Hybrid Simulator for Projecting Synthetic Households in Unforeseen Events

M. Kukic[a]*, M. Bierlaire[a]

[a] Transport and Mobility Laboratory, EPFL, Lausanne, Switzerland
marija.kukic@epfl.ch, michel.bierlaire@epfl.ch

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 25, 2024

Keywords: Population Dynamic Model, Simulation and Modeling, Synthetic Data Generation

1 INTRODUCTION

State-of-the-art methods for generating synthetic population create synthetic data at a given point in time, i.e., a synthetic snapshot, that quickly becomes outdated due to demographic changes like births, deaths, and migrations. This limits their usefulness in long-term transport demand forecasting, which relies on up-to-date demographic projections to accurately assess future infrastructure demand or analyze the impact of transportation policies on the population (Lomax *et al.*, 2022). In the literature, three methods for projecting synthetic individuals into the future stand out: static projection, dynamic projection and resampling (Lomax *et al.*, 2022, Geard *et al.*, 2013, Prédhumeau & Manley, 2023). Static projection uses iterative proportional fitting to simulate the evolution of individuals based on past trends and historical data, assuming they remain constant over time. While this simplification is useful for short-term analysis, it may not accurately capture long-term trends. In contrast, dynamic projection is often used for long-term forecasts as it simulates the effects of different demographic events based on the demographic annual rates (e.g., death, birth, couple formation, couple dissolution, and leaving home) on the socio-demographic characteristics of individuals (e.g., age, gender). However, if the used rates are not frequently updated, unexpected events like COVID-19 can result in projections that do not represent the real population accurately.

Our previous work introduced the hybrid simulator that combines dynamic projection with resampling for projecting synthetic individuals described by age, gender, and employment. We show that by using resampling, we can correct the propagated biases and accumulated errors of dynamic projection over extended projection periods, which results in improving the fit between the projected synthetic data and the most recent real sample when projecting far into the future (Kukic *et al.*, 2023). The resampling procedure adjusts the projected sample by comparing age category frequencies with real data, and then randomly duplicating or removing individuals in each age group to align them. The random selection can decrease population heterogenity by adding or removing individuals with similar characteristics which impacts the representativity of the sample. This paper extends the existing hybrid simulator from the individual level to the level of households, which entails two main contributions: (i) we broaden the range of simulated events and specify their influence on household attributes and member characteristics, and (ii) we replace the previous resampling procedure with the Gibbs sampler. Unlike the prior resampling method, the Gibbs sampler ensures added data replicates the joint distribution of all attributes, preserving population heterogeneity. Additionally, we evaluate the robustness of dynamic projection and hybrid simulator to unforeseen events by designing two scenarios in which we use pre and post-pandemic demographic rates for projection. We want to test if the intermediate step of hybrid simulator makes projections more robust to unusual social events compared to dynamic projection.

2 METHODOLOGY

We describe necessary modification to expand the hybrid simulator to the level of households using the Swiss Mobility and Transport microcensus disaggregated data (MTMC) from 2010, 2015, and 2021 (Swiss Federal Office of Statistics, 2012), and pre and post-pandemic estimated demographic rates for births, deaths and migration provided by the Swiss Federal Statistical Office (BFS) (Swiss Federal Office of Statistics, 2010,2020). Let t_0 be the starting year for generating the baseline synthetic sample (e.g., 2010), and t_{end}, where $t_{end} > t_0$, be the final projection year (e.g., 2021). Using the one-step Gibbs sampler proposed by Kukic & Bierlaire (2023), we employ Markov Chain Monte Carlo simulation to generate the baseline synthetic population. The synthetic sample contains vectors X that represent households defined by discrete random variables such as household size (X_{hs}), type (X_t), number of cars (X_c), with a set of individuals described by their own set of characteristics containing age (X_a), gender (X_g), employment (X_e), marital status (X_m) and driving license (X_ℓ). When a disaggregated real sample is unavailable (e.g., 2010 – 2015), for each year t_n, such that $t_0 < t_n < t_{end}$, we increment population age by 1 and update the synthetic sample by simulating births, deaths, migrations, divorces, marriages, young adults leaving parental homes, and employment changes, as explained in Section 2.1. Once the disaggregated real sample is available (e.g., 2015), we decrease the accumulated error of simulation and correct the projected sample using resampling method described in 2.2. Finally, we continue the projection to 2021 and compare it to the real data. Note that we project the generated synthetic sample from 2010 to 2021 in two scenarios: using the pre and post-pandemic rates. In the pre-pandemic scenario, demographic rates are calculated between 2010 and 2021 without considering the influence of COVID-19 on events such as births, deaths, and migrations. In the post-pandemic scenario, adjustments to these rates are made based on actual data reflecting the impact of the pandemic on these demographic factors.

2.1 Dynamic projection of all household attributes

For each female individual between 20 and 49, we randomly simulate giving birth using the probability with respect to the woman's age. We compute fertility rates for each age class by dividing the number of births by the number of women. The gender of the newborn is randomly attributed with uniform probability and other characteristics are attributed based on household attributes. Using age and gender specific mortality rates, we randomly remove individuals from each household in the dataset, adjusting household size and type accordingly to reflect these demographic changes. We simulate annual immigration and emigration by using net migration data, which reflects the difference between immigrant and emigrant numbers. For positive net migration, we duplicate individuals of specific ages and genders, creating new households, while for negative net migration, we apply the same process used in death simulations.

The marriage simulation utilizes marriage counts categorized by year and the ages of spouses. We extract partners of a certain age from a synthetic sample and form a new couple that we add as new households to the sample. When forming new households, we include all dependent children from previous households of either parent, while simultaneously adjusting the household type and size. Divorces redistribute household members by randomly assigning one partner to form a new household, turning the previous one into a single or single-parent household. Candidates are selected using age-weighted probabilities, with deterministic weights assigned to age groups. We also simulate young adults aged 15-29 leaving their parental homes based on fixed percentages, affecting household structures in a manner similar to deaths and divorces. We assume stable employment distribution over time and do not simulate annual employment changes. Instead, we use a contingency table to calculate and apply conditional probabilities of employment transitions based on age and gender at the end of the projection period.

2.2 Resampling using Gibbs sampler

We redefine the resampling procedure by employing the one-step Gibbs sampler that is only capable of creating additional data. In order to deal with both overrepresented and underrepresented categories, we have to determine the exact quantities of data that need to be added to each household size category to align the projected data's marginal distribution with the real data. We resample based on household size, as the one-step Gibbs sampler allows specifying the number of households to generate for each household size category. Since the conditionals are formed based on the real data from 2015, the generated subsample replicates the joint distributions of this dataset.

Let λ and λ' be the given marginal probability distributions of household size for the projected and real datasets, respectively. Correspondingly, let $\mathbf{c} = (c_1, \ldots, c_n)$ and $\mathbf{c}' = (c'_1, \ldots, c'_n)$ be the vectors of exact frequency counts for each household size category in the projected and real datasets. The task is to determine the vector $\mathbf{x} = (x_1, \ldots, x_n)$, such that $x_i \geq 0$ represents a small nonnegative quantity corresponding to the number of additional observations required for each household size category to adjust the projected dataset. The objective is to solve N equations, where the j-th equation for $j = 1, \ldots, N$ is given by:

$$c_j + x_j = \lambda'_j \sum_{i=1}^{N} c_i + \lambda'_j \sum_{i=1}^{N} x_i \tag{1}$$

Observe that for any $\alpha > 0$, choosing $x_j = -c_j + \alpha \cdot c'_j$ satisfies equation (1) since $c'_j = \lambda'_j \sum_{k=1}^{N} c'_k$ by definition. Therefore, to ensure that $x_i \geq 0$ for all i, it suffices to choose $\alpha = \max_{j \in \{1,\ldots,N\}} \frac{c_j}{c'_j}$. For this α, the corresponding vector $\mathbf{x}$ when added to the existing counts $\mathbf{c}$ will result in a new distribution that aligns with the real data probability distribution λ'. We maintain the population size by uniformly removing households per category while preserving the obtained probability distribution. The vector $\mathbf{x}$ is provided as an input to the one-step Gibbs sampler.

3 RESULTS

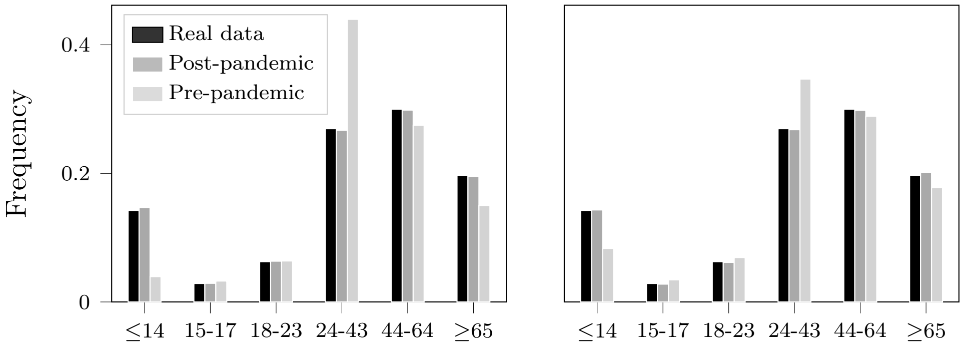
In Table 1, we compare SRMSE scores between real marginals from 2021 and projected marginals of dynamic projection and hybrid simulator, where a lower score indicates a better fit.

Table 1 – *SRMSE of marginal distributions between real data from 2021 and projected samples (2010–2021) using dynamic projection and hybrid simulator: pre and post-pandemic scenarios*

	Pre-pandemic scenario		**Post-pandemic scenario**	
Variable	**Dynamic projection**	**Hybrid**	**Dynamic projection**	**Hybrid**
Household size	0.22	0.15	0.19	0.12
Household type	0.24	0.10	0.15	0.08
Number of cars	0.32	0.18	0.24	0.12
Age	0.24	0.07	0.04	0.02
Gender	0.01	0.01	0.01	0.01
Driving licence	0.10	0.10	0.10	0.10
Marital status	0.07	0.06	0.07	0.06
Employment	0.26	0.25	0.16	0.15
Average SRMSE	0.18	0.11	0.12	0.08

First, both scenarios show that the hybrid simulator fits real data better for each attribute, indicating that its intermediate resampling step corrects errors of dynamic projection. Some attributes, like driving licenses and gender, remain unaffected by unforeseen events due to their stability over time. Second, based on the average SRMSE for both methods, we conclude that using post-pandemic rates contributes to achieving a closer fit to real data than by using the pre-pandemic rates. Finally, we see that the average SRMSE difference between the dynamic projection using pre and post-pandemic rates is 0.06, compared to 0.03 for the hybrid simulator. Figure 1 shows the differences between the two scenarios using the dynamic projection and hybrid simulator while projecting the age. For most age categories, the gaps are smaller between pre and post-pandemic scenarios when using the hybrid simulator, indicating that the intermediate resampling step makes the projections more robust and less dependent on the rates.

Figure 1 – *Marginal distribution of the projected age (2010-2021) using pre and post-pandemic rates compared to the real data from 2021 - (left) dynamic projection; (right) hybrid simulator*



4 CONCLUSION

This paper extends the existing hybrid simulator from the individual level to level of the households by defining procedures that evolve all household attributes and redefining the resampling procedure. We assess the robustness of dynamic projection and hybrid simulators by testing pre and post-pandemic scenarios, demonstrating the effectiveness of the intermediate resampling step in reducing projection errors, especially over long timeframes with unforeseen events. Since the hybrid simulator can serve as a method for updating synthetic datasets, in future research, we aim to explore the efficiency and accuracy of incremental generation using a hybrid simulator based on different factors (e.g., number of generated attributes, size, sparsity of initial sample) compared to the complete regeneration.

References

Geard, Nicholas, McCaw, James M, Dorin, Alan, Korb, Kevin B, & McVernon, Jodie. 2013. Synthetic Population Dynamics: A Model of Household Demography. *Journal of Artificial Societies and Social Simulation*, **16**(1), 8.

Kukic, Marija, & Bierlaire, Michel. 2023. *Divide-and-conquer one-step simulator for the generation of synthetic households.* Technical Report TRANSP-OR 230408. Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Kukic, Marija, Benchelabi, Salim, & Bierlaire, Michel. 2023. Hybrid Simulator for Capturing Dynamics of Synthetic Populations. *In: 2023 IEEE International Intelligent Transportation Systems Conference.*

Lomax, Nik, Smith, Andrew, Archer, Luke, Ford, Alistair, & Virgo, James. 2022. An Open-Source Model for Projecting Small Area Demographic and Land-Use Change. *Geographical Analysis*, **54**(02).

Prédhumeau, Manon, & Manley, Ed. 2023. A synthetic population for agent based modelling in Canada. *Scientific Data*, **10**(03).

Swiss Federal Office of Statistics. 2010,2020. *Les scénarios de l'évolution de la population de la Suisse 2010-2060 et 2020-2050.* Neuchâtel: Bundesamt für Statistik (BFS).

Swiss Federal Office of Statistics. 2012, 2018, 2023. *Comportement de la population en matière de mobilité.* Neuchâtel: Bundesamt für Statistik (BFS).