Generation of a synthetic population for agent-based transport modelling with small sample travel survey data using statistical raster census data

Samuel Felbermair\textsuperscript{a,b,*}, Florian Lammer\textsuperscript{a,c}, Eva Trausinger-Binder\textsuperscript{d}, Cornelia Hebenstreit\textsuperscript{a,c}

\textsuperscript{a} JOANNEUM RESEARCH, Graz, Austria
\textsuperscript{b} University of Natural Resources and Life Sciences (BOKU), Vienna, Austria
\textsuperscript{c} Institute of Highway Engineering and Transport Planning, Graz University of Technology, Graz, Austria

Abstract

This paper presents a step-by-step method to generate a synthetic population for agent-based transport modelling as input to MATSim software, which requires an activity chain for each agent. We make use of high spatial resolution statistical raster (250 m) census data, applying all calculations at this scale. Due to the small sample, size of travel survey data an Iterative Proportional Fitting method is not suitable. Therefore, we devise a method utilizing Bayesian networks, maximum likelihood and Markov Chain Monte Carlo simulation to reproduce attribute distribution and fit to raster margins. Stratified sampling along households is employed to generate activity chains for the synthetic population.

Keywords: agent-based modelling; transport modelling; MATSim; population synthesis; Markov Chain Monte Carlo; Bayesian Networks; data integration;

1. Introduction

Agent-based transport modelling is based on the idea of modelling the behavior of human agents, who take individual decisions as to when, how and where they move while being influenced by the decisions of other agents [1], [2]. It incorporates the interaction of several choices taken in the process like route, mode, time and destination better than classical four-step transport modelling [2]. Person and household characteristics such as age, gender and household composition inform these decisions. Therefore, it is important to consider these influencing factors in the synthetic population for agent-based transport modelling. The spatial resolution of agent-based models is higher due to working with coordinates as opposed to traffic zones. Therefore, in our method we make use of high-resolution population data.

We build a model using the “Multi-Agent Transport Simulation” MATSim [2] software, which has specific requirements for the demand input. Trip generation and trip distribution are not part of the software but must be supplied exogenously. Each agent must be assigned (one or more) activities (such as “work” or “home”) with topological coordinates (or an associated link in the network), an initial travel mode (which may be changed during simulation by the replanning tool) and a start or end time for each activity (see chapter 2.2.2 in [2]). This paper deals only with population generation and does not present the simulation results.

The data needed to provide these attributes in a synthetic population is twofold: The first are observations on people’s travel behavior that include person characteristics, usually travel surveys or censuses. The second is a geographical distribution of persons and households and their attributes, since they are not evenly distributed across space. In only very rare cases, one dataset can fulfill both requirements. Most travel surveys do not supply a sample size large enough to scale to the full population. Moreover, the spatial resolution of available data is increasing. Population statistics become available on a high-resolution raster scale as opposed to larger (sub)community boundaries. With these precise locations at hand in agent-based modelling, more accuracy in terms of traffic origins and destinations as well as trip lengths and durations can be achieved. Therefore, methods are needed to combine these two datasets in order to generate a synthetic population [3].

A typical method in population synthesis for agent-based modelling are various implementations of Iterative Proportional Fitting (IPF) (cf. [4]–[8]). However, IPF has three major problems: (a) integerization, (b) synthesizing at only one level and the so-called (c) zero-cell problem [8]. Since agent-based models work with individual agents, integer results are necessary. IPF does not yield integer results; therefore, some...
kind of rounding approach is required subsequently. Several authors ([3], [4], [8]) have developed methods for integerization. In our case, with very small population numbers per raster zone, the error from such procedures is still non-negligible. IPF in its traditional form works only on one level, i.e. either household or person marginals, although methods have been developed to tackle that problem [9]. Third, there are several “zero-cell” problems in IPF (cf. [4], [7], [8]). The reference table, from which probabilities are obtained, may contain zeroes either in a cell, leading to slow convergence, or the margins, leading to undefined results (division by zero). These cases occur increasingly more often with smaller sample size of travel survey, from which the reference table is obtained, and increasing number of attributes and categories considered. The zero-cell problem also occurs significantly more frequently with smaller zones, i.e. higher spatial resolution. In order to avoid this problem the suggestion for using IPF would be to aggregate zones and thereby abandon the use of this disaggregated information [8]. This counteracts the benefit of increasingly more information being available at a higher spatial resolution, such as the 250 m raster data in our case. We consider the quality of the census raster data, in terms of true numbers and spatial distribution, to be very good. It would be a loss in quality of the model not to use this information. The sample size of travel survey we are working with is small. Therefore, we conclude that in our case IPF is no suitable methodical choice.

The simulation-based approach proposed by several authors [10], [11] also does not take into account a high-resolution spatial distribution as we have in our data. Rather than distribute the population geographically after synthesis we aim to keep the accuracy of records and location throughout. Incorporating the spatial distribution into population synthesis is an advantage in transport modelling since many mobility decisions are taken based on the local transport infrastructure. Most papers on population synthesis unfortunately do not reveal their level of spatial resolution or discuss how geographical distribution is achieved. Therefore, we suggest a different combination of methods to make best use of available data, which we perform on 250 m raster zones.

In this paper, we present a step-by-step approach to generate a synthetic population, which fulfills all requirements for a so-called “plans” file in MATSim. It does not only generate the population in households with person attributes but also includes assignments of activity and trip chains and their destinations. Our procedure consists of the three main parts (1) household generation (2) assignment of activity and travel chains and (3) destination choice in a total of seven steps. We use Bayesian networks to obtain a probability table from a small survey sample. Subsequently a Markov Chain Monte Carlo simulation is applied in each raster cell. By stratified sampling activities and trips are assigned, for which destinations are looked up in a commuter matrix or matched with OpenStreetMap facilities. The seven steps are explained in detail in chapter 3, in chapter 2 we describe our data.

2. Data

The proposed step-by-step method was devised while setting up an agent-based transport model for Carinthia, a state of Austria that is classified as ‘intermediate’ and ‘predominantly rural’ using the urban-rural typology of Eurostat [12]. The available data is the main factor to determine which methods can be applied. As mentioned in the previous chapter we need travel survey information as well as a geographical distribution of persons, households and their attributes for generating a synthetic population. A third important dataset are origin-destination (O-D) matrices. Trip distribution (or destination assignment as we call it) is significantly improved by O-D matrices based on observations, such as are often provided by commuter matrices, compared to calculations based on production and attraction of trips.

2.1. Austrian travel survey “Österreich unterwegs 2013/2014”

The activity and trip information is obtained from travel survey data. The most recent is “Österreich unterwegs 2013/2014” [13] with a sample size of 17,070 households and 38,220 persons. However, for the state of Carinthia there are only 782 households and 1,733 persons in the sample (even less after excluding missing values and households who were only questioned on weekends). This is about 0.3 % of the overall population (548,562 in 2016) of Carinthia, which is a quite small sample (cf. [14, p. 3]). We apply weighting factors present in the data to scale to the full population that originate from the register-based census of 2012 and are thus comparable to the data in statistical raster units. The weights take into account totals along age, work status and household size, among other attributes [15, p. 36], which we make use of in population synthesis.

2.2. Population data in regional statistical raster units

Population and household statistics from register-based census data are available by Statistics Austria, on a yearly basis, at a 250 m raster spatial resolution. However, there is no information which persons form a household together. The attributes are given in separate tables, so we are lacking a joint distribution (with the exception of age groups and work status by gender). In the state of Carinthia, there are 23,931 inhabited raster cells. Being a partly mountainous region there are many cells with a very low population. Because of data privacy there is no information available if there are less than four persons or households in a cell. Twenty-four per cent of all cells are excluded from calculation due to this restriction. These however, constitute only 2 % of the population and 3 % of households. Table 1 shows the quantiles and mean of the total number of persons and households per raster cell in the data for Carinthia. We use the attributes: age, gender, work status and household size from 2016.

<table>
<thead>
<tr>
<th>Quantiles</th>
<th>0 %</th>
<th>25 %</th>
<th>50 %</th>
<th>75 %</th>
<th>100 %</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>22</td>
<td>1045</td>
<td>23.4</td>
</tr>
<tr>
<td>Households</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>598</td>
<td>10.4</td>
</tr>
</tbody>
</table>

2.3. Commuter matrix

Commuter origin-destination matrices for employed persons and students are available on the same 250 m raster. Its basis is labor market data, namely all employment registered, regardless of working hours, for the reference date, 31 October 2016. It does not represent a daily mobility pattern. It also does not contain any person or household attributes. Residence and being employed or in education are the only attributes how an origin-destination relation may be assigned to a specific person. Due to its high spatial resolution, it offers valuable information on real origin-destination relations.
3. Methodology

Figure 1 shows an overview of the consecutive steps in the generation process: the data are represented by the rhombus shaped dark grey boxes, the methods applied are the white curved boxes and the intermediate results are denoted by the light grey rectangular boxes.

Part (1) household generation consists of steps 1 to 3. We first extract a joint probability of person and household characteristics using Bayesian networks (BN). Then we solve a Maximum Likelihood equation in each raster cell by its margins, to obtain a starting distribution for a Markov Chain Monte Carlo (MCMC) simulation to generate households. Further household persons are sampled using the BN partial probabilities and cell margins, similar to [10]. Part (2) assignment of activity and travel chains consists of step 4 only. Stratified sampling is used for specifying trip information for each person, which also includes parameters such as car availability. In part (3) destination choice steps 5 to 7 are undertaken to assign destinations for activities, depending on data availability based either on commuter matrix or facilities at raster cell level. In the final step, all are converted to coordinates, which MATSim associates with the closest network link.

3.1. Household generation

The population for the Carinthia MATSim model is constructed on the basis of households, since many relevant mobility decisions, such as bringing kids to school or which person of the household does the daily shopping, are made at the household level [9], [16]. Since the Austrian register-based census data does not group persons into households (as opposed to Swiss data, cf. [6]) this relation has to be synthesized. In order to generate the synthetic households, a household reference person needs to be identified in the travel survey data. The reference person is selected using a hierarchical definition from Statistics Austria. Active participation in the labor market is the first defining condition, with education status and age sequentially being applied to break ties or designate a person if the previous conditions are not met. The assignment is modelled as close as possible in the travel survey to the definition used by Statistics Austria (with some deviation in categories of attributes in the two datasets).

For the statistical matching, joint probabilities of the attributes are necessary. We obtain them from the weighted travel survey data through Bayesian networks (Step 1 in Figure 1). This method has been applied in population synthesis for cases with limited data (cf. [17], [14]). We are interested in joint probability distributions of several attributes for different household members leading to a high number of combinations. Bayesian networks are a graphical method for categorical attributes, to identify a hierarchical structure of relations. Its advantage in computing a joint probability table is in reducing the amount of contingency tables computed. Since the graph is hierarchically organized, the joint probability can be obtained by multiplication of partial probabilities. It saves computation time when many attributes are combined. Another advantage of this method is that we can interpret the results graphically. This allows running the same method for varying numbers of attributes and compare the results for decisions on which attributes to include. We use the same search algorithms and score function as proposed by Sun and Erath [17]: tabu search with Akaike information criterion (AIC) to assess the goodness of fit working with bnlearn R package [18], [19]. We calculate three models using the attributes age groups, gender, work status and household size for (1) household reference persons, (2) second household persons and (3) further household persons, each containing also the previous person(s) attributes. Networks may be expanded with further attributes. After revealing the model structure with unweighted sample data, the conditional probabilities are extracted using weighted survey data.

Step 2 comprises of the application of two methods in sequence: maximum likelihood and Markov Chain Monte Carlo (MCMC). First, a maximization problem using linear programming is run for each raster cell. The joint probability distribution \( p(\text{household size}(i), \text{age groups}(j), \text{gender}(k)) \) of household reference person serves as the objective function. The census data provide constraint margins \( c \), which have to be met exactly by population \( x \). In mathematical expression:

\[
\text{max} \sum_{i,j,k} (p_{i,j,k} \cdot x_{i,j,k})
\]

\[
\sum_{j,k} (x_{i,j,k}) = c_i; \ i = 1…3
\]

\[
\sum_{i,k} (x_{i,j,k}) = c_j; \ j = 1…5
\]

\[
\sum_{i,j} (x_{i,j,k}) = c_k; \ k = 1…2
\]
The Maximum Likelihood (ML) result is used as starting seed for a Markov Chain Monte Carlo (MCMC) simulation. In the ML result, the most likely combinations are overrepresented when the population is too small to represent also less likely combinations, which is often the case in our data. However, it is fast to compute and is used to reduce the number of steps in MCMC, which is computationally expensive. The MCMC approach applied here differentiates from those employed by [10], [11] and others using similar approaches. We do not synthesize a population that is then adjusted to margins. Instead, we employ MCMC to look for a different combination of attributes compared to the Maximum Likelihood result, which already fits the margins and that constraint as well as that of an integer result, is maintained during Markov simulation. In 100 steps, a different result fitting within the margins of each cell is simulated. It takes about 15 hours on a 4-core 1.8 GHz office laptop with 32 GB RAM to generate 241,669 household reference persons with only three attributes in a two-dimensional distribution (age groups and gender are cross tabulated, work status is added later separately). Depending on computational resources, more attributes can be assigned in this step through MCMC simulation. Figure 2 shows the resulting distribution of household reference persons with the attributes age groups and gender (2M to 6F, age group 1 are underage persons and cannot be household reference persons) and household size (1, 2, 3 or more persons). The left bar is the weighted count from the travel survey, which reflects the true distribution since the weights were calculated using census data. The middle bar is the initial maximum likelihood result, whereas the right bar shows the MCMC result. It is visible that MCMC improves the fit. In multi-person households, further persons equalize the deviations especially regarding to gender.

In step 3, we add further persons with attributes age group, gender and work status to the synthetic households. Given the attributes of the reference person there exists a probabilities vector towards the attributes of the further household persons (if household size is greater one). This is used to draw persons’ attributes from the vector of undistributed persons’ attributes in that cell. If the probability of drawing from existing attributes is zero, as may occur due to not all existing combinations being present in the sample, an attribute from the available pool is drawn. This procedure is repeated for as many attributes as there are marginal raster data. Table 2 shows the most recurrent reference and second household person pairs by age group, gender and work status and their respective total numbers and percentages in the synthetic population as well as the travel survey. Although there are differences in the resulting number of households of a specific type, the most recurrent combinations of the travel survey appear among the top in the synthetic population. The differences are also explained by the fact that the synthetic population contains more than twice as many combinations due to its margins adjusted to the raster census data. Underrepresented and even missing household combinations in the travel survey are present in the synthetic population. This can be seen specifically in the case of single mothers with underage children in lines 10 and 11 of Table 2. The total in the synthetic population is close to the number in the census population of Carinthia despite being severely underrepresented in the sample of the travel survey. The cell margins led to creation of these despite their very low occurrence in the sample, which could not be fully compensated for by weighting. We conclude that the best-fit method works quite well to achieve plausible combinations of reference persons and second household persons.
allow the formation of six groups with at least 59 sample households as can be seen in Table 3. Unfortunately, the sample is too small to also consider a regional typology (like urban/rural). It is evident that households with children are underrepresented since they constitute only about 10% of our sample but about 30% of the population. This is partly explained by the fact that children under 6 years are not included in the travel survey.

For trip generation the data from the travel survey, as well as the synthetic population, are split into these groups. For each household in the synthetic population a household of the same behaviourally homogeneous household group is selected randomly from the travel survey. Each household in the synthetic population has exactly one equivalent household from the travel survey. Inversely, trips of one household from the travel survey are assigned to multiple households in the synthetic population. As there is already a socio-demographic preselection through the classification of the household categories, within the households only a distinction between children and adults is conducted, although gender could be considered for adults. For each household member of the synthetic population a whole trip chain is picked from the equivalent household. The trip characteristics assigned are activity type, activity duration, activity end time (required by MATSim, equals trip start time), mode, trip length, as well as the availability of car, bike and public transport, denoted by the ownership of a transit pass. No origins or destinations are assigned in this step; therefore, multiple selections of households or persons within do not affect unique trip chains of the final synthetic population.

There are spatially induced differences in the mobility behavior between cities and rural areas, especially as regards trip lengths and transport modes. Since we use trip lengths to determine destinations for activities in following steps, their distribution is important. Among the 554 Carinthian households in our sample, urban areas are underrepresented. If we sample the trips of Klagenfurt’s residents from the rural population, they are too long on average, leading to a compromised destination choice outside the city. Therefore, we apply the same method as described above on travel survey data from the five major cities of Austria with a population between 100,000 and 300,000, of which Klagenfurt is the smallest at the low end of the scale. The trip length distribution among these cities is reasonably similar and compares better than to rural Carinthian areas. Other attributes such as mode choice and car availability

### Table 2: most recurrent combinations of reference and second household person in synthetic population and travel survey

<table>
<thead>
<tr>
<th>reference person</th>
<th>second household person</th>
<th>synthetic population</th>
<th>travel survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>age gender</td>
<td>employment status</td>
<td>no. of hh</td>
<td>percent</td>
</tr>
<tr>
<td>30-49 M employed</td>
<td>30-49 F employed</td>
<td>16054</td>
<td>10.5%</td>
</tr>
<tr>
<td>30-49 F employed</td>
<td>30-49 M employed</td>
<td>9766</td>
<td>6.4%</td>
</tr>
<tr>
<td>65-84 M retired</td>
<td>65-84 F retired</td>
<td>9154</td>
<td>6.0%</td>
</tr>
<tr>
<td>30-49 M employed</td>
<td>30-49 F</td>
<td>6348</td>
<td>4.2%</td>
</tr>
<tr>
<td>50-64 M retired</td>
<td>50-64 F retired</td>
<td>5239</td>
<td>3.4%</td>
</tr>
<tr>
<td>50-64 M retired</td>
<td>50-64 F</td>
<td>5038</td>
<td>3.3%</td>
</tr>
<tr>
<td>50-64 M employed</td>
<td>50-64 F employed</td>
<td>4661</td>
<td>3.1%</td>
</tr>
<tr>
<td>65-84 M retired</td>
<td>65-84 F other</td>
<td>3289</td>
<td>2.2%</td>
</tr>
<tr>
<td>50-64 M retired</td>
<td>50-64 F other</td>
<td>3263</td>
<td>2.1%</td>
</tr>
<tr>
<td>30-49 F employed</td>
<td>0-14 F student</td>
<td>2945</td>
<td>1.9%</td>
</tr>
<tr>
<td>30-49 F employed</td>
<td>0-14 M student</td>
<td>2447</td>
<td>1.6%</td>
</tr>
</tbody>
</table>

### Table 3: homogeneous household groups and their sample sizes in Austrian travel survey

<table>
<thead>
<tr>
<th>household size</th>
<th>work status combinations</th>
<th>kids</th>
<th>[n]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Single</td>
<td>Education/working/other</td>
<td>no</td>
<td>67</td>
</tr>
<tr>
<td>2 Single</td>
<td>Retired</td>
<td>no</td>
<td>91</td>
</tr>
<tr>
<td>3 Min. 2 persons</td>
<td>Education/working/other</td>
<td>no</td>
<td>153</td>
</tr>
<tr>
<td>4 Min. 2 persons</td>
<td>Retired and Education/working</td>
<td>no</td>
<td>67</td>
</tr>
<tr>
<td>5 Min. 2 Persons</td>
<td>Retired</td>
<td>no</td>
<td>117</td>
</tr>
<tr>
<td>6 Min. 2 Persons</td>
<td>Education/working/retired/other</td>
<td>yes</td>
<td>59</td>
</tr>
</tbody>
</table>

**Sum of households**: 554

Sample size has to be considered in the definition of the groups since we are drawing from those households. If there were too few in one group, the characteristics of this small sample could become overrepresented in the synthetic population. Limiting the data to only working-day trips reduces the sample size to 554 households. The three chosen household characteristics

3.2. Assignment of activity and travel chains

Agents in MATSim need activity and travel chains. There are two different methods that have been applied in activity assignment to synthetic populations. Discrete choice modelling [11], [20], [21] and statistical matching techniques [6], [22]–[24] We use the latter approach because it allows us to preserve the combinations of activities within households, which also enables us to identify ride sharing within households. Activity and trip characteristics correlate with the socio-demographic data synthesized in the steps 1 to 3. In step 4 stratified sampling, a statistical matching technique, is performed along behaviourally homogenous household groups to assign trip attributes. This approach is similar to [6] but on a household basis as opposed to individuals. However, due to the limited sample size we use less attributes.

### Table 3: homogeneous household groups and their sample sizes in Austrian travel survey

<table>
<thead>
<tr>
<th>household size</th>
<th>work status combinations</th>
<th>kids</th>
<th>[n]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Single</td>
<td>Education/working/other</td>
<td>no</td>
<td>67</td>
</tr>
<tr>
<td>2 Single</td>
<td>Retired</td>
<td>no</td>
<td>91</td>
</tr>
<tr>
<td>3 Min. 2 persons</td>
<td>Education/working/other</td>
<td>no</td>
<td>153</td>
</tr>
<tr>
<td>4 Min. 2 persons</td>
<td>Retired and Education/working</td>
<td>no</td>
<td>67</td>
</tr>
<tr>
<td>5 Min. 2 Persons</td>
<td>Retired</td>
<td>no</td>
<td>117</td>
</tr>
<tr>
<td>6 Min. 2 Persons</td>
<td>Education/working/retired/other</td>
<td>yes</td>
<td>59</td>
</tr>
</tbody>
</table>

**Sum of households**: 554

Sample size has to be considered in the definition of the groups since we are drawing from those households. If there were too few in one group, the characteristics of this small sample could become overrepresented in the synthetic population. Limiting the data to only working-day trips reduces the sample size to 554 households. The three chosen household characteristics
become biased by this procedure since other cities offer better public transport services. Mode choice is only an initial information that may be changed during the simulation and therefore does not constitute a significant error. Car availability would be biased in the opposite direction when sampled from rural households. There are similar deviations for the second biggest city of the province; however, there is no suitable spatial category to select data from. Generally the mobility choices might be better captured with a different classification of regional typology as shown by Lammer [25]. Due to limited sample size, it cannot be applied to the data of Carinthia. A possibility to overcome the sample size issue would be to include data from the travel survey of other provinces of Austria, provided follow a similar pattern.

3.3. Assignment of destinations

In step 5 destinations for the paired activities home-work and home-education (considered as primary activities in the model) are assigned based on known origin–destination relations from the commuter matrix. This approach is similar to [6], however there are no socio-demographics associated with the Austrian commuter matrix. For each person the destination with the best travel distance fit is allocated among the multiple relations from a home raster cell to different work or education raster cells. Therefore, we calculate the Euclidean distance to all destination cells for all persons living in a raster cell and multiply it by a detour factor of 1.4 ($\sqrt{2}$, see taxicab geometry) to estimate the travel length on the road network. Compared with the assigned trip length from the travel survey the destination with the smallest difference in length is assigned. The commuter matrix shows for each relation how many commuters use this relation. Based on this information and keeping count of already assigned destinations we make sure that no relation is assigned disproportionately.

In step 6 destination assignment of secondary activities (all other than work and education) is conducted. In accordance with the previously assigned activity chains for each agent, we generate cell potentials for each activity. This total activity potential for each cell is derived from the amount and type of activities present in OSM data if possible; otherwise, a cell potential is estimated by the preassigned activity chains and the cell potential. For distances above >30 km we assume that the cell distance is less well defined with the best travel distance fit defined by the detection factor 1.4 for comparison. In addition, a routing between cell center coordinates is carried out to improve the destination assignment and account for obstacles like mountains or lakes.

This is done on the network specified in the Graph Integration Platform (GIP) [27] only for trip chain distances less than 30 km. For distances above >30 km we assume that the percentual error in travel length introduced by obstacles is negligible. For the routing we allow a constant offset of half the diagonal of the cell (±177 m), and a 10% error in distance that is increased stepwise up to 50%. If no cells in the defined torus match the activity criteria, meaning no cell with corresponding potential is available for the specified travel distances, the area of the torus is increased stepwise up to limits defined for each cell search method separately in order to satisfy runtime efficiency. Above these limits, activity potentials are neglected and a larger municipality area matching the given travel distance is selected.

During the destination assignment, we also check for shared activities and trips within households, in the form that agents with mode car passenger are matched to car drivers when having the same departure time, travel distance as well as origin and destination activity, for which the activity “escort” is assumed to match every other defined activity. Via this procedure, we are able to assign same origin or destination coordinates for 37% of all car passenger trips.

Finally, in step 7, we match home coordinates to facilities present in an address register obtained from the Government of Carinthia. All other coordinates are assigned to suitable facilities present in OSM data if possible; otherwise, a random coordinate is assigned.

4. Results analysis

Results comparisons refer to the initial demand fed into the model not simulation results. During the simulation, some of these parameters may be changed, especially modes, distances and durations. Figure 4 shows trip demand distributions of the synthetic population against the travel survey data. There are slightly more shorter trips in the synthetic population, which is good since it is known that short trips have been underestimated in the travel survey [28].

Figure 5 shows the resulting activity distribution between the synthetic population and the travel survey. There are only slight deviations in the total sum of activity distribution; work is underrepresented while escort is overrepresented, however both are within the 95% confidence interval.

Figure 6 shows that the initial modal split (mode choice is part of the simulation process in MATSim) of the synthetic population fits reasonably well with the modal split of the travel survey.

The resulting synthetic population displays a good geographic distribution through the approach with the raster cells. As regards destination choice, a reasonably good fit was attained between the simulation and traffic count data given that transit traffic was neglected.
Figure 3: trip distance matching with raster cells for secondary activities destination assignment

Figure 4: travel distances by activity for travel survey and synthetic population

Figure 5: Comparison of activity distribution between travel survey and synthetic population for Carinthia on a working day

Figure 6: Comparison of modal split between travel survey and synthetic population for Carinthia on a working day
5. Conclusion

It is a common problem to be faced with limited travel survey data. We show how this can be compensated if spatially high-resolution population data are available. A higher emphasis is put on the more robust data. We present a method for generating a synthetic population when the data is not suitable for the most-used IPF procedure. This is quite common, since sample sizes of travel surveys are often small and high-resolution spatial data are available in many places. Aiming to produce a good geographical distribution an emphasis is put on doing computations on a raster cell level throughout. The method produces satisfying results for the most important attributes such as activity distribution and trip lengths. Partly the method is able to make up for underrepresentation in the travel survey, such as single mother households. The method is presented as an easy to follow seven step approach that might assist others in producing all the input needed for MATSim agent-based transport modelling software including trip generation. This approach may be used in cases where sample size of travel survey is small and a high resolution on the spatial scale even in a rural area shall be achieved.

References


[26] OpenstreetMap (OSM), overpass turbo: a web based data mining tool for OpenStreetMap using Overpass