Synthetic Population Generation and Georeferenced Household Allocation via Iterative Proportional Fitting and Integer Programming

Maria Eduarda Pinheiro, João Vitor Pamplona, Vitória Caroline Blau, Douglas Soares Gonçalves, Luiz-Rafael Santos, Hugo Jose Lara Urdaneta

Abstract. Georeferenced synthetic populations are essential inputs for agent-based simulations in epidemiology, transportation, and urban planning, yet existing methods for spatial allocation of households to residences lack formal optimality guarantees. We present a two-stage mathematical framework for generating such populations from publicly available census data. The first stage combines Iterative Proportional Fitting (IPF) with a mixed-integer linear program (MILP) to synthesize households that preserve observed demographic correlations from census microdata. The second stage formulates the household-to-residence allocation as a MILP that maximizes a composite objective balancing allocation volume, housing-quality matching, and adherence to census tract targets Our numerical experiments for the state of Santa Catarina (SC) show a realistic demographic distribution, demonstrating the practical applicability and effectiveness of our approach.

1. Introduction

Synthetic populations are foundational inputs for agent-based models in epidemiology (Eubank et al. 2004; Zhu et al. 2024), transportation planning, urban simulation, and policy evaluation (Chapuis, Taillandier, and Drogoul 2022). Unlike aggregate statistical summaries, synthetic populations preserve the micro-level correlations between demographic attributes (age, gender, ethnicity, household composition) and spatial location, enabling simulations that capture heterogeneous behavioral responses across subpopulations. The quality of such simulations depends critically on the realism of the underlying population: household composition shapes contact patterns in epidemiological models, and georeferenced placement determines accessibility to services and exposure to environmental risks.

A growing body of work has addressed the challenge of generating synthetic populations from census data. Sample-based approaches, originating with Beckman, Baggerly, and McKay (1996), apply Iterative Proportional Fitting (IPF) to reweight microdata samples so that marginal distributions match known census totals (Müller and Axhausen 2011; Pritchard and Miller 2012). When a representative sample is unavailable, sample-free methods construct individuals directly from aggregated data: Barthélemy and Toint (2013) proposed a pipeline combining estimated joint distributions with entropy maximization and Tabu search for household formation, validated on the Belgian population. Harland et al. (2012) compared IPF-based and combinatorial optimization approaches at varying spatial

Date: August 20, 2026.

2020 *Mathematics Subject Classification.* 90C10, 90C05, 62D05, 91D20 .

Key words and phrases. Synthetic population generation, iterative proportional fitting, household allocation, mixed-integer linear programming.

scales, while Lenormand and Deffuant (2013) showed that sample-free methods can match the accuracy of sample-based ones. More recently, Yameogo et al. (2021) evaluated four reconstruction methods for French municipalities, providing systematic comparisons of their ability to reproduce household-level and individual-level distributions.

A distinct line of work has addressed population synthesis without relying on IPF. Mooij et al. (2024) introduced GenSynthPop, a deterministic, sample-free approach in which attributes are iteratively added to the synthetic population, conditioned on previously assigned attributes, so that the population itself represents the estimated joint distribution at each stage. Their method avoids the integerization step inherent to IPF and was validated for a district of The Hague (Netherlands).

The spatial allocation step, assigning synthesized households to specific georeferenced dwelling units, has received comparatively less attention. Friedrich et al. (2026) formulated the household-to-address assignment as a maximum-weight matching with side constraints, solving it via Lagrangian relaxation for the city of Trier (Germany). In the European context, the German MikroSim project (Münnich et al. 2021) built a dynamic microsimulation of the entire German population, and the GesyLand framework (Burgard, Pamplona, and Pinheiro 2026) produced georeferenced synthetic populations for epidemiological modeling, combining register data with statistical disclosure techniques.

For countries such as Brazil, where individual-level census microdata are released only for a statistical sample of households, constructing georeferenced synthetic populations at full scale poses specific challenges: the microdata sample is drawn from a prior census wave (2010), the georeferenced address registry and the aggregate dissemination tables use different census tract codings, and the resulting data products must be reconciled across incompatible spatial and demographic resolutions.

In this paper, we present a two-stage framework that addresses these challenges. The first stage combines IPF with integer programming to synthesize households that preserve observed intra-household demographic correlations, using census microdata as a structural seed. Unlike approaches that assign independently synthesized individuals to household shells, our method constructs explicit ordered relationships between the household head and each subsequent member ($k = 2, \ldots, 15$), respecting the declining stock of available individuals at each step. The second stage formulates the household-to-residence allocation as a mixed-integer linear program that maximizes a composite objective balancing allocation volume, housing-quality matching, and adherence to census tract targets.

We apply the framework to the state of Santa Catarina, Brazil, generating a population of approximately 8.4 million individuals organized into 2.8 million households across 295 municipalities, and allocating them to georeferenced residences derived from the national address registry.

The remainder of the paper is organized as follows. Section 2 details the family composition methodology. Section 3 formulates the allocation model. Sections 4 and 5 report computational results and validation for each stage. Section 6 concludes the paper. Appendixes A, B,C describe the data preparation procedure.

2. Family Composition Methodology

We consider a geographic region partitioned into municipalities. The family composition phase requires three input sets, described below. In practice, these are derived from official census products; for the Brazilian case, they are constructed from information published by Brazilian Institute of Geography and Statistics (IBGE) (Instituto Brasileiro de Geografia e Estatística 2023), as detailed in Section 4.

- `(P)` – **Population register:** The complete enumeration of individuals in the region. Each element of `(P)` contains a unique identifier and a vector of demographic attributes such as gender, age, and ethnicity.
- `(F)` – **Family structure:** The set of households. Each element of `(F)` contains a unique household identifier, the number of household members, and the demographic profile of the household head (including gender, ethnicity, and age).
- `(D)` – **Census microdata sample:** A probability sample of households from a prior census wave. Each element contains the same demographic attributes as in `(P)`, plus a household identifier, a householder role indicator, and a sampling weight.

The family composition phase constructs households by sequentially assigning individuals to family roles. The process begins with the identification of household heads ($k = 1$) and proceeds iteratively to assign members of increasing household order ($k = 2, 3, \ldots$). In what follows, we fix an arbitrary municipality to describe the procedure; the full population is then obtained by repeating this process across all municipalities independently.

2.1. **Household Head Assignment.** Each household in `(F)` has been assigned a head (order $k = 1$) drawn from `(P)`, such that the head's exact age, gender, and ethnic group match the demographic profile specified in `(F)`. Since multiple individuals in `(P)` may share the same demographic attributes, the selection among them is arbitrary, any individual with a matching profile is an equally valid head for a given household.

After this step, every household has exactly one designated head. The individuals in `(P)` not yet assigned to any household form the set of candidates for the subsequent allocation of members of order $k \geq 2$.

2.2. **Sequential Member Allocation via IPF.** Let K denote the maximum number of household members. The allocation of members of order $k \in \{2, \ldots, K\}$ must preserve the intra-household demographic correlations observed in the census sample `(D)`. For example, the tendency for heads of a certain age to cohabit with individuals of specific age groups and ethnic backgrounds. To capture this structure, we introduce the notion of demographic types.

A *demographic type* $t \in T$ is a unique combination of gender, ethnic group, and age group. The age groups are defined by a fixed partition of non-overlapping intervals covering the full age range, and $N := |T|$. The same partition is applied to the individuals in `(D)`.

For each order k, the goal is to determine a matrix $X(k) \in \mathbb{Z}_{\geq 0}^{N \times N}$, where each entry $X_{ij}(k)$ represents the number of households in which the head is of demographic type i and the k-th member is of demographic type j. This matrix must satisfy three conditions simultaneously:

(1) Preservation of correlation structure: The internal distribution of $X(k)$ must reflect the cohabitation patterns observed in `(D)`.
(2) Row marginal consistency: The row sums of $X(k)$ must match the demand from `(F)`: for each head type i, the total $\sum_{j=1}^{N} X_{ij}(k)$ equals the number of heads of type i in households with at least k members.
(3) Column marginal consistency: The overall column total must equal the overall row total, i.e., $\sum_{j=1}^{N} C_j(k) = \sum_{i=1}^{N} R_i(k)$, so that exactly one member of order k is supplied for each head that requires one.

    Additionally, for each type j, the column sum $\sum_{i=1}^{N} X_{ij}(k)$ cannot exceed the number of unassigned individuals of type j in `(P)`.

To solve this collection of problems (one for each k, independently), we adopt the *Iterative Proportional Fitting* (IPF) method (Deming and Stephan 1940), which iteratively adjusts a seed matrix to match prescribed row and column marginals while preserving the internal correlation structure. Since IPF produces continuous values, an integer programming model is subsequently applied to recover an integer solution. We now describe each component.

2.2.1. *Seed Matrices.* For each k, the seed matrix $S(k) \in \mathbb{R}_{\geq 0}^{N \times N}$ encodes the cohabitation patterns observed in `(D)`. Within each sampled household, the head's demographic type (row index i) is paired with the k-th member's type (column index j), and the sampling weight of the record is added to $S_{ij}(k)$. Unobserved pairs receive a small positive value $\varepsilon > 0$ to avoid excluding any combination *a priori*, while rows corresponding to heads below the minimum age for household responsibility are set to zero.

2.2.2. *Row and Column Marginals.* The *row marginals* $R(k) \in \mathbb{Z}_{\geq 0}^{N}$ are determined directly from `(F)`: for each head type i, the entry $R_i(k)$ counts the number of households with at least k members whose head is of type i. The *column marginals* $C(k) \in \mathbb{Z}_{\geq 0}^{N}$ require a more careful construction.

Let $C_j^{\text{stock}}(k)$ denote the number of individuals of type j in `(P)` not yet assigned to any household. The column marginals are obtained by projecting the seed structure onto the current demand and capping by the available stock, as described in Algorithm 1. When the stock of some type is insufficient to meet its projected demand, the residual units are redistributed cyclically to the types with the largest remaining stock.

Algorithm 1: Column marginal adjustment by demand and stock

Require: Seed matrix $S(k) \in \mathbb{R}^{N \times N}$, row marginals $R(k)$, available stock $C^{\text{stock}}(k)$.
Ensure: Adjusted column marginals $C(k)$ satisfying $\sum_{j=1}^{N} C_j(k) = \sum_{i=1}^{N} R_i(k)$.

$V_j \leftarrow \sum_{i=1}^{N} S_{ij}(k)$ for all j {Seed value per type}
$T_R \leftarrow \sum_{i=1}^{N} R_i(k)$ {Total demand}
$T_S \leftarrow \sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij}(k)$ {Total seed value}
$F_E \leftarrow T_R \,/\, T_S$ {Expansion factor}
$D_j \leftarrow V_j \cdot F_E$ for all j {Expanded demand per type}
for $j = 1, \ldots, N$ **do**
    $C_j(k) \leftarrow \lfloor \min(D_j, C_j^{\text{stock}}(k)) \rfloor$ {Initial allocation capped by stock}
end for
$\Delta \leftarrow T_R - \sum_{j=1}^{N} C_j(k)$ {Unallocated residual}
if $\Delta > 0$ **then**
    for $j = 1, \ldots, N$ **do**
        $G_j \leftarrow C_j^{\text{stock}}(k) - C_j(k)$ {Remaining stock per type}
    end for
    Sort indices by G_j in decreasing order
    for $\ell = 1, \ldots, \Delta$ **do**
        $idx \leftarrow$ index at position $\big((\ell - 1) \bmod N\big) + 1$ in the sorted list {Cyclic selection}
        if $G_{idx} > 0$ **then**
            $C_{idx}(k) \leftarrow C_{idx}(k) + 1$
            $G_{idx} \leftarrow G_{idx} - 1$
        end if
    end for
end if
return $C(k)$

The key idea is that the seed matrix $S(k)$ already encodes which member types tend to cohabit with which head types. The expansion factor F_E scales this distribution to match the total demand T_R from the current census, and the stock constraint ensures that no type is allocated more individuals than are actually

available. The cyclic redistribution in lines 15–21 guarantees that the balance condition $\sum_{j=1}^{N} C_j(k) = \sum_{i=1}^{N} R_i(k)$ is satisfied exactly.

Let us present an example of Algorithm 1.

Example 1. *Consider* $N = 3$ *and*

$$S(k) = \begin{bmatrix} 4 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \quad R(k) = \begin{bmatrix} 12 \\ 6 \\ 6 \end{bmatrix}, \quad C^{\text{stock}}(k) = \begin{bmatrix} 7 \\ 15 \\ 10 \end{bmatrix}.$$

The seed total is $T_S = 16$ *and the total demand is* $T_R = 24$*, giving expansion factor* $F_E = 3/2$*. Then,*

$$V = \begin{bmatrix} 6 \\ 6 \\ 4 \end{bmatrix}, \quad D = \begin{bmatrix} 9 \\ 9 \\ 6 \end{bmatrix}, \quad C(k) = \begin{bmatrix} \min\{9, 7\} \\ \min\{9, 15\} \\ \min\{6, 10\} \end{bmatrix} = \begin{bmatrix} 7 \\ 9 \\ 6 \end{bmatrix},$$

with residual $\Delta = 24 - 22 = 2$ *and remaining stock* $G = [0\ 6\ 4]^\top$*. The cyclic redistribution assigns one unit to type 2 and one to type 3, yielding*

$$C(k) = \begin{bmatrix} 7 \\ 10 \\ 7 \end{bmatrix}, \qquad \sum_{j=1}^{3} C_j(k) = 24 = \sum_{i=1}^{3} R_i(k).$$

Since each subproblem is solved independently for a fixed k, we suppress the dependence on k in the notation for the remainder of this section, writing S, R, C, and X in place of $S(k)$, $R(k)$, $C(k)$, and $X(k)$.

2.2.3. *Iterative Proportional Fitting.* With the seed matrix S and marginals R, C, we apply the standard IPF algorithm (Deming and Stephan 1940). Starting from $X^{(0)} = S$, the algorithm alternates between scaling rows to match R and scaling columns to match C, so that at each iteration the row sums of $X^{(n)}$ approach R and the column sums approach C:

$$X_{ij}^{(n,\text{row})} = X_{ij}^{(n-1)} \cdot \frac{R_i}{\sum_{j'=1}^{N} X_{ij'}^{(n-1)}}, \quad \forall i \in \{1, \ldots, N\}$$

$$X_{ij}^{(n)} = X_{ij}^{(n,\text{row})} \cdot \frac{C_j}{\sum_{i'=1}^{N} X_{i'j}^{(n,\text{row})}}, \forall j \in \{1, \ldots, N\}$$

The procedure terminates once

$$\max\Big(\max_i \Big|\sum_j X_{ij}^{(n)} - R_i\Big|,\ \max_j \Big|\sum_i X_{ij}^{(n)} - C_j\Big|\Big) < \varepsilon_{\text{IPF}}.$$

It is well known that IPF converges to the unique distribution that minimizes the Kullback–Leibler divergence $D_{\text{KL}}(X \,\|\, S)$ subject to the prescribed marginals (Csiszár 1975), making it the maximum relative-entropy estimate consistent with the current census totals. This property ensures that the correlation structure encoded in the seed is preserved as faithfully as possible under the marginal constraints.

2.3. **Integer Rounding via MILP.** The matrix X^* produced by IPF has continuous entries. Since each entry represents a count of individuals, an integer solution is required. We obtain one by solving the following mixed-integer linear program,

which minimizes the total absolute deviation from X^* while preserving the marginal totals exactly:

$$\min_{Y,\,t} \quad \sum_{i=1}^{N}\sum_{j=1}^{N} t_{ij} \tag{P1a}$$

$$\text{s.t.} \quad \sum_{j=1}^{N} Y_{ij} = R_i, \qquad \forall\, i \in \{1,\dots,N\}, \tag{P1b}$$

$$\sum_{i=1}^{N} Y_{ij} = C_j, \qquad \forall\, j \in \{1,\dots,N\}, \tag{P1c}$$

$$-t_{ij} \le Y_{ij} - X^*_{ij} \le t_{ij}, \qquad \forall\, i,j, \tag{P1d}$$

$$Y_{ij} \in \mathbb{Z}_{\ge 0},\ t_{ij} \ge 0, \qquad \forall\, i,j. \tag{P1e}$$

Constraints (P1b)–(P1c) enforce exact marginal consistency. Constraints (P1d) linearize the absolute value $|Y_{ij} - X^*_{ij}|$ through the auxiliary variables t_{ij}, and the objective (P1a) minimizes the total deviation. The integrality requirement on Y in (P1e) ensures that each entry represents a whole number of individuals.

Example 2. *Returning to Example 1, the IPF is initialized with $X^{(0)} = S$ and marginals $R = [12\ 6\ 6]^\top$, $C = [7\ 10\ 7]^\top$. After convergence:*

$$X^* \approx \begin{bmatrix} 5.26 & 4.45 & 2.29 \\ 0.77 & 3.90 & 1.33 \\ 0.97 & 1.65 & 3.38 \end{bmatrix}.$$

Solving (P1) yields

$$Y^* = \begin{bmatrix} 5 & 4 & 3 \\ 1 & 4 & 1 \\ 1 & 2 & 3 \end{bmatrix}.$$

The correlation structure of S is preserved: the dominant entry in row 3 remains in column 3, reflecting the propensity of type-3 heads toward same-type cohabitation.

Given the optimal integer matrix Y^*, for each cell $Y^*_{ij} > 0$ we select Y^*_{ij} heads of type i and Y^*_{ij} unassigned individuals of type j, pairing them one-to-one. Each paired individual receives the household identifier of the corresponding head and the householder role indicator is set to order k. As in Section 2.1, the selection among individuals of the same type is arbitrary.

With all household members assigned, the next step is to place each synthesized family into a specific georeferenced residence. We formulate this allocation problem in the following section.

3. Household Allocation Model

With the family composition complete, each household is characterized by a vector of attributes derived from `(F)` and the assigned head. The allocation phase requires one additional input.

- `(R)` – **Residence registry.:** The set of georeferenced dwelling units in the region. Each element of `(R)` contains a unique identifier, geographic coordinates, a spatial stratum code (e.g., census tract), and a vector of structural and demographic attributes such as dwelling subtype (house, apartment), infrastructure indicators, and the ethnicity of the occupant head.

As in Section 2, the procedure is applied to each municipality independently. For each combination of spatial stratum and attribute levels, the total number of occupied residences is assumed to be known from official tabulations. We denote this

count by $M_\ell \in \mathbb{N}$, where ℓ indexes the distinct combinations (e.g., stratum × infrastructure level × head gender). These counts serve as upper bounds, since not every residence in a stratum need receive a family, moreover the synthesized population may lack compatible households for some strata.

3.1. **Problem Formulation.** Households and residences with identical attribute vectors are grouped into *classes*. For example, a family class may collect all households of a given dwelling subtype, household size, head ethnicity, and head gender; a residence class may collect all dwellings in the same stratum with the same subtype, infrastructure level, and head ethnicity. Let $\mathcal{F} = \{1, \dots, F\}$ and $\mathcal{R} = \{1, \dots, R\}$ denote the sets of family classes and residence classes, respectively.

A family of class f may be allocated to a residence of class r only if their attributes satisfy a prescribed compatibility condition, for instance, matching dwelling subtype and head ethnicity. The set of admissible pairings is

$$P = \big\{(f, r) \in \mathcal{F} \times \mathcal{R} \;\big|\; f \text{ and } r \text{ are compatible}\big\}. \tag{1}$$

Let $y_f^r \in \mathbb{Z}_{\geq 0}$ denote the number of families of class f allocated to residences of class r, for $(f, r) \in P$, and let $\mathcal{C}_f$ and $\mathcal{K}_r$ denote the cardinality of classes f and r, respectively. For each index $\ell \in \mathcal{L}$, let $P_\ell \subseteq P$ denote the subset of admissible pairs whose family and residence attributes match the combination indexed by ℓ. The allocation model reads:

$$\max_{y,\,\delta} \quad \sum_{(f,r)\in P} w_f^r\, y_f^r \;-\; \beta \sum_{\ell \in \mathcal{L}} \delta_\ell \tag{P2a}$$

$$\text{s.t.} \quad \sum_{r:\,(f,r)\in P} y_f^r \leq \mathcal{C}_f, \qquad \forall\, f \in \mathcal{F}, \tag{P2b}$$

$$\sum_{f:\,(f,r)\in P} y_f^r \leq \mathcal{K}_r, \qquad \forall\, r \in \mathcal{R}, \tag{P2c}$$

$$\sum_{(f,r)\in P_\ell} y_f^r \leq M_\ell + \delta_\ell, \qquad \forall\, \ell \in \mathcal{L}, \tag{P2d}$$

$$y_f^r \in \mathbb{Z}_{\geq 0},\ \delta_\ell \geq 0. \tag{P2e}$$

Constraints (P2b) and (P2c) ensure that the number of allocated families and residences does not exceed the cardinality of each class. Constraint (P2d) enforces adherence to the spatial targets: the total allocation within each attribute combination ℓ is bounded by M_ℓ, with the slack variable $\delta_\ell \geq 0$ absorbing any excess. At optimality, $\delta_\ell^* = \max\{0, \sum_{(f,r)\in P_\ell} y_f^r - M_\ell\}$.

The objective (P2a) combines two terms. The weights $w_f^r > 0$ reflect the compatibility between a family of class f and a residence of class r. The penalty term $\beta \sum_\ell \delta_\ell$, with $\beta > 0$, discourages deviations from the spatial targets. Problem (P2) is a mixed-integer linear program (MILP) due to the integrality requirement on y_f^r in (P2e). The formulation belongs to the family of generalized assignment problems (Burkard, Dell'Amico, and Martello 2009). As the number of family and residence classes grows, the number of integer variables $|P|$ in (P2) can become very large, making the direct solution of the MILP computationally expensive.

4. Application to Santa Catarina: Family Composition

We apply the proposed framework to the state of Santa Catarina, Brazil, which comprises 295 municipalities and approximately 8.4 million inhabitants according to the 2022 Census. All data are publicly available from the Brazilian Institute of Geography and Statistics (IBGE) (Instituto Brasileiro de Geografia e Estatística

2023), and all computational procedures were implemented in `Julia` (Bezanson et al. 2017).

The complete source code and generated datasets are publicly available at https://github.com/mariaepinheiro/SigeSC to support reproducibility and reuse by the social simulation community.

4.1. **Data Instantiation.** The three input sets introduced in Section 2 are constructed from the following IBGE sources.

- `(P)` – **Population register:** From the SIDRA system (IBGE's public data retrieval platform; (Instituto Brasileiro de Geografia e Estatística – IBGE 2024c)), based on the 2022 Census, we obtain, for each of the 295 municipalities, the number of residents stratified by:
  - Gender: 1 = Male, 2 = Female.
  - Age: integer, from 0 to 100.
  - Ethnicity: 1 = White (*Branca*), 2 = Black (*Preta*), 3 = Asian (*Amarela*), 4 = Mixed (*Parda*), 5 = Indigenous (*Indígena*).

  Each combination with a positive count generates the corresponding number of individual records in `(P)`, yielding 7 610 178 elements.
- `(F)` – **Family structure.:** Also from SIDRA, based on the 2022 Census (Instituto Brasileiro de Geografia e Estatística – IBGE 2024d), we obtain, for each municipality, the number of households stratified by the following attributes:
  - Dwelling subtype: 101=detached house, 102= condominium house, 103= apartment.
  - Number of household members: $N_M \in \{1, \ldots, 6\}$, where 6 denotes six or more.
  - Ethnicity of the head: $\{1, \ldots, 5\}$, same encoding as `(P)`.
  - Gender of the head: $\{1, 2\}$.
  - Age range of the head: ≤ 17, 18–24, 25–39, 40–59, ≥ 60.

  Each row in the source table generates a household record in `(F)`, with 2 801 165 units in total.
- `(D)` – **Census microdata sample.:** We use the individual-level microdata from the 2010 Census sample (Instituto Brasileiro de Geografia e Estatística – IBGE 2011). Each record contains gender, age, ethnicity, intra-household order, and the corresponding sampling weight.

  As of July 2026, the microdata sample from the 2022 Census had not been released by IBGE; hence we use the 2010 sample. Two municipalities created after 2010,-Balneário Rincão and Pescaria Brava, enherit the distributions of their parent municipalities (Içara and Laguna, respectively).

4.2. **Family Composition.**

4.2.1. *Household Head Assignment.* Since `(F)` provides only age ranges for the household head rather than exact ages, the disaggregation into exact ages relies on the microdata sample `(D)`: for each combination of municipality, gender, ethnicity, age range, and household size, the weighted frequencies of heads in `(D)` are used to estimate the proportion of heads at each exact age within the range. The full procedure is described in Appendix A. The minimum age for household responsibility was set to 14 years and the maximum to 100 years, giving the effective age ranges $\mathcal{FE} = \{14\text{–}17,\ 18\text{–}24,\ 25\text{–}39,\ 40\text{–}59,\ 60\text{–}100\}$. After this step, every household in `(F)` has exactly one designated head drawn from `(P)`.

4.2.2. *IPF Configuration.* As described in Section 2.2, the sequential member allocation requires partitioning the population into demographic types. Each type is a unique combination of gender ($\{1, 2\}$), ethnicity ($\{1, \ldots, 5\}$), and age group. The age groups are defined by the partition

$$\big\{[0, 13],\ [14, 17],\ [18, 28],\ [29, 39],\ [40, 49],\ [50, 59],\ [60, 69],\ [70, 79],\ [80, 100]\big\}.$$

The choice of intervals reflects demographic considerations. From age 18 onward, intervals of approximately 10 years provide sufficient granularity to distinguish generational roles within households, such as young adults, working-age parents, and retirees. The intervals $[0, 13]$ and $[14, 17]$ are narrower to separate children from adolescents, who may assume distinct household responsibilities.

The seed matrix residual was set to $\varepsilon = 0.001$, the IPF convergence tolerance to $\epsilon_{\text{IPF}} = 10^{-3}$. Under these parameters, IPF converged for all 295 municipalities.

The integer rounding problems (P1) were solved using HiGHS (Huangfu and Hall 2018) via JuMP (Dunning, Huchette, and Lubin 2017).

4.2.3. *Extension to Large Households.* As mentioned, in the IBGE table, $N_M = 6$ denotes households with *six or more* members. The IPF-based allocation of Section 2.2 is therefore first applied for orders $k = 2, \ldots, 6$, so that the remaining unassigned population stock in each municipality is known precisely.

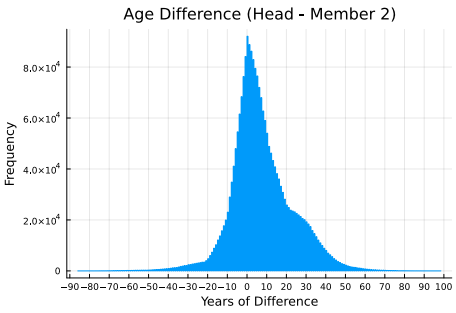
With the stock determined, we identify large-household profiles ($d = 7, \ldots, 15$ members) in the 2010 microdata `(D)`. For each such profile, the head's demographic attributes and sampling weight indicate how many similar households should exist in the current period. A household already recorded with $N_M = 6$ is converted to size d if its head matches the reference profile from `(D)` and the municipality's unassigned stock can supply the additional $d-6$ members. Once all conversions are finalized, the IPF-based allocation is applied for the remaining orders $k = 7, \ldots, 15$.

4.3. **Results.** After the complete family composition, 33563 individuals in `(P)` remain unassigned to any household. These correspond to residents of collective dwellings (such as shelters, prisons, and nursing homes) or other non-household arrangements. The official count reported by IBGE is 33556 (Instituto Brasileiro de Geografia e Estatística 2022a), yielding an agreement of 99.98%.

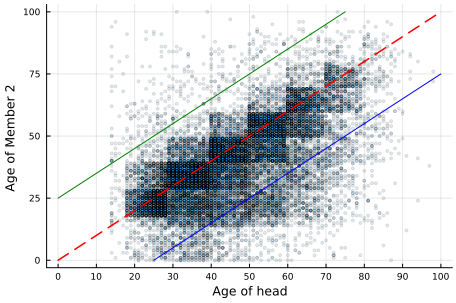
4.3.1. *Age correlation.* To assess the realism of intra-household age relationships, we computed the age-difference histograms and scatter plots between the head and members of order $k = 2, 3, 4$, using a sample of 100000 families (3.57% of the total). The histograms show the distribution of age differences between the head and the k-th member, while the scatter plots illustrate the correlation between the head's age and the k-th member's age. The results are presented in Figure 1.

For the relationship between the household head and the second member, the histogram (Figure 1a) reveals a primary mode near zero, indicating partners of similar ages, alongside a secondary mode at 20–30 years, which is consistent with head-child pairs or single-parent arrangements. The corresponding scatter plot (Figure 1b) supports these findings: the highest data density is concentrated along the diagonal $y = x$ (partners of similar age) and below the line $y = x - 25$ (children who are roughly 25 years younger than the head).

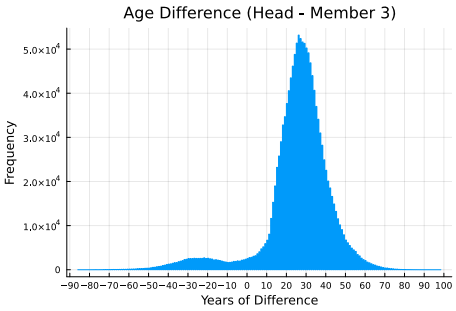
When examining the third and fourth members (orders 3 and 4), the primary modes in the histograms (Figures 1c and 1e) shift toward an age difference of 20–40 years, reflecting co-residence between parents and children. Additionally, a secondary peak appears between -20 and -30 years, becoming noticeably more pronounced for the fourth member (Figure 1e); this pattern points to the presence of ascending relatives, such as the head's own parents, within extended family



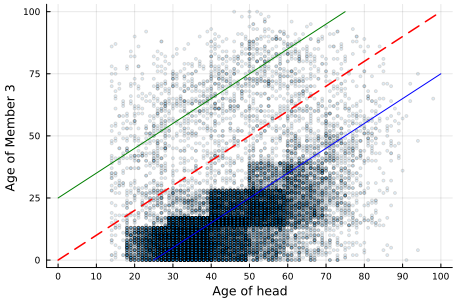
(A) Histogram of age difference (order 2).



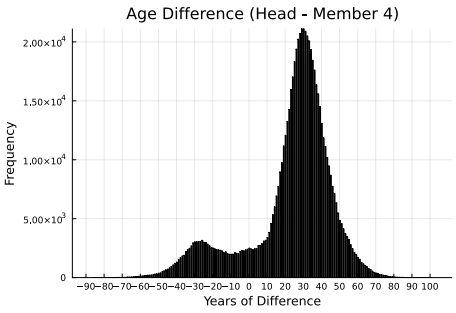
(B) Scatter plot: age of head vs. second member.



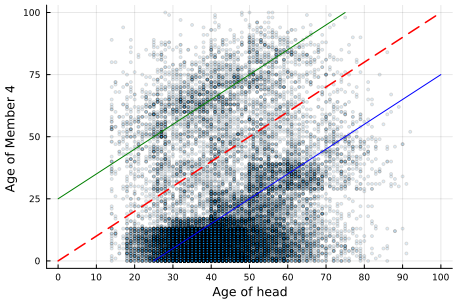
(C) Histogram of age difference (order 3).



(D) Scatter plot: age of head vs. third member.



(E) Histogram of age difference (order 4).



(F) Scatter plot: age of head vs. fourth member.

FIGURE 1. Age correlations and differences between the household head and members of orders 2, 3, and 4.

households, a distribution further supported by the alignment in the respective scatter plots (Figures 1d and 1f).

4.3.2. *Gender composition.* We analyze the gender pairing between the household head and the second member across all 295 municipalities (Figure 2). Male–Female pairs dominate, with a median proportional frequency of approximately 0.60 and an interquartile range concentrated between 0.55 and 0.65. Female–Male pairs account for the second largest share, with a median around 0.25. Together, these two mixed-gender configurations represent roughly 85% of all head–second-member pairings in a typical municipality.

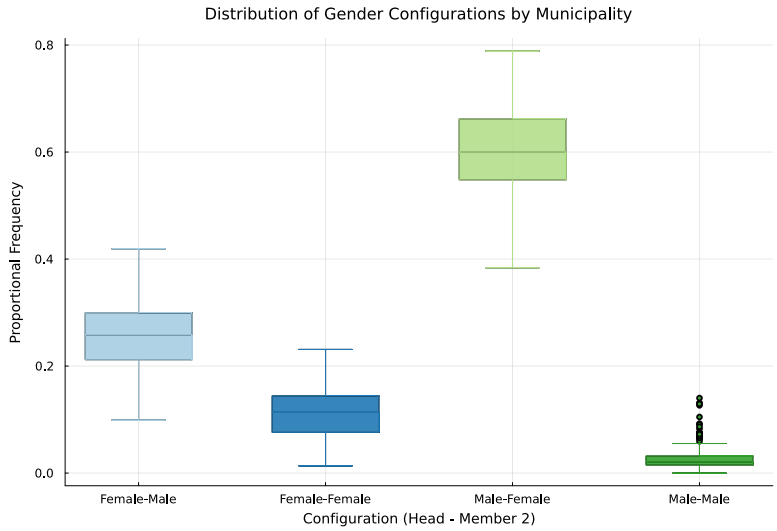
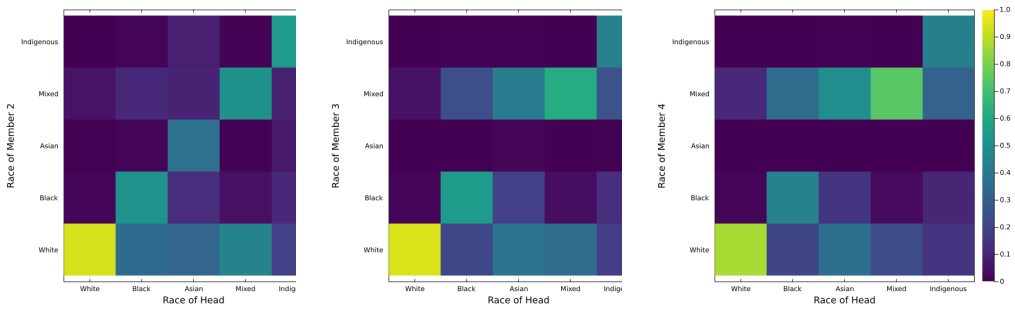


FIGURE 2. Distribution of household head and second-member gender configurations across the 295 municipalities.

Same-gender pairs are markedly less frequent. Female–Female configurations have a median near 0.10, approximately three times higher than Male–Male, whose median lies close to 0.03. This asymmetry is largely driven by single-parent households: in 2022, Santa Catarina had 286 888 households headed by women without a spouse but with children, compared to only 53 269 for men,a ratio of approximately 5:1 (Instituto Brasileiro de Geografia e Estatística 2022b). Since roughly half of these children are female, single-mother households contribute substantially more Female–Female pairings than single-father households contribute Male–Male ones.

4.3.3. *Ethnic composition.* Since approximately 76.3% of the population in Santa Catarina self-identifies as White, absolute frequencies would visually mask minority groups. To ensure a meaningful comparison across all ethnic categories, we employ row-normalized heatmaps: each cell represents the proportion of household members conditional on the ethnicity of the head, so that each row sums to one.

The resulting heatmaps for members of orders $k = 2, 3, 4$ (Figure 3) reveal strong ethnic homogamy, characterized by diagonal dominance across all household orders. This pattern is expected in nuclear family structures, where parents and children typically share the same ethnic classification. The only notable exception is the Yellow category in the $k = 4$ heatmap (Figure 3c), which displays zero entries due to the very low demographic representation of this group ($< 0.2\%$ of the population), resulting in an absence of four-or-more-member households headed by individuals of this category.



(A) Second member. (B) Third member. (C) Fourth member.

FIGURE 3. Row-normalized heatmaps of ethnic composition between the household head and members of orders $k = 2, 3, 4$.

5. Application to Santa Catarina: Household Allocation

With the family composition of Section 4 complete, we now apply the allocation model of Section 3 to assign families to georeferenced residences in Santa Catarina. Both the population and the residence registry were updated to 2025: the population was adjusted using IBGE's municipal population estimates for 2025 (Instituto Brasileiro de Geografia e Estatística 2025), removing individuals in declining municipalities via demographically weighted sampling and cloning existing families in growing ones; the residence registry was expanded using construction records available in the CNEFE (Instituto Brasileiro de Geografia e Estatística – IBGE 2024a). The full procedures are described in Appendix C.

5.1. Data Information.

(R) – **Residence registry.:** From the CNEFE (National Address Registry for Statistical Purposes; (Instituto Brasileiro de Geografia e Estatística – IBGE 2024a)), we extract all dwelling units in Santa Catarina. The CNEFE contains a wide range of attributes per dwelling; for our purposes, we retain the following:

- Unique residence identifier.
- Census tract code.
- Latitude and longitude.
- Dwelling type: 1 = private, 2 = collective, 3 = other.
- Dwelling subtype: 101 = detached house, 102=condominium house, 103 = apartment.

Census aggregates.: From the 2022 Census aggregates by census tract (Instituto Brasileiro de Geografia e Estatística – IBGE 2024b), we obtain, for each tract, the following stratified counts of occupied private dwellings:

- By dwelling subtype and number of bathrooms (N_B): 1, 2, 3 indicate the number of bathrooms, 4 = four or more, 5 = shared bathroom, 6 = rudimentary sanitation, 7 = no bathroom.
- By N_B: the ethnicity ($\{1, \ldots, 5\}$) and gender ($\{1, 2\}$) of the head.

With the above information, we assign N_B and the head's ethnicity to each residence in (R) following a hierarchical allocation procedure, and k-nearest neighbors over geographic coordinates. The full procedure is detailed in Appendix B.

The gender information is used to construct the spatial targets. Let $\mathcal{S}$ denote the set of census tracts, $\mathcal{B} = \{1, \ldots, 7\}$ the set of bathroom categories, and $\mathcal{G} = \{1, 2\}$ the set of genders. The index set of spatial targets is

$$\mathcal{L} = \big\{(s, b, g) \in \mathcal{S} \times \mathcal{B} \times \mathcal{G} \;\big|\; M_\ell > 0\big\},$$

where M_ℓ denotes the number of occupied dwellings corresponding to the combination $\ell = (s, b, g)$. For example, M_ℓ may represent the number of detached houses in a given tract with $N_B = 2$ and a male head.

5.2. **Configuration.** The allocation model (P2) is solved independently for each of the 295 municipalities. The specific choices for the abstract components introduced in Section 3 are as follows.

Classes. Family classes are defined by the attribute tuple

$$(\text{subtype},\ N_M,\ \text{head ethnicity},\ \text{head gender}),$$

and residence classes by

$$(\text{subtype},\ \text{tract},\ N_B,\ \text{head ethnicity}).$$

Compatibility. A family of class f and a residence of class r are compatible if and only if they share the same dwelling subtype and the same head ethnicity:

$$(f, r) \in P \iff \text{subtype}(f) = \text{subtype}(r) \wedge \text{head ethnicity}(f) = \text{head ethnicity}(r).$$

The weight function in the objective function (P2a) is defined as

$$w_{fr} = \alpha + q(N_M(f), \tilde{b}(r)),$$

where the term $\alpha > 0$ ensures that allocating any family is always preferable to leaving it unallocated.

The quality term q evaluates the suitability of an allocation based on the ratio between the household size (N_M) and the available sanitary infrastructure (NB). Because the census categories for infrastructure include qualitative conditions rather than strict bathroom counts for higher indices, specifically, NB $\in \{5, 6, 7\}$ representing shared, rudimentary, or absent facilities, we introduce an effective capacity mapping, denoted by $\tilde{b}(k)$. This function maps each census category k to a positive real value as follows:

$$\tilde{b}(k) = \begin{cases} k, & \text{if } k \in \{1, 2, 3, 4\}, \\ 0.5, & \text{if } k = 5, \\ 0.25, & \text{if } k = 6, \\ 0.1, & \text{if } k = 7. \end{cases}$$

Consequently, the quality of the match is modeled by the function $q(N_M, \tilde{b})$, defined as:

$$q(N_M, \tilde{b}) = \frac{1}{1 + \left(\frac{N_M}{\tilde{b}} - \tau_s\right)^2} \in (0, 1]$$

where τ_s is a census-tract-level target ratio calibrated to the local sanitary infrastructure of tract s. Hence, each τ_s captures the local infrastructure profile of its tract, ensuring that the quality function q evaluates allocations against the sanitary conditions actually observed in that tract.

Regarding the families, let μ_{N_M} denote the average household size across all families in the municipality, and let $\mu_{\tilde{b}}(s)$ denote the average effective capacity computed over the valid regular dwellings in tract s. The tract-level target is then defined as:

$$\tau_s = \frac{\mu_{N_M}}{\mu_{\tilde{b}}(s)}.$$

The numerator is kept at the municipal level because families have not yet been assigned to tracts at the time the weights are computed; only the housing stock is spatially resolved. The function q attains its maximum value of 1 when $N_M/\tilde{b} = \tau_s$ and decays symmetrically as the ratio deviates from the target, penalizing both overcrowded and underutilized allocations. Since $q \leq 1$, we set $\alpha = 1$ and $\beta = 1$ so that all terms in the objective operate on the same scale. Regarding the spatial targets, the census aggregates provide, for each census tract s, the number of occupied private dwellings stratified by number of bathrooms $b \in \{1, \ldots, 7\}$ and head gender $g \in \{1, 2\}$. For each combination with a positive count, we define an index $\ell = (s, b, g) \in \mathcal{L}$ and denote by M_ℓ the corresponding count.

5.3. **Results.** A perfect allocation would achieve $q = 1$ everywhere, indicating an exact match between household size and sanitary infrastructure relative to the local baseline. Figure 4 shows how this metric varies across the state, with municipalities grouped by size.

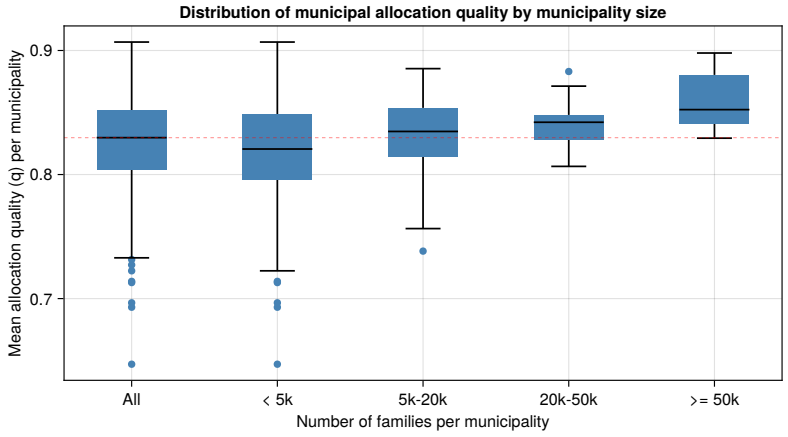


FIGURE 4. Distribution of mean allocation quality q per municipality, grouped by number of families. Each box summarizes the distribution across municipalities within that size group.

We observe a clear positive relationship between municipality size and allocation quality: larger municipalities consistently achieve higher and less dispersed values of q. The explanation is combinatorial—larger municipalities offer a more diverse housing stock in terms of sanitary categories, dwelling types, and census tracts, giving the solver a richer feasible set from which to construct high-quality matchings. In smaller municipalities, the limited variety of available dwellings restricts the admissible pairings, and the model must accept lower-quality allocations when no better alternative exists.

Beyond allocation quality, we ask how well the model respects the demographic targets from the census marginal counts. Recall that for each combination of census tract, sanitary category, and head gender $\ell = (s, b, g)$, the model introduces a slack variable $\delta_\ell \geq 0$ whenever the synthetic population cannot exactly meet the target M_ℓ. To quantify these deviations in a scale-invariant manner, we compute the relative violation $\delta_\ell / M_\ell \times 100\%$, expressing the slack as a percentage of the census target. Figure 5 presents this metric under the same size-based grouping. The pattern is consistent with the quality analysis: smaller municipalities exhibit both higher and more dispersed violations, reflecting the structural scarcity of certain household-infrastructure profiles in their housing stock. Larger municipalities, by

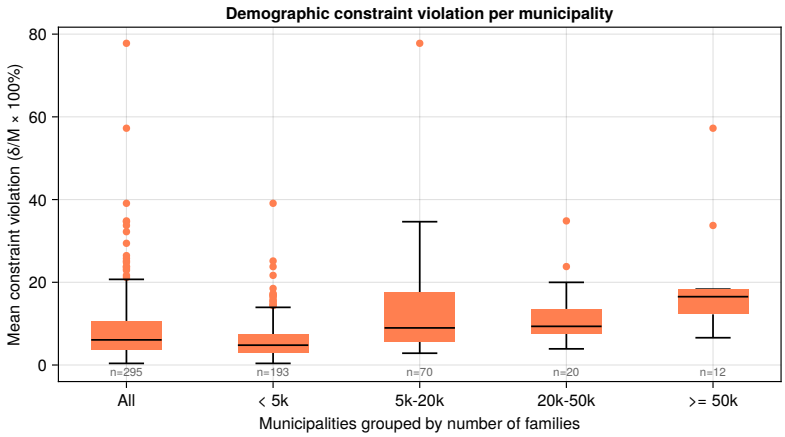


FIGURE 5. Distribution of mean demographic constraint violation ($\delta/M \times 100\%$) per municipality, grouped by number of families. Each box summarizes the distribution across municipalities within that size group.

contrast, present lower and more concentrated violations, as their more diverse housing stock allows the demographic targets to be met more closely.

6. Conclusion

We have presented a two-stage framework for generating georeferenced synthetic populations from publicly available census data. The first stage combines Iterative Proportional Fitting with integer programming to synthesize households that preserve observed demographic correlations, constructing explicit intra-household relationships member by member while respecting the available population stock at each step. The second stage assigns the resulting households to georeferenced residences via a mixed-integer linear program that balances allocation volume, housing-quality matching, and adherence to census tract targets.

The framework was applied to the state of Santa Catarina, Brazil, synthesizing approximately 8.4 million individuals organized into 2.8 million households and allocating them to georeferenced residences across 295 municipalities. Intra-household age, gender, and ethnic correlations reproduce expected sociological patterns, including spousal age proximity, parent–child age gaps, and strong ethnic homogamy.

The resulting dataset differs from existing synthetic populations in two respects that are relevant for agent-based social simulation. First, each household contains explicitly constructed ordered relationships, rather than independently sampled individuals assigned to household shells. This structure enables simulations in which intra-household contact patterns,central to epidemiological and social network models, are grounded in realistic demographic correlations. Second, each household is placed at a specific georeferenced address, enabling spatially explicit analyses at sub-municipal resolution: accessibility to health services, school catchment modeling, exposure to environmental risks, and urban vulnerability assessment all require population data at this level of geographic detail.

The methodology has limitations that suggest further research directions. First, the family composition stage relies on 2010 census microdata because the 2022 microdata sample had not been released at the time of writing; updating to the 2022 sample once available should improve the fidelity of intra-household correlations, particularly for household structures that have changed over the intervening decade. Second, the synthetic population currently includes only demographic attributes; enrichment with socioeconomic variables (income, education, employment) and health indicators using national survey microdata (IBGE 2020, 2024) would broaden the range of simulations the population can support. Finally, integrating activity schedules and mobility patterns would transform the static population into a dynamic input suitable for transport microsimulation and epidemic modeling.

Acknowledgements

This research was supported by the National Council for Scientific and Technological Development (CNPq), under the project CNPq/MCTI/FNDCT 444264/2024-8. M.E.P. was funded by CNPq grants 314788/2025-5 and partially by 409837/2025-3. V.C.B was funded by CNPq grant 198186/2025-8.

References

Barthélemy, Johan and Philippe L. Toint (2013). "Synthetic Population Generation Without a Sample." In: *Transportation Science* 47.2, pp. 266–279. DOI: 10.1287/trsc.1120.0408.

Beckman, Richard J, Keith A Baggerly, and Michael D McKay (1996). "Creating synthetic baseline populations." In: *Transportation Research Part A: Policy and Practice* 30.6, pp. 415–429. DOI: https://doi.org/10.1016/0965-8564(96)00004-3.

Bezanson, Jeff et al. (2017). "Julia: A fresh approach to numerical computing." In: *SIAM review* 59.1, pp. 65–98. DOI: 10.1137/141000671.

Burgard, Jan Pablo, João Vitor Pamplona, and Maria Eduarda Pinheiro (2026). "An optimization-based algorithm for fair and calibrated synthetic data generation." In: *Computers & Operations Research* 187. ISSN: 0305-0548. DOI: https://doi.org/10.1016/j.cor.2025.107337. URL: https://www.sciencedirect.com/science/article/pii/S0305054825003661.

Burkard, Rainer E., Mauro Dell'Amico, and Silvano Martello (2009). *Assignment Problems*. Philadelphia: SIAM. ISBN: 978-0-89871-663-4. DOI: 10.1137/1.9781611972238. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611972238.

Chapuis, Kevin, Patrick Taillandier, and Alexis Drogoul (2022). "Generation of Synthetic Populations in Social Simulations: A Review of Methods and Practices." In: *Journal of Artificial Societies and Social Simulation* 25.2, p. 6. DOI: 10.18564/jasss.4762. URL: https://www.jasss.org/25/2/6.html.

Csiszár, Imre (1975). "I-Divergence Geometry of Probability Distributions and Minimization Problems." In: *The Annals of Probability* 3.1, pp. 146–158. DOI: 10.1214/aop/1176996454.

Deming, W Edwards and Frederick F Stephan (1940). "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." In: *The Annals of Mathematical Statistics* 11.4, pp. 427–444.

Dunning, Iain, Joey Huchette, and Miles Lubin (2017). "JuMP: A Modeling Language for Mathematical Optimization." In: *SIAM Review* 59.2, pp. 295–320. DOI: 10.1137/15M1020575.

Eubank, Stephen et al. (2004). "Modelling Disease Outbreaks in Realistic Urban Social Networks." In: *Nature* 429.6988, pp. 180–184. DOI: 10.1038/nature02541.

Friedrich, Ulf et al. (2026). "Computational Methods for the Household Assignment Problem." In: *Mathematical Methods of Operations Research*. To appear. Preprint available at https://optimization-online.org/2024/07/computational-methods-for-the-household-assignment-problem/.

Harland, Kirk et al. (2012). "Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques." In: *Journal of Artificial Societies and Social Simulation* 15.1, p. 1. DOI: 10.18564/jasss.1909.

Huangfu, Qi and JA Julian Hall (2018). "Parallelizing the dual revised simplex method." In: *Mathematical Programming Computation* 10.1, pp. 119–142.

IBGE (2020). *Pesquisa Nacional de Saúde (PNS)*. Acessado em: 2026. Instituto Brasileiro de Geografia e Estatística. URL: https://www.ibge.gov.br.

– (2024). *Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua)*. Acessado em: 2026. Instituto Brasileiro de Geografia e Estatística. URL: https://www.ibge.gov.br.

Instituto Brasileiro de Geografia e Estatística (2022a). *Tabela 9869 – População residente, por situação do domicílio e condição em relação ao domicílio*. Sistema SIDRA. Censo Demográfico 2022. URL: https://sidra.ibge.gov.br/tabela/9869.

– (2022b). *Tabela 9882 – Domicílios particulares, por presença de cônjuge da pessoa responsável e filhos*. Sistema SIDRA. Censo Demográfico 2022. URL: https://sidra.ibge.gov.br/tabela/9882.

– (2023). *Censo Demográfico 2022: Resultados do Universo*. URL: https://www.ibge.gov.br/estatisticas/sociais/populacao/22827-censo-demografico-2022.html.

Instituto Brasileiro de Geografia e Estatística (2025). *Estimativas da População 2025*. IBGE. URL: https://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2025/.

Instituto Brasileiro de Geografia e Estatística – IBGE (2011). *Microdados da Amostra do Censo Demográfico 2010*. Acessado em: 18 jun. 2026. Rio de Janeiro. URL: https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html.

– (2024a). *Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE): Censo Demográfico 2022*. URL: https://www.ibge.gov.br/estatisticas/sociais/populacao/38734-cadastro-nacional-de-enderecos-para-fins-estatisticos.html.

– (2024b). *Censo Demográfico 2022: Agregados por Setores Censitários: Resultados do Universo*. URL: https://www.ibge.gov.br/estatisticas/sociais/saude/22827-censo-demografico-2022.html.

– (2024c). *Tabela 9606: População residente, por idade e sexo, segundo a forma de declaração da cor ou raça e a cor ou raça - Resultados do Universo*. Sistema IBGE de Recuperação Automática (SIDRA). URL: https://sidra.ibge.gov.br/tabela/9606.

– (2024d). *Tabela 9880: Domicílios particulares permanentes ocupados, por número de moradores, segundo o tipo de domicílio, a cor ou raça, o sexo e a idade da pessoa responsável - Resultados do Universo*. Censo Demográfico 2022. Acessado em: 18 jun. 2026. Sistema IBGE de Recuperação Automática (SIDRA). Rio de Janeiro. URL: https://sidra.ibge.gov.br/tabela/9880.

Lenormand, Maxime and Guillaume Deffuant (2013). "Generating a Synthetic Population of Individuals in Households: Sample-Free vs Sample-Based Methods." In: *Journal of Artificial Societies and Social Simulation* 16.4, p. 12. DOI: 10.18564/jasss.2319.

Mooij, Jan de et al. (2024). "GenSynthPop: Generating a Spatially Explicit Synthetic Population of Individuals and Households from Aggregated Data." In: *Autonomous Agents and Multi-Agent Systems* 38.2, p. 48. DOI: 10.1007/s10458-024-09680-7.

Müller, Kirill and Kay W. Axhausen (2011). "Hierarchical IPF: Generating a Synthetic Population for Switzerland." In: *Transportation*. Presented at the 51st Congress of the European Regional Science Association.

Münnich, Ralf et al. (2021). "A population based regional Dynamic microsimulation of Germany: the MikroSim model." In: *methods, data, analyses* 15.2, p. 24.

Pritchard, David R. and Eric J. Miller (2012). "Advances in Population Synthesis: Fitting Many Attributes per Agent and Fitting to Household and Person Margins Simultaneously." In: *Transportation* 39.3, pp. 685–704. DOI: 10.1007/s11116-011-9367-4.

Yameogo, Basile F. et al. (2021). "Generating a Two-Layered Synthetic Population for French Municipalities: Results and Evaluation of Four Synthetic Reconstruction Methods." In: *Journal of Artificial Societies and Social Simulation* 24.2, p. 5. DOI: 10.18564/jasss.4482.

Zhu, Kemin et al. (2024). "Generating Synthetic Population for Simulating the Spatiotemporal Dynamics of Epidemics." In: *PLoS Computational Biology* 20.2, e1011810. DOI: 10.1371/journal.pcbi.1011810.

Appendix A. Household Head Assignment Procedure

This appendix details the procedure for assigning household heads (order $k = 1$), summarized in Section 4.2.1. The procedure consists of two steps: estimating the proportion of heads at each exact age within a given age range (Algorithm 2),

and using these proportions to pair individuals from (P) with households in (F) (Algorithm 3).

Two aspects of the procedure deserve attention. In Algorithm 2, ages with zero observed frequency receive a small positive fill value (the minimum positive proportion divided by the number of ages in the range; see lines 24–27), so that no age is excluded *a priori*. In Algorithm 3, when the population stock at a given age is insufficient to meet the computed allocation, a redistribution mechanism (lines 19–47) transfers the excess demand to ages with available surplus, ensuring that every household receives a head.

Algorithm 2: Head proportions by exact age

Require: Population register (P), census microdata (D), municipality m, gender g
Ensure: Proportion vector P of household heads by exact age within each age range, ethnicity, and household size.

1: {**Step 1: Demographic variation**}
2: **for** each ethnicity r, age i **do**
3: $\mathcal{P}_r^i \leftarrow \{p \in \texttt{(P)} \mid \text{mun}(p) = m,\ \text{eth}(p) = r,\ \text{age}(p) = i,\ \text{gen}(p) = g\}$ {Current Information}
4: $\mathcal{D}_r^i \leftarrow \{d \in \texttt{(D)} \mid \text{mun}(d) = m,\ \text{eth}(d) = r,\ \text{age}(d) = i,\ \text{gen}(d) = g\}$ {Prior census sample}
5: $W_r^i \leftarrow \sum_{d \in \mathcal{D}_r^i} \text{weight}(d)$ {Weighted count in prior census}
6: **end for**
7: $W_{\text{tot}} \leftarrow \sum_{r,i} W_r^i$ {Total weighted prior population}
8: $P_{\text{tot}} \leftarrow \sum_{r,i} |\mathcal{P}_r^i|$ {Total current population}
9: {**Step 2: Variation factor**}
10: **for** each ethnicity r, age i **do**
11: $C_r^i \leftarrow \dfrac{|\mathcal{P}_r^i|/P_{\text{tot}}}{W_r^i/W_{\text{tot}}}$ {Ratio of current to prior population shares}
12: **end for**
13: {**Step 3: Age-specific proportions**}
14: **for** each age range f, ethnicity r **do**
15: $\mathcal{B} \leftarrow \sum_{i \in \mathcal{FE}(f)} W_r^i$ {Prior population in this range}
16: **for** each household size N_M **do**
17: **for** each age $i \in \mathcal{FE}(f)$ **do**
18: $\mathcal{H}_i \leftarrow \{d \in \texttt{(D)} \mid \text{mun}(d) = m,\ \text{eth}(d) = r,\ \text{age}(d) = i,$
19: $\text{gen}(d) = g,\ \text{size}(d) = N_M,\ \text{role}(d) = \text{head}\}$
{Heads of exact age i in prior census}
20: $P_i \leftarrow C_r^i \cdot \dfrac{\sum_{h \in \mathcal{H}_i} \text{weight}(h)}{\mathcal{B}}$ {Adjusted proportion for age i}
21: **end for**
22: $P \leftarrow P/\|P\|_1$ {Normalize to sum to 1}
23: {**Step 4: Smoothing zero entries**}
24: **if** $\min(P) = 0$ **then**
25: $v^+ \leftarrow \min\{P_i : P_i > 0\}/|\mathcal{FE}(f)|$ {Small positive fill value}
26: **for** each i with $P_i = 0$ **do**
27: $P_i \leftarrow v^+$ {Replace zero with fill}
28: **end for**
29: $P \leftarrow P/\|P\|_1$ {Renormalize}
30: **end if**
31: **end for**
32: **end for**

Remark 3. *In two municipalities, the above procedure required allocating individuals below the minimum age of 14 due to mismatches between the 2010 microdata (D) and the 2022 population counts. In Joaçaba (IBGE code 4209003), one female of Yellow ethnicity aged 12 was assigned as head; in Doutor Pedrinho (IBGE code 4205159), one male of Black ethnicity aged 13 was assigned as head. These cases arise in small municipalities where specific age, ethnicity, gender combinations have very few individuals, and the demographic structure of (D) does not perfectly match that of (P) and (F).*

Algorithm 3: Household head assignment

Require: Population register (P), family structure (F), census microdata (D), age ranges $\mathcal{FE}$.
Ensure: Each household in (F) is paired with one individual from (P) designated as head (order $k = 1$).

for each municipality m, gender g **do**
  for each age range f, ethnicity r **do**
    for each household size N_M **do**
      $P \leftarrow$ proportion vector from Algorithm 2 {Head proportions by exact age}
      {**Largest-remainder rounding**}
      $\mathcal{LF} \leftarrow \{h \in$ (F) $\mid \text{mun}(h) = m,\ \text{eth}(h) = r,\ \text{gen}(h) = g,$
        $\text{range}(h) = f,\ \text{size}(h) = N_M\}$ {Households matching this profile}
      for each age $i \in \mathcal{FE}(f)$ **do**
        $V_i \leftarrow |\mathcal{LF}| \cdot P_i$ {Fractional allocation for age i}
        $A_i \leftarrow \lfloor V_i \rfloor$ {Integer part}
        $E_i \leftarrow V_i - A_i$ {Rounding residual}
      end for
      Sort indices by E_i descending $\rightarrow IE$ {Prioritize largest residuals}
      $B \leftarrow |\mathcal{LF}| - \sum_i A_i$ {Units still to distribute}
      for $b = 1, \ldots, B$ **do**
        $A_{IE(b)} \leftarrow A_{IE(b)} + 1$ {Assign one extra unit to largest residual}
      end for
      {**Redistribution when stock is insufficient**}
      for each age $i \in \mathcal{FE}(f)$ **do**
        $\mathcal{L}_i \leftarrow \{p \in$ (P) $\mid \text{mun}(p) = m,\ \text{eth}(p) = r,$
          $\text{gen}(p) = g,\ \text{age}(p) = i,\ \text{unassigned}\}$ {Available individuals of age i}
        $\Delta_i \leftarrow |\mathcal{L}_i| - A_i$ {Surplus (> 0) or deficit (< 0)}
      end for
      while $\min_i \Delta_i < 0$ **do**
        $\delta^- \leftarrow \{i : \Delta_i < 0\}$ {Deficit ages}
        $\delta^+ \leftarrow \{i : \Delta_i > 0\}$ {Surplus ages}
        for each $i \in \delta^-$ **do**
          $A_i \leftarrow |\mathcal{L}_i|$ {Cap at available stock}
        end for
        if $|\delta^+| > 0$ **then**
          $\mathcal{S} \leftarrow \sum_{i \in \delta^-} |\Delta_i|$ {Total excess demand}
          $q \leftarrow \lfloor \mathcal{S} / |\delta^+| \rfloor$ {Uniform share per surplus age}
          $\rho \leftarrow \mathcal{S} - q \cdot |\delta^+|$ {Remainder after uniform split}
          Sort δ^+ by Δ_i descending $\rightarrow I\Delta$ {Prioritize ages with most stock}
          for each $i \in \delta^+$ **do**
            $A_i \leftarrow A_i + q$ {Distribute uniform share}
          end for
          for $j = 1, \ldots, \rho$ **do**
            $A_{I\Delta(j)} \leftarrow A_{I\Delta(j)} + 1$ {Distribute remainder one by one}
          end for
          for each $i \in \mathcal{FE}(f)$ **do**
            $\Delta_i \leftarrow |\mathcal{L}_i| - A_i$ {Recompute surplus/deficit}
          end for
        else
          break {No surplus ages left}
        end if
      end while
      {**Head-to-household pairing**}
      for each age $i \in \mathcal{FE}(f)$ **do**
        if $A_i > 0$ **then**
          Select the first A_i individuals from $\mathcal{L}_i$ {Draw heads from population}
          Select the first A_i households from $\mathcal{LF}$; call this subset ℓ {Draw households to fill}
          Assign each selected individual the household identifier of the corresponding element of ℓ and set order $k \leftarrow 1$ {Pair and mark as head}
          $\mathcal{LF} \leftarrow \mathcal{LF} \setminus \ell$ {Remove filled households}
        end if
      end for
    end for
  end for
end for

Appendix B. Residence Data Preparation

The residence registry (`R`) from the CNEFE (Instituto Brasileiro de Geografia e Estatística – IBGE 2024a) provides georeferenced dwelling units but lacks the demographic and infrastructure attributes available only in the census aggregates.

However, the aggregates employ a different tract coding system from the CNEFE. Of the 15 239 CNEFE tracts in Santa Catarina, 1 936 have no direct match in the aggregates, a consequence of the IBGE redrawing tract boundaries between the field-collection stage recorded in the CNEFE and the post-enumeration consolidation used in the published aggregates.

To recover the missing linkages we apply a hierarchical matching procedure. For each unmatched CNEFE tract s we compute the number of occupied private and collective dwellings (OPD_s and OCD_s) and search for the closest aggregate tract within progressively wider geographic scopes—subdistrict, district, municipality, and finally state—using a squared-distance metric:

$$D(s,c) \;=\; \big(\mathrm{OPD}_s - \mathrm{OPD}_c\big)^2 \;+\; \big(\mathrm{OCD}_s - \mathrm{OCD}_c\big)^2. \tag{2}$$

At each scope the candidate $c^* = \arg\min_c D(s,c)$ is selected, and both s and c^* are removed from their respective pools so that the mapping is one-to-one. The procedure terminates when every CNEFE tract has been paired with an aggregate tract.

After this step, we assign the number/class of bathrooms (N_B). The census aggregate provides, for each census tract, the distribution of N_B by structural type (101–103), which we use to populate (`R`) in three hierarchical stages.

1 -Tract level: For each census tract, a pool is built by replicating each bathroom class ($N_B \in \{1,\dots,7\}$) according to its frequency in the aggregate data. The pool is then randomly shuffled.

   For detached houses and townhouses (types 101–102), dwellings are assigned a bathroom class one-to-one from the shuffled pool; when the tract contains more dwellings than pool entries, only as many as available are assigned. For apartments (type 103), units sharing the same geographic coordinates,i.e. located in the same building,are grouped and receive a single value drawn at random from $\mathcal{B}$.

2 -Municipality and state levels: Dwellings still unassigned after Stage 1 are processed at successively coarser geographic levels (municipality, then state). At each level the pool is constructed as before, but the counts from the census aggregates are reduced by the number of units of the same type and class already assigned in earlier stages.

3 -k-NN imputation Any dwellings still lacking bathroom information are imputed via k-nearest neighbors ($k = \min\{10, |L|\}$) over latitude and longitude, where L is the set of already-assigned dwellings of the same structural type.

Again focusing on private dwellings, the census aggregates provides the count of dwellings by ethnicity of the household head (`HEAD_RACE` $\in \{1,\dots,5\}$) cross-tabulated with the number of bathrooms for each strata. For these dwellings, `HEAD_RACE` is assigned in three stages:

1 -Tract level: For each census tract and bathroom class, a pool is built by replicating each ethnicity code according to its frequency in the census aggregates and then randomly shuffled. Unassigned dwellings of the corresponding type are filled one-to-one from the pool, up to the number of available entries.

2 -Municipality level: Dwellings still unassigned after Stage 1 are processed at the municipality level. The pool is constructed as before, but the counts from the census aggregates are reduced by the number of units of the same bathroom class and ethnicity already assigned in Stage 1.

3 -Demand-based safeguard. After the two distribution-based stages, a deterministic correction ensures consistency with the household dataset `(F)`. This correction is crucial to guarantee that the ethnic distribution of families matches the available dwellings. For each municipality and ethnicity, the *deficit* is computed as $\max\{0,\ d_r - s_r\}$, where d_r is the number of families of ethnicity r in `(F)` and s_r is the number of dwellings already carrying ethnicity r. The deficit values are collected into a vector, shuffled with a fixed seed (the municipality code) for reproducibility, and assigned to the remaining unoccupied dwellings (`HEAD_RACE` $= 0$). Dwellings that still have no assignment after this step are treated as vacant.

Appendix C. Population and Residence Registry Update to 2025

While originally sourced from the 2022 Census, the family composition and residence registry have been updated to reflect 2025 demographics. The revised details are shown below.

C.1. **Population.** Using IBGE's municipal population estimates for 2025 (Instituto Brasileiro de Geografia e Estatística 2025). Of the 295 municipalities in Santa Catarina, 18 experienced population decline and 277 registered growth relative to 2022.

In the 18 declining municipalities, individuals were removed via weighted sampling without replacement. The sampling weights reflect stylized demographic dynamics: individuals aged above 70 receive weight 3.0 (higher mortality), those aged 17–30 receive weight 2.0 (higher emigration propensity), those below 17 receive weight 0.5 (lower removal likelihood), and all others retain the baseline weight 1.0. To preserve household structure, heads whose second member is a minor (age < 17) receive weight 0, preventing their removal.

After the removal step, intra-household orders are recomputed: for each removed individual, all members of the same household with a higher order have their order decremented by one, maintaining a contiguous sequence.

In the 277 growing municipalities, new individuals were created by cloning entire existing families. Families are drawn at random and replicated sequentially until the municipal population target is reached. If the remaining gap is smaller than the size of the last drawn family, only the first members (by household order) are retained, preserving the hierarchical structure. Cloned individuals receive a temporary null identifier, and their household identifier is set to the negative of the original family's identifier, allowing subsequent distinction between original and cloned records.

After processing all 295 municipalities, unique individual and household identifiers are reassigned globally. The reassignment prioritizes reusing identifiers vacated during the removal step, minimizing gaps in the identifier sequence. The consolidated result constitutes the population register `(P)` projected to 2025.

C.2. **Residence.** The Census records the residences that existed in 2022. To update the stock to 2025, we rely on the residence registry `(R)`, which includes addresses flagged as "under construction" at the time of the census. For each such address, the registry provides the number of buildings (a single unit or multiple units, fewer or more than ten) and the intended purpose (residential, non-residential, mixed, or undetermined).

We estimate from the data itself how likely each situation is to represent a new household. We compute the fraction of constructions that are purely residential among those with a known purpose, and the fraction of already-built addresses that are actually occupied households. These two proportions serve as parameters for Bernoulli trials: a mixed-purpose construction is added to the registry if a random draw falls below the second proportion, and an undetermined-purpose construction is added only if it passes two independent draws, one for each proportion.

Once a construction is accepted as a residence, its dwelling type must be determined. For a single construction with no direct information, we examine the ten nearest neighbours among already-registered residences, first within the same census tract and then within the same municipality, and assign the most frequent type. In the absence of any local reference, a standard house is assumed. For addresses with multiple buildings, we treat them as apartment buildings and estimate the number of units by counting registered apartments within a 300-metre radius. A single mixed-purpose construction is handled as an apartment building as well, since these typically consist of commercial space on the ground floor with residential units above.

Reconciling housing supply and family demand. After projecting the 2025 population, the number of families assigned to a given dwelling type in a municipality may exceed the available stock of residences of that type. Since the simulation requires every family to be allocated to a residence, these mismatches must be resolved before proceeding.

To this end, we perform a constraint-satisfaction step at the municipal level. For each municipality, we compute the available supply of individual dwellings by type (codes 101, 102, and 103) from the combined stock of 2022 residences and new constructions. Families that retained their original type from the 2022 Census are allocated first. The remaining capacity is then distributed among newly created families: each new family keeps its inherited type if a vacancy exists; otherwise, it is reassigned to the type with the largest remaining surplus in the same municipality. When the inherited type is unavailable, a small number of existing residences are reclassified between types 101, 102, and 103 to accommodate the demand. This reclassification reflects real intercensal dynamics, as dwellings are routinely converted between houses, row houses, and apartments through subdivision and renovation.

(M. E. Pinheiro, J. V. Pamplona, V. C. Blau, L.-R. Santos) Departamento de Matemática, Universidade Federal de Santa Catarina, Campus Blumenau, Blumenau, SC 89065-300, Brazil

Email address: maria.eduarda.pinheiro@ufsc.br

Email address: joao.vitor.pamplona@ufsc.br

Email address: vitoria.c.blau@hotmail.com

Email address: l.r.santos@ufsc.br

(H. J. Lara Urdaneta) Departamento de Engenharia de Controle, Automação e Computação, Universidade Federal de Santa Catarina, Campus Blumenau, Blumenau, SC 89065-300, Brazil

Email address: hugo.lara.urdaneta@ufsc.br

(D. S. Gonçalves) Departamento de Matemática,, Universidade Federal de Santa Catarina,, Campus Florianópolis,, Florianópolis, SC 88040-900,Brazil

Email address: douglas.goncalves@ufsc.br