Synthetic population for the state of California based on open-data: examples of San Francisco Bay area and San Diego County

Milos Balac, Sebastian Hörl

Synthetic population for the state of California based on open data: Examples of the San Francisco Bay Area and San Diego County

Milos Balac[a,*], Sebastian Hörl[a,b]

[a] *Institute for Transport Planning and Systems, ETH Zurich, Switzerland*
[b] *Institut de Recherche Technologique SystemX, Palaiseau 91120, France*

Abstract

This paper describes the steps to create a synthetic population of any region in California. By using only open data, and an open-source population synthesis pipeline, we ensure that the whole process can be easily repeated by others. This not only ensures reproducibility and transparency of the synthesis process, but also allows that studies using this population can be easily replicated. As agent-based models are gaining in popularity in recent times due to the rapid developments in the transportation sector, the need for convenient ways to generate synthetic individuals and their daily patterns has grown as well.

We present our approach for two regions: nine-county San Francisco Bay area and San Diego County. The validation results show that the methodology used is suitable to replicate socio-demograpahics and activity patterns of the population. However, it also points to some limitations due to the lack of data and the methods used. Nevertheless, the approach could be a good complement to the local and regional transportation models, as it allows easy access and can be readily used in agent-based models.

Keywords: agent-based, syntehtic, population, California, reproducible, open-source

This paper was presented at the *100th Annual Meeting of Transportation Research Bord (TRB), January 2021.*

1. Introduction

Transport planners and practitioners have been challenged with many questions around emerging transportation solutions throughout the past decades. Those include car-sharing, bike-sharing, automated vehicles, air taxis, e-scooter sharing, and new concepts like Mobility as a Service (MaaS) and increased inter-modality. All of these transportation solutions and concepts are dynamic in nature. On the one hand, the availability of the service can vary in space and time, and the price structure can depend on the demand (i.e. ride-hailing services); on the other hand, the operator can control its service by adjusting the price, relocating bikes or vehicles or dispatching certain vehicles for pick-up. Consequently, complex interactions between transport supply and demand arise. Such new emergent behavior is hard to understand with traditional transport forecasting models, which led to the development of agent-based models.

While there are different approaches to implement agent-based models, what most of them have in common is the necessary input data. These data sets can be divided in two parts: transport supply and transport demand. Transport supply usually consists of the road and public transport network, while more detailed models also contain walking and cycling networks, or information on other transportation modes like shared, ride-hailing or taxi systems. Transport demand consists of the population that performs activities and uses the transportation system to reach those activities in the study area. The population consists of persons that are grouped into households. Each person and household is furthermore characterized by attributes such as age, gender, and mobility tool ownership.

In the literature, the process of obtaining a representative population of the study area is called population synthesis. After the population synthesis step, each person is assigned an ordered set of activities that he or she wants to perform during a predefined period of time. To perform this step, usually an activity-based model is used. A synthetic population with activity chains (from now on referred to as *synthetic population* for brevity) can then be used within an agent-based model. However,

*Corresponding author
Email address: milos.balac@ivt.baug.ethz.ch (Milos Balac)

the synthesis process is usually lengthy, requires a considerable amount of calibration, and usually relies on data that is not publicly accessible. While the first two issues can be overcome with time, the third limitation is a large stumbling block to make synthetic populations accessible to larger academic and practice oriented community, and to ensure that later studies are reproducible.

This paper, therefore, presents an approach to generate a synthetic population of the state of California in USA based completely on open-data that is reproducible and easily extendable, and readily available for use. It must be stressed that the work presented here does not try to replace the work of local or regional planning agencies that maintain sophisticated activity-based models in the region. It however, should provide an agent-based model to those in the need of a less detailed model, but easily accessible and reproducible.

The rest of the paper is organized as follows. The following section briefly covers the state of the art of modelling human behavior in general and in the State of California. It is followed by a section that describes the input data. Next the population synthesis methodology is presented. Some of the validation analysis are then shown, before the paper is finished with the discussion of the presented approach and concluding remarks respectively.

2. Background

A long-established approach to forecast transportation demand is to use four-step models. These aggregated models are the traditional way of evaluating policies for large infrastructure investments and focus on large car and transit flows. Activity-based models emerged (for early reviews of activity-based models, see (Recker, 1995; Axhausen and Gärling, 1992; Kitamura, 1988)) to overcome the aggregation drawback of these models. They use methods to schedule distinct activities for individuals and make mode or destination choices, within a single framework. These models were the answer to the aggregation drawback of four-step models and the inability to model tour constraints.

Moreover, activity-based models were able to provide policy implications on many more dimensions than four-step models. However, these models usually involve a range of econometric sub-models that need to be estimated, and later calibrated to fit the data. Unfortunately, many activity-based models only focus on a small number of regions, are not easily extendable, not open-source, or lack documentation to achieve reproducibility of studies. Examples of activity-based models are CEMDAP (Bhat et al., 2008), which is based on the Dallas-Fort Worth region in the USA, or the rule-based model TASHA (Hao, 2009), which is specifically designed for the Greater Toronto area in Canada. Another notable example of activity-based models is ActivitySim (ActivitySim, 2020). It is being developed as an open-source platform for activity-based travel modeling by multiple transportation agencies in the USA. Another framework that was developed through the same collaboration is PopulationSim. It creates a synthetic population based on the marginal data available from the USA census, which creates the basis for ActivitySim.

Agent-based models appeared as the need to model interactions between individuals became important. Today, this need becomes evident as many different transportation services co-exist, and they are used in a competing, but also in a complementary fashion. Often, these forms of transport are highly dynamic as vehicles are managed on a minute-by-minute or second-by-second basis and therefore require modeling on a shorter time-scale than activity-based models usually provide. Some examples of agent-based transport models are POLARIS (Auld et al., 2016), SimMobility (Adnan et al., 2016), SUMO (Lopez et al., 2018), or MATSim (Horni et al., 2016).

There are different ways of creating the necessary input demand for agent-based models. Some of the approaches include pairing existing activity-based models with agent-based models (Ziemke et al. (2015); Hao (2009); Diogu (2019); Auld and Mohammadian (2009)) or developing new approaches to population synthesisViegas and Martínez (2010); Erath et al. (2012); Ziemke and Nagel (2017); Kickhofer et al. (2016); Hörl and Balać (2020). Some of these approaches have been used to make openly available scenarios, like for Berlin Ziemke and Nagel (2017) or Santiago de Chile Kickhofer et al. (2016). Thus, open-data-based models exist, yet they are mostly only documented as part of a more applied, larger-scope case study, whereas the details of the synthesis process are only described briefly. The framework developed by Hörl and Balać (2020), called eqasim, provides open-source pipelines that can be used to create open and reproducible scenarios directly from the raw data. This ensures that given the same inputs everyone can be able to generate the same demand input for later use in the agent-based model. This paper adds to this pipeline by providing the steps to generate regional agent-based scenarios for the state of California, without any or only small calibration effort.

Transport modeling in California has a long tradition. A number of models exist, such as the activity-based SANDAG model for San Diego (SANDAG, 2020), or a range of models by the Southern California

Association of Governments, including a TRANSCAD-based sub-regional model for Los Angeles, and a regional model based on the SimAGENT framework (Goulias et al., 2011), which makes use of PopGen for population synthesis (Pendyala et al., 2013) (as we do in this paper), and has already been coupled experimentally with MATSim (Goulias et al., 2012).

For San Francisco and the Bay Area, the activity-based model SF-CHAMP exists (SFCTA, 2020), as well as the frequently updated "Bay Area Metro Travel Model One" (MTC, 2020) by the Metropolitan Transport Commission, which is based on ActivitySim (ActivitySim, 2020). Based on those outputs, Rodier et al. (2018) present a simulation study of the Bay Area with automated vehicles, in which the travel demand is simulated in MATSim. Further applications of MATSim were presented by Pozdnoukhov et al. (2016) and are developed in the scope of the MATSim-based BEAM framework (see, e.g., (Sheppard et al., 2017) for a study of estimating electric vehicle charging demand).

The work presented in this paper aims to provide methods that can facilitate open-source and reproducible research using agent-based models in the State of California. It provides a pipeline that minimizes the calibration effort, but one that can still provide representative population and their mobility behavior. The approach is completely based on open-data, which makes the studies conducted using the pipeline transparent and reproducible.

3. Input Data

In this section, insights will be given on the different sources that were used in the context of the creation of the synthetic population. Those sources are:

- American Census Data
- American Community Survey (ACS)
- California Household Travel Survey (CHTS)
- OpenStreetMap (OSM)
- Open-data from the Ministry of Education

3.1. American Census Data

The USA census data contains information on the socio-demographic characteristics of people and households. The information is anonymized by providing only marginals (i.e. number of people in each age group) on a zonal level. Different divisions of the regions are available, with block level being the smallest. As this level usually contains a high margin of error, and not all information is available, usually a higher level of aggregation is used - the census tract. This is also the case in this study. The census data used in this study is based on the data collected in 2010 that was officially updated in 2017 with 5-year estimates.

Marginals can be useful when one wants to analyze a single variable or maybe a correlation between two variables. However, in the case where more information is required about households, their structures, and people within, additional data is required. This is provided by the *Public Use Microdata Sample* (PUMS) data set Bureau (2020b). This data provides detailed samples of households within a *Public Use Microdata Areas* (PUMAs), which contain at least 100,000 people. This information is also provided by the USA Census Bureau.

3.2. American Community Survey

The *American Community Survey* (ACS) is an on-going survey conducted by the USA Census Bureau Bureau (2020a). Among other information, it provides information on the commuting patterns of the population, their mode choices, and travel times to work. The information that is used in this work is based on the ACS 2012-2016 data. While the data can be obtained with different levels of aggregation, again the census tract level was used.

3.3. California Household Travel Survey

The *California Household Travel Survey* (CHTS) was conducted between 2010 and 2012 Laboratory (2020). At the time, it was the largest state-wide travel behavior survey ever conducted in the USA. The CHTS is used here to obtain relevant information on daily travel behavior of those individuals living in the study area.

3.4. Pre-processing the input data

While most of the data-sets are used in their original form, some of the information from the CHTS have to be adapted, in order to reduce complexity. These adaptations are presented in the following.

3.4.1. Allowed transport modes

Interviewees had a choice between 29 different transportation modes when reporting their daily travel behavior. Even though this detailed division is important when studying the usage of different transportation modes, for our purposes, it is not necessary. Therefore, we merge all modes to five main modes: *car (driver)*, *car (passenger)*, *public transport*, *bike*, and *walk*. Not only does this simplify the population synthesis process, but it also enables estimation of mode-choice models.

3.4.2. Activities

In the CHTS, respondents had a choice of 39 different activity types. As is the case for transportation modes, also here we group activities. All activities were grouped into seven categories: *home*, *work*, *leisure*, *shopping*, *education*, *errands*, *business* and *other*.

The CHTS provides information on daily behavior on a stage level. Therefore, some of these activities, like transfers, are not relevant for the population synthesis. These activities are merged with the first activity that is the main reason of travel (i.e. going to work). To define the main mode of a trip, we assign a priority to each stage: 4 - public transport; 3 - car; 2 - bike; 1 - walk. The main mode of the trip is then defined as the mode with the highest priority among the ones used in all stages composing the trip. This is both important for population synthesis, but also for later analysis.

3.5. Education locations

The state of California keeps record of all public and private schools, which are made available as open-data of Education (2020). Each data-set contains information on the address and type of school (pre-school, kindergarten, middle, junior and high school). While the public schools contain geographical coordinates, private do not.

The Ministry of Education also keeps track of all colleges and universities in the USA. This data set among many other attributes also contains information on geographical coordinates, which are vital for our approach to population synthesis.

3.6. OpenStreetMap

As we did not find an open-data resource for the state of California that contains work, shopping or leisure places, we use *OpenStreetMap* (OSM) data to obtain these locations. OSM also contains information on residential buildings, however, it is only of reasonable quality in high density areas. Therefore, we used the sequence of coordinates of residential and living streets from OSM as possible dwelling locations. Using this approach, 2% of census tracts did not contain any workplace and only one census tract out of approximately 1600 did not have a home location. As this would cause problems for later stages of the pipeline, a centroid of each census tract that did not contain work or home location was used as a possible location for work and home activities, respectively.

4. Synthetic Population

The open data described in the previous section allows to create a synthetic population of any region in the State of California. Creation of the synthetic population follows five steps:

1. Create synthetic persons and households and attach socio-demographic and mobility tool ownership attributes
2. Attach daily activity chains to individuals
3. Impute home location for each household
4. Impute work and education zones and locations to individuals
5. Assign locations to non-mandatory activities

The steps described above are executed sequentially within the pipeline. Each stage can be further extended and enriched with additional data. Furthermore, new stages within the pipeline can be added if desired.

Table 1: Control variables for population synthesis.

Variable	Atrribute levels
Age	0-5, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85+
Gender	male, female
Employed	yes, no
Household Type	family, non-family
Household Income [x1000 USD]	0-10, 10-15, 15-20, 20-25, 25-30, 35-40, 40-45, 45-50, 50-60, 60-75, 75-100, 100-135, 125-150, 150-200, 200+
Vehicle Ownership	0, 1, 2, 3+

4.1. Individuals and Households

In order to create a representative population of the study area, a tool called PopGen Group (2020) is used. PopGen uses iterative proportional fitting and updating methods to match household and person level marginals using the sample data from the PUMS data-set. The marginals are fitted on two spatial levels: census tract and PUMA. The marginals used here are age (18 levels), gender (2 levels), employment (2 levels), household type (2 levels), household income (16 levels), and household vehicle ownership (4 levels). The final output of the tool are households that have household yearly income, household type (family or non-family), household size, and number of cars attributes. Within the households, individuals have age, gender and employment (unemployed or employed) attributes. After the individuals and households are generated using the PUMS samples, additional attributes from the PUMS data are added: exact age, income, and school attendance information.

4.1.1. Adding pt accessibility

For each census tract in the modeled region we define a public transport (pt) accessibility measure as a binary variable:

$$pt_{access} = 1, \text{ if there are more than 4 pt stops per } km^2 \tag{1}$$

$$pt_{access} = 0, \text{ otherwise} \tag{2}$$

4.2. Statistical matching

In this step each synthetic individual is matched to an observation from the household travel survey, using hot-deck matching D'Orazio et al. (2012). The sampling from the household travel survey is done from a probability distribution:

$$\pi_t(s) = \begin{cases} w_s / \sum_{s' \in \mathcal{S}_t^*} w_{s'} & \text{if } s \in \mathcal{S}_t^* \\ 0 & \text{else} \end{cases} \tag{3}$$

where w_s is a weight of each observation from HTS, and $\mathcal{S}_t^*$ is the set of CHTS observations that match to the synthetic observations for k matching attributes. The attributes that are taken into consideration to perform matching are the age class, the gender, the employment status, the availability of a car inside the household, and, if possible, accessibility to public transport service. Age class was defined to replicate the one used in CHTS (where 8 age classes exist). This makes it, without public transport accessibility, a total of 96 combination of attributes. For regions where fewer observations exist, we suggest merging some of the age classes to reduce the possible number of attribute combinations. Therefore, some of the correlation between age and activity chain might be lost, especially for those below 16 years of age.

After the matching, some of the variables present in CHTS, but missing in the socio-demographic information obtained from census data is added to the synthetic individuals. Those are driver's license, public transport season ticket and bicycle ownership.

4.3. Imputing activity locations

Once the agents are assigned a daily plan based on the CHTS, a location for each of the activities has to be defined. This step is split into four parts: imputing home, work, education and finally secondary (non-mandatory) activity locations.

4.3.1. Imputing home locations

The first step consists of assigning each synthesized household to a home location. The census tract in which each agent lives is known from the step 1 of the pipeline. The possible home locations within each census tract are taken from OSM. All residential buildings and roads that are classified as a residential or a living street are taken from the OSM and were used as a possible place of living. A home place for each synthesized household is then imputed by randomly selecting a home place among all available locations within the census tract.

4.3.2. Imputing work locations

Once the individuals are assigned to a home location, a work location is assigned to those that are employed. For this purpose, the Origin-Destination (OD) matrices obtained from ACS are used. For each origin census tract we sample the workplace census tract from the multinomial distribution:

$$(f_{k,1}, ..., f_{k,\cdot}) \sim \text{Multinomial}(O_k; \pi_{k,\cdot}) \tag{4}$$

where $f_{k,k'}$ are the trips counts between origin and destination census tracts and $\sum_{k'} f_{k,k'} = O_k$ where O_k is the demand coming from census tract k.‘

Once the workplace census tract have been sampled they still have to be assigned to the synthetic individuals. In order to do this, a heuristic is used. The heuristic uses the *commute distance* of the individual in CHTS that was matched to the synthetic individual in the previous step, in order to minimize the differences between this distance and the commuting distances obtained in this step. The pipeline provides different approaches to do this, with each of them trading off between speed and accuracy. Here, we use a heuristic that minimizes the difference between the *commute distance* from CHTS and possible commute distance obtained from the sampling procedure above, for each synthetic individual, one by one. When a synthetic individual receives a workplace census tract from the sample, that workplace sample is removed from the set and the procedure continues. While this procedure is fast, it produces certain discrepancies for individuals that are getting matched later in the process. A better, although slower, approach is also available in the pipeline that minimizes total difference between all commuting samples and *commute distances* of the individuals for the given origin census tract Hörl and Balać (2020).

Once each individual is assigned a work zone, what is left to do is to assign the exact work location. This is performed in the following way:

- If the individual lives and works in the same census tract, the workplace is chosen among the available locations based on the distribution of distance to workplace for persons living and working in the same zone as captured by the CHTS
- If the individual does not work and live in the same census tract, the workplace location is drawn base on the heuristic ordering.

4.3.3. Imputing education locations

Since the ACS only contains information on commuters to a workplace, the imputation of the education locations follows a different approach.

All education-related trips from the household travel survey are first split into two groups based on the age group reported in the CHTS: those younger then 16, and those older than 16 years of age. Then, they are further divided into two subgroups: (1) those that used *car_passenger*, (2) those that did not, to reach their place of education. Since CHTS groups all individuals younger than 16 in the same age group, the mode of transportation to education activity is used to approximate the differences between those going to kindergarten or elementary school and those attending a high school. For each of these groups, it is then possible to construct the histogram of distances separating the education place to the home of the individual samples. Finally, a probability density function corresponding to each histogram is obtained.

For each agent, a target distance is drawn from the probability function related to the group (age and type of residence area) the agent belongs to. Using a bi-dimensional k-d tree, an education place is then selected such that the distance separating it from the agent's home location is as near to the

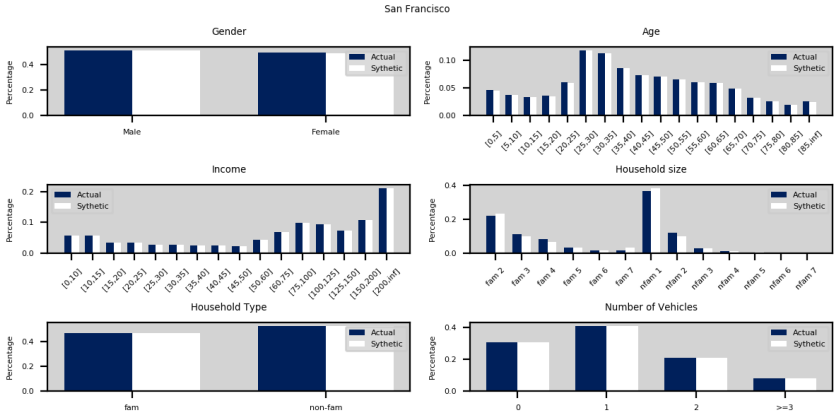

Figure 1: Socio-demographics comparison between synthetic and census population marginals for San Francisco County.

target distance as possible. As the data from the Ministry of Education allows to distinguish between kindergartens, primary schools, secondary schools, colleges and universities, an education place fitting the age of the person closest to the sampled distance from home is assigned. By doing this, we ensure that everybody is assigned to a school that fits their age, location and preferred mode of travel.

4.3.4. Imputing secondary locations

The imputation of secondary locations, which means places in which leisure, shopping or other discretionary activities are performed, is done using the method developed in Hörl and Axhausen (2020).

As a preparation for the algorithm proposed by Hörl and Axhausen (2020), all trips in the CHTS are analyzed and divided into bins of modes and travel times such that each combined bin contains at least 200 observations. This algorithm assigns discrete locations to all secondary activities while maintaining realistic distance distributions given the travel times and modes in the activity chains. This is the only step in the pipeline process where slight calibration might be necessary. The reason behind this lies in the structural constraints of the algorithm, which causes oversampling of short distances. To counteract this behavior, weights for each travel time bin/mode can be adjusted.

5. Validation

Here, we present some of the validation metrics of the synthetic population and their daily plans. While the pipeline allows to create synthetic populations of any region in the State of California, we will here present two examples:

- Synthetic population and activity patterns of the nine-county San Francisco Bay area (Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano and Sonoma)
- Synthetic population and activity patterns of the San Diego county

All nine counties around the Bay Area are chosen as each of them contributes substantial traffic to other counties in the region. San Diego was chosen as more than 95% of people both live and work in the San Diego County. Therefore, it is convenient to build a model of the region. To the knowledge of the authors, an agent-based model of the San Diego County was never built before, which makes another motivating fact to have it as one of the examples for this paper.

5.0.1. Comparison of the socio-demographics

Figure 1 and Figure 2 show the validation of some of the socio-demographics for the San Francisco and San Diego Counties, between the census data and the synthesized population. We have avoided showing the rest of the counties in the San francisco Bay area as they all show similar results.

A perfect match can be observed for most of the variables (age, gender, income, household type, and vehicle ownership), except for the household size. We have observed that the household size marginals do not fit the total number of people living in different regions, therefore, we have decided not to use household size as one of the control variables. Therefore, some of the differences, though small, do exist.

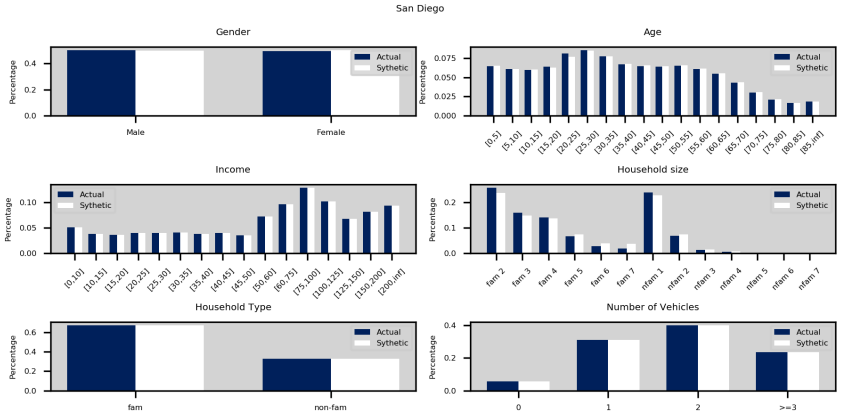

Figure 2: Socio-demographics comparison between synthetic and census population marginals for San Diego County.

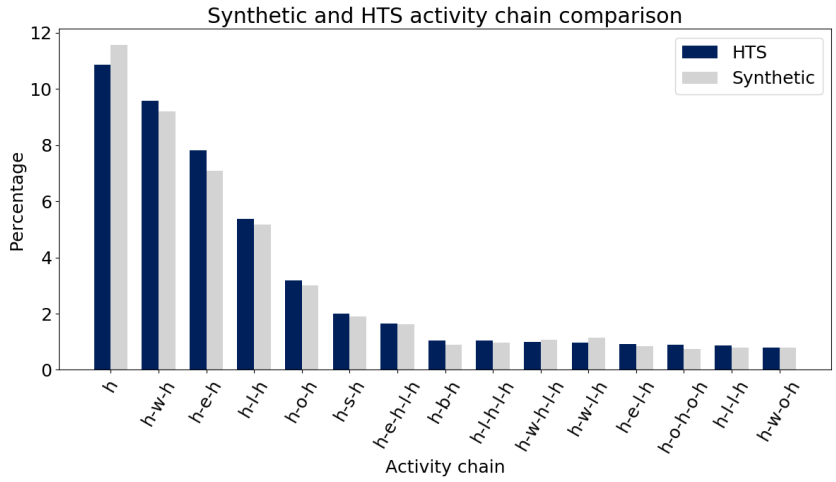

Figure 3: Activity chains comparison.
`h` stands for "home", `w` for "work", `e` for "education", `l` for "leisure", `s` for "shopping", `b` for "business" and `o` for "other".

5.0.2. Comparison of the activity chains

Figure 3, shows for the San Francisco Bay Area the distribution of activity chains in the synthesized population and compares it to the observed distribution obtained from the household travel survey.

The activity chains are present in the correct order and the observed differences between the actual population and the synthesized one are always lower than one percentage point.

Figure 4, presents the same for the San Diego County. It is clear that the activity chains are not matching, and have up to 2% difference. This is especially evident for h-e-h and h-w-h chains. These two chains would suggest that there are in general more students and less workers in the CHTS than in the census dataset. This is actually the case, as in CHTS 36% of the population goes to school while in the census it is only 30%. Although employment in San Diego county has risen between 2012 and 2017, it might not completely explain the substantial difference. Further discussion on this can be found in the following section.

Figure 5 and Figure 6 show the number of out of home activities. Most of the individuals have one out of home activity, which is also suggested by the analysis of the activity-chains.

5.0.3. Comparison of the number of activities

Figure 7 and Figure 8 show a comparison of the number of activities that women perform in these two modeled regions. The HTS and synthetic distributions match reasonably well. However, it is interesting to notice that in the area of San Francisco there are less female individuals that do not perform any out of home activities, and when they do perform activities, they also perform more activities on average. While the activity counts are shown here for women the same patterns can be observed for men as well.

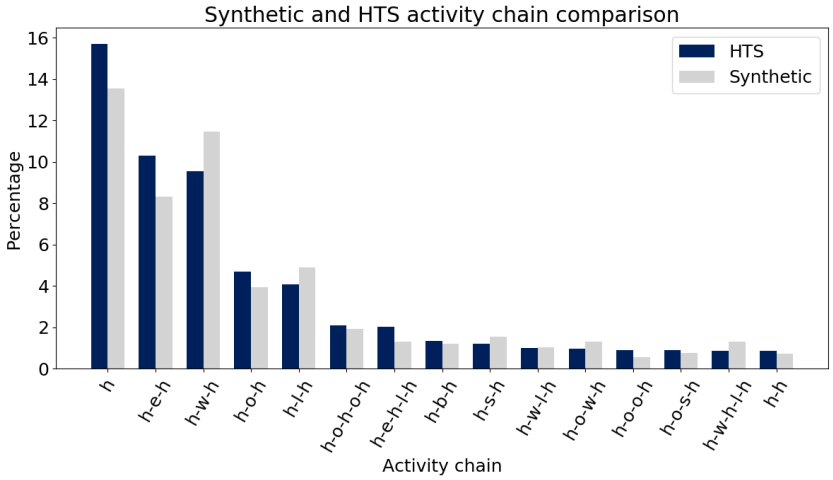

Figure 4: Activity chains comparison.
`h` stands for "home", `w` for "work", `e` for "education", `l` for "leisure", `s` for "shopping", `b` for "business" and `o` for "other".

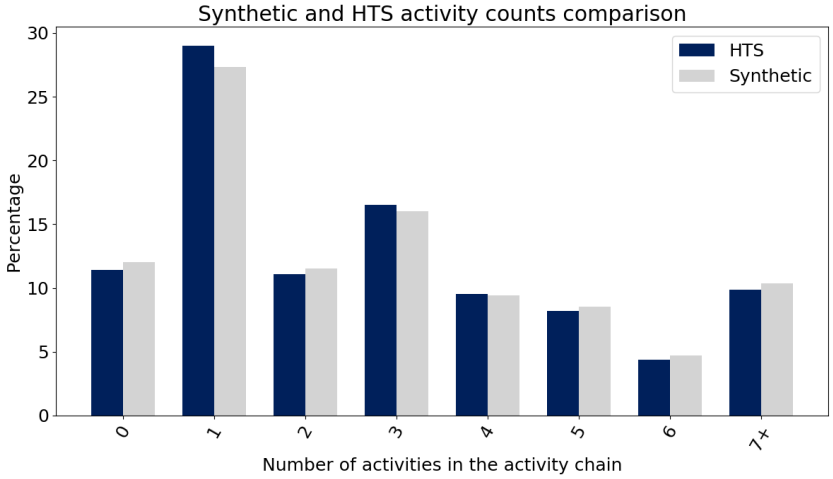

Figure 5: The number of activities that individuals perform in the San Francisco Bay Area.

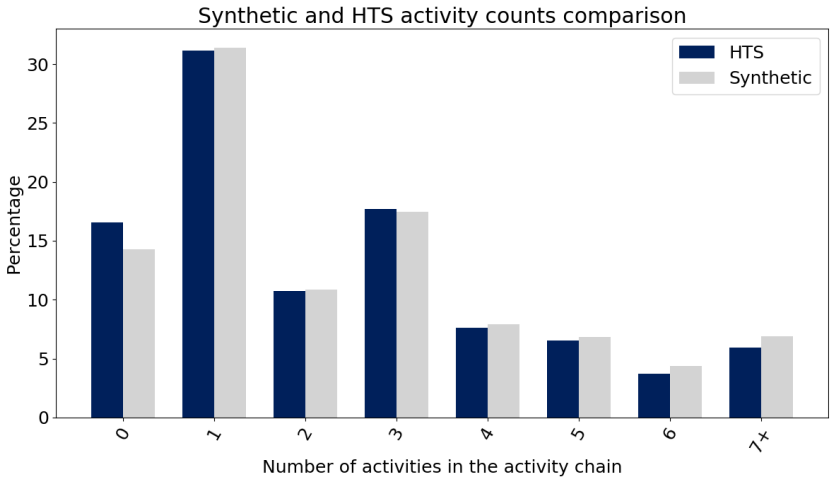

Figure 6: The number of activities that individuals perform in the San Diego County.

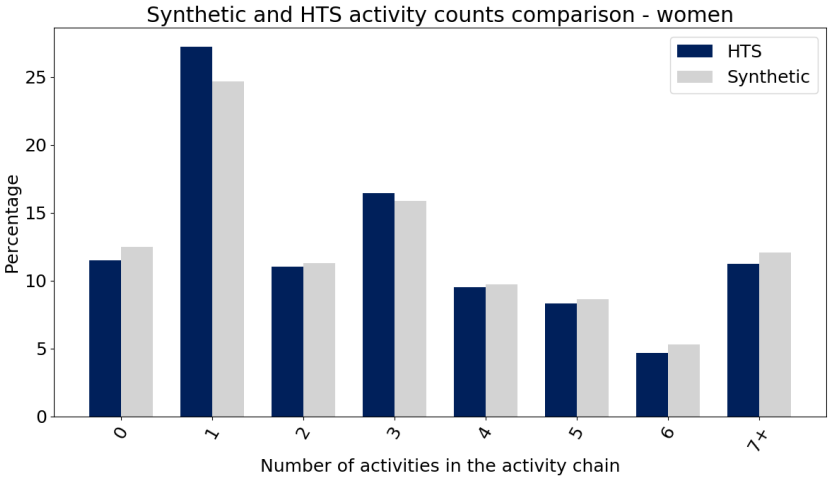

Figure 7: The number of activities that women perform in San Francisco Bay Area.

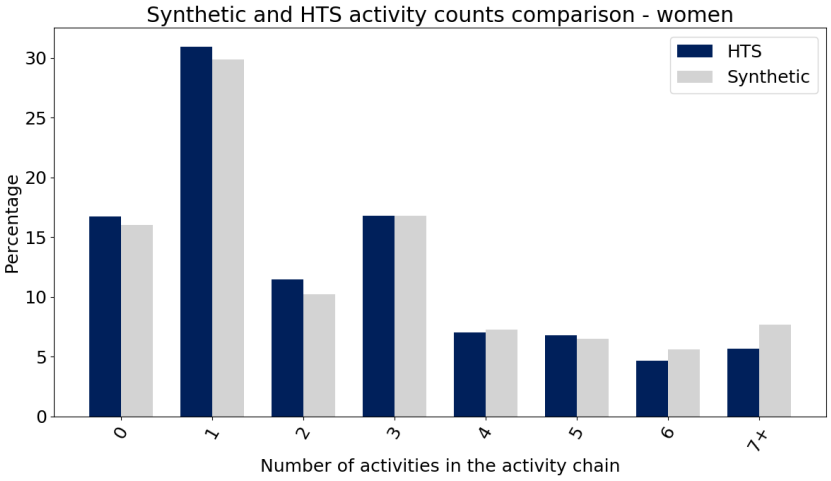

Figure 8: The number of activities that women perform in San Diego County.

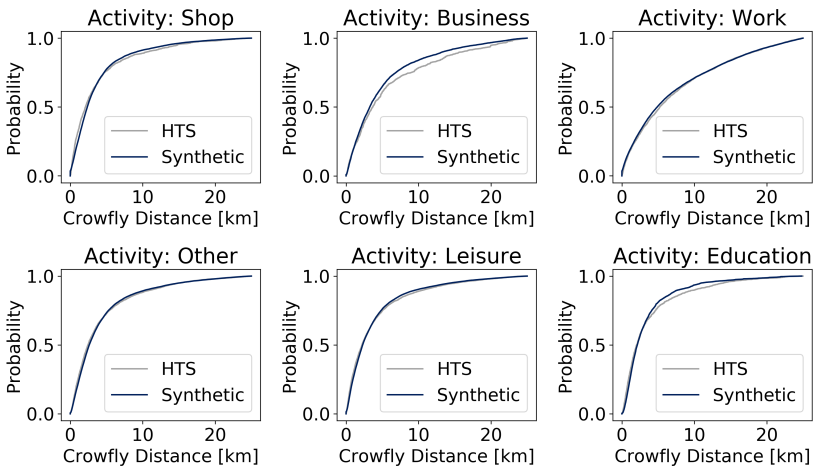

Figure 9: Cumulative distribution functions of distances to different activities in HTS and synthetic population for the San Francisco Bay area

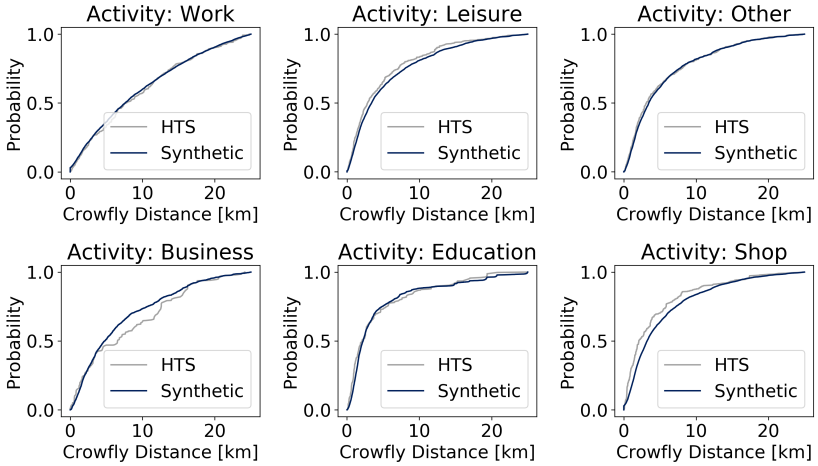

Figure 10: Cumulative distribution functions of distances to different activities in HTS and synthetic population for the San Diego County.

5.0.4. Comparison of the travel distance distribution to activities

How far individuals travel, on average, to perform a given activity in the synthetic population and in CHTS is presented in Figure 9 and Figure 10. Besides for business activities, distributions match very well for both regions. This suggests that the methods presented above are able to realistically replicate this aspect of the travel behavior. The reason behind this inconsistency in the travel distance distribution lies in the small sample of business related trips, and the distribution of the reported travel distances in CHTS. In Figure 11 and Figure 12 we can see the density distribution of distances, of business related trips, for San Francisco Bay are and San Diego County for the synthetic population and CHTS data sets. While San Francisco area has considerable number of observations for business related trips, San Diego County does not. This does not come as a surprise, considering the difference in the population size. This, however, creates spikes in the distribution of the distance distribution, which are hard to replicate. Nevertheless, the distance distribution to business activities, of the synthetic population of San Diego County in general follows the same shape as the HTS data. Therefore, one can safely assume that the inconsistency mostly lies in the small sample, and not in the methods used.

Looking at the figures, it is noticeable that commuting distances to work are on average considerably higher than for discretionary or education trips. A similar pattern is observed for both San Francisco Bay Area and San Diego County.

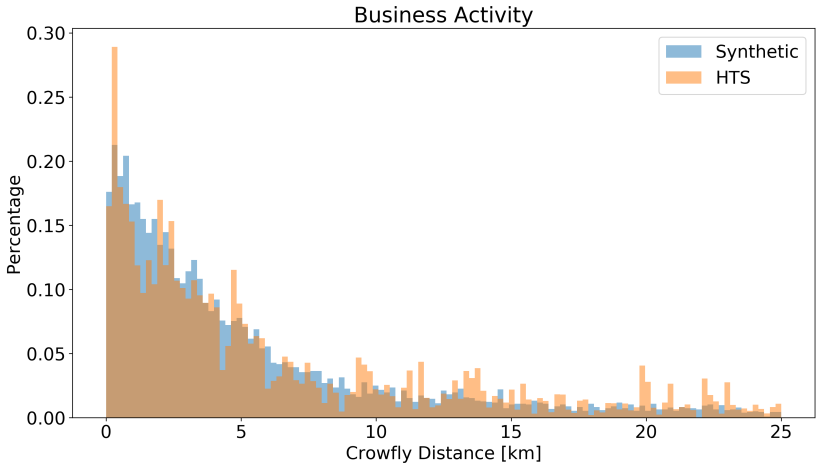

Figure 11: Density histogram of trip distances with "business" as a purpose for the San Francisco Bay area.

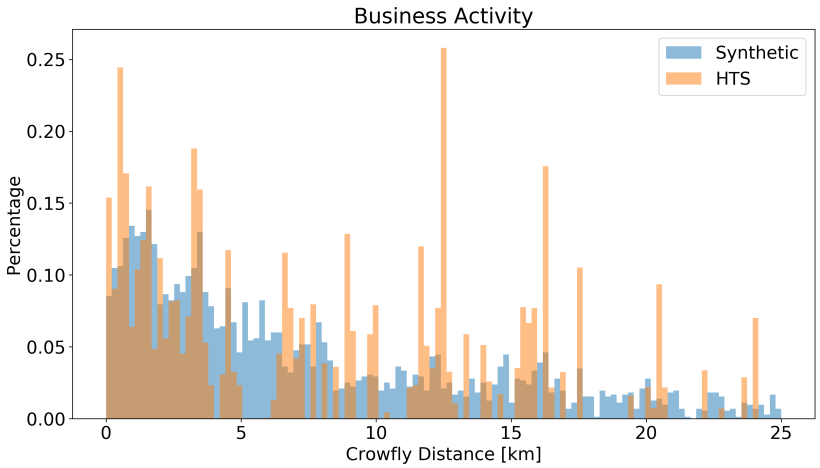

Figure 12: Density histogram of trip distances with "business" as a purpose for the San Diego County.

6. Discussion

The previous section has shown that the population synthesis pipeline is able to reproduce socio-demographics of the population and their daily plans, based on the validation measures presented. However, certain gaps and limitations exist and are discussed here.

6.1. Data

Data has a strong impact on the quality of the generated synthetic population. Therefore, data gaps listed below should be addressed, if possible, in the future:

- The data used in the pipeline has been collected at different time periods. While commuting and census estimates are for a similar point in time, CHTS was conducted six years prior. How, and if, the activity patterns have been changing during this period is difficult to say, as no other data is available. However, one might assume that even though socio-demographics of the population might have changed, the activity patterns for specific socio-demographic groups, captured by CHTS might have not. This inconsistency of the dataset has some implications on validation results as has been seen in the case of the San Diego County. Once the new household travel survey, for the region, becomes available, it would be worthwhile to verify this.

- Potential activity locations are based on the data available from the OSM. Even though, OSM data is very detailed in the urbanized areas, outside of larger urban cores this is not the case. Therefore, some of the shopping or leisure locations are currently not available for secondary activities. Additional data sources would be helpful, either on detailed location or on land use, in order to improve the reliability of the approach.

6.2. Methodology

From the methodological side, there are several points for further improvement:

- Current methodology does not consider household interactions. Therefore, activity patterns of individuals living in the same household in the synthesized population are most certainly not correlated. This would make difficult to model sharing of mobility tools in later stages, where mode-choices are for instance estimated within an agent-based model.

- The secondary location assignment procedure currently does not take into account the attractiveness of the locations. This fact can lead to over/under-estimation of the number of individuals performing activities at particular shopping or leisure locations. The attractiveness of the place could be measured by the size of the shop/leisure place or its capacity. Such data could be obtained from either OSM or other sources, which would ensure that shopping malls or large supermarkets are attracting more individuals than smaller shops.

7. Conclusion

This paper presented the process of generating a synthetic population with socio-demographic attributes and activity schedules for the state of California. Two examples are presented: nine-county San Francisco Bay area and San Diego County.

The generated models are validated against socio-demographic marginal data and travel patterns from California Household Travel Survey. While in general the models fit well, some discrepancies and limitations are pointed out in the previous chapter. Some of those can limit the approach, for certain policy studies, for example how travelers plan for joint trips or how attractiveness of a destination can influence travel decisions. However, the approach can be easily extended, as the framework is open-source and each stage of the pipeline can be individually adapted.

The pipeline also already provides converters of the synthetic population to the MATSim input format. Agent-based scenarios of the nine-county San Francisco Bay Area and five-county Los Angeles area have already been implemented and calibrated using the approach presented here to provide the synthetic population. The presentation of these scenarios was out of scope for this paper, but code and data are readily available for usage and are accessible through the same GitHub repository as the population synthesis pipeline eqasim (2020).

It must be pointed out that this modeling effort should not be considered as a try to replace the efforts of local or regional planning agencies in the state of California, but to complement them. The proposed framework should be considered as a straightforward and reproducible pipeline to generate

mobility patterns of the specific region that can, among other things, be readily used for agent-based modeling studies, based on open-data. The approach is, in addition, interesting because it can also be easily transferred to other regions in the USA.

References

ActivitySim, 2020. An open platform for activity-based travel modeling. URL: https://activitysim.github.io/. Accessed 27 Apr 2020.

Adnan, M., Pereira, F.C., Azevedo, C.M.L., Basak, K., Lovric, M., Raveau, S., Zhu, Y., Ferreira, J., Zegras, C., Ben-Akiva, M., 2016. Simmobility: A multi-scale integrated agent-based simulation platform, in: 95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record.

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., Zhang, K., 2016. Polaris: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. Transportation Research Part C: Emerging Technologies 64, 101 – 116.

Auld, J., Mohammadian, A., 2009. Framework for the development of the agent-based dynamic activity planning and travel scheduling (ADAPTS) model. Transportation Letters 1, 245–255.

Axhausen, K.W., Gärling, T., 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. Transport reviews 12, 323–341.

Bhat, C., Guo, J., Srinivasan, S., Sivakumar, A., 2008. CEMDAP User's Manual. Center for Transportation Research, University of Texas 3.

Bureau, U.S.C., 2020a. American community survey (acs). URL: https://www.census.gov/programs-surveys/acs. Accessed 22 Jul 2020.

Bureau, U.S.C., 2020b. Public use microdata samples. URL: https://www.census.gov/programs-surveys/acs/data/pums.html. Accessed 22 Jul 2020.

Diogu, W.O., 2019. Towards the Implementation of an Activity-based Travel Demand Model for Emerging Cities: Integrating TASHA and MATSim. Master's thesis.

D'Orazio, M., Di Zio, M., Scanu, M., 2012. Statistical matching of data from complex sample surveys, in: Proceedings of the European Conference on Quality in Official Statistics-Q2012.

of Education, C.D., 2020. Public schools and districts data files. URL: https://www.cde.ca.gov/ds/si/ds/pubschls.asp. Accessed 22 Jul 2020.

eqasim, 2020. Population synthesis pipeline repository. URL: https://github.com/eqasim-org. Accessed 22 Jul 2020.

Erath, A., Fourie, P.J., van Eggermond, M.A., Ordonez Medina, S.A., Chakirov, A., Axhausen, K.W., 2012. Large-scale agent-based transport demand model for singapore. Arbeitsberichte Verkehrs-und Raumplanung 790.

Goulias, K., Isbell, N., Tang, D., Balmer, M., Chen, Y., Bhat, C., Pendyala, R., 2012. TRANSIMS and MATSIM Experiments in SimAGENT. Technical Report.

Goulias, K.G., Bhat, C.R., Pendyala, R.M., Chen, Y., Paleti, R., Konduri, K.C., Huang, G., Hu, H.H., 2011. Simulator of activities, greenhouse emissions, networks, and travel (simagent) in southern california: Design, implementation, preliminary findings, and integration plans, in: 2011 IEEE Forum on Integrated and Sustainable Transportation Systems, Vienna.

Group, M.A.R., 2020. Popgen: Synthetic population generator. URL: http://www.mobilityanalytics.org/popgen.html. Accessed 22 Jul 2020.

Hao, J.Y., 2009. TASHA-MATSim integration and its application in emission modelling. Master's thesis.

Hörl, S., Axhausen, K.W., 2020. Relaxation-discretization algorithm for spatially constrained secondary location assignment, in: 99th Annual Meeting of the Transportation Research Board.

Hörl, S., Balać, M., 2020. Reproducible scenarios for agent-based transport simulation: A case study for Paris and Île-de-France .

Horni, A., Nagel, K., Axhausen, K.W., 2016. The Multi-Agent Transport Simulation MATSim. Ubiquity Press, London.

Kickhofer, B., Hosse, D., Turnera, K., Tirachinic, A., 2016. Creating an open matsim scenario from open data: The case of santiago de chile. http://www. vsp. tuberline. de/publication: TU Berlin, Transport System Planning and Transport Telematics .

Kitamura, R., 1988. An evaluation of activity-based travel analysis. Transportation 15, 9–34.

Laboratory, N.R.E., 2020. Transportation secure data center. URL: `www.nrel.gov/tsdc`. Accessed 22 Jul 2020.

Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wießner, E., 2018. Microscopic Traffic Simulation using SUMO, in: 21st IEEE International Conference on Intelligent Transportation Systems, IEEE.

MTC, 2020. travel-model-one. URL: `https://github.com/BayAreaMetro/travel-model-one`. Accessed 29 Jul 2020.

Pendyala, R., Bhat, C., Goulias, K., Paleti, R., Konduri, K., Sidhartan, R., Christian, K., 2013. SimAGENT Population Synthesis. Technical Report.

Pozdnoukhov, A., Campbell, A., Feygin, S., Yin, M., Mohanty, S., 2016. San Francisco Bay Area: The SmartBay Project - Connected Mobility. Ubiquity Press, London. pp. 485–490.

Recker, W.W., 1995. The household activity pattern problem: general formulation and solution. Transportation Research Part B: Methodological 29, 61–77.

Rodier, C., Jaller, M., Pourrahmani, E., Bischoff, J., Freedman, J., Pahwa, A., 2018. Automated Vehicle Scenarios: Simulation of System-Level Travel Effects Using Agent-Based Demand and Supply Models in the San Francisco Bay Area. Technical Report.

SANDAG, 2020. Regional models. URL: `https://www.sandag.org/index.asp?classid=32&fuseaction=home.classhome`. Accessed 29 Jul 2020.

SFCTA, 2020. Sf-champ modeling. URL: `https://www.sfcta.org/sf-champ-modeling`. Accessed 29 Jul 2020.

Sheppard, C., Waraich, R., Campbell, A., Pozdnukhov, A., Gopal, A., 2017. Modeling plug-in electric vehicle charging demand with BEAM, the framework for behavior energy autonomy mobility. Technical Report.

Viegas, J.M., Martínez, L.M., 2010. Generating the universe of urban trips from a mobility survey sample with minimum recourse to behavioural assumptions, in: Proceedings of the 12th World Conference on Transport Research.

Ziemke, D., Nagel, K., 2017. Development of a fully synthetic and open scenario for agent-based transport simulations–The MATSim Open Berlin Scenario. Transport Systems Planning and Transport Telematics–Technische Universität Berlin, Tech. Rep .

Ziemke, D., Nagel, K., Bhat, C., 2015. Integrating CEMDAP and MATSim to increase the transferability of transport demand models. Transportation Research Record 2493, 117–125.