Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 170 (2020) 273–280

Procedia Computer Science

www.elsevier.com/locate/procedia

The 11th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 6-9, 2020, Warsaw, Poland

Generating synthetic population with activity chains as agent-based model input using statistical raster census data

Samuel Felbermair[a]*, Florian Lammer[a], Eva Trausinger-Binder[b], Cornelia Hebenstreit[bc]

[a] *LIFE Institute for Climate, Energy and Society, Joanneum Research, 8020 Graz, Austria*
[b] *LIFE Institute for Climate, Energy and Society, Joanneum Research, 9020 Klagenfurt, Austria*
[cb] *Institute of Highway Engineering and Transport Planning, Graz University of Technology, 8010 Graz, Austria*

Abstract

Agent-based transport modelling needs more detail on the synthetic population compared to conventional transport models, as activity chains are required. In many cases, however the sample size of travel surveys from which to gain activity chains is small. Using Bayesian networks and Markov Chain Monte Carlo as well as stratified sampling, we show how a population with activities plans can be generated using limited survey data.
Moreover, this paper presents a method for using statistical raster (250 m) census data for all activities and facilities, which guarantees a high spatial resolution. The synthetic population was developed for the predominantly rural to intermediately urban state of Carinthia in Austria. Realistic travel plans were assigned to each agent, considering trip dependencies between household members as well as correlations between socio-demographic attributes and travel behaviour. The resulting synthetic population includes agents with a sequence of activities for 24 hours. The activities and trip length distributions of the simulated population fit the survey data well. The simulation results fit the traffic counts.

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

* Corresponding author. Tel.: +43 316 876 7673; fax: +43 316 876 7699
E-mail address: samuel.felbermair@joanneum.at

10.1016/j.procs.2020.03.040

1. Introduction

The synthetic population was developed for Carinthia, a state of Austria that is classified as 'intermediate' and 'predominantly rural' using the urban-rural typology of Eurostat [1]. The chosen simulation tool is the "Multi-Agent-Transport Simulation" (MATSim) [2]; however, this paper deals only with the population generation. MATSim has specific requirements for the demand that contains both trip generation and trip distribution. Each agent must be assigned (one or more) activities (such as "work" or "home") with topical coordinates (or an associated link in the network), an initial travel mode (which may be changed during the simulation by the replanning tool) and a start or end time for each activity (see chapter 2.2.2 in [2]).

Data availability is the main limiting factor in population synthesis, leading to the development of different methods. A typical method are various implementations of Iterative Proportional Fitting (IPF) (cf. [3] [4] [5] [6]). We decided against such an approach due to the limited sample size of travel survey data, which leads to a zero-cell problem (cf. [3] [6]). Instead we employ a Markov Chain Monte Carlo (MCMC) method and also use a hierarchical approach as suggested by [7]. Bayesian networks, which recently were applied in the field of population synthesis (cf. [8] [9]), were used to estimate the joint probability distribution.

Conversely to the aforementioned approaches we first generated a distribution that fits the marginal totals on socio-demographic attributes. This is because the reliability of register-based census data is greater than of travel survey data. We then proceeded to adjust the fit of the initial distribution, obtained through maximum likelihood estimation, to the joint probability distribution by MCMC simulation.

2. Data

We briefly characterize the available data for our model here while the next chapter describes the methods applied.

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. Population and household data are not linked; there is no information which persons form a household together. The attributes are not given in a joint distribution but separate tables (with the exception of age and work status each separated by gender). Because of data privacy there is no information available if there are less than four persons or households in a cell. In the state of Carinthia, there are 23,931 inhabited raster cells. We used data for the attributes: age, gender, work status and household size from 2016.

Commuter matrix

A commuter origin-destination matrix for employed persons and students is available on the same 250 m raster. Its basis is labour market data, namely all employment registered by social security number, regardless of working hours, for the reference date, 31 October, every year. Therefore, 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 variables how an origin-destination relation may be attributed to a specific person.

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" [10] 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. This is about 0.3 % of the overall population (548,562 in 2016) of Carinthia, which is a quite small sample (cf. [9, p. 3]). This leads to limitations in the assignment of trip characteristics since we are looking at combinations of variables. We applied weighting factors to scale the data to the full population. The weighting totals originate from the register-based census of 2012 and are therefore comparable to the data in statistical raster units. The weights take into account totals along age, work status and household size, among other attributes, which we make use of in population synthesis [11, p. 36].

3. Methodology

We included the spatial distribution into the first step of population synthesis. This is an advantage in transport modelling, since we do not have to distribute our population to home locations later on. We also incorporate activity choice and its associated destination choice in the population synthesis to provide all required attributes for direct use in MATSim software.

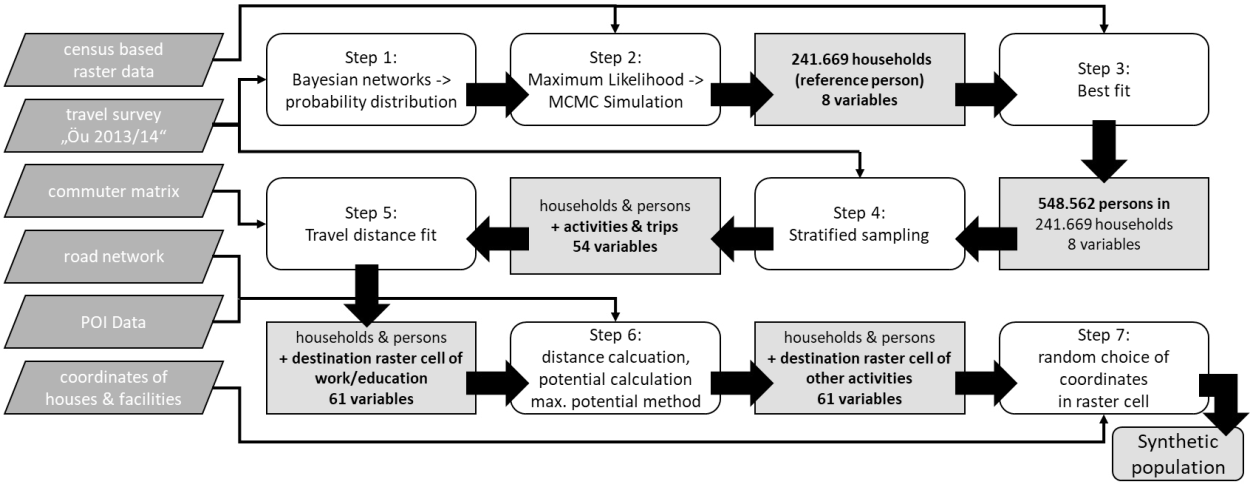

Figure 1: overview of methods applied to create demand for MATSim modelling

Figure 1 shows an overview of 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.

3.1. Household generation

The population for the Carinthia MATSim model was constructed on the basis of households, since many relevant mobility decisions, such as bringing kids to school or only one person of the household doing the daily shopping, are made at the household level. Since the Austrian register-based census data does not group persons into households (as opposed to Swiss data, cf. [5]) this relation has to be synthesized.

In order to generate the synthetic population, a reference person needed to be identified in the travel survey data. The reference person is selected using a stringent hierarchical definition from Statistics Austria, which takes into account several attributes. Active participation in the labour 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 was modelled as close as possible in the travel survey after obtaining the definition used by Statistics Austria.

For the statistical matching, joint probabilities of the attributes are necessary. We obtained them from the weighted travel survey data (Step 1 in Figure 1). Because we are interested in joint probability distributions of several attributes for different household members we had to analyse a series of interactions. We employed Bayesian networks because this method can identify which categorical variables are directly related to one another. This method has been applied in population synthesis for cases with limited data [8] [9]. Another advantage of this method is that we can interpret the results graphically. This allowed us to quickly run the same method for varying numbers of variables and compare the results. Tabu search (as in [8]) was used with Akaike information criterion (AIC) score to assess the goodness of fit. Three models were calculated for (1) household reference persons, (2) second household persons and (3) further household persons. In Figure 2 the third network, which includes the attributes age and work status by gender for up

to four persons in a household, is presented. Networks may be expanded with further attributes like income or driving license. After revealing the model structure with unweighted survey data it is easy to extract conditional probabilities with the weighted sample. These were used in the following steps.

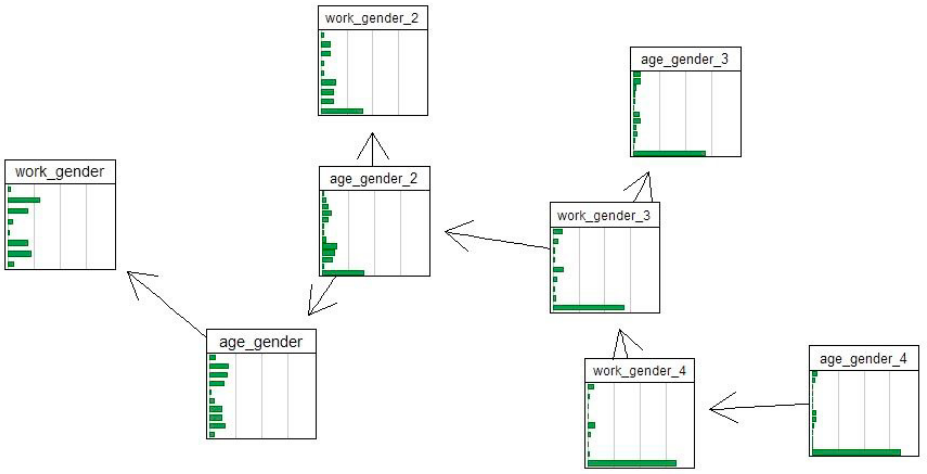

Figure 2: Bayesian network of age, gender and work status for up to four household persons

The marginal distribution of age, gender, work status and household size for the household reference persons in each raster cell is known. The combination of these attributes was calculated using Bayesian networks. A method is needed to achieve a realistic combination given the margins.

Two mathematical methods were applied in step 2. First, a maximization problem using linear programming was run for each raster cell. The joint probability (p) distribution of household size (i), age (j) and gender (k) of household reference person served as the objective function. The census data were used as constraint margins (c), which have to be met exactly by our population (x), representing the true population and its attributes. It can be stated like this:

$$\max \sum_{i,j,k} (p_{i,j,k} * x_{i,j,k})$$

where

$$\sum_{j,k} (x_{i,j,k}) = c_i \qquad i = 1\ldots3$$
$$\sum_{i,k} (x_{i,j,k}) = c_j \qquad j = 1\ldots5$$
$$\sum_{i,j} (x_{i,j,k}) = c_k \qquad k = 1\ldots2$$

This does not lead to a good fit to the joint distribution of these attributes. Therefore a Markov Chain Monte Carlo (MCMC) simulation was run that draws a realization of the conditional distribution given raster cell margins. The maximum likelihood distribution as a starting point reduces computation time for MCMC. This method leads to a significantly better fit as can be seen in **Fehler! Verweisquelle konnte nicht gefunden werden.**.

Step 3 was to add further persons with attributes age, 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 was used to draw persons' attributes from the vector of undistributed persons' attributes in that cell. If the probability of drawing from existing attributes was zero, an attribute from the available pool was drawn. This procedure can be repeated for as many attributes as there are marginal raster data.

The distribution of age, gender, work status and household size was controlled by their spatial location during the generation process. High accuracy of data stemming from a register-based census provides for valid distributions along both socio-demographic attributes as well as locations. Depending on computational resources, more attributes could be assigned in the first step through MCMC simulation rather than sampling.

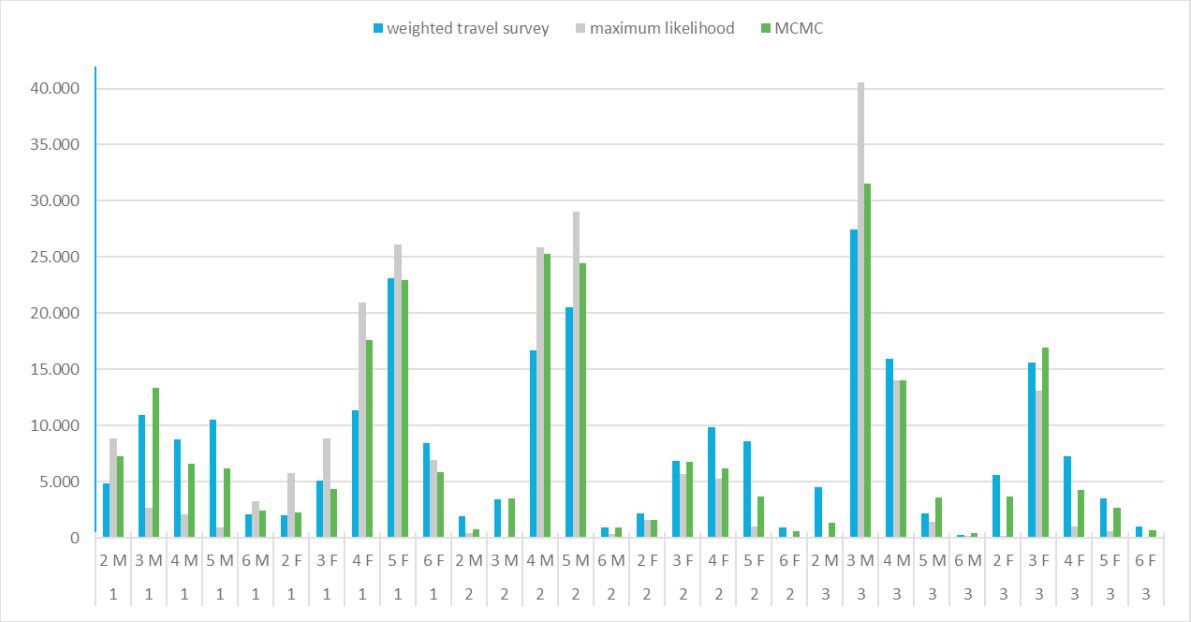

Figure 3: distribution of reference persons' attributes in travel survey, maximum likelihood estimation and MCMC simulation

3.2. Assignment of activity and travel plans

Agents in MATSim need activity and travel plans. These correlate with the socio-demographic data that we simulated in the previous steps. We used stratified sampling along behaviourally homogenous household groups to assign trip attributes. This approach is similar to [5] but on a household basis and using less attributes because of limited sample size.

The mobility behaviours for members of a household are interrelated. To model a realistic synthetic population these dependencies within a household have to be considered. Since there are different socio-demographic types of households, not all households can be considered equal concerning the typical mobility behaviour of the inhabitants. Therefore, six household groups were defined using the following three characteristics:

- household size (single or multi-person)
- employment status (mainly separated along the line of employed vs. retired)
- presence of children in the household (yes, no)

The basic data from the travel survey, as well as the synthetic population, were divided into these groups. The sample size had to be considered during definition of the groups. The sample size was reduced to 554 households for trip generation by limiting the data to only working-day trips. The three chosen household characteristics allowed the formation of six behaviourally homogeneous groups with at least 59 households. Unfortunately, the sample was too small to consider a regional typology.

The household groups were the basis for a stratified sampling (step 4 in Figure 1) to assign trip chains to the synthetic population. For each household in the synthetic population a randomly selected household of the same behaviourally homogeneous household group was selected from the travel survey. Therefore, each household in the synthetic population had exactly one equivalent household from the travel survey. Inversely, trips of one household from the travel survey could be assigned to multiple households in the synthetic population. The trip characteristics assigned were activity type, activity duration, activity end time, mode, trip length, as well as the availability for car, bike and public transport, denoted by the ownership of a transit pass. It has to be noted, that no origins or destinations were assigned in this step, therefore multiple selections of households did not affect unique trip chains for the final synthetic population. As there was already a socio-demographic preselection through the classification of the

household categories, within the households only a distinction between children and adults was conducted. Considering this distinction for each household member of the synthetic population a whole trip chain was picked from the equivalent household. Figure 4 and 5 show the resulting activity and travel distance distributions respectively, of the synthetic population against the travel survey data. Both have a reasonably good fit.

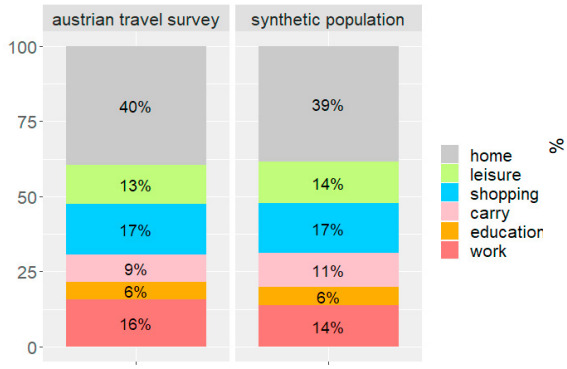

Figure 4: Comparison of activity distribution between travel survey and synthetic population for Carinthia on a working day

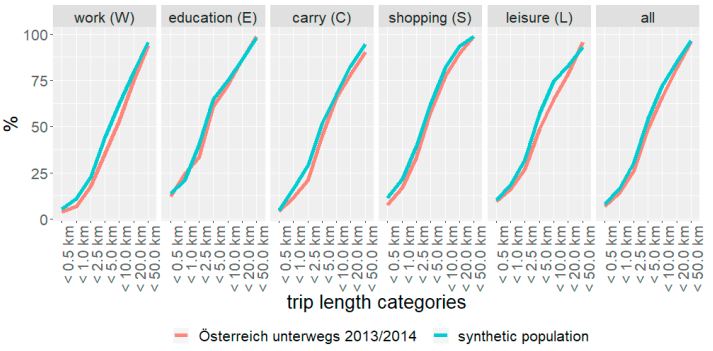

Figure 5: travel distances by activity for travel survey (red) and synthetic population (blue)

Because there are known spatially caused differences in the mobility behaviour between cities and rural areas, the households of the capital city Klagenfurt received special attention. This is especially important since destinations were assigned using the selected trip lengths. Among the 554 Carinthian households, urban areas were underrepresented. Therefore, for Klagenfurt the same method was applied on travel survey data along regional typology, taking as sample all households in the five major cities of Austria with a population between 100,000 and 400,000.

In step 5 destinations for the paired activities home-work and home-education were assigned based on known origin – destination relations from the commuter matrix. For each person the destination with the best travel distance fit was allocated among the multiple relations from a home raster cell to different work or education raster cells. Therefore, for all persons living in a raster cell the beeline distance to all destination cells was calculated and multiplied by a detour factor of 1.4 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 was 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 it was made sure that no relation was assigned disproportionately.

To allow for a valid destination assignment of secondary activities in accordance with the previously assigned activity chains for each agent, we generated cell potentials for each activity kind in step 6. This total activity potential for each cell was derived from the amount and type of facilities obtained from Open Street Map (OSM) [12] data. According to the facility type, we manually assigned different potentials to the facilities. The decision on their potential, based on classes like supermarkets, was done qualitatively. Then we summed up the potentials within a cell to calculate a total potential for each secondary activity. While this was done for the activities shopping and leisure separately, the activity ‘escort is less well defined and hence, besides education facilities and health care centres, also shopping and leisure facilities within the given cell were considered. Some manual adjustments had to be made, since some cells containing e.g. shopping centres were slightly underrepresented.

In the following step, we carried out the actual destination assignment of cells to the predefined secondary activity using the travel distance given in the preassigned activity chains and the cell potential. Between two already known cells of work or education activities, our algorithm searched for all possible destination cells in accordance with the given distances from the activity chains. Using the activity potential as weights one of those cells was chosen randomly. In the case of more than one secondary activity between two known cells, the same was done for whole cell combinations for which the potential was simply summed up. In the case, that a secondary activity was the start or the end of an activity chain, the cell distance could only be matched to the first or last known destination cell. The cell search could also be divided into two steps. First, the algorithm searched for cells using beeline distances plus a

variable offset. The given travel distances from the activity chains were divided by the detour factor 1.4 for comparison. In order to improve the destination assignment and account for obstacles like mountains or lakes, a routing between cell centre coordinates was carried out. This was done for trip chain distances less than 30 km on the network specified in the Graph Integration Platform (GIP) [13]. For the routing we allowed a constant offset of half the diagonal of the cell (177 m), and a 10 % error in distance.

Finally, in step 7, we matched home coordinates to facilities present in an address register obtained from the Government of Carinthia, while all other coordinates were assigned to suitable facilities present in OSM data if possible; otherwise, a random coordinate was assigned. Furthermore, we also checked for shared activities and trips within households, in the form that agents with mode car passenger were 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 could assign similar origin or destination coordinates for 2/3 of all car passenger trips.

4. Results

The resulting synthetic population displays a very good geographic distribution while exactly meeting the known margins along different attributes. The small sample in the travel survey would not have allowed a scaling to full population. The combination of Bayesian networks to extract probability distributions and Markov Chain Monte Carlo simulation lead to a satisfactory result. The distributions of activities and trip lengths in the synthetic population, match well with those of the travel survey. As regards location choice, a reasonably good fit was found between the modelled and traffic count data given that transit traffic was neglected.

5. Conclusion

Since the available data for the generation of a synthetic population varies, different methods have to be applied. It is a common problem to be faced with limited travel survey data to gain detailed information about activity chains, which are needed for agent-based simulation. 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. This approach may be used for other cases with a small travel survey sample and detailed census data. It may also aide in assigning movements obtained from mobile phone data that do not have any socio-demographic variables attached.

References

[1] Eurostat, „Statistics Explained: Urban-rural typology," 7 3 2018. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php/Archive:Urban-rural_typology. [Zugriff am 10 12 2019].

[2] A. Horni, K. Nagel und K. W. Axhausen, „The Multi-Agent Transport Simulation MATSim.," 2016.

[3] K. Müller und K. W. Axhausen, „Hierarchical IPF: Generating a synthetic population for Switzerland," in *ERSA 2011*, 2011.

[4] K. Müller und K. W. Axhausen, „Multi-level fitting algorithms for population synthesis," in *Working Paper*, Zurich, 2012.

[5] P. M. Bösch, K. Müller und F. Ciari, „The IVT 2015 Baseline Scenario," in *Swiss Transport Research Conference*, Monta Veritá/Ascona, 2016.

[6] J. Y. Guo und C. R. Bhat, „Population Synthesis for Microsimulating Travel Behavior," *Transportation Research Record,* pp. 92-101, 2014.

[7] D. Casati, K. Müller, P. J. Fourie, A. Erath und K. W. Axhausen, „Synthetic Population Generation by Combining a Hierarchical, Simulation-Based Approach with Reweighting by Generalized Raking," in *94th Annual Meeting of Transportation Research Board*, Washington, D.C., 2015.

[8] L. Sun und A. Erath, „A Bayesian network approach for population synthesis," *Transportation Research Part C,* Nr. 61, pp. 49-62, 2015.

[9] A. Ilahi und K. W. Axhausen, „Implementing Bayesian Network and Generalized Raking Multilevel IPF for Constructing Population Synthesis in Megacities," in *Swiss Transport Research Conference*, Monte Veritá/Ascona, 2018.

[10] Bundesministerium für Verkehr, Innovation und Technologie (bmvit), „Österreich unterwegs 2013/2014," [Online]. Available: https://www.bmvit.gv.at/themen/verkehrsplanung/statistik/oesterreich_unterwegs.html. [Zugriff am 26 11 2019].

[11] R. Tomschy und M. Herry, „Österreich unterwegs 2013/2014: Methodenbericht zum Arbeitspaket „Datenverarbeitung, Hochrechnung und Analyse"," Wien, 2016.

[12] "Open Street Map (OSM), Overpass Turbo: A web based data mining tool for OpenStreetMap using Overpass API.," [Online]. Available: https://overpass-turbo.eu. [Accessed 18 January 2019].

[13] "Graph Integration Platform GIP: The reference system of Austrian public authorities for transport infrastructure data," ÖVDAT - Österreichisches Institut für Verkehrsdateninfrastruktur, [Online]. Available: https://gip.gv.at. [Accessed 21 November 2018].