentence is that, when the disease spread is the matter of discussion, only social interactions having spatial aspects are of interest. For example, an online meeting between different individuals is also a social interaction, but is irrelevant from the perspective of a disease spread modelling. Despite the mentioned difficulties in the implementation process, on the other hand, this class of models are more probable to produce better results for the discussed scenario, since they have a potential to unveil hidden social and spatiotemporal patterns in the population, unlike the temporal models. All in all, the models discussed under Section 2.2.2 can be seen as more comprehensive alternatives to their temporal counterparts analysed under Section 2.2.1, independent of which methodology is used to build the model.

The simplest way to integrate spatial aspects into an epidemic model is to associate the disease transmission with a distance between the individuals. A model using a MC based algorithm [42], adapts this technique where the displacement in the population is distributed normally between the individuals based on the principles of Brownian motion. The model also considers the community structure in the population by adding households in order to be able to represent household contacts alongside non-household contacts. Another Bayesian MC approach-based probabilistic model builds a spatiotemporal kernel by combining various temporal and region-specific demographic, political features with the aim of giving county-level predictions for reported cases of an epidemic [43].

ML techniques can as well be utilized for creating a spatiotemporal infectious disease spread model. For example, a spatiotemporal model based on a sampling algorithm and Bayesian optimization proposed in [44] aims to detect where the infections occur most frequently, in other words which spatial entities can be counted as hotspots for the disease spread. The mobility model of Bern, Switzerland which includes site locations of probable hotspots like schools, research institutes, workplaces, social places, supermarkets and etc. is built and the visits of the individuals generated based on demographic data in the simulation are recorded to detect the correlation between the hotspots and disease spread. The study also explores how contact tracing, testing and containment measures affect the course of the coronavirus outbreak.

Spatiotemporal city-level simulation for COVID-19 described in [45] uses the agent-based modelling approach. Actual geospatial data of the simulated city including residential areas, business areas, roads, schools, population density etc. are used in order to create the environment for the simulation. For integrating the concept of households into the simulation and to create static residence addresses for the members of those households, the authors have utilized static agents representing houses in the simulation alongside mobile agents representing the individuals. Distribution of the human agents to households is implemented based on the real statistic data and the selection of workplaces and education places for the individuals is done randomly. The role of the public transportation in disease spread is also considered in this study.

A geospatial agent-based simulator presented in [46] makes it possible to create city-level disease spread models. The simulation framework offers numerous features like built-in location graph construction tools, realistic disease transmission algorithm and templates for modelling infectious disease which allows non-programmers to benefit from using the simulator. However, the framework has several drawbacks such as the distribution of the households to houses are done based on the assumption that all houses are of identical structure and each house accommodates two households by default; the number of members for each household is randomly assigned between 1 and 4, using a uniform distribution; people only visit the amenities closest to them and the workplaces are generated randomly on the map.

There are different approaches conducted in several studies for constructing a scale-free network-based model in viral pathogen spread modelling. A study inspecting the role of migration in the spread, builds a scale-free network, where different regions of China during early outbreak of COVID-19 form the nodes of the network, and the edges connecting nodes represent migration between the regions [47]. Meanwhile in [48], where the global spatiotemporal transmission of avian influenza is under investigation, each particular outbreak of the disease is a single node in the network. The links between the nodes representing potential transmission pathways are formed by spatiotemporal proximity based on the assumption that transmission can only take place over a local area.

Small-scale real-world social network built on GPS data collected from town of Haslemere is analysed in [49], with the purpose of measuring the efficiency of control strategies like case isolation and contact tracing in reducing the impact of COVID-19 outbreak. Another interesting aspect about this model, besides being constructed on a real-world GPS data is that, the network implemented in this study is a weighted network. The contacts between individuals are defined on a daily basis as at least one 5-min period in which the distance between the individuals must be within 4 meters. The number of days two individuals have made contact satisfying the mentioned criteria defines the weight of the contact edge between those two persons in the model.

Another compulsive study [50] proposes a hybrid model for describing spatiotemporal spread characteristics of a vector-borne chikungunya outbreak, where the macro-scale dynamics of the spread are described with a scale-free urban network and the micro-scale factors are handled using agent-based model. The network nodes are towns and villages where the host agents live and the edges connecting the nodes reflect transportation channels between those geographical entities. Majority of the host agents don’t have the capability to move between nodes, while a minor group of agents can travel to a neighbouring node in one simulation iteration, but must return back to its home node in the next iteration. Thus, disease transmission between distinct spatial units is made possible through urban network, while the internal spread of the infection inside the node is described with an agent-based modelling approach.

2.3 Synthetic population generation

This subchapter discusses synthetic population generation in terms of generic implementation steps, solution methods offered for common problems faced during the implementation and further employment examples.

There is an established methodology used in generating a synthetic population. Three common steps for the implementation are mentioned in multiple studies [51], [52]:

1. Population synthesis – Generating a synthetic population for a specific geographical region starts with creating individuals and households and assigning those individuals to the households based on real-world aggregate socio-economic data of the region and its child entities. Individual and household level characteristics are also attributed at this step.
2. Activity assignment – Each individual generated at the previous step is assigned a set of activities like residence, work, education etc. based on the collected statistical data.
3. Location assignment – At this step, a group of locations are constructed based on the activity needs assigned previously and the activity habits of the individuals are associated with these locations.

Since generating a synthetic population is a complex process which needs disaggregation of an aggregate statistical data to individual level, there are several obstacles faced in the implementation process. Two of the commonly acknowledged problems are discussed and three solution methods are proposed in [53]. First problem is related to the fact that, usually administrative registers provide one-dimensional distribution data, meanwhile, generating a synthetic population demands multi-dimensional distributions. The proposed solution in [53], namely, iterative proportional fitting method, is utilized in the majority of the analysed studies [51], [52] for composing multi-dimensional distributions by using several one-dimensional distributions in conjunction. For example, [51] uses IPF for sampling joint distribution from three disjoint data distributed by age of householder, household income and household size. The authors of the study suggesting IPF for these kind of situations [53] propose Monte Carlo sampling as an alternate solution for the cases, when generating a more complex set of attributes is intended or the features which have to be generated are not available among the collected data. An iterative semi-stochastic algorithm offered by [54] can also be thought as replacement for IPF for the cases when the application of IPF method is not possible due to lack of individual-level data to create cross-joint with the aggregate household characteristics data. This algorithm is also employed in some other synthetic population generators [55]. Another difficulty met in the implementation process is related to localization of the activities assigned to the individuals. Sometimes, the exact spatial location data for activities like residence, work, education etc. is not available, and the assignment of locations to them must be done based on spatially aggregated data. Under given circumstances, the distribution of data spatially aggregated by zones is done by first creating raster map of the zone and then distributing data to those pixels based on density [53].

3 Methodology

The methodology section is organized as follows. Section 3.1 describes the methods used in creating the spatial hierarchy of the Republic of Estonia. The approaches followed for population synthesis, which is the essential part of the generator software are discussed in Section 3.2. Finally, Section 3.3 reports the techniques utilized for assigning different activities to the generated population.

3.1 Spatial hierarchy

As the initial step of the population generator, the spatial hierarchy of the Republic of Estonia has been constructed based on official Estonian Administrative and Settlement Classification (EHAK) [56] maintained by Statistics Estonia. According to EHAK, the country is composed of fifteen state administrative units – counties (maakonnad) and seventy-nine local governments (omavalitsused) forming those counties. Local governments themselves can also be classified as fifteen cities (linnad) and sixty-four rural municipalities (vallad). Other than the administrative division, there are settlement units like villages (külad), small towns (alevikud), towns (alevid), and cities without municipal status (vallasisesed linnad). Two cities, namely, Tallinn and Kohtla-Järve have several city districts (linnaosad).

All the geographical entities mentioned above are represented in the generator with a hierarchical tree structure, where the Republic of Estonia is the root node and the maximum level for leaf nodes is three. Cities which don’t have city districts are the only group of leaf nodes whose level in the tree is equal to two. The size of the tree is equal to 4801, including the country itself, administrative units and settlement units. The tree structure for the spatial hierarchy is presented in Figure 2.

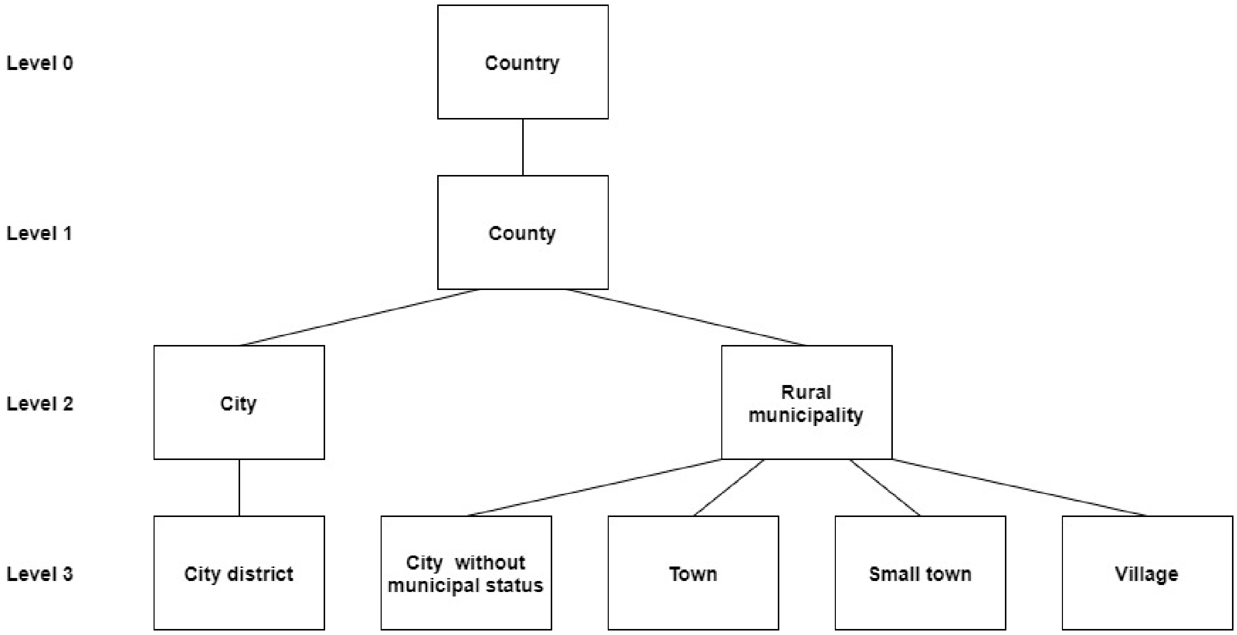
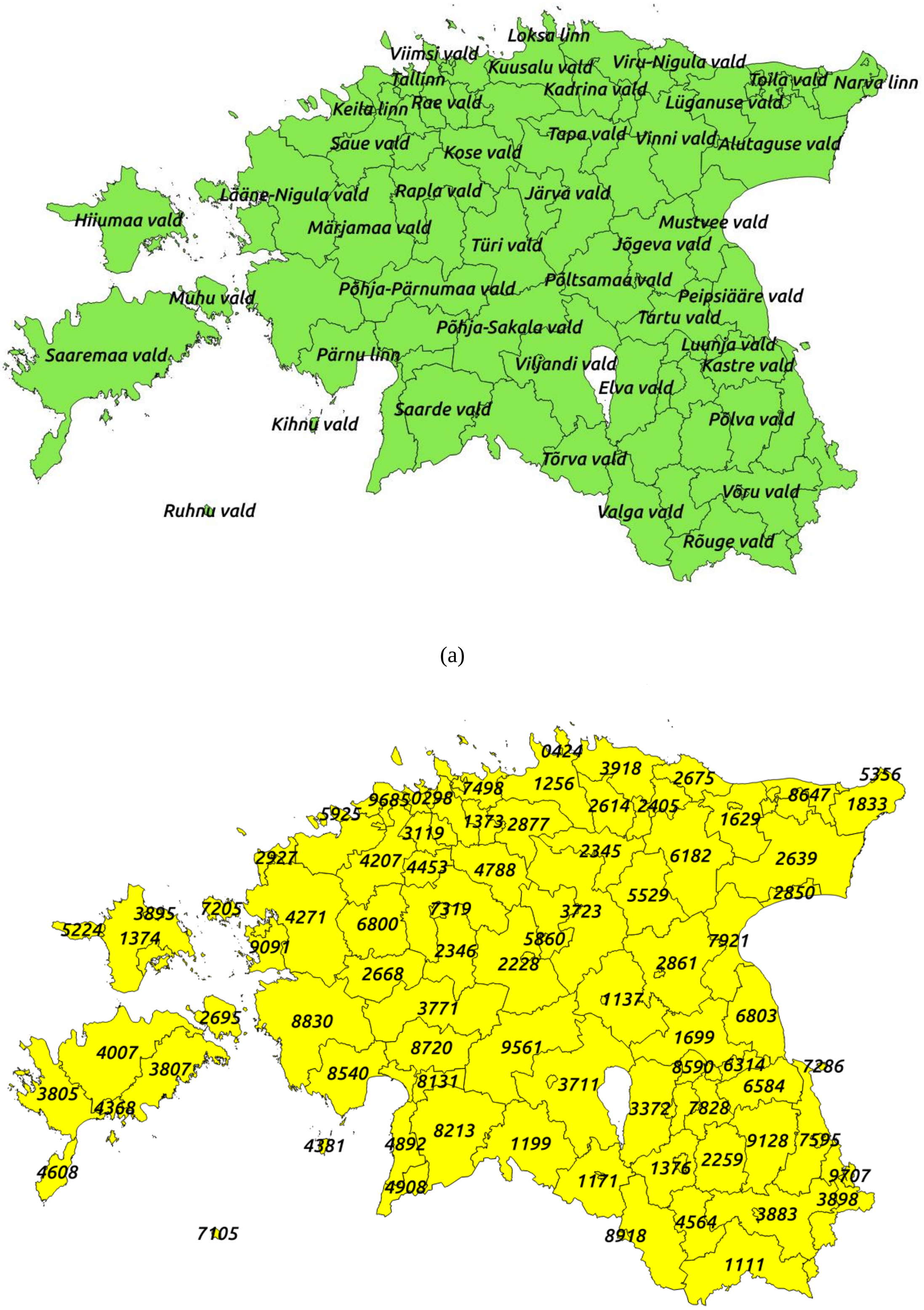


Figure 2. Spatial hierarchy in the population generator.

Since from the programming perspective, the nodes placed at the same level in the tree have very few minor differences, for the sake of simplicity, the phrases country, counties, municipalities and settlements are used in the following sections of this thesis work to refer to the total of all nodes of zeroth, first, second and third level, respectively, in the cases when differentiating nodes at the same level has no particular importance.

After the creation of the tree structure spatial hierarchy, geospatial vector data for all counties, municipalities and settlements have been retrieved from Estonian Land Board Geoportal [57] and integrated to the software (please refer to Figure 9 – Figure 12 for examples).

Some of the data provided for this project by the Geography Department of the University of Tartu are distributed by a new zone system where the country is divided into 113 distinct zones, each having a unique zone code. This new zone system can be thought of as a level two spatial segregation of the country, similar to official cities and rural municipalities division, where some of those cities and municipalities are further split into several zones or some new zones are created with merging territories taken from multiple municipalities. Figure 3 juxtaposes the official cities and rural municipalities division with this new zone division for visualising the differences between them.



(a)

(b)

Figure 3. Level 2 division of the Republic of Estonia: (a) official cities and rural municipalities division, (b) new zone division (Plotted based on data collected from [57] and Geography Department of the University of Tartu).

In order to be able to distribute data available for the zones to next levels of the spatial hierarchy and eventually to micro-scale, it must be determined inside which unique zone, each settlement is located. The Estonian Address Data System (ADS) data retrieved from the Geoportal of Estonian Land Board [58], which contains addresses of cadastral units, buildings and apartments, have been utilized for performing this task. The core usage purposes of the ADS data in this project are related to the distribution of households to the settlements and the assignment of the residence addresses to them, which are covered in Section 3.2.2 and Section 3.3, respectively. The zone code allocation to the settlements are done after those steps chronologically, but discussed here due to fact that, the nature of the task is more related to building the spatial hierarchy of the country. Therefore, the ADS data itself and how the addresses have been linked with the settlements are explained more detailly in the aforementioned sections, and at this point the step following the association of the addresses with the settlements is provided. In order to assign zone codes to the settlements, I have checked if all the addresses associated with each settlement are located in a single zone or not. As a result, it has been observed that there is a unique zone polygon for each settlement containing all of its address points, in other words, each settlement unit is completely contained by a unique zone and consequently, the code of that zone has been appointed to the settlement.

3.2 Population Synthesis

Population is synthesized in three steps. Section 3.2.1 and Section 3.2.2 describe the methodologies utilized in generating the individuals and the households, respectively. Distribution process of the individuals to the households is explained in Section 3.2.3

3.2.1 Generation of the individuals

Individuals have been generated based on two major individual characteristics: age and gender. The highest-level age-gender distribution data available for the Republic of Estonia except for the last population and housing census conducted by Statistics Estonia in 2011 are the municipality-level distribution published by the same organization in [59]. Despite the fact that, the census provides more diverse range of data and those data are generally distributed at a higher spatial level compared to statistics published in recent years (for example, age-gender distribution data are provided at settlement-level in the census), no dataset from the census is used in this project, since the synthetic population generated based on socio-demographic data collected ten years ago cannot be seen as reliable for representing the dynamics of a modern day society. Another important point to mention about datasets employed in this project is that, although Statistics Estonia has already published several socio-demographic datasets for year 2021, in order to preserve the consistency in data usage as much as possible (since year 2020 was the latest date available for some datasets while the generator was implemented), only data reported for year 2020 have been used for building the generator.

The table in [59] provides information about the number of individuals in all municipalities, city districts and cities without municipal status for each age-gender tuple possible. For instance, there were 351 newly born baby boys in Mustamäe city district of Tallinn, and the number of forty-years old females residing in Järve rural municipality was equal to 34 as of 1 January, 2020. Dataset also contains totally 2070 individuals; whose residence address is unknown. So, first of all, 2 × 101 *population_by_gender_age* matrices representing number of individuals needed to be generated, have been created for the country, all counties, municipalities, city districts and cities without municipal status, based on age-distribution data. Adding the identical matrix for towns, small towns and villages is not possible, since the same information is not available for those type of spatial regions. Then, the individuals with unknown residence addresses have been distributed from country-level to county-level and form county-level to municipality-level with a proportional distribution, where weight coefficient for each child region has been calculated as the sum of its *population_by_gender_age* matrix before the distribution.

Individuals are initialized with a bottom-up approach, based on the *population_by_gender_age* matrices of the nodes having the highest level in the spatial tree, for which age-gender distribution data are available, namely, the city districts of Tallinn and Kohtla-Järve, other thirteen cities which don’t have district division, the cities without municipal status and the rural municipalities. After an individual is generated, the values associated with its age-gender pair in the *population_by_gender_age* matrices of the initialization node and all of its ancestor nodes are decremented. For example, if a thirty years old female is generated in Kehra city, the value in the thirty first column of the second row in the matrices of Kehra city, Anija rural municipality, Harju county and Estonia are decremented once. The generation of individuals continue till all the values in the matrices of all regions reach zero. Creating the individuals with the methodology described above has several advantages. Firstly, it guarantees avoiding multiple generation of the same individual represented in aggregated age-distribution data of different spatial levels. Furthermore, the generation process can easily be validated by checking the matrices after it has finished.

Six parameters are assigned to the individuals at the initialization: *index*, *gender*, *age*, *settlement_code*, *municipality_code* and *county_code*. *Index* is an integer in the range of [1, 1328890], used as a unique identifier for the individuals. The values of *settlement_code*, *municipality_code*, and *county_code* parameters depend on the type of the node individuals are initialized at, as explained in Table 2. As it can be seen from the table, the only thing remaining after the generation process of the population, is to assign a settlement for those who have been initialized based on the matrices of the rural municipalities. In other words, the people living in the settlements of a rural municipality, except for the cities without municipal status, must be distributed to the towns, the small towns and the villages of the corresponding rural municipality. Since, the exact numbers for the population living in those settlements are not provided in the statistical data, this distribution process is done after the generation of households, based on the number of households in the settlements.

Table 2. The values of the settlement_code, municipality_code and country_code parameters based on the type of the initialization node.

Type of the initialization node	settlement_code	municipality_code	county_code
City district	EHAK code of the city district.	EHAK code of the parent city.	EHAK code of the ancestor county.
City without districts	EHAK code of the city itself, since it doesn’t have child nodes.	EHAK code of the city.	EHAK code of the parent county.
City without municipal status	EHAK code of the city without municipal status.	EHAK code of the parent rural municipality.	EHAK code of the ancestor county.
Rural municipality	Not assigned at the initialization, since a child settlement for the individual is not selected yet.	EHAK code of the rural municipality.	EHAK code of the parent county.

3.2.2 Generation of the households

The households are initialized in county-level, based on households by county statistics of the year 2020, published by Statistics Estonia in [60]. Two parameters, namely, *index* and *county_code* are assigned to the households at this step, where *index* is an integer in the range of [1, 606000], uniquely identifying the household in the software, and *county_code* represents the EHAK code of the county the household is initialized in. Table 3 shows the number of households initialized for each county.

Table 3. Total number of households by county (Retrieved from [60]).

List of counties	Total number of households
Harju	285500
Hiiu	5000
Ida-Viru	66600
Jõgeva	13100
Järva	14200
Lääne	9800
Lääne-Viru	27900
Põlva	11000
Pärnu	39500
Rapla	15500
Saare	15100
Tartu	70100
Valga	13100
Viljandi	20900
Võru	18700
TOTAL	626000

Further distribution of the households from county-level to municipality-level and form municipality-level to settlement-level is done by using two different methodologies. Firstly, it is trivial that there is a positive correlation between the population size of a region and the number of households in that region. Considering that the number of people residing in each city and rural municipality is already known after the generation of the individuals, the households are disaggregated to the municipalities with a proportional distribution based on the population size of the municipalities. An important point to take into consideration about this distribution is that using total number of individuals in the municipalities as the weight coefficient can lead to a scenario in which, regions possessing relatively larger proportion of children than the average of the country will have more households than adult individuals. Since assuming that all households contain at least one adult individual is a logical common approach in synthetic population generation [53], all regions must have at least as many adults as the number of households. Therefore, the number of adults in the municipalities is used as the weight coefficient for the distribution, instead of total number of individuals.

Since information about the population size in majority of the settlements (towns, small towns and villages) is not present at this step, the same methodology can not be applied for disaggregation of the households from municipality-level to settlement-level. Thus, another parameter having positive correlation with number of households in a region, namely, number of dwellings is used instead of population size to distribute households to the settlements. Key point here is that statistics about the number of buildings in the settlements, or even the number of residential buildings are not enough alone to make correct deductions about the relative proportion of the number of households among the settlements. For example, two settlements having the same number of residential buildings can accommodate completely different number of households, in the case one of them is an urban centre and the other one is a rural settlement. The reason is that, in urban centres the majority of the residential buildings are apartments, while houses are the dominant residential building type in rural places. Despite, the difference in accommodation capacity, both an apartment and a house are reflected as a single residential building in aggregated statistical data. Therefore, the only justified accommodational indicator, which can be used to determine the proportions in household distribution is the data about the number of dwellings in the settlements.

The number of dwellings for each settlement is determined after applying a series of operations on Estonian Address Data System (ADS) data maintained by Estonian Land Board [58]. The data provided in CSV format contain address information about more than two million spatial objects including residential buildings, non-residential buildings, cadastral units, traffic units, as well as, residential and non-residential premises, which are parts of the buildings having a separate address. So, firstly, the data have been filtered by object type, leaving only rows classified as residential buildings and residential premises. Each residential premise in the data is associated with a residential building, in a way that *HOONE_OID* column of a residential premise shows *ADS_OID* value of the building it is located in. As the next step, residential premise rows also have been cleared from the dataset after assigning *DWELLINGS* column for each residential building row based on the number of associated premises.

Another issue which has to be addressed about ADS data is the fact that, the buildings intended for daily working activities are also classified as the residential buildings, alongside the buildings used for permanent or temporary accommodation. In other words, the filtered data at this step contains all the buildings in the Republic of Estonia where the people's presence is expected on a daily basis. Therefore, all the non-accommodational building types like shops, office buildings, hospitals, schools, university buildings, public service buildings etc., as well as the buildings used for temporary accommodation purposes like hotels, hostels, motels, guest houses etc. must be cleared from the dataset, before it can be used in household distribution process, based on the fact that the vast majority of the households live in permanent accommodation type buildings in daily life. The only exception for the temporary accommodation facilities which must not be filtered from the dataset is the dormitory buildings which although being the temporary accommodation facilities, are used for long-term residence purposes by a small group of households.

Openly published data from OpenStreetMap (OSM) project [61] have been used for performing the required filtering operations described in the previous paragraph. Since the process of building the geospatial environment of the country has been done in parallel with filtering of the residential buildings, firstly the footprints of all buildings in the country, alongside with the street networks, and other networked infrastructure types have been retrieved from OSM and integrated to the software. Next, the geometries of the geospatial entities tagged as non-residential building or amenity types in OSM are retrieved, grouped and added to the population generator.

The indication points for the addresses are presented in L-EST97 coordinate system in ADS data, which is the main plane coordinate system used in Estonia by the official government registries. L-EST97 coordinates are calculated by applying Lambert's two-dimensional conical map projection on the geodetic coordinates [62]. Since the coordinate system used both in OSM and in the geospatial environment built as part of the population generator software is a geodetic coordinate system, the coordinates given for the buildings in address system data had to be transformed to the WGS before the OSM data could be used in the filtering process. Therefore, I have developed a small Python program which takes a file containing L-EST coordinate pairs and converts them to the geodetic coordinate pairs, based on the inverse projection formulas for Lambert conformal conic projection given in [63] and the initial parameter values for LAMBERT-EST provided in [62]. The program is tested on a subset of ten L-EST97 coordinates randomly sampled from the ADS data and validated with Google Maps, before the coordinates of all buildings remaining in ADS data have been transformed to the geodetic coordinate system. After the coordinate system consistency in the used datasets has been assured, the buildings whose indication point lay inside the polygon of any geospatial entity tagged with non-accommodational values in OSM have been cleared from the data.

Finally, *TAISAADRESS* column, containing the full addresses of the buildings has been processed and the county, the municipality and the settlement each building belongs to have been determined and assigned as *STLMNT_1*, *STLMNT_2*, and *STLMNT_3* columns, respectively. As a result, the number of dwellings for each spatial region has been calculated as the sum of the number of dwellings in all residential buildings located in that region. Afterwards, the households have been distributed from the municipality level to the settlement level proportionally based on the number of dwellings each settlement possess.

3.2.3 Distribution of the individuals to the households

At this step, the households have already been created based on the number of total households by county and then have been distributed till the settlement level. The generated individuals, as well, have fully been distributed till the municipality level, whereas a proportion of the population have also been assigned the settlement code, except for the people living in the towns, the small-towns and the villages. Therefore, the further development of the population generator continues with associating the individuals with the households. In order to distribute the population to the households in the right way, the household characteristics revealing their internal structures must be known. There are three statistical tables published by Statistics Estonia, providing information about the members of the households in Estonia. The tables presented in [64] and [65], group the households in all of the country by different household structure types and shows the total population size living in each household structure type, respectively. Additionally, the households are grouped by the number of members in [66]. Table 4 groups data published in [64] and [65] for the year 2020, whereas Table 5 represents the households aggregated by the member count for the year 2020, as it is provided in [66].

Table 4. Number of households and total population size by household structure type (Retrieved from [64] and [65]).

Household structure type	Number of households	Total population size
HOUSEHOLD WITHOUT CHILDREN	470500	731600
Single person aged under 65	154600	154600
Single person aged 65 and over	111400	111400
Couple without children, at least one partner is aged under 65	91400	182800
Couple aged 65 and over without children	42500	85000
Other household without children	70600	197800
HOUSEHOLD WITH CHILDREN	155500	584400
Adult and child(ren)	20100	48100
Couple with one child	42900	128700
Couple with two children	40500	162000
Couple with three or more children	17200	88900
Couple with minor and adult children	19100	86400
Other household with children	15700	70300
TOTAL	626000	1316000

Table 5. Number of households by household size type (Retrieved from [66]).

Household size type	Number of households
One member	26600
Two members	175100
Three members	84700
Four members	67000
Five or more members	33200
TOTAL	626000

As it can be seen from Table 4 and Table 5, the data about household structures are provided as one-dimensional tables by Statistics Estonia. Additionally, another one-dimensional data, which is the number of households in each settlement must also be considered, while distributing the population to the households. None of the one-dimensional data mentioned above is enough on its own for the distribution process. For example, doing the distribution based only on household structure data can end up with violating the preservation of household distribution statistics by household member count in the final result. It is a very common problem in population generation as discussed in Section 2.3. The most widely used [51], [52] methodology for solving the analogical problems, namely, IPF, is not applicable in our particular case, since some of the rows in Table 4 and Table 5 are not actual household types, but rather aggregation of numerous distinct household types. For instance, *Adult and child(ren)* row in Table 4, is actually the aggregation of an unknown number of disjoint household structure types like *Adult and a child*, *Adult and two children*, *Adult and three children*, etc. for which the information about the number of representative households is not available. The same can be said about *Five or more members* row in Table 5. Furthermore, IPF method is not eligible enough when more complex set of data are planned to be generated, which represent the correlation of the individual-level characteristics of the members in the households both logically and realistically. The reason why both the words logically and realistically are emphasized in the previous sentence can be explained with an example of a hypothetic household generated as a representative of *Couple with one child* household structure type. If the age difference between one of the individuals assigned to the family as a parent and the individual playing the role of the child is, let's say equal to five, it implies that the logical integrity is not preserved in the data produced as the result of the population generation process. On the other hand, the scenario when the age difference between the parents are more than twenty in half of the generated households, although not being evaluated as the violation of the logical integrity in the data, but at the same time cannot be accepted as a realistic population generation process, since it contradicts the real-life statistical data.

Taking all the factors mentioned above into the consideration, Monte Carlo sampling methodology proposed as an alternative for IPF in [53] is utilized for assigning the individuals to the households. As a preliminary step, the statistical data related to the classification of the households by structure types and number of members given in Table

4 and Table 5 are disaggregated till the settlement-level with the same methodology used for distributing the total number of households in Section 3.2.2. After that, the individuals are distributed to the households by consecutive execution of three Monte Carlo sampling procedures.

Before describing the MC sampling algorithms used, there are some assumptions and parameters which have to be explained. Firstly, it is assumed that a household can have representatives from three different generations maximum. In other words, each household produced by the population generator can contain only children, parents and grandparents. The status of the individual in the household (child, parent, grandparent) is determined with age ranges controlled by several parameters. The parameters *min_age_difference_gen* and *max_age_difference_gen* regulate the age difference limits between the parents and the children and also, between the grandparents and the parents. The age difference between the adult individuals of the same generation (the parents couple and the grandparents couple) is controlled by the variables *max_age_difference_male* and *max_age_difference_female*. The parameters mentioned in the previous sentence define the maximum possible age difference in the couples for the cases when the male individual is older and when the female individual is older, respectively. The parameters *min_child_age*, *max_child_age*, *min_parent_age*, *max_parent_age*, *min_grandparent_age*, *max_grandparent_age*, are used for setting the age boundaries in the assignment process of the individuals. These parameters have the initial values provided in X, but their values are updated after every individual assignment, based on the age of the selected individual.

Table 6. Age boundary parameters in the households and their initial values.

Age boundary parameters	**Initial values**
min_child_age	0
max_child_age	17
min_parent_age	min_child_age + min_age_difference_gen
max_parent_age	max_child_age + max_age_difference_gen
min_grandparent_age	min_parent_age + min_age_difference_gen
max_grandparent_age	max_parent_age + max_age_difference_gen

Another assumption is that, each household has at least one adult individual, regardless of its status in the family. This necessity is represented with the *adult_set* parameter, which initially has zero as the value for all households, and set to one with the assignment of the first adult to the household. Additionally, in order to be able to represent the family nucleus concept in the households, which holds for the majority of the families in any society, the assignment of the adults is regulated such that, the second individual of the same generation must have a different gender value than the first one. The parameters *first_parent_gender* and *first_grandparent_gender* get their values after the assignment of the first parent and the first grandparent, respectively and if the second individual of the same generation must be assigned to the household, it is selected from the subgroup of the population having different gender value with these parameters.

As it is already mentioned, the overall process of the distribution of the individuals to the households is handled with three sequential Monte Carlo sampling algorithms. Since the correlative age difference between the members of a household plays an important role in the assignment of new members, and also considering that the individuals younger than 18 years old are the only subgroup of the population whose statuses in the families are known from the prior, the first algorithm deals with the distribution of those individuals to the families as children. It can be noticed from Table 5 that, for some household structure types the exact number of children in the household is given, while for some others we only know the minimum number of children a representative household must contain. Therefore, the algorithm assigning the children to the households is composed of two parts. In the first part, the household structure types are assigned to the households and only the mandatory children assignments are done for every household. For example, the number of children assigned for the households with *Coupe with two children*, *Adult and child(ren)* and *Couple with three or more children* structure types at the end of this step are equal to two, one and three respectively. After the mandatory assignment process is finished, the remaining individuals under 18 are distributed among the households for which further assignment of children is possible. For instance, an individual can be allocated to a household with *Other household with children* family type during this stage, meanwhile further assignments for the specimen of *Couple with one child* structure are disallowed.

One of the most frequently encountered problems in using the Monte Carlo sampling for constructing the internal structures of the households is that, as the sampling procedure advances to the final stages, the selection set gets more and more narrow, and there is a possibility that at some point the number of adults left in the selection set will be less than the number of households whose members are not assigned yet. In this case, the generation of households which don't contain any adult member is unavoidable. Running an additional subprogram, which takes an adult individual from one of the already inhabited households containing more than one adult, each time the program reaches the point discussed above, and replaces it with one of the children left in the selection set is offered as a solution to this problem in [53]. But, the point to be considered here is that, the intention of using the Monte Carlo sampling while disaggregating the population to the families in our particular case is not to produce a fully-probabilistic distribution mechanism, but rather, to utilize a semi-stochastic methodology which besides being more capable that the deterministic approaches, also preserves the consistency with the real-world household statistics at the same time. Therefore, the algorithm implemented for conducting the distribution process employs numerous validation mechanisms working based on the aggregated household statistics, at each step. Since, the solution method discussed above proposes to replace the members in the households which already have been generated and populated, with some other individuals having completely different personal characteristics (age, gender), adopting this solution could have created lots of unnecessary complexity in tracking the validity of the generation process. Thus, in order to address the issue in a more convenient and direct way, the second MC sampling procedure appoints an adult individual to each household before the distribution of the remaining adults are handled with a third sampling mechanism. Generally, an adult individual can be assigned all three possible statuses in the family (child, parent, grandparent), but the individual selected to be appointed to a household at this phase, must have an age value corresponding to either [*min_parent_age*, *max_parent_age*] or [*min_grandparent_age*, *max_grandparent_age*] age intervals created based on the parameter values of the household. That is to say, the individual must be suitable to play the role of either a parent or a grandparent in the family. Owing to the fact that, during the distribution of the remaining adults, some households will not have any further adult assignment, by doing so, it is guaranteed that every family has at least one parent or one grandparent in the final result.

Eventually, the final sampling algorithm distributes the unsettled adult population in such a way that the maximum possible consistency between the generated households and the statistical data is established in terms of the household structure types and the number of households by the member count. The algorithm follows an analogical approach used in the process of children assignment to the households. The mandatory assignments based on the structure type of the family are conducted for all the households in the first place, and then, the adults which are not appointed to any family yet are distributed among the households having a structure type for which the exact number of adult members is not known.

3.3 Activity and location assignment

There are three types of activities assigned to the generated individuals: residence, work, and education. The common feature of these activities in real life is that, they all generate repetitive spatial patterns. That is to say, these activities are happening periodically, with a known time interval, and also the activity locations are permanent, at least in short term. The concept of the work activity and the education activity is clear and easy to understood, and the residence activity can be explained as the time every individual is spending with the other members of its household. If we would think of a single day in life of an individual as partitioned into two parts, namely, daytime and nighttime, then the work and the education activities would cover the daytime and the residence activity would correspond to nighttime activity.

As have been discussed in Section 3.2.3, the residence activity has already been assigned to the individuals by distributing them into the households, and *household_index* attribute for every individual instance, points to the *index* attribute of the household instance it belongs to. Also, the locations of the residential buildings and the number of dwellings for each residential building are already known after the steps performed in Section 3.2.2. Therefore, the task at this step is to distribute the households to the residential buildings. This distribution process is done at settlement-level, where the households residing in each settlement are distributed to the residential buildings located in that settlement with a weighted random distribution, where the weight coefficient for a building is equal to the number of dwellings it has.

The work activity assignment for the individuals are done based on two datasets. Firstly, the workplaces are generated by using the county-level data showing the total number of enterprises for the year 2020 by the employee count, published by Statistics Estonia in [67]. At this step the member variables *min_employees* and *max_employees* are also set for the generated enterprises. The data provided in [67] includes all private enterprises, as well as, state and local government organizations and can be found in Table 7.

Table 7. The number of enterprises in the counties by member count (Retrieved from [67]).

County	Less than 10 employees	10-49 employees	50-249 employees	250 and more employees	Total
Harju	72753	3701	724	122	77300
Hiiu	966	22	9	0	997
Ida-Viru	6075	340	57	11	6483
Jõgeva	2179	107	12	0	2298
Järva	2172	125	19	3	2319
Lääne	1906	65	16	0	1987
Lääne-Viru	4549	265	33	5	4852
Põlva	2004	83	9	2	2098
Pärnu	7829	363	54	4	8250
Rapla	2910	136	16	0	3062
Saare	3341	127	24	2	3494
Tartu	13940	794	104	15	14853
Valga	1959	82	21	3	2065
Viljandi	3789	177	36	2	4004
Võru	3038	121	22	3	3184
TOTAL	129410	6508	1156	172	137246

The overall process of assigning working individuals to the enterprises is very similar to the allocation of total population to the households in characteristics. Therefore, almost the same procedures applied in Section 3.2 are also utilized in the distribution of the employees to the enterprises with some differences in the constraint parameters.

The disaggregation of the enterprises till the settlement-level is done with a methodology analogous to the one used for the distribution of the households in Section 3.2.2, with only difference being that, this time before distributing the workplaces from municipality-level to the settlements, the number of active individuals who are eligible to work is calculated for each settlement by filtering its population with *min_work_age* and *max_work_age* parameters, and used as the weight coefficient instead of the number of adults. The reason is that, while the number of households in a settlement depends on the number of adults in that settlement, the proportional distribution of the enterprises can only be performed based on the size of the active population in each spatial region.

Next, the county-level statistics about the number of employed persons by gender are retrieved from [68] and disaggregated till the settlement-level in a similar manner to the distribution of the size of total population. The difference between these two processes is that, this time, the number of enterprises in the settlements is used as the weight coefficient in the proportional distribution from municipality-level to the settlement-level, instead of the number of households.

Finally, the assignment of working individuals to the enterprises is performed in two steps with a Monte Carlo sampling methodology alike the one used for distribution of the children to the households. First, the mandatory assignments for the enterprises based on the value of the *min_employees* member variable are done. Then, the remaining working individuals who have not yet been assigned to any enterprise after the initial step, are randomly distributed to the enterprises with the selection criteria that the number of assigned employees for the selected enterprise must be less than the value of its *max_employees* member variable before a new employee is appointed.

The localization of the workplaces is not performed, since it has not been possible to collect sufficient data about the location coordinates of the enterprises existing in the country.

The educational activities are assigned to the individuals based on the data provided for the project by the Geography Department of the University of Tartu. The dataset contains information about the location coordinates and the number of students for all educational facilities located in the Republic of Estonia, which are classified under six groups, namely, kindergartens, basic or secondary schools, colleges, universities, vocational schools and hobby schools. Since, the location of the both residence addresses and the educational facilities are known at this step, first, the candidates list is created for each educational institution based on the age constraints and the distance between the school and the residence addresses, then the students are randomly sampled from the candidates list. Considering that the methodology determining the candidates list is relatively complex, further elaboration can best be given over a particular example. Let's say that forty individuals must be assigned to a kindergarten. There is a *min_age* and *max_age* parameters for each facility type, which is determined as two and six, respectively, in the case of kindergartens in the generator software. Also, considering that the proportion of the students for each unique age in [*min_age*, *max_age*] range cannot be the same for the most of the educational institutions, there is a list of weight coefficients for each school type defining the proportion of students by unique ages, which is selected as [0.1,0.15,0.2, 0.25,0.3] for this particular example. In other words, by selecting the weight coefficients list as it is given in the previous sentence, it is assumed that 10% of the students at kindergartens are two years old, 15% are three years old and so on. Another parameter used in selecting the candidate list is the spatial distribution range parameter which defines how sparsely the residence addresses of the candidates will be distributed, which is let's say equal to 20 in this example. Firstly, the selection list is cleared from the individuals whose age value is out of [*min_age*, *max_age*] range or for whom an education activity is already assigned. Then, for an age *m* in [*min_age*, *max_age*] range, the number of the representatives which must be added to the candidates list, let's say denoted by *P(m)* is calculated as:

$$P(m) = \frac{c_m}{\sum_{n\,=\,\min\_age}^{\max\_age} c_n} \times t\ \times d \qquad (14)$$

where, *c* is the weight coefficient, *t* is the total number of students must be assigned to the school and *d* is the spatial distribution range parameter. In our particular example, the number of two years old candidates must be 80, the number of three years old candidates must be 120 etc. In the next step, the selection list is sorted by the distance between the residence location of the individual and the location of the educational facility and first *P(m)* individuals for the age *m* in [*min_age*, *max_age*] are added to the candidates list. In this way, the candidates list of size $t\ \times d$ is created for the school with the capacity of t and, finally, the students for the facility are randomly sampled from the candidates list.

4 Implementation

This section is organized in the following manner. Section 4.1 lists the data used in the thesis work and the providers. Software dependencies are discussed in Section 4.2, meanwhile Section 4.3 explains the software architecture. Finally, Section 4.4 describes the Monte Carlo sampling algorithms used in this thesis work in terms of the flowchart diagrams and the source code.

4.1 Data

One of the biggest advantages of the software is that, all input data used in the implementation process is open access. Two types of data have been collected for building the synthetic population generator, which are geospatial vector data and aggregated socio-demographical statistics. The full list of the dataset utilized in the software grouped by the providers is presented below:

1. The Department of Geography, University of Tartu.
   - Geospatial vector data for the zone distribution.
   - Geospatial vector and statistical data for the educational facilities located in the Republic of Estonia.
2. Estonian Land Board.
   - Address Data System (ADS) [58].
   - Geospatial vector data for administrative and settlement division (EHAK) of the Republic of Estonia [57].
3. Statistics Estonia.
   - Classification of Estonian administrative units and settlements (EHAK) [56].
   - RV0240: Population by sex, age and place of residence after the 2017 administrative reform [59].
   - LEM02: Households by county [60].

- LEM01: Households by structure [64].
- LEM05: Population in households by household structure [65].
- LEM04: Households by size [66].
- ER028: Enterprises in the statistical profile by year, county and number of employees [67].
- TT243: Employed persons by sex, county and type of employer [68].

4. OpenStreetMap project [61].

- Building footprints and street network.

4.2 Software dependencies

The software is developed in Python programming language with the version 3.9.1. The main reason for selecting Python for the implementation is that it is a dynamic programming language containing numerous convenient libraries that can straightforwardly serve the needs of the synthetic population generator. Furthermore, Python provides an object-oriented approach with a faster development process and rather easily understandable syntax.

One of the most essential frameworks utilized in this thesis work is Pandas (version 1.1.5), which has been extensively used in data pre-processing and manipulation operations. Specifically, it supplies wieldy tools for bidirectional conversion between the *.csv* and *.xlsx* files and data frame objects for processing and output generation, which are frequently used during the implementation. Considering that the geospatial vector data are employed rather often in the program, the GeoPandas (version 0.8.1) extension is particularly significant to be able to manipulate the spatial datasets. OSMnx presented in [69], allows the automated data download from OpenStreetMap project geodatabases. The library (version 1.0.1) is utilized in this thesis work to retrieve the street networks as well as building footprints grouped by the tags attached to them in OpenStreetMap, which later are used in the filtering of the residential buildings. Moreover, other libraries such as NumPy (version 1.19.4), Shapely (version 1.7.1) and Python standard library modules like random and copy are also employed to address various implementation requirements of the project.

Meanwhile, the visualization of the geospatial vector data is performed through a GIS software called QGIS (version 2.8.6). As the debugging of the geospatial data is relatively difficult, especially when the number of the geometries in the data are of thousands, QGIS has been particularly handy in detecting differences between the raw and the processed vector data by letting them to add them in a layered structure.

4.3 Software architecture

The synthetic population generator is built by adopting an object-oriented programming approach. There are five classes used in the program which are: *Settlement*, *Individual*, *Household*, *Enterprise*, and *Educational_Facility*. The mentioned classes are defined together with their member attributes and methods in *Settlement.py*, *Individual.py*, *Household.py*, *Enterprise.py* and *Educational_Facility.py* program files, respectively.

PopGen.py is the main program file of the software, meanwhile *PopGen_methods.py* and *Debug_methods.py* contain the functions used in the development and the validation of the synthetic population generator. The interaction of the software with the input data is organized through *Data.py* program file and *Extract_results.py* produces the outputs under the *Results* directory. *Parameters.py* and *Enums.py* files define all the parameters and enumerated data types used in the software. *Spatial_data_preprocess.py* file is not actually the integral part of the population generator architecture, but rather includes all the functions conducting necessary pre-processing steps for the geospatial input dataset.

4.4 Key algorithms

There are five MC sampling algorithms in the program composing the population synthesis and activity assignment procedures, which are the core parts of the population generator. Three of them are used in the assignment of the population into the households as discussed in Section 3.2.3, one is utilized for appointing the working individuals to the enterprises (Section 3.3) and the final MC sampling algorithm selects students for the educational institutions. Since, the last one has already been explained comprehensively over an example in Section 3.3, only the function performing the most crucial task of the algorithm, which is selecting *n* individuals from the selection list residing closest to the location coordinate of the educational institution is provided in Figure 8. The flowchart diagrams explaining the overall structure of the remaining MC sampling algorithms are presented in Figure 4, Figure 5, Figure 6, and Figure 7.

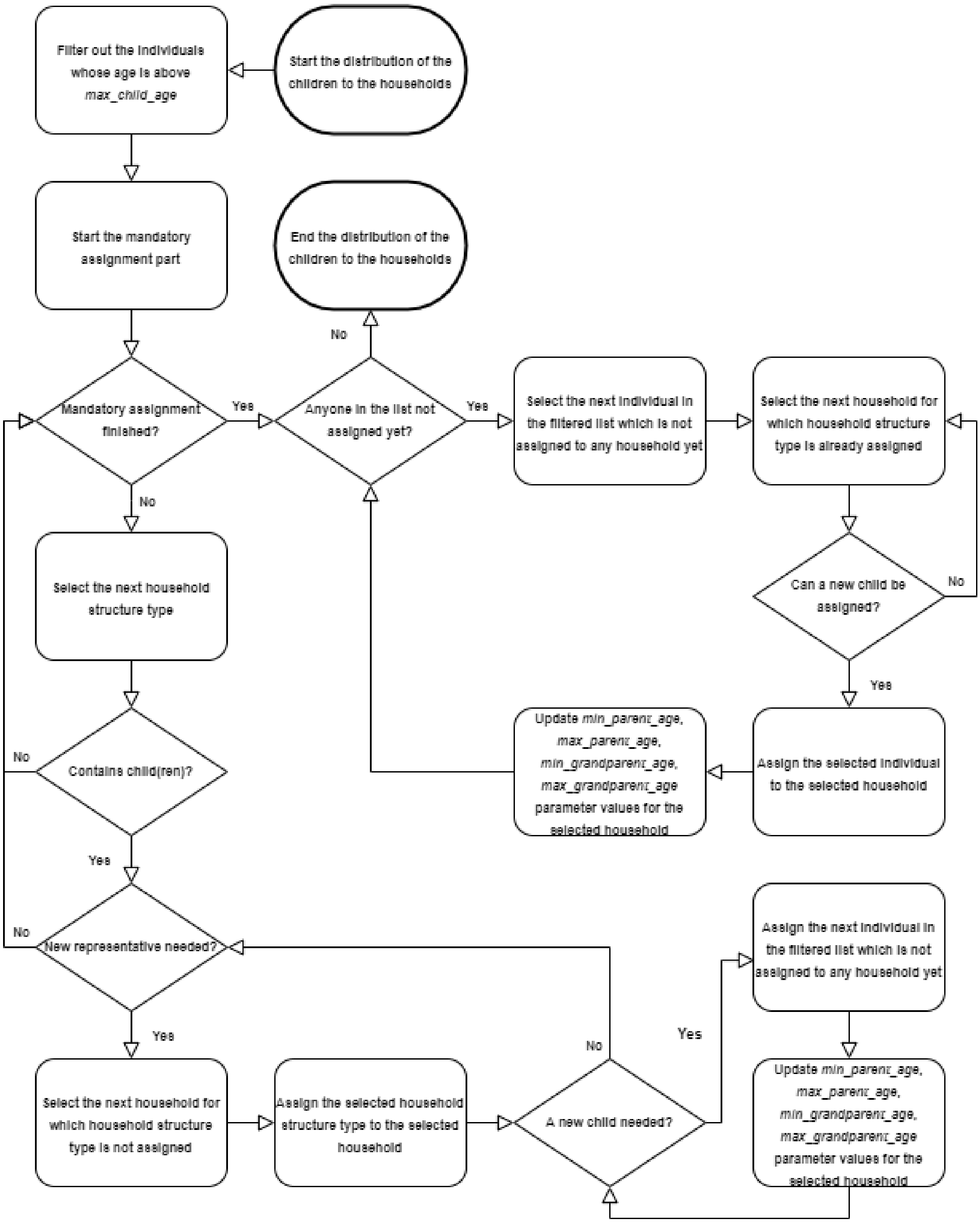

Filter out the individuals whose age is above *max_child_age*
Start the distribution of the children to the households
Start the mandatory assignment part
End the distribution of the children to the households
No
Mandatory assignment finished?
Yes
Anyone in the list not assigned yet?
Yes
Select the next individual in the filtered list which is not assigned to any household yet
Select the next household for which household structure type is already assigned
No
Select the next household structure type
Can a new child be assigned?
No
Yes
No
Contains child(ren)?
Update *min_parent_age*, *max_parent_age*, *min_grandparent_age*, *max_grandparent_age* parameter values for the selected household
Assign the selected individual to the selected household
Yes
No
New representative needed?
Assign the next individual in the filtered list which is not assigned to any household yet
Yes
No
Yes
Select the next household for which household structure type is not assigned
Assign the selected household structure type to the selected household
A new child needed?
Update *min_parent_age*, *max_parent_age*, *min_grandparent_age*, *max_grandparent_age* parameter values for the selected household

Figure 4. The flowchart diagram of the MC sampling algorithm distributing the children to the households.

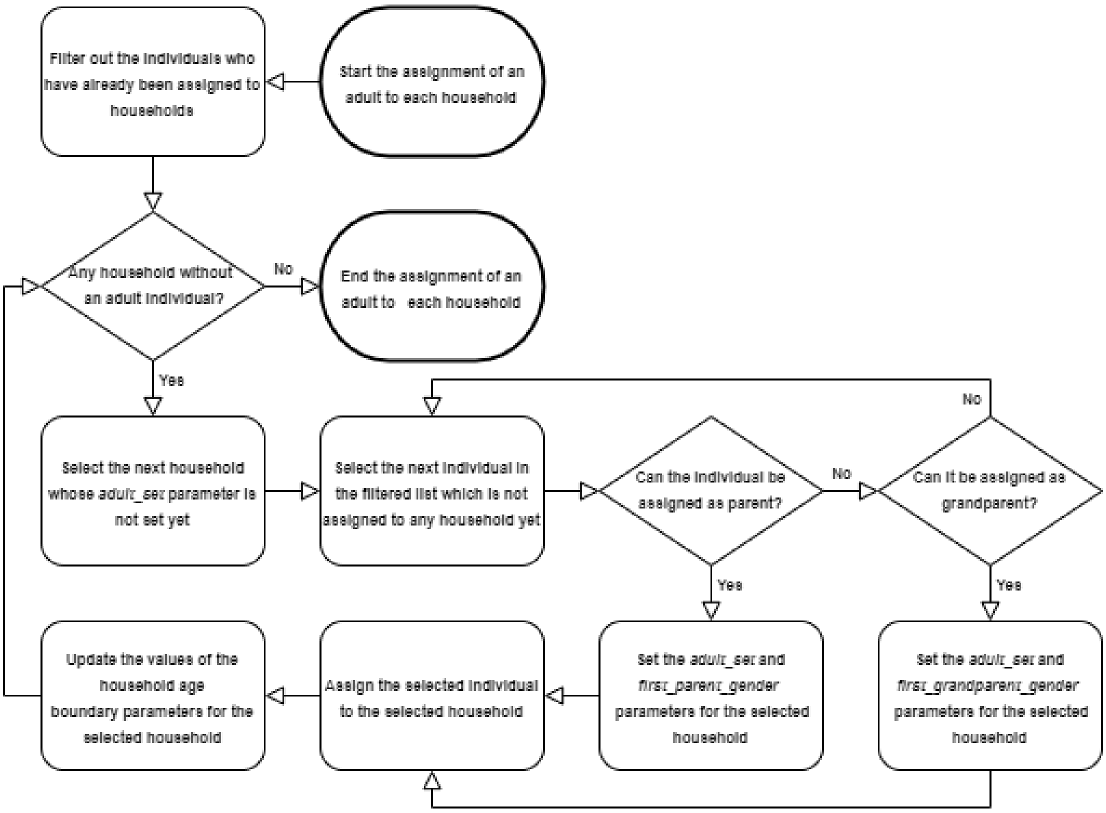

Start the assignment of an adult to each household
Filter out the individuals who have already been assigned to households
Any household without an adult individual?
No
End the assignment of an adult to each household
Yes
Select the next household whose adult_sex parameter is not set yet
Select the next individual in the filtered list which is not assigned to any household yet
Can the individual be assigned as parent?
No
Can it be assigned as grandparent?
No
Yes
Yes
Set the adult_sex and first_parent_gender parameters for the selected household
Set the adult_sex and first_grandparent_gender parameters for the selected household
Assign the selected individual to the selected household
Update the values of the household age boundary parameters for the selected household

Figure 5. The flowchart diagram of the MC sampling algorithm making the first adult assignments for the households.

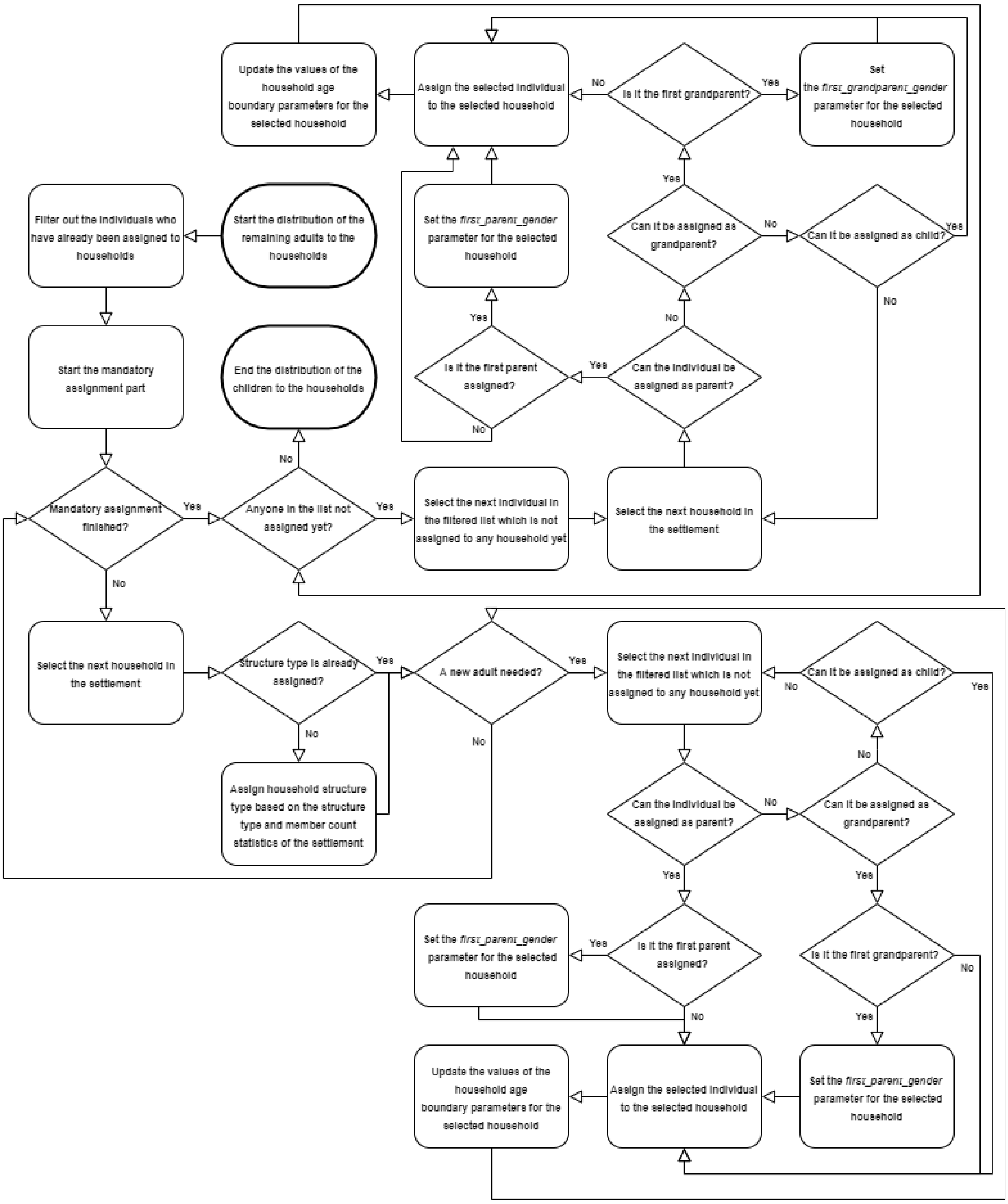

Update the values of the household age boundary parameters for the selected household
Assign the selected individual to the selected household
No
Is it the first grandparent?
Yes
Set the *first_grandparent_gender* parameter for the selected household
Filter out the individuals who have already been assigned to households
Start the distribution of the remaining adults to the households
Set the *first_parent_gender* parameter for the selected household
Yes
Can it be assigned as grandparent?
No
Can it be assigned as child?
Yes
No
Start the mandatory assignment part
End the distribution of the children to the households
Yes
Is it the first parent assigned?
Yes
No
Can the individual be assigned as parent?
No
No
Mandatory assignment finished?
Yes
Anyone in the list not assigned yet?
Yes
Select the next individual in the filtered list which is not assigned to any household yet
Select the next household in the settlement
No
Select the next household in the settlement
Structure type is already assigned?
Yes
No
Assign household structure type based on the structure type and member count statistics of the settlement
A new adult needed?
Yes
No
Select the next individual in the filtered list which is not assigned to any household yet
Can it be assigned as child?
No
Yes
No
Can the individual be assigned as parent?
No
Can it be assigned as grandparent?
Yes
Yes
Set the *first_parent_gender* parameter for the selected household
Yes
Is it the first parent assigned?
No
Is it the first grandparent?
No
Yes
Update the values of the household age boundary parameters for the selected household
Assign the selected individual to the selected household
Set the *first_parent_gender* parameter for the selected household

Figure 6. The flowchart diagram of the MC sampling algorithm distributing the remaining adults to the households.

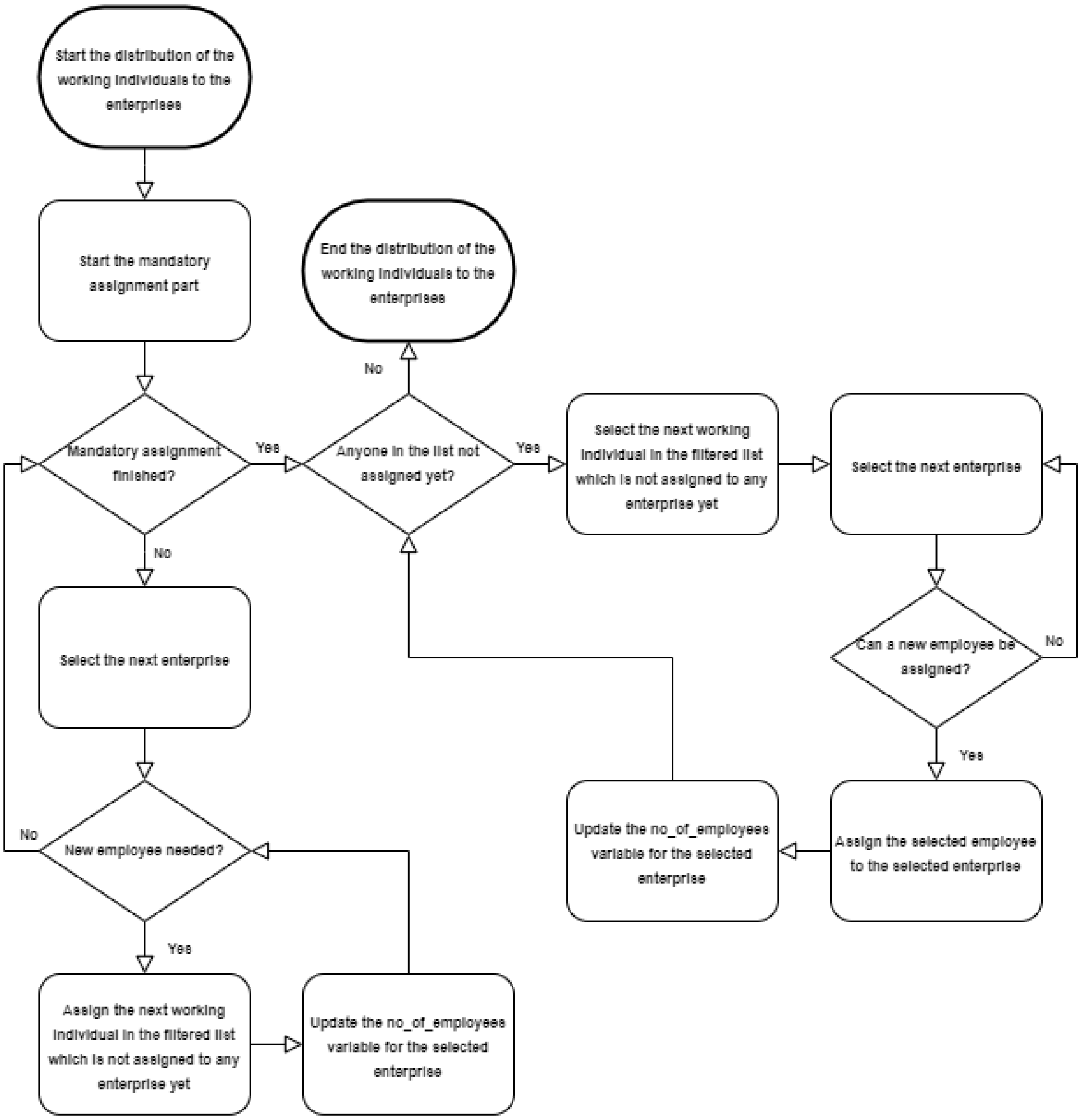

Start the distribution of the working individuals to the enterprises
Start the mandatory assignment part
End the distribution of the working individuals to the enterprises
No
Mandatory assignment finished?
Yes
Anyone in the list not assigned yet?
Yes
Select the next working individual in the filtered list which is not assigned to any enterprise yet
Select the next enterprise
No
Select the next enterprise
Can a new employee be assigned?
No
Yes
No
New employee needed?
Update the no_of_employees variable for the selected enterprise
Assign the selected employee to the selected enterprise
Yes
Assign the next working individual in the filtered list which is not assigned to any enterprise yet
Update the no_of_employees variable for the selected enterprise

Figure 7. The flowchart diagram of the MC sampling algorithm assigning the working individuals to the enterprises.

```
get_nearest_n_people(point_of_interest, selection_list, n):
      for individual in selection_list:
            individual.distance =
            individual.address_point.distance(point_of_interest)

      sorted_selection_list = sorted(selection_list, key=lambda x:
      x.distance, reverse = False)

      nearest_n = sorted_selection_list[:n]
      return nearest_n
```

Figure 8. Algorithm selecting n individuals residing closest to the point_of_interest.

As it can be seen from Figure 8, the choice of the individuals which must be added to the candidates list is done by first setting the *distance* variable of all individuals in the selection list based on the distance between their residence address coordinates and the point of interest, which is the address coordinate of the educational facility in the case of the fifth MC sampling algorithm described in Section 3.3. Then, the *selection_list* which contains instances of the *Individual* class is sorted in the ascending order by the value of the *distance* variable of its elements. Finally, first *n* members of the list are returned as the *n* individuals

5 Results

This section presents the results of the steps followed in the development of the synthetic population generator, validates them with the aggregated statistical data and describes the final outcome of the software.

First of all, by building the spatial hierarchy of the Republic of Estonia and integrating the geospatial vector data for all distinct spatial regions, it is ensured that all spatial regions at level zero (Figure 9), level one (Figure 10) and level two (Figure 11 and Figure 12) can be represented as the aggregate of their child regions and therefore, the spatial data collected at each level can also be disaggregated till the settlement level, just like the socio-demographical data.

Then, as the initial step of the population generation process, the individuals given in [59] with unknown residence addresses have been distributed from country-level till the municipality level in order to guarantee the integrity among all instances of the Individual class, before proceeding to the disaggregation from the municipality-level to the settlement-level. Table 8 presents the total population size in county-level as it is published in [59], in comparison with the number of individuals generated for each county by the software, after the individuals with unspecified residence information are distributed.

Table 8. Comparison of the county-level population size before and after the proportional distribution of the individuals with unspecified residence information.

List of counties	Total number of individuals as it is provided in [59]	Number of generated individuals after the proportional distribution
Harju	605029	605973
Hiiu	9315	9330
Ida-Viru	134259	134469
Jõgeva	28442	28486
Järva	30174	30221
Lääne	20444	20476

List of counties	Total number of individuals as it is provided in [59]	Number of generated individuals after the proportional distribution
Lääne-Viru	58862	58954
Põlva	24647	24685
Pärnu	86185	86319
Rapla	33282	33334
Saare	33083	33135
Tartu	153317	153556
Valga	28204	28248
Viljandi	46161	46233
Võru	35415	35470
County unknown	2070	0
TOTAL	1328889	1328889

As it is mentioned in Section 3.2.2, since it has been needed for the second-layer filtering of the residential buildings, the footprints of all buildings in the country, alongside with the street networks have been retrieved from OSM [61] and integrated to the software and while doing that, the geospatial entities tagged as non-residential building or amenity types in OSM have been grouped together. Figure 13 demonstrates the result of the process with the map of Mustamäe district where some non-residential building type groups are presented in different colors.

The final results of the two-layered filtering discussed in Section 3.2.2 are demonstrated in Figure 14, again on the example of the Mustamäe district, so that what has been done at this step can be better understood as the continuation of the Figure 13.

Since, the distribution of the population to the households is performed by three sequentially executed semi-stochastic Monte Carlo sampling algorithms and four independent parameters, namely, *min_age_difference_gen*, *max_age_difference_gen*, *max_age_difference_male* and *max_age_difference_female* have been utilized in the distribution procedure, the overall process gives different results for different variations of the values assigned to the mentioned parameters. Therefore, experiments have been conducted with several value set configurations, and as the result of the tests done, the optimal set of values for the aforementioned independent parameters have been determined as [19,46,6,2], respectively. The outcome of the distribution process, still contain slight differences, even when the values of the independent variables taken as constant, due to random sampling behaviour. Table 9 compares the average results of the four different generations done with the optimal parameter value sets with the aggregated statistical data presented in [65].

Table 9. Comparison of the total population by household structure type generated by the software with aggregated statistical data provided in [65].

Household structure type	Total population as the average result of four different generations	Total population size as it is in [65]
HOUSEHOLD WITHOUT CHILDREN	736043	731600
Single person aged under 65	154600	154600
Single person aged 65 and over	111400	111400
Couple without children, at least one partner is aged under 65	182800	182800
Couple aged 65 and over without children	85000	85000
Other household without children	202243	197800
HOUSEHOLD WITH CHILDREN	592846	584400
Adult and child(ren)	50207	48100
Couple with one child	128700	128700
Couple with two children	162000	162000
Couple with three or more children	90955	88900
Couple with minor and adult children	88911	86400
Other household with children	72073	70300
TOTAL	1328889	1316000

The reason why the total population size generated by the program is not same with the statistical data is that, the individuals are generated based on the data published in [59]. But, the fact that the remaining population which are not taken into consideration in [65] have been distributed to only the household structure types which accept additional assignments, besides the mandatory ones in the MC sampling algorithm, validates that the distribution of the population is conducted as it has been intended.

With the same analogy, the country-level results of the MC sampling process used in the distribution of the employees to the enterprises is compared with the aggregated statistical data provided in [67] in Table 10.

Table 10. Comparison of the number of the enterprises in the country by member count between the aggregated data and the results of the generator.

	Less than 10 employees	10-49 employees	50-249 employees	250 and more employees	Total
Aggregated data	129410	6508	1156	172	137246
Generator	129407	6511	1156	172	137246

Table 10 shows that MC methodology adopted for distribution of the working individuals to the enterprises produces very similar results to the real-world statistical data it has been built on.

After the population synthesis is conducted, totally 606000 households have been generated based on the statistical data and the sum of dwellings in all residential buildings after two-layered filtering is equal to 804862, which makes sense, considering the cases of uninhabited dwellings and multiple dwellings being owned by the same household. Another point which must be noted is that, the number of dwellings located at each settlement is greater than or equal to the number of households, which let’s to assign a residence location coordinates for all households, and of course some dwellings are left inhabited. The chance of a residential building having 50 dwellings (which makes it an apartment) left inhabited after the distribution of the households to the dwellings is negligible, but a residential building having single dwelling (a house) can be unpopulated after the distribution process, which totally conforms to the real-life scenarios. The outcome of the location assignment to the residence activity can be observed from Figure 15 with the example of Mustamäe district.

The results of the educational activity assignment to the population based on the location of the educational facilities can be seen in Figure 16, based on the example of Tallinna Mustamäe Gümnaasium. The spatial_distribution_range parameter is chosen as 10 in the particular example, meaning that 854 students for the school have been sampled from 8540 nearest-living proper-aged individuals.

As the final result, the generator featuring all the aspects mentioned during this section, produces a dataset covering a synthetic population and the majority of the permanent social links among the synthetic individuals. The dataset also allows to construct some repetitive spatial patterns based on the location coordinates assigned to the activities.

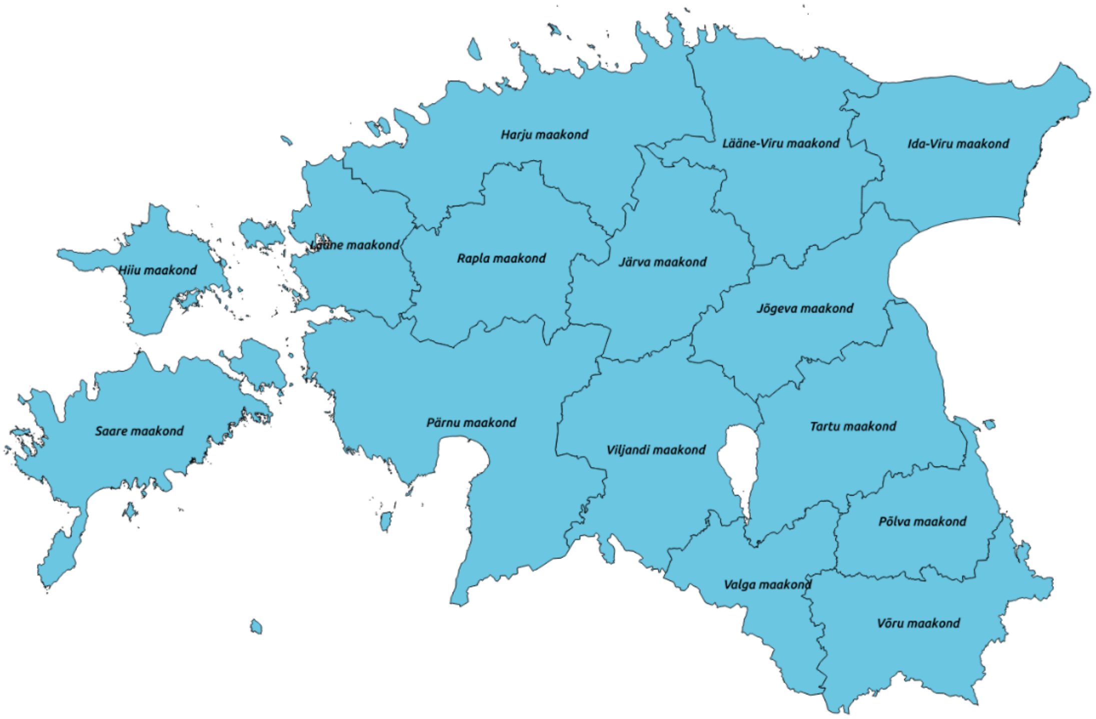


Figure 9. The Republic of Estonia as the aggregation of the counties (Plotted based on data retrieved from [57]).

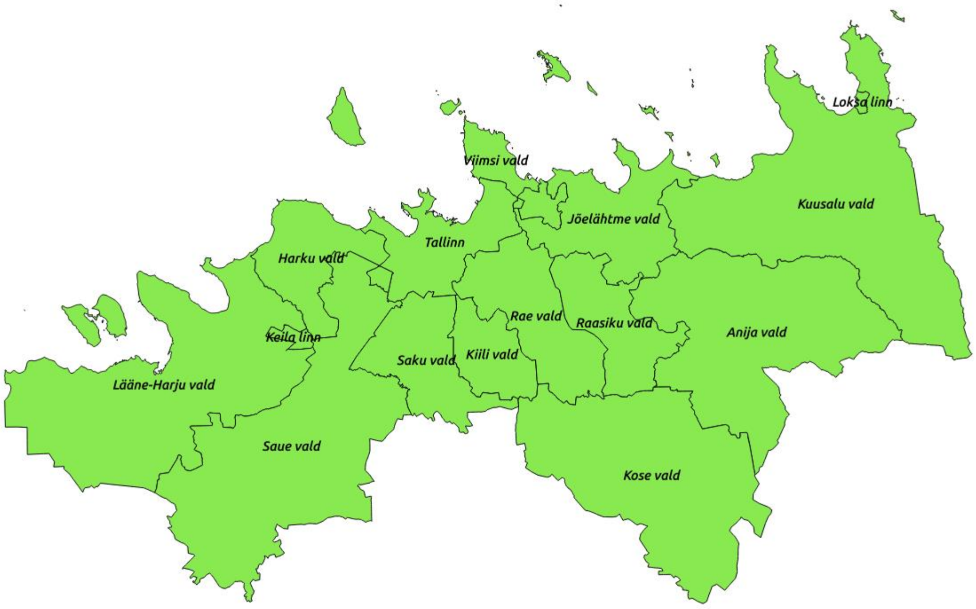


Figure 10. Harju county as the aggregation of its cities and rural municipalities (Plotted based on data retrieved from [57]).

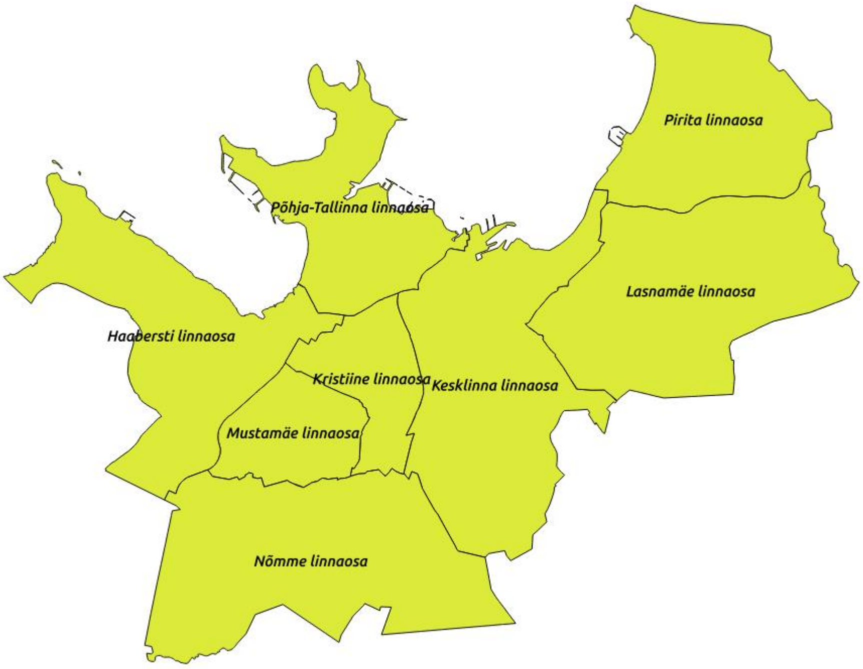


Figure 11. Tallinn city as the aggregate of its city districts (Plotted based on data retrieved from [57]).

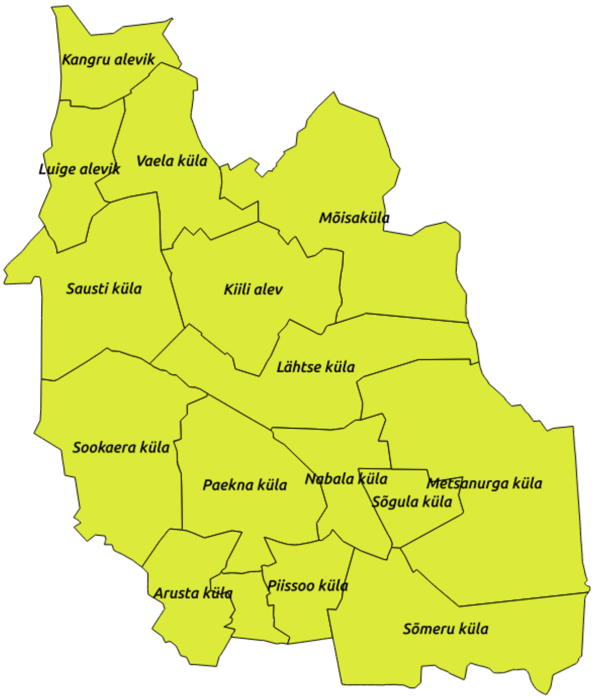


Figure 12. Kiili rural municipality as the aggregate of its settlements (Plotted based on data retrieved from [57]).

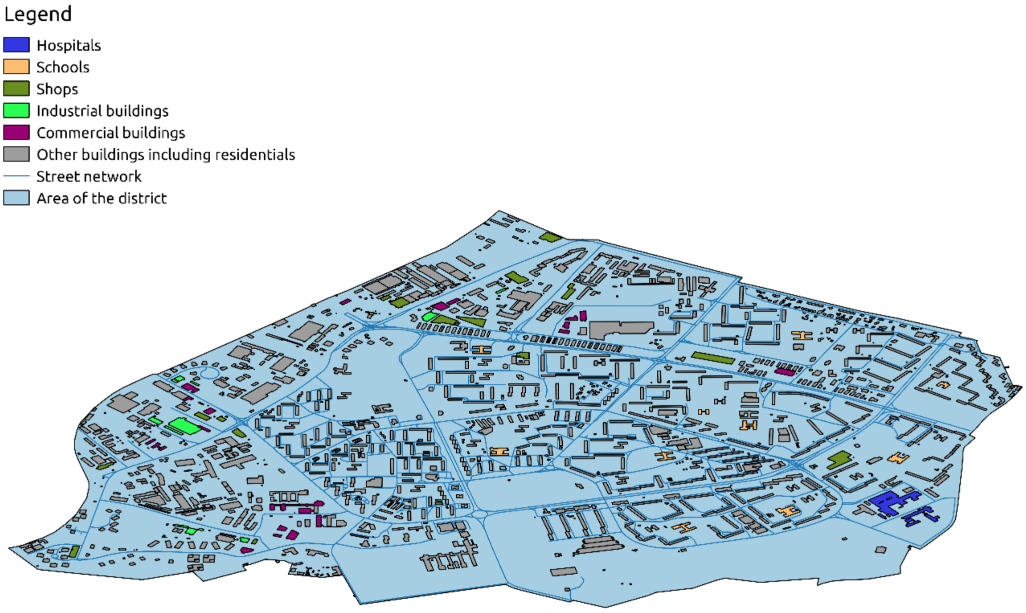


Figure 13. Map of the Mustamäe district with some non-residential buildings being grouped by OSM tags (Composed based on data retrieved from [61]).

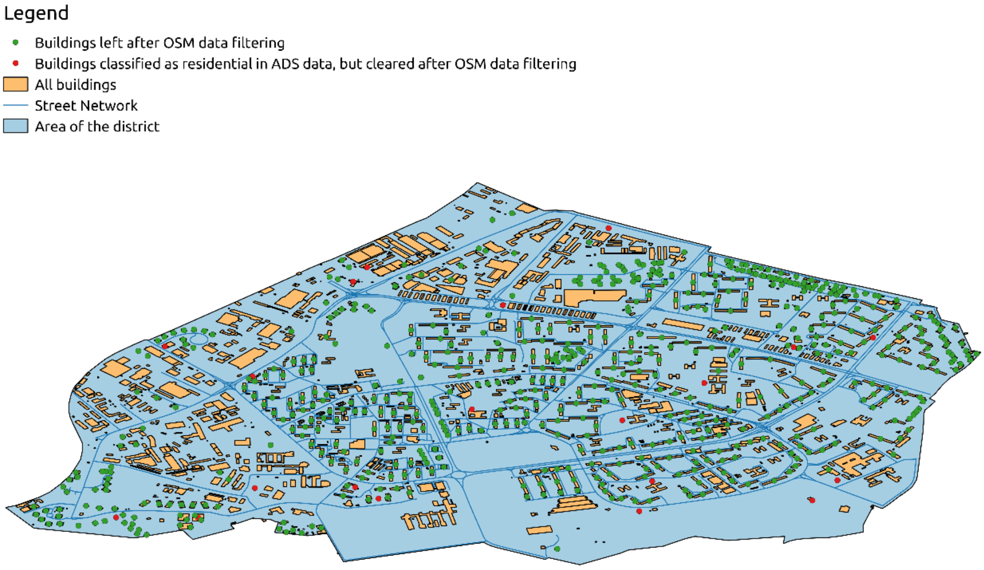

Legend
Buildings left after OSM data filtering
Buildings classified as residential in ADS data, but cleared after OSM data filtering
All buildings
Street Network
Area of the district

Figure 14. Results of the two-layered filtering of the residential buildings with the example of Mustamäe district (Composed based on data retrieved from [58], [61]).

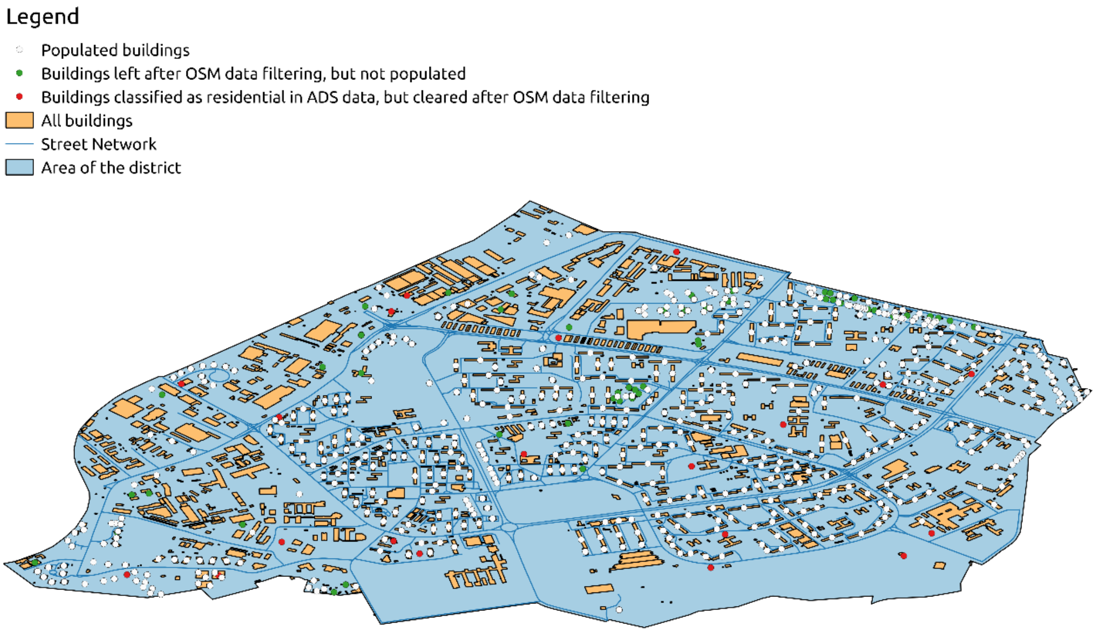

Legend
Populated buildings
Buildings left after OSM data filtering, but not populated
Buildings classified as residential in ADS data, but cleared after OSM data filtering
All buildings
Street Network
Area of the district

Figure 15. Results of the location assignment for the residence activity with the example of Mustamäe district.

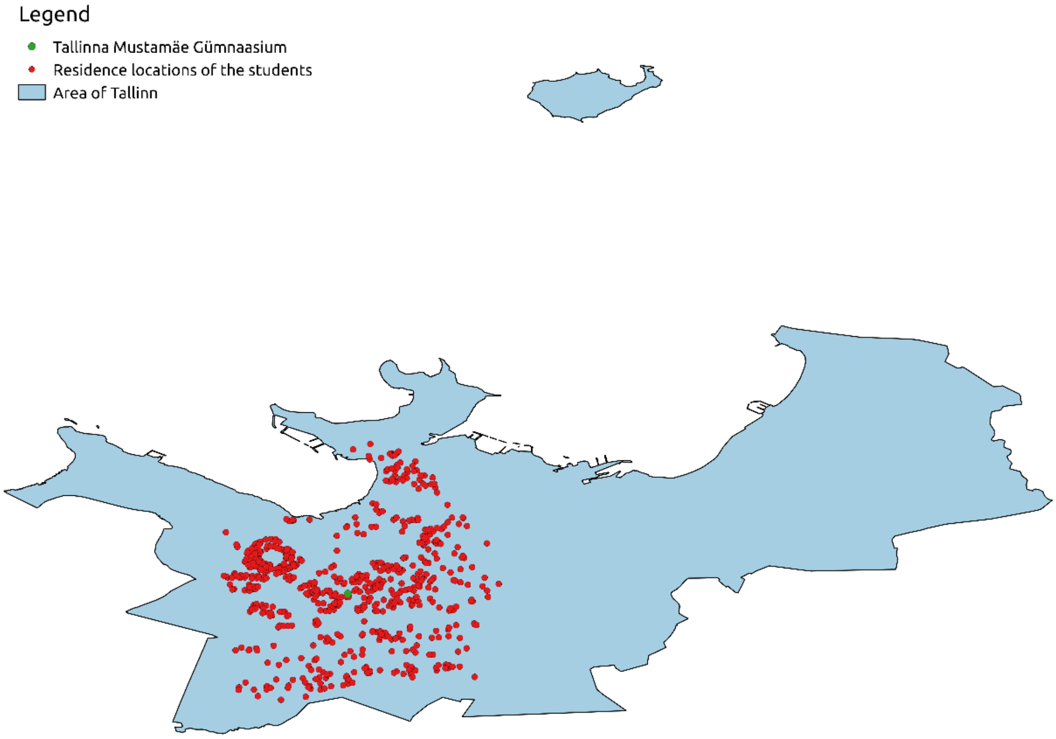

Legend
Tallinna Mustamäe Gümnaasium
Residence locations of the students
Area of Tallinn

Figure 16. The spatial distribution of the residence addresses of the individuals assigned to Tallinna Mustamäe Gümnaasium, when spatial_distribution_range parameter chosen as 10.

6 Summary

This thesis work started with a comprehensive literature study on the epidemic modelling and the synthetic population generation fields, analyzing the key factors determining the behaviour of a disease spread, the methodologies utilized in modelling a viral pathogen transmission, the datasets used in the existing models, and possible approaches for producing a dataset convenient for conducting epidemic simulations. As the result, a Monte Carlo sampling based semi-stochastic synthetic population generator, which is the main contribution of this work, was developed with an object-oriented approach employing the open-access aggregated statistics and geospatial vector data. The synthetic population produced by the software has proved itself to be capable to represent the overall socio-demographic characteristics of the Estonian population, during the validation process with the real-world statistical data, which makes it capable to imitate the behavioural characteristics of the society in large-scale simulations. The generated dataset covers the majority of the long-term social connections and interactions existing among the synthetic individuals composing the population. The spatial aspects integrated to the software lets it possible to make deductions to some extend about collective repetitive spatial patterns emerging in the society, which is of the crucial importance in the epidemic modelling. The original usage intention for the software is to supply the free-scale network-based disease spread simulations conducted in the scope of COVSG22 project with a real-world based dataset. Therefore, being able to reflect the social connections in the population in a realistic way was prioritized during the development process. Meanwhile, the parametrized structure of the population generator makes it also useful for other type of Monte Carlo simulations and since, the geospatial environment has been built in parallel with the population, another possible adaptation choice for the software is to be converted to the large-scale agent-based simulator in the future.

Since, synthetic population generation is a broad concept, there are limitless further improvement possibilities. For example, one of the feasible additions to the software is to integrate other type of activities assignment for the generated individuals in order to be able to create temporary social interactions besides, the long-term connections existing in the current version.

References

[1] “COVID-19 CORONAVIRUS PANDEMIC,” COVID Live Update, 29-Jul-2021. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 29-Jul-2021].

[2] Hethcote, H. W. (2000). The mathematics of infectious diseases. SIAM review, 42(4), 599-653.

[3] Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.

[4] Keeling, M. J., & Eames, K. T. (2005). Networks and epidemic models. Journal of the royal society interface, 2(4), 295-307.

[5] Huang, W., & Li, C. (2007). Epidemic spreading in scale-free networks with community structure. Journal of Statistical Mechanics: Theory and Experiment, 2007(01), P01014.

[6] Nasution, H., Jusuf, H., Ramadhani, E., & Husein, I. (2020). Model of Spread of Infectious Disease. Systematic Reviews in Pharmacy, 11(2).

[7] Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., & Hens, N. (2020). Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. Eurosurveillance, 25(17), 2000257.

[8] Leavitt, S. V., Lee, R. S., Sebastiani, P., Horsburgh Jr, C. R., Jenkins, H. E., & White, L. F. (2020). Estimating the relative probability of direct transmission between infectious disease patients. International journal of epidemiology, 49(3), 764-775.

[9] Chen, Y., Jiao, J., Bai, S., & Lindquist, J. (2020). Modeling the spatial factors of COVID-19 in New York City. Available at SSRN 3606719.

[10] Fortaleza, C. M., Guimarães, R. B., de Almeida, G. B., Pronunciate, M., & Ferreira, C. P. (2020). Taking the inner route: spatial and demographic factors affecting vulnerability to COVID-19 among 604 cities from inner São Paulo State, Brazil. Epidemiology & Infection, 148.

[11] Nishiura, H., Oshitani, H., Kobayashi, T., Saito, T., Sunagawa, T., Matsui, T., ... & Suzuki, M. (2020). Closed environments facilitate secondary transmission of coronavirus disease 2019 (COVID-19). MedRxiv.

[12] Goldstein, J. R., & Lee, R. D. (2020). Demographic perspectives on the mortality of COVID-19 and other epidemics. Proceedings of the National Academy of Sciences, 117(36), 22035-22041.

[13] Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., ... & Mills, M. C. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. Proceedings of the National Academy of Sciences, 117(18), 9696-9698.

[14] Onder, G., Rezza, G., & Brusaferro, S. (2020). Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. Jama, 323(18), 1775-1776.

[15] Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R., & Klein, S. L. (2020). Impact of sex and gender on COVID-19 outcomes in Europe. Biology of sex differences, 11, 1-13.

[16] Jin, J. M., Bai, P., He, W., Wu, F., Liu, X. F., Han, D. M., ... & Yang, J. K. (2020). Gender differences in patients with COVID-19: focus on severity and mortality. Frontiers in public health, 8, 152.

[17] Mikhael, E. M., & Al-Jumaili, A. A. (2020). Can developing countries face novel coronavirus outbreak alone? The Iraqi situation. Public Health in Practice, 1, 100004.

[18] Tanne, J. H., Hayasaki, E., Zastrow, M., Pulla, P., Smith, P., & Rada, A. G. (2020). Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. Bmj, 368.

[19] Stojkoski, V., Utkovski, Z., Jolakoski, P., Tevdovski, D., & Kocarev, L. (2020). The socio-economic determinants of the coronavirus disease (COVID-19) pandemic. Available at SSRN 3576037.

[20] Qiu, Y., Chen, X., & Shi, W. (2020). Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China. Journal of Population Economics, 33(4), 1127-1172.

[21] Clouston, S. A., Natale, G., & Link, B. G. (2021). Socioeconomic inequalities in the spread of coronavirus-19 in the United States: a examination of the emergence of social inequalities. Social Science & Medicine, 268, 113554.

[22] Chaudhry, R., Dranitsaris, G., Mubashir, T., Bartoszko, J., & Riazi, S. (2020). A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. EClinicalMedicine, 25, 100464.

[23] UKEssays. (November 2018). Mathematical Modeling of Disease Epidemics. Retrieved from https://www.ukessays.com/essays/sciences/mathematical-modeling-of-disease-epidemics.php?vref=1

[24] Brauer, F. (2008). Compartmental models in epidemiology. In Mathematical epidemiology (pp. 19-79). Springer, Berlin, Heidelberg.

[25] Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the national academy of sciences, 99(suppl 3), 7280-7287.

[26] Gharakhanlou, N. M., & Hooshangi, N. (2020). Spatio-temporal simulation of the novel coronavirus (COVID-19) outbreak using the agent-based modeling approach (case study: Urmia, Iran). Informatics in Medicine Unlocked, 20, 100403.

[27] Columbia University Irving Medical Center. (2019). Spatiotemporal Analysis. Retrieved from https://www.publichealth.columbia.edu/research/population-health-methods/spatiotemporal-analysis#Overview.

[28] Nasution, H., Jusuf, H., Ramadhani, E., & Husein, I. (2020). Model of Spread of Infectious Disease. Systematic Reviews in Pharmacy, 11(2).

[29] Beckley, R., Weatherspoon, C., Alexander, M., Chandler, M., Johnson, A., & Bhatt, G. S. (2013). Modeling epidemics with differential equations. Tennessee State University Internal Report.

[30] Hethcote, H. W. (2000). The mathematics of infectious diseases. SIAM review, 42(4), 599-653.

[31] Batistela, C. M., Correa, D. P., Bueno, Á. M., & Piqueira, J. R. C. (2021). SIRSi compartmental model for COVID-19 pandemic with immunity loss. Chaos, Solitons & Fractals, 142, 110388.

[32] Mahajan, A., Sivadas, N. A., & Solanki, R. (2020). An epidemic model SIPHERD and its application for prediction of the spread of COVID-19 infection in India. Chaos, Solitons & Fractals, 140, 110156.

[33] He, J., Guo, Y., Mao, R., & Zhang, J. (2021). Proportion of asymptomatic coronavirus disease 2019: A systematic review and meta-analysis. Journal of medical virology, 93(2), 820-830.

[34] Ivorra, B., Ferrández, M. R., Vela-Pérez, M., & Ramos, A. M. (2020). Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. Communications in nonlinear science and numerical simulation, 88, 105303.

[35] Arino, J., & Van den Driessche, P. (2003). A multi-city epidemic model. Mathematical Population Studies, 10(3), 175-193.

[36] Szabó, G. M. (2020). Propagation and mitigation of epidemics in a scale-free network. arXiv preprint arXiv:2004.00067.

[37] Herrmann, H. A., & Schwartz, J. M. (2020). Why COVID-19 models should incorporate the network of social interactions. Physical Biology, 17(6), 065008.

[38] Lv, W., Ke, Q., & Li, K. (2020). Dynamical analysis and control strategies of an SIVS epidemic model with imperfect vaccination on scale-free networks. Nonlinear dynamics, 99(2), 1507-1523.

[39] Xie, G. (2020). A novel Monte Carlo simulation procedure for modelling COVID-19 spread over time. Scientific reports, 10(1), 1-9.

[40] Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals, 135, 109864.

[41] Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos, Solitons & Fractals, 135, 109850.

[42] De-Leon, H., & Pederiva, F. (2020). Particle modeling of the spreading of coronavirus disease (COVID-19). Physics of Fluids, 32(8), 087113.

[43] Stojanović, O., Leugering, J., Pipa, G., Ghozzi, S., & Ullrich, A. (2019). A Bayesian Monte Carlo approach for predicting the spread of infectious diseases. PloS one, 14(12), e0225838.

[44] Lorch, L., Kremer, H., Trouleau, W., Tsirtsis, S., Szanto, A., Schölkopf, B., & Gomez-Rodriguez, M. (2020). Quantifying the effects of contact tracing, testing, and containment. arXiv e-prints, arXiv-2004.

[45] Gharakhanlou, N. M., & Hooshangi, N. (2020). Spatio-temporal simulation of the novel coronavirus (COVID-19) outbreak using the agent-based modeling approach (case study: Urmia, Iran). Informatics in Medicine Unlocked, 20, 100403.

[46] Mahmood, I., Arabnejad, H., Suleimenova, D., Sassoon, I., Marshan, A., Serrano-Rico, A., ... & Groen, D. (2020). FACS: a geospatial agent-based simulator for analysing COVID-19 spread and public health measures on local regions. Journal of Simulation, 1-19.

[47] Song, W. Y., Zang, P., Ding, Z. X., Fang, X. Y., Zhu, L. G., Zhu, Y., ... & Peng, Z. H. (2020). Massive migration promotes the early spread of COVID-19 in China: a study based on a scale-free network. Infectious Diseases of Poverty, 9(1), 1-8.

[48] Small, M., Walker, D. M., & Chi, K. T. (2007). Scale-free model for spatio-temporal distribution of outbreaks of avian influenza. IEICE Proceedings Series, 41(18PM2-A-2).

[49] Firth, J. A., Hellewell, J., Klepac, P., Kissler, S., Kucharski, A. J., & Spurgin, L. G. (2020). Using a real-world network to model localized COVID-19 control strategies. Nature medicine, 26(10), 1616-1622.

[50] Dommar, C. J., Lowe, R., Robinson, M., & Rodó, X. (2014). An agent-based model driven by tropical rainfall to understand the spatio-temporal heterogeneity of a chikungunya outbreak. Acta tropica, 129, 61-73.

[51] Wang, K., Zhang, W., Mortveit, H., & Swarup, S. (2020, May). Improved travel demand modeling with synthetic populations. In International Workshop on Multi-Agent Systems and Agent-Based Simulation (pp. 94-105). Springer, Cham.

[52] Adigaa, A., Agashea, A., Arifuzzamana, S., Barretta, C. L., Beckmana, R., Bisseta, K., ... & Xiea, D. (2015). Generating a synthetic population of the United States∗.

[53] Moeckel, R., Spiekermann, K., & Wegener, M. (2003, May). Creating a synthetic population. In Proceedings of the 8th international conference on computers in urban planning and urban management (CUPUM) (pp. 1-18).

[54] Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. PloS one, 5(1), e8828.

[55] Xu, Z., Glass, K., Lau, C. L., Geard, N., Graves, P., & Clements, A. (2017). A synthetic population for modelling the dynamics of infectious disease transmission in American Samoa. Scientific reports, 7(1), 1-9.

[56] "Classification of Estonian administrative units and settlements 2020v3," Hierarchical view. [Online]. Available: http://metaweb.stat.ee/view_xml.htm?id=4601370&amp;siteLanguage=en. [Accessed: 10-Jul-2021].

[57] "Administrative and Settlement Division," Administrative and Settlement Division | Geoportal | Estonian Land Board, 05-Nov-2020. [Online]. Available: https://geoportaal.maaamet.ee/eng/Spatial-Data/Administrative-and-Settlement-Division-p312.html. [Accessed: 10-Jul-2021].

[58] "Address Data," Address Data | Geoportal | Estonian Land Board. [Online]. Available: https://geoportaal.maaamet.ee/eng/Spatial-Data/Address-Data-p313.html. [Accessed: 11-Jul-2021].

[59] "RV0240: POPULATION BY SEX, AGE AND PLACE OF RESIDENCE AFTER THE 2017 ADMINISTRATIVE REFORM, 1 JANUARY," PX-Web. [Online]. Available: https://andmed.stat.ee/en/stat/rahvastik__rahvastikunaitajad-ja-koosseis__rahvaarv-ja-rahvastiku-koosseis/RV0240. [Accessed: 11-Jul-2021].

[60] "LEM02: HOUSEHOLDS BY COUNTY," LEM02: HOUSEHOLDS BY COUNTY. Statistical database, 11-May-2020. [Online]. Available: https://andmed.stat.ee/en/stat/sotsiaalelu__leibkonnad__leibkondade-uldandmed/LEM02. [Accessed: 13-Jul-2021].

[61] OpenStreetMap contributors, OpenStreetMap. [Online]. Available: https://www.openstreetmap.org. [Accessed: 15-Jul-2021].

[62] "Geodetic System," Geodetic System | Geoportal | Estonian Land Board. [Online]. Available: https://geoportaal.maaamet.ee/eng/Spatial-Data/Geodetic-Data/Geodetic-System-p668.html. [Accessed: 16-Jul-2021].

[63] J. P. Snyder, "15. Lambert Conformal Conic projection," in Map projections - a working manual, Washington: US Gov. Print. Off., 1987.

[64] “LEM01: HOUSEHOLDS BY STRUCTURE,” LEM01: HOUSEHOLDS BY STRUCTURE. Statistical database, 11-May-2020. [Online]. Available: https://andmed.stat.ee/en/stat/sotsiaalelu__leibkonnad__leibkondade-uldandmed/LEM01. [Accessed: 17-Jul-2021].

[65] “LEM05: POPULATION IN HOUSEHOLDS BY HOUSEHOLD STRUCTURE,” LEM05: POPULATION IN HOUSEHOLDS BY HOUSEHOLD STRUCTURE. Statistical database, 11-May-2020. [Online]. Available: https://andmed.stat.ee/en/stat/sotsiaalelu__leibkonnad__leibkondade-uldandmed/LEM05. [Accessed: 17-Jul-2021].

[66] “LEM04: HOUSEHOLDS BY SIZE,” LEM04: HOUSEHOLDS BY SIZE. Statistical database, 11-May-2020. [Online]. Available: https://andmed.stat.ee/en/stat/sotsiaalelu__leibkonnad__leibkondade-uldandmed/LEM04. [Accessed: 17-Jul-2021].

[67] “ER028: ENTERPRISES IN THE STATISTICAL PROFILE by Year, County and Number of employees,” 25-Jan-2021. [Online]. Available: https://andmed.stat.ee/en/stat/majandus__majandusuksused__ettevetjad/ER028. [Accessed: 20-Jul-2021].

[68] “TT243: EMPLOYED PERSONS BY SEX, COUNTY AND TYPE OF EMPLOYER,” 15-Feb-2021. [Online]. Available: https://andmed.stat.ee/en/stat/sotsiaalelu__tooturg__heivatud__aastastatistika/TT243. [Accessed: 22-Jul-2021].

[69] Boeing, G. 2017. “OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks.” Computers, Environment and Urban Systems. 65, 126-139.

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Azer Ramazanli

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Disease Spread Modelling”, supervised by Aleksei Tepljakov
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

02.08.2021

[1] The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.