Handout: Synthetic population generator software

Open Data Forum

02.11.2021

1 Data

The current version of the SPG software (*version 2.0*) benefits from the usage of the aggregated interzonal human mobility statistics datasets provided for the COVSG22 project by Positium and Mobility Lab, University of Tartu for more realistic modeling of the human mobility patterns. The software version discussed in the "Valuing Open Data" workshop (*version 1.0*) is the initial open-access version, meaning that, all input data used in the implementation process of this edition are either open-access publications by the official administrative registers of the Republic of Estonia or retrieved from open data sources licensed under the Open Database License (ODbL). Two types of data have been collected for building the *version 1.0*, which are geospatial vector data and aggregated socio-demographical statistics. The full list of the datasets utilized in the software grouped by the providers is presented below:

1. Mobility Lab, 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 Geoportal
    - Address Data System (ADS) [1]
    - Geospatial vector data for administrative and settlement division (EHAK) of the Republic of Estonia [2]
3. Statistics Estonia
    - Classification of Estonian administrative units and settlements (EHAK) [3]

- RV0240: Population by sex, age and place of residence after the 2017 administrative reform [4]
- LEM02: Households by county [5]
- LEM01: Households by structure [6]
- LEM05: Population in households by household structure [7]
- LEM04: Households by size [8]
- ER028: Enterprises in the statistical profile by year, county and number of employees [9]
- TT243: Employed persons by sex, county and type of employer [10]

4. OpenStreetMap project [11]

    - Street network, building footprints and metadata containing tags

2 ADS and OSM data for determining micro-scale density

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, 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 cannot 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 [1]. 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 [11] 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.

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.

As the next step, *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.

Finally, the location assignment for the residence activity of the individuals is done by distributing the households to the residential buildings. This distribution process is

performed 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.

2.1 OSMNx for retrieving data from OSM

OSMnx presented in [12], allows the automated data download from OpenStreetMap project geodatabases. The reason why I decided to include OSMNx as a subchapter in the handout is that, the package is especially useful for a project of any scale containing OSM data processing and analysing steps, by letting to download, model, visualize and analyse real-world street-networks and other geospatial entities only with a few lines of code. The library (version 1.0.1) is utilized in the project 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.

For example, the code piece provided in Figure 1 produces the result shown in Figure 2.

```
import osmnx as ox
import matplotlib.pyplot as plt

place_name = "Mustamae, Tallinn, Estonia"
# Retrieve the street network
G = ox.graph_from_place(place_name, network_type='drive')
nodes, edges = ox.graph_to_gdfs(G)
# Retrieve the area footprint of Mustamäe
area = ox.geocode_to_gdf(place_name)
# Retrieve all building footprints, school and hospital amenities separately
all_buildings = ox.geometries_from_place(place_name, {'building': True})
schools = ox.geometries_from_place(place_name, {'amenity': 'school'})
hospitals = ox.geometries_from_place(place_name, {'amenity':'hospital'})
# Plot the graph
fig, ax = plt.subplots(figsize=(15,15))
area.plot(ax=ax, color='blue')
edges.plot(ax=ax, linewidth=1, edgecolor = 'black')
all_buildings.plot(ax=ax, color='khaki', alpha=0.7, markersize=10)
schools.plot(ax=ax, color='yellow', alpha=0.7, markersize=10)
hospitals.plot(ax=ax, color='green', alpha=0.7, markersize=10)
plt.tight_layout()
plt.show()
```

Figure 1. Usage example for the OSMNx package.

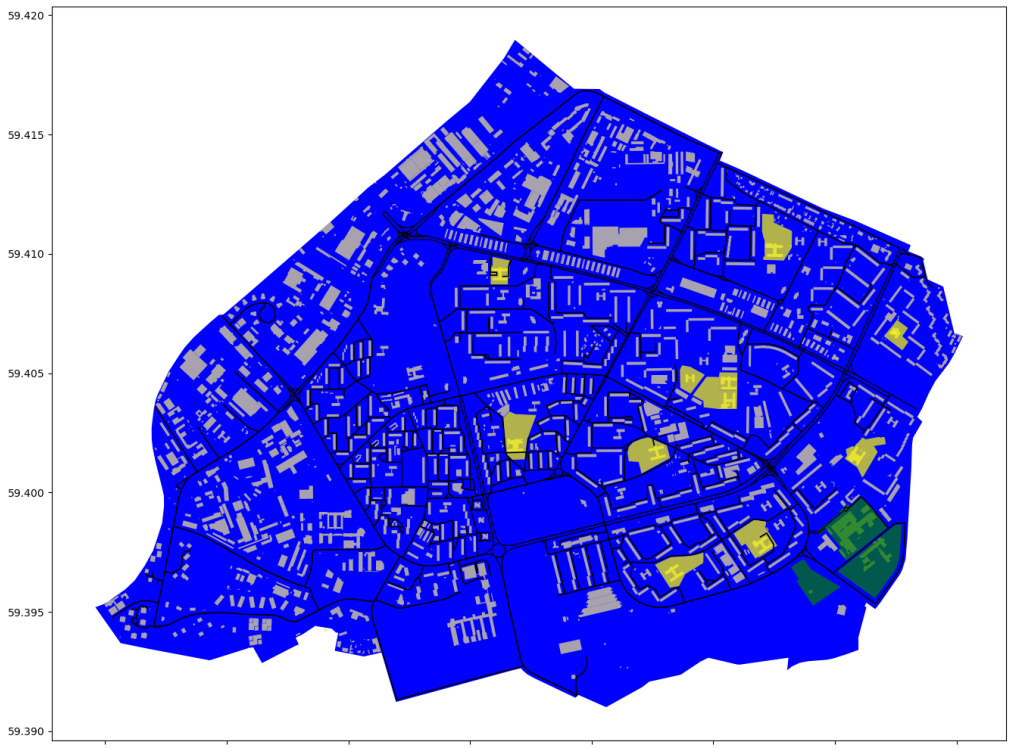

Figure 2. The result of the example code provided in Figure 1.

3 Monte Carlo sampling 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, one is utilized for appointing the working individuals to the enterprises and the final MC sampling algorithm selects students for the educational institutions.

3.1 Distribution of the population into the households

As a preliminary step, the statistical data related to the classification of the households by structure types and number of members are disaggregated till the settlement-level. 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 Table 1, but their values are updated after every individual assignment, based on the age of the selected individual.

Table 1. 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. In [6], 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 the most popular solution offered for solving this particular problem in literature. 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.

The flowchart diagrams explaining the overall structure of the MC sampling algorithms used for distributing the population to the households are presented in Figure 3, Figure 4, and Figure 5, respectively.

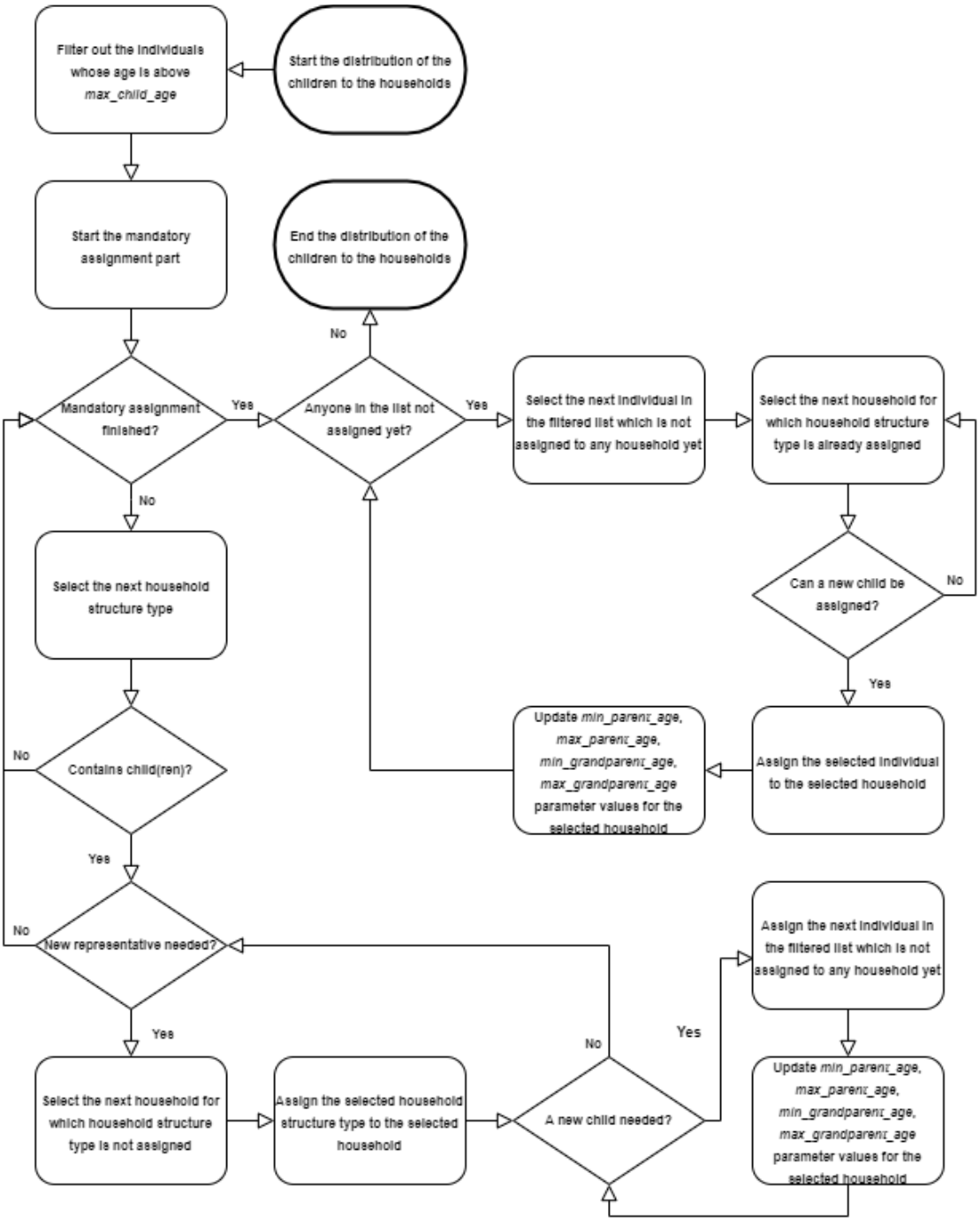

Figure 3. The flowchart diagram of the MC sampling algorithm distributing the children to the households.

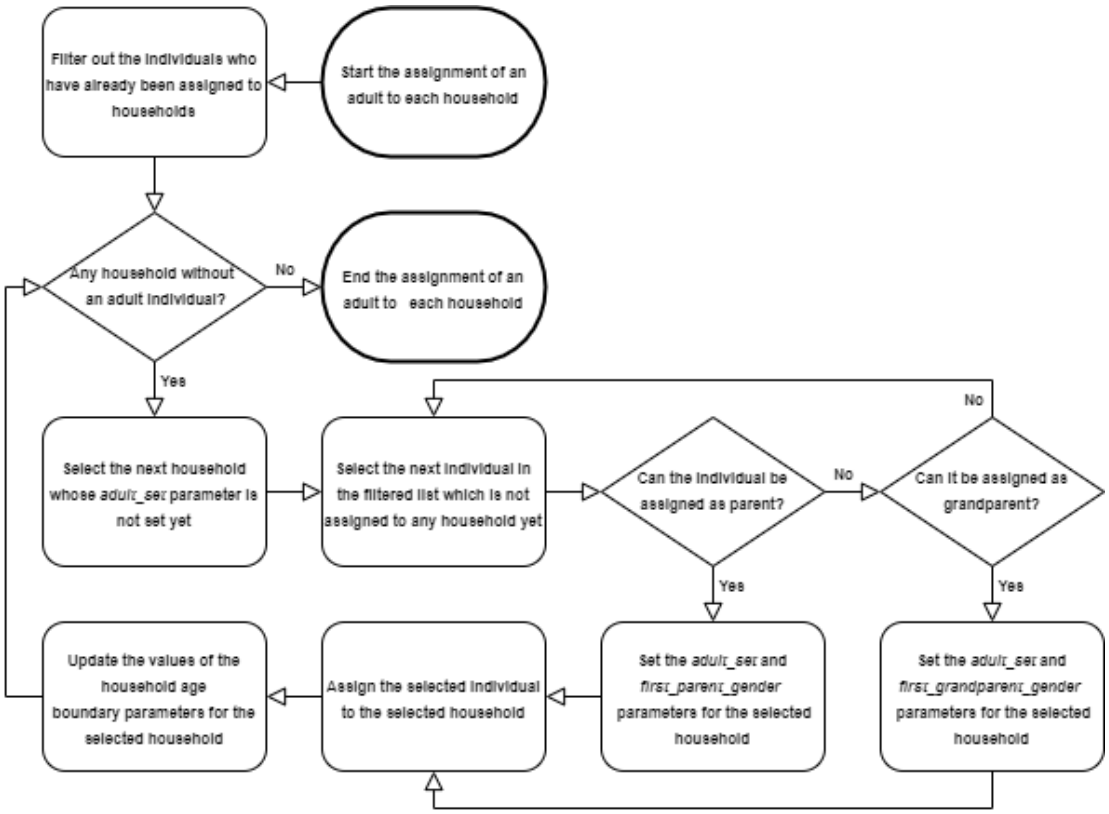

Figure 4. The flowchart diagram of the MC sampling algorithm making the first adult assignments for the households.

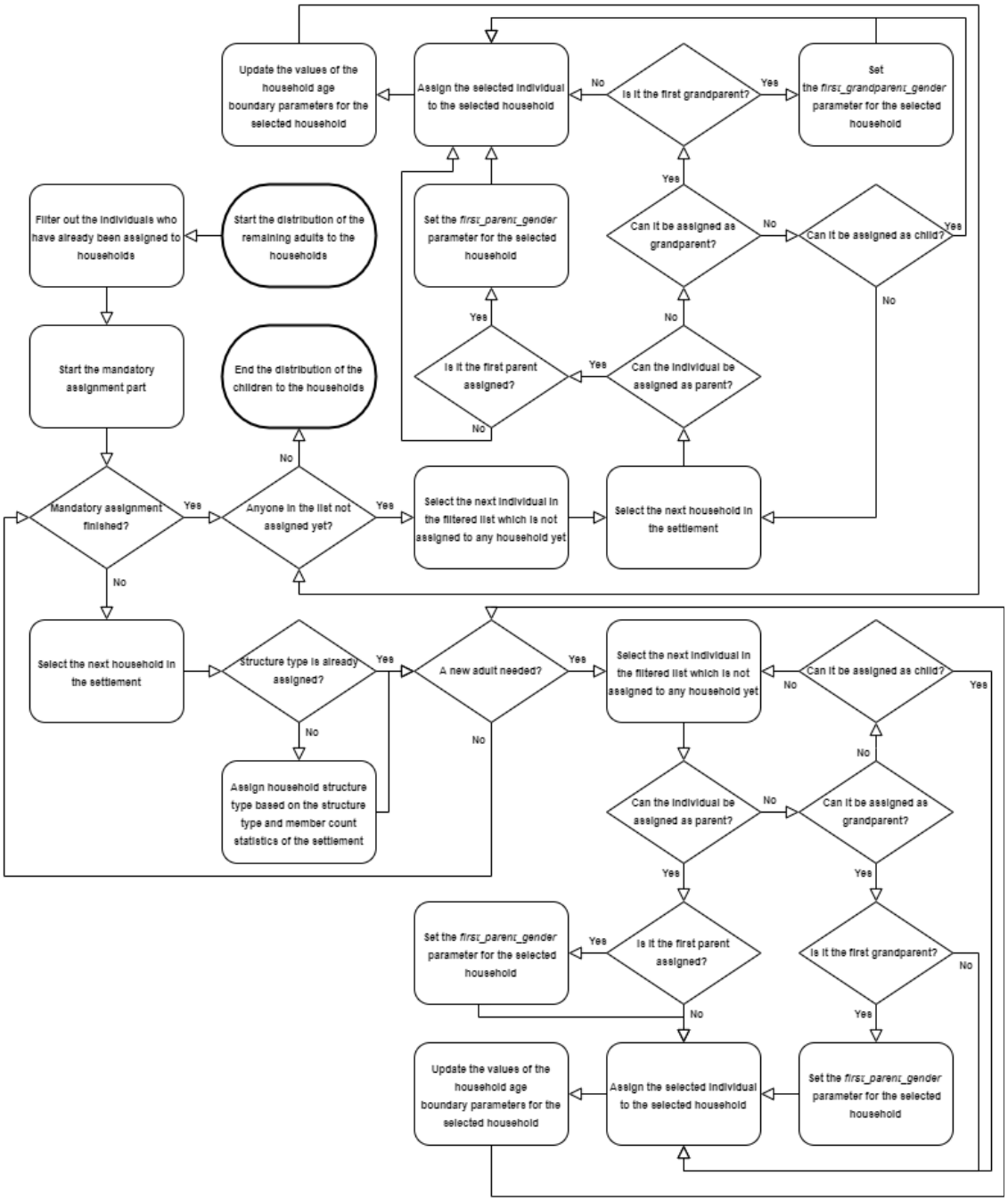

Figure 5. The flowchart diagram of the MC sampling algorithm distributing the remaining adults to the households.

3.2 Distribution of the working individuals into the enterprises

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 [9]. At this step the member variables *min_employees* and *max_employees* are also set for

the generated enterprises. The data provided in [9] includes all private enterprises, as well as, state and local government organizations.

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, 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 [10] 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 flowchart diagram of the MC sampling procedure assigning the work activity to the individuals is presented in Figure 6.

Figure 6. The flowchart diagram of the MC sampling algorithm assigning the working individuals to the enterprises.

3.3 Distribution of the students into the educational facilities

The educational activities are assigned to the individuals based on the data provided for the project by the Mobility Lab 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:

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

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. 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 7.

```
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 7. Algorithm selecting n individuals residing closest to the point_of_interest.

References

[1] “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].

[2] “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].

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

[4] “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].

[5] “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].

[6] “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].

[7] “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].

[8] “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].

[9] “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].

[10] “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].

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

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