Methodology for Adding a Variable to a Synthetic Population from Aggregate Data: Example of the Income Variable

Boyam Fabrice Yaméogo, Pierre-Olivier Vandanjon, Pierre Hankach, Pascal Gastineau

This paper is the preprint version of the research article submitted to The Journal of Artificial Societies and Social Simulation Submission date: January 29, 2021. It is currently under review.

Methodology for Adding a Variable to a Synthetic Population from Aggregate Data: Example of the Income Variable

Boyam Fabrice Yaméogo[1,2,3], **Pierre-Olivier Vandanjon**[1], **Pierre Hankach**[4], **Pascal Gastineau**[1]

[1]*AME-EASE, Univ Gustave Eiffel, IFSTTAR, Bouguenais, France*
[2]*Agency for ecological transition (ADEME), Angers, France*
[3]*SNCF TER Mobilités Pays de la Loire, Nantes, France*
[4]*MAST-LAMES, Univ Gustave Eiffel, IFSTTAR, Bouguenais, France*
Correspondence should be addressed to *boyam-fabrice.yameogo@univ-eiffel.fr*

Abstract: This paper presents a framework to tackle the problem, which has received little attention in the literature, of adding variables to a synthetic population from aggregate data. The work herein thus enriches the existing literature by proposing a new and efficient methodology to meet this practical need. The methodology integrates three distinct stages, the first of which theoretically models the problem as a multinomial distribution. The addition of a new variable is formulated as an entropy maximization using the variables available in both the synthetic population and aggregate data. Solving this problem (in our specific case study) is not possible due to the large number of constraints involved. The second stage then presents a heuristic yielding a practical solution to the problem. This heuristic combines Bayes' theorem with the cross-entropy minimization algorithm. However, given the large number of parameters to be estimated by the proposed heuristic, some of the results obtained prove to be invalid. To rectify this shortcoming, a post-processing method is applied during a third stage to ensure the consistency of our results. The methodology is described in great detail, and examples are provided for a better understanding of these three stages. Also, this methodology is applied to a real-world case study. An income is allocated to each of the 157 000 households in the French city of Nantes based on aggregate data from the FiLoSoFi database. Income constitutes an essential microsimulation variable for taking many social and economic aspects into account (e.g. household purchasing power, redistribution policy, tax policy). Special attention is also paid to the reproducibility of our results with the databases and R-scripts used, all of which are freely available. This method remains general and is indeed applicable to other variables with available aggregate data.

Keywords: Microsimulation, Agent based model, Synthetic Population Generation, Entropy, Aggregate data

1 Introduction

In recent years, microsimulation models have been applied to many fields: health (Tomintz et al. 2008; Edwards & Clarke 2013), economic policy evaluation (Avram et al. 2013; Sutherland & Figari 2013), geography (O'Sullivan 2008), and transportation (Saadi et al. 2016; Rich 2018). This approach relies on two basic premises: 1) agent behavior is determined by agents' attributes, and 2) a more realistic picture of aggregate behavior can be drawn by examining individual behavior (Tanton & Edwards 2013).

For human agents, these models require detailed attributes of individuals and households in terms of socioeconomic characteristics. Due to privacy, resources and time concerns, no comprehensive dataset containing these characteristics exists at a small geographical scale. The recommended solution consists of generating "synthetic population" representative of the actual population derived from available data. The generation of synthetic populations refers to the methods and tools employed to assemble populations of entities that: fulfill

the requirements of both the model and simulation, and 2) fit the data or hypotheses available on the target population (Thiriot & Sevenet 2020).

Population synthesis has been approached from two different methodological perspectives: sample-less and sample-based methods. The former category of methods (Gargiulo et al. 2010; Barthelemy & Toint 2013; Huynh et al. 2016) does not require a sample and only considers aggregate data. In the latter category, a sample of individuals and/or households is used to control the joint distribution of attributes while generating the population.

Given their greater accuracy, sample-based methods are typically used whenever sample data are available (Müller & Axhausen 2012; Yaméogo et al. 2020). Most statistical institutes offer the public a population sample in addition to an aggregate dataset. Such a sample is commonly obtained from the census or a specific representative survey like the household travel survey (HTS) in the transportation field. However, when applying these methods, it is only possible to generate a synthetic population from the variables already existing in the sample. For example, if an income variable is not included in the sample, then a synthetic population with an income cannot be directly generated. Depending on the data generation needs, it is entirely possible that not all variables of interest are present in the sample. In such situations, it may often become necessary to enrich the synthetic population with additional information, a process usually performed in two steps. During a first step, the synthetic population is generated from the sample, while the second step adds the desired variables from another data source (most often another sample) based on a set of variables common to both samples. This latter step is known as statistical matching. According to D'Orazio et al. (2006), statistical matching: "aims to integrate two (or more) datasets characterized by the fact that: 1) the different datasets contain information on a set of common variables and variables that are not jointly observed, and 2) the units observed in the datasets are different". Census and HTS data can thus be combined (He et al. 2020; Sallard et al. 2020).

However, a second sample that includes the additional socioeconomic variables of interest is not always available. Instead, practitioners often have access to other databases providing aggregate data for these variables, e.g. income, level of education. Yet when adding variables to a synthetic population from aggregate data, the statistical matching method can no longer be used and another approach is required. Applications of this type are less common in the academic literature. This paper proposes a new method to meet this practical need, namely by means of adding variables to a synthetic population generated from aggregate data. Our study makes the following contributions:

- Introduction of a comprehensive framework based on an entropy formulation;
- Proposal of a general algorithm, which however leads to numerical failure;
- Development and application of an efficient heuristic to easily allocate a global constant income to a synthetic population of 157 000 households;
- Dissemination of our codes for any interested reader (R-scripts).

The remainder of this paper is organized as follows. The second section provides a brief literature review on the allocation of new variables to a generated synthetic population of individuals and households. Section 3 is devoted to presenting the problem and relevant data. The fourth section formally describes the theoretical framework and proposed solution. The final two sections display and discuss the results of our analyses followed by a conclusion offering some future perspectives.

2 Literature review

To generate a synthetic population (households and/or individuals), the most widely used methods are sample-based. After producing a synthetic population, some studies focus on adding a new variable to the population from another data source. The traditional approach involves statistical matching which is a flexible method that draws on information available from different data sources. This approach is based on a selection and matching scheme.

In the selection operation, a list of attributes common to both synthetic population data and the additional sample is defined. These common attributes are referred to as matching attributes and can be: gender, age, profession, or household size. Following this step, a matching operation is conducted between the synthetic population and the additional sample over a set common attributes: for each synthetic person, only the additional sample observations found to be equal over these attributes are eligible for matching (Bösch et al. 2016).

At the end of this second step, the synthetic individuals inherit other attributes from the additional sample of individuals. The higher the number of common attributes, the more accurate the matching operation.

This selection and matching scheme has been used to assign additional attributes, i.e. daily activity-travel patterns for each member of the synthetic population generated for Jakarta (Indonesia) (Ilahi & Axhausen 2019), Zurich (Switzerland) (Hackl & Dubernet 2019), Rouen (France) (Vosooghi et al. 2019), Carinthia (Austria) (Felbermair et al. 2020), New York (USA) (He et al. 2020) and Sao Paulo (Brazil) (Sallard et al. 2020). The activity-travel patterns were derived from household travel surveys (HTS), and each person in the synthetic population was matched to an observation from the HTS.

Zhang et al. (2019) used another method to assign a community (a set of individuals possessing stronger ties within a group) for each person in a synthetic population. Their framework utilizes both census data and a sample of Call Detail Records (CDR), which is a standardized format of call logs collected by cell phone network operators. However, due to privacy protection, no personal information is available in CDR data; therefore, the statistical matching method could not be used. To assign individuals to communities, these authors formulated community assignment as an integer programming problem.

Nonetheless, the two approaches described above, i.e statistical matching and integer programming, were only designed for situations where the new data sources introduced contained disaggregate data (sample data). This requirement serves to limit the application scope of such approaches.

A handful of approaches have been developed to add new attributes to a synthetic population from aggregate data sources. Bösch et al. (2016) and Hackl & Dubernet (2019) assigned an activity location (places of work and education) from aggregate commuter statistics. For each employed synthetic person, the workplace is derived as follows: for all persons from the same municipality with the same mode of commuting, workplace municipalities are sampled at random from the observed commuter matrix, as weighted by their relative frequencies. Educational locations are chosen from the closest ones of the appropriate type.

Murata et al. (2017) assigned an income to each worker in a synthetic population from a Japanese aggregate data source using a two-step procedure. First, they used a simulated annealing method to assign a working status (i.e. employed or unemployed) to each member of synthesized households according to three distinct aggregate statistics showing the relationship between gender, family type and age in a prefecture or city. They also assigned an industry type for employed individuals. Depending on both the working status and industry type, the second step assigned an average wage to each worker using a different aggregate data source. The allocation of average wage is controlled by gender, five-year age cohort and industry type.

More recently, Hörl & Balac (2020) developed an approach to associate an income with each household of the synthetic population by means of aggregate statistics. Income distribution by deciles is provided at the municipal level, and for each synthetic household, the municipality of residence is known. The two-step assignment process is actually quite simple: each household is assigned to a decile with a 10 % probability; then, a random income value is sampled with a uniform probability from the range between the lower and upper bounds of each household's decile.

Our study is based on the following analysis of this state-of-the-art. The allocation of variables to a synthetic population from aggregate data is relatively unexplored in the literature. In most cases, this allocation process relies on a random sampling distribution, which does not guarantee data consistency. Some agents (whether individuals or households) may indeed receive irrelevant attributes. We have thus designed an efficient and consistent method that allocates a variable to a synthetic population from aggregate data. We first solve inconsistency problems through a conditional probability estimation using a combination of synthetic household attributes. Next, we optimize the conditional probabilities obtained in order to assign the appropriate values to the synthetic households. As a practical example herein, we assign an income to each synthetic household. This method remains quite general and easily applicable to other variables for which aggregate data are available. To the best of our knowledge, no clear methodology for adding variables to a synthetic population from aggregate data already exists.

3 Problem description

We begin by generating a synthetic population of households with data drawn from the French census. These data are provided by the French National Institute of Statistics and Economic Studies (INSEE); more specifically a sample [1] of census data for the city of Nantes has been used. This sample included approximately included

[1] The sample used is available via the following link: https://www.insee.fr/fr/statistiques/3625223?sommaire=3558417, consulted on November 27, 2020.

62 000 households; data were collected from 2013 to 2017 and adjusted to the reference year of 2015. Each observation in the sample represents a unique household and its main residence characteristics (household size, family composition, floor area of the dwelling unit, etc.). A statistical weight is also assigned to each household.

Since the objective of this paper is not to develop or test a particular generation method, the population was generated as simply as possible. The sample was weighted according to the weight of each household. However, since the weights are not integers, we applied the Truncate, Replicate Sample (TRS) method (Lovelace & Ballas 2013) to convert these fractional weights into integer weights. The TRS method uses both deterministic and probabilistic sampling through a three-step process. We first selected the households with a weight above 1 and retained the integer part of these weights (**Truncation step**). The selected households were then replicated based on their integer weights (**Replication step**). In the last step (**Sampling**), the remaining households were randomly chosen with selection probabilities equal to the decimal part of the household weights. At the end of this process, we obtained a synthetic population of approx. 157 000 households.

Next, we sought to allocate an income to each household of the synthetic population. For the city of Nantes, we only had access to aggregate information on household living conditions (income, inequality and poverty indicators), from a database called FiLoSoFi (for "localized disposable income system"). According to the FiLoSoFi protocol, the unit surveyed is the tax household. This term refers to all individuals included on the same income tax declaration.

The income provided in FiLoSoFi is an annual income per consumption unit, calculated as income divided by a coefficient, called consumption unit (CU), which is a weighting system that assigns a coefficient to each member of a household according to both household size and ages of its members. This system is used to compare standards of living across households of different sizes and compositions. The following weights (known as the OECD scale) have been applied herein:

- 1 CU for the first adult in the household;
- 0.5 CU for every other person in the household aged 14 years or older;
- 0.3 CU for each child under 14.

For our purposes, the distributions of annual disposable income per CU (annual income available to a household for consumption and savings divided by the number of CU in that household) were used. This income measures households' standard of living, and its distributions are given in deciles (from first to ninth). For the city of Nantes, these deciles are provided for the entire population but also for certain specific variables, namely: [2]

- number of persons in the tax household;
- family composition of the tax household;
- ownership status of the tax household;
- age of the reference person in the tax household.

Table 1 shows the different modalities of variables.

[2]The database used is available via the following link: https://www.insee.fr/fr/statistiques/3560118, consulted on November 3O, 2020.

Table 1: Attributes in the FiLoSoFi database

Variable	Definition [number of modalities]	Modalities
S	Number of persons in the tax household [5]	1 person; 2 persons; 3 persons; 4 persons; 5 persons and more
C	Family composition of the tax household [6]	Single woman ; Single man; Couple without children; Couple with children; Single-parent family; Complex household
A	Age of the reference person (RP) in the tax household [6]	Under 30 ; 30-39 ; 40-49 ; 50-59 ; 60-74 ; 75/+
O	Ownership status of the tax household [2]	Owner; Tenant

A number of differences exist between FiLoSoFi data and aggregate census data. For example, in the FiLoSoFi base, the term "person" is used (rather than inhabitant) in order to highlight the fact that a person affiliated with a household from a tax perspective is not necessarily a resident of that household (e.g. students affiliated with their parents for tax purposes but living in a separate dwelling). For these specific reasons, the number of persons does not always equal the household population in the census.[3]

Let's now estimate the annual disposable income per CU with the four variables in Table 1. For the sake of simplification, this variable will be referred to as income (I). Table 2 lists the distribution of income deciles for the city of Nantes.

[3] https://www.insee.fr/en/metadonnees/definition/c1285, consulted on November 30, 2020.

Table 2: Distribution of income deciles in the FiLoSoFi database

Modalities	Deciles (euros)								
	D1	D2	D3	D4	D5	D6	D7	D8	D9
Entire population	10 303	13 336	16 024	18 631	21 263	24 188	27 774	32 620	41 308
1 person	9 794	12 961	14 914	16 865	18 687	20 763	23 357	27 069	33 514
2 persons	12 176	15 553	18 356	20 919	23 435	26 331	30 140	35 136	44 134
3 persons	10 584	13 656	16 489	19 145	21 893	24 891	28 440	33 432	42 079
4 persons	10 740	14 130	17 207	20 138	22 955	26 148	29 644	34 238	42 998
5 persons/+	8 758	10 990	12 879	15 467	18 991	23 164	27 638	33 238	43 292
Single woman	10 714	13 334	15 332	17 186	19 031	21 111	23 715	27 360	33 480
Single man	9 016	12 224	14 288	16 388	18 268	20 305	22 908	26 696	33 551
Couple without children	14 417	18 066	20 791	23 225	25 785	28 911	32 718	37 961	47 273
Couple with children	10 822	14 238	17 646	20 665	23 596	26 837	30 528	35 573	44 977
Single-parent family	8 702	10 367	11 915	13 557	15 179	17 135	19 370	22 761	28 733
Complex households	8 692	11 052	13 063	15 207	17 648	20 452	23 853	27 843	35 179
Under 30	8 371	11 117	13 501	15 678	17 572	19 557	21 803	24 513	28 920
30-39	9 985	12 872	15 539	18 122	20 688	23 463	26 704	30 771	37 300
40-49	9 827	12 733	15 226	17 993	20 839	24 055	27 842	32 837	42 018
50-59	10 371	13 512	16 617	19 509	22 561	26 030	30 095	35 710	46 658
60-74	12 474	15 582	18 641	21 412	24 360	27 778	32 049	37 751	48 548
75/+	14 005	16 389	18 583	20 869	23 275	26 028	29 648	34 849	43 945
Owner	16 543	19 966	22 545	25 022	27 626	30 612	34 336	39 727	50 060
Tenant	8 764	10 912	12 748	14 433	16 265	18 266	20 615	23 870	29 860

Note: The first decile (D1) is the income below which 10% of incomes are situated; The ninth decile (D9) is the income below which 90% of incomes are situated.

An estimated household income (I) distribution is sought based on:

- Household size (S);
- Family composition (C);
- Age of the reference person (A);
- Household ownership status (O).

4 Problem-solving heuristic

The objective here is to assign an income to all the various cross-modalities of the four variables $S_iC_jA_kO_l$, with:

- $i = 1, 2, ..., 5$: the number of modalities of household size;
- $j = 1, 2, ..., 6$: the number of modalities of household composition;
- $k = 1, 2, ..., 6$: the number of modalities of age of the reference person (RP);
- $l = 1, 2$: the number of modalities of ownership status.

As an example, $S_4C_5A_2O_2$ represents the following cross-modality in the synthetic population: household size - 4, family composition - single-parent, age of the RP - 30-39 and ownership status - tenant. The synthetic population contains a total of 19 modalities ($5+6+6+2$) and 360 potential cross-modalities ($5 \times 6 \times 6 \times 2$). Among these cross-modalities, some are obviously infeasible, eg. $S_1C_3A_kO_l$ (household size = 1, family composition = couple with children). Therefore, $\mathrm{P}(S_1C_3A_kO_l) = 0$ regardless of the age and ownership status modalities. Of the 360 potential cross-modalities, 187 have nonzero probabilities. The income assignment consists of estimating the probability distribution $\mathrm{P}(I|S_iC_jA_kO_l)$.

Since no information is available between two income deciles, the estimation of continuous probability $\mathrm{P}(I)$ is replaced by an estimation of M discrete probabilities $\mathrm{P}(I_{m-1} < I < I_m)$. I_m is a growing sequence of all income deciles ranging from $m = 1$ to $M = 171$ (19 modalities in the synthetic population $\times$ 9 deciles). The decile of the entire is not taken into account because this information is redundant with the information for each modality. For extrapolation purposes, let's assume a linear distribution of income and then set for each modality a minimum income $I_0 = 0$ and a maximum income I_{max} equal to 1.5 times the 9^{th} decile. Column vector $\mathbf{I}$ is defined whereby the m component is the interval $]I_{m-1}, I_m[$, with $m = 1$ to $M = 190$. This vector includes the previous 171 deciles and the 19 maximum incomes, thus in all 190 modalities for the income are defined.

We therefore estimate $\mathrm{P}(I_{m-1} < I < I_m|S_iC_jA_kO_l)$ which represents 35 530 (187 $\times$ 190) cross-modalities with income. This probability can be written in two distinct ways:

$$\mathrm{P}(I_{m-1} < I < I_m|S_iC_jA_kO_l) = \mathrm{P}((I_{m-1} < I < I_m) \cap (S_iC_jA_kO_l)) \times \frac{1}{\mathrm{P}(S_iC_jA_kO_l)} \quad (1)$$

By applying Bayes' theorem:

$$\mathrm{P}(I_{m-1} < I < I_m|S_iC_jA_kO_l) = \mathrm{P}(S_iC_jA_kO_l|I_{m-1} < I < I_m) \times \frac{\mathrm{P}(I_{m-1} < I < I_m)}{\mathrm{P}(S_iC_jA_kO_l)} \quad (2)$$

with :
$i = 1, 2, ..., 5$
$j = 1, 2, ..., 6$
$k = 1, 2, ..., 6$
$l = 1, 2$

The remainder of this section presents a model, based on Equation 1, that is capable of theoretically solving this problem. For our particular case study however, the practical resolution of this problem fails. In a second step, we will propose a heuristic based on Equation 2 to solve the problem in practice.

4.1 Theoretical framework

We have separately estimated the two parts of Equation 1 as follows.

4.1.1 Step 1 : The $\mathrm{P}(S_iC_jA_kO_l)$ estimate

$\mathrm{P}(S_iC_jA_kO_l)$ represents the probabilities corresponding to the various cross-modalities in the synthetic population. These joint probabilities are calculated from the synthetic population. For example:
$\mathrm{P}(S_4C_5A_2O_2) = \frac{\text{number of synthetic households } S_4C_5A_2O_2}{\text{total number of synthetic households}}$

4.1.2 Step 2 : The $\mathrm{P}((I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l))$ estimate

This probability is estimated by the quantity:

$$\mathrm{P}((I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)) = \frac{\text{number of households}(I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)}{\text{total number of synthetic households}} \quad (3)$$

Number of households $(I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)$ is not known but can be estimated as demonstrated in the following.

$(I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)$ is one cross-modality among 35 530 and shall be numbered it k. The number of households whose cross-modality with income is $(I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)$ can then be denoted n_k, and the probability of this cross-modality being denoted $\mathbf{p}_k$ is the k[th] component of column vector $\mathbf{p}$. These notations serve to simplify Equation 3 as follows:

$$\mathrm{P}((I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)) = \mathbf{p}_k = \frac{n_k}{N} \quad (4)$$

Let's now introduce the random vector $\mathbf{n} = (n_1 \dots n_s)^t$, which describes a state of the population where n_1 households have the cross-modality with income 1,... and n_s households have the cross-modality with income s. The probability distribution of $\mathbf{n}$ is thus described by a multinomial distribution:

$$\mathrm{P}(\mathbf{n}|N, s, \mathbf{q}) = N! \prod_{k=1}^{s} \frac{\mathbf{q}_k^{n_k}}{n_k!} \quad (5)$$

where $\mathbf{q}_k$ is the prior probability of the cross-modality with income k. For the sake of simplicity, let's replace $\mathrm{P}(\mathbf{n}|N, s, \mathbf{q})$ by $\mathrm{P}(\mathbf{n})$ in the following.

$\mathbf{n}$ is constrained in order to ensure its consistency with the total number of households, the joint frequency calculated from the synthetic population, and the deciles.

- $\sum_k n_k = N$ is the natural constraint.
- $\sum_{k \in M_{S_i C_j A_k O_l}} n_k =$ the number of synthetic households whose cross-modality in the synthetic population is $S_i C_j A_k O_l$.

  $M_{S_i C_j A_k O_l}$ is the subset of indices corresponding to the cross-modality $S_i C_j A_k O_l$ regardless of the income modalities $]I_{m-1}, I_m[$. These constraints ensure consistency between $\mathbf{n}$ and the synthetic population.
- $\sum_{k \in M_{(I_{d-1,S_i} < I < I_{d,S_i})}} n_k = 10\% \times n_{S_i}$

  I_{d,S_i} is the income value corresponding to decile d for modality S_i. Table 2 displays these values.
  $M_{(I_{d-1,S_i} < I < I_{d,S_i})}$ is the subset of indices corresponding to the modality S_i, with $I_{d-1,S_i} < I < I_{d,S_i}$ regardless of the values of other modalities of the synthetic population $C_j A_k O_l$. n_{S_i} is the number of households whose modality is S_i. This constraint is repeated for all modalities of all variables and for all deciles. This set of constraints ensures the compatibility of $\mathbf{n}$ with the income information.

These constraints are linear, and summarized by:

$$\mathbf{C} \cdot \mathbf{n} = \mathbf{b} \quad (6)$$

By convention, the first row of $\mathbf{C}$ corresponds to the natural constraint; its components sum to equal 1. This property will be used subsequently.

The best solution to our problem is the most likely realization of the random vector $\mathbf{n}$:

$$\mathbf{n}^* = \underset{\mathbf{C} \cdot \mathbf{n} = \mathbf{b}}{\arg\max}\, \mathrm{P}(\mathbf{n}) \quad (7)$$

In accordance with Niven (2005), we will demonstrate that this optimization problem can be approximated by a cross-entropy minimization problem, which is much easier to solve.

The logarithm of $\mathrm{P}(\mathbf{n})$ is introduced in order to transform the multiplications $\prod$ from Equation 5 into a sum Σ. The referenced optimization problem then becomes:

$$\mathbf{n}^* = \underset{\mathbf{C}\cdot\mathbf{n}=\mathbf{b}}{\arg\max} \log(\mathrm{P}(\mathbf{n})) \tag{8}$$

In the following, $\mathbf{n}$ is considered to be a real random vector and not an integer random vector, i.e. components of this vector can be non-integer.

$\mathbf{n}^*$ is found by applying Fermat's Rule to the Lagrangian of this last maximization, i.e. the differentiation of the Lagrangian L at $\mathbf{n}^*$ is zero. Let's begin by differentiating the Lagrangian:

$$\frac{\partial L}{\partial n_k} = \frac{\partial}{\partial n_k}\left(\log(N!) - \log(n_k!)\right) + \log(\mathbf{q}_k) + \boldsymbol{\lambda}\cdot\mathbf{C}_k \tag{9}$$

with $\boldsymbol{\lambda}$ being Lagrange's multipliers, and $\mathbf{C}_k$ the k$^{\text{th}}$ column of Matrix $\mathbf{C}$.

According to Equation 9, this differentiation requires differentiating $\log(n_k!)$ with respect to n_k, a step that can be accomplished analytically using Stirling's Approximation.

$$\log(n_k!) \approx n_k\log(n_k) - n_k$$

From this approximation and by recalling that $N = \sum_k n_k$, Equation 9 becomes:

$$\frac{\partial L}{\partial n_k} \approx \log(N) - \log(n_k) + \log(\mathbf{q}_k) + \boldsymbol{\lambda}\cdot\mathbf{C}_k$$

Fermat's Rule yields:

$$\log(\mathbf{q}_k) - \log(\frac{n_k^*}{N}) + \boldsymbol{\lambda}\cdot\mathbf{C}_k = 0 \tag{10}$$

The first component of vector $\mathbf{C}_k$, i.e $\mathbf{C}_{k1}$, equals 1 according to the comment in Equation 6, which implies that for one $\boldsymbol{\lambda}$, we are able to define $\boldsymbol{\lambda}' = \boldsymbol{\lambda}$ except for its fist component. Therefore $\boldsymbol{\lambda}'_1 = \boldsymbol{\lambda}_1 + 1$, such that:

$$\log(\mathbf{q}_k) - \log(\frac{n_k^*}{N}) + \boldsymbol{\lambda}'\cdot\mathbf{C}_k = 1$$

This equation defines the solution to the following minimization problem:

$$\mathbf{p}^* = \underset{\mathbf{C}\cdot\mathbf{p}=\mathbf{b}/N}{\arg\min}\ \mathbf{p}^t\cdot\log(\frac{\mathbf{p}}{\mathbf{q}}) \tag{11}$$

$\mathbf{p}$ is a column vector whose component k is $\mathbf{p}_k$; $\log(\frac{\mathbf{p}}{\mathbf{q}})$ must be read as the column vector whose component k is $\log(\frac{\mathbf{p}_k}{\mathbf{q}_k})$.

By changing the variable of interest n_k into $\mathbf{p}_k = \frac{n_k}{N}$, Optimization Problem 11 is a minimization of the Kullback-Leibler Divergence from $\mathbf{q}$, often called the cross-entropy minimization problem and denoted MinxEnt in the literature. Its solution is then the solution to the multinomial maximization problem 8 subject to Stirling's Approximation. Since the solution to Problem 8 is the most likely realization, the MinxEnt solution will be called the most probable distribution subject to Stirling's Approximation.

In the absence of prior information on the probability of a given cross-modality with income, according to Laplace's Principle of Insufficient Reason, all prior probabilities are then equal, i.e. $\mathbf{q} = \frac{1}{s}$, with s being the number of cross-modalities with income. In this case, cross-entropy minimization Problem 11 becomes the entropy maximization problem, as denoted by MaxEnt in the literature:

$$\mathbf{p}^* = \underset{\mathbf{C}\cdot\mathbf{p}=\mathbf{b}/N}{\arg\max} -\mathbf{p}\log(\mathbf{p}) \tag{12}$$

In the following discussion, this analysis of the root of MaxEnt is applied to our specific study, given our lack of prior information on discrete probabilities $\mathrm{P}((I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l))$. However, two types of constraints are present, one on cross-modalities (Equation 15) in the synthetic population the other on income deciles (Equation 16). Let's now use the maximum entropy method:

$$Max \; - \sum_{i,j,k,l} \mathrm{P}((I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)). \ln \mathrm{P}((I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)) \tag{13}$$

subject to the following constraints.

The natural constraint:

$$\sum_{i,j,k,l} \mathrm{P}((I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)) = 1 \tag{14}$$

Constraints derived from the synthetic population:

$$\sum_{m} \mathrm{P}((I_{m-1} < I < I_m) \cap (S_i C_j A_k O_l)) = \mathrm{P}(S_i C_j A_k O_l) \tag{15}$$

This constraint means that for a given cross-modality, the sum of all income probabilities is equal to the probability of the cross-modality. Since 187 cross-modalities exist, 187 constraints of this type can be derived, with those from the deciles being:

$$\begin{aligned}
&\sum_{j,k,l} \mathrm{P}((S_i C_j A_k O_l) \cap (I < I_{d,S_i})) = d \\
&\sum_{i,k,l} \mathrm{P}((S_i C_j A_k O_l) \cap (I < I_{d,C_j})) = d \\
&\sum_{i,j,l} \mathrm{P}((S_i C_j A_k O_l) \cap (I < I_{d,A_k})) = d \\
&\sum_{i,j,k} \mathrm{P}((S_i C_j A_k O_l) \cap (I < I_{d,O_l})) = d
\end{aligned} \tag{16}$$

where d= 0.1, 0.2, ..., 0.9 and I_d is the income value corresponding to decile d for a given modality of the synthetic population.

In all, 19 modalities are present in the synthetic population; for each modality, there are 9 deciles and one maximum income, hence yielding 190 constraints of this type.

Let's rely on Kapur & Kesavan (1992) and Mattos & Veiga (2004) to numerically solve optimization Problem 12. The numerical algorithm is based on the solution to Equation 10:

$$\mathbf{p}_k = \mathbf{q}_k \exp(\boldsymbol{\lambda} \cdot \mathbf{C}_k) \tag{17}$$

This equation is the key to rewriting optimization Problem12 in its dual form:

$$\boldsymbol{\lambda}^* = \arg\max \sum_k (\log(\mathbf{q}_k) + \boldsymbol{\lambda} \cdot \mathbf{C}_k) \mathbf{q}_k \exp(\boldsymbol{\lambda} \cdot \mathbf{C}_k) + \boldsymbol{\lambda} \cdot (\mathbf{C} \cdot \mathbf{p} - \mathbf{b}/N) \tag{18}$$

The variables of interest in this dual optimization are the Lagrange multipliers whose number is considerably less than the number of variables of interest in the primal optimization, i.e. the number of constraints is far less than the dimension of the probability distribution. This dual optimization problem is solved by a Newton's algorithm, which is well suited and efficient by virtue of having demonstrated that the gradient and the Hessian of the objective function can be calculated analytically (Kapur & Kesavan 1992).

Before launching the optimization algorithm, it is advisable to verify that the constraints are consistent; otherwise every optimization algorithm would fail. This step implies searching for an initial probability distribution $\mathbf{p}^0$ consistent with the constraints. The initial distribution was found by solving the linear programming problem:

$$(\mathbf{p}^0 \; \mathbf{0}) = \underset{\mathbf{C} \cdot \mathbf{p} + \boldsymbol{\Delta} = \mathbf{b}/N, \boldsymbol{\Delta} \geq 0}{\arg\min} \mathbf{1} \cdot \boldsymbol{\Delta} \tag{19}$$

where: $\mathbf{\Delta}$ is a column vector whose dimension is equal to the number of constraints, $\mathbf{1}$ is a row vector of one whose dimension is the same as $\mathbf{\Delta}$, and $\mathbf{0}$ is a column vector of zero with the same dimension as $\mathbf{\Delta}$. $(\mathbf{p}^0\ \mathbf{0})$ is the concatenation of $\mathbf{p}^0$ and $\mathbf{0}$. The variable of interest in this Linear Programming set-up are $\mathbf{p}$ and $\mathbf{\Delta}$. The initialization is set by $\mathbf{p} = 0$ and $\mathbf{\Delta} = \mathbf{b}/N$. If the system is consistent, the minimum is reached with probability distribution $\mathbf{p}^0$ that verifies: $\mathbf{C} \cdot \mathbf{p}^0 = \mathbf{b}/N$ and $\mathbf{\Delta} = \mathbf{0}$. Linear programming Problem 19 is easy to solve numerically.

Some implementations of MaxEnt algorithm require that matrix $\mathbf{C}$ be full rank. If such is not the case, a subset of the rows of $\mathbf{C}$ is selected by using the QR decomposition with column pivoting of the transpose Matrix $\mathbf{C}^t$ (Golub et al. 1996). This point is not fundamental to understanding our approach and is being mentioned here in order to facilitate the understanding of the R-script accompanying this paper.

Our specific study, entails a numerical problem raised by virtue of using this classical numerical approach. Our optimization problem includes 378 (187+190+1) constraints (actually slightly less since all constraints are not independent) and 35 530 variables of interest. Equation 17 becomes numerically unstable due to the exponential function, which depends on too many of the various $\boldsymbol{\lambda}$ component values. The failure to compute this equation leads to failure of the algorithm. A heuristic yielding a practical solution is proposed below. The global optimization problem will be divided into several optimization sub-problems in order to lower the number of variables of interests and constraints.

4.2 Problem-solving heuristic

Let's proceed starting from Equation 2 (Bayes' theorem):

$$\mathrm{P}(I_{m-1} < I < I_m | S_i C_j A_k O_l) = \mathrm{P}(S_i C_j A_k O_l | I_{m-1} < I < I_m) \times \frac{\mathrm{P}(I_{m-1} < I < I_m)}{\mathrm{P}(S_i C_j A_k O_l)}$$

with :
$i = 1, 2, ..., 5$
$j = 1, 2, ..., 6$
$k = 1, 2, ..., 6$
$l = 1, 2$

4.2.1 Step 1 : The $\mathrm{P}(S_i C_j A_k O_l)$ estimate, see Section 4.1.1.

4.2.2 Step 2 : The $\mathrm{P}(I_{m-1} < I < I_m)$ estimate

I_m is a growing sequence of all income deciles (171 total) and the 19 maximum incomes (see Section 4). Let's now proceed with a linear extrapolation of these 190 incomes from the total population deciles (first row of Table 2). For example, if I_m lies between the I_{d-1} and I_d deciles of the total population, we estimate:

$$\mathrm{P}(I < I_m) = \mathrm{P}(I < I_{d-1}) + \frac{I_m - I_{d-1}}{I_d - I_{d-1}} (\mathrm{P}(I_d) - \mathrm{P}(I_{d-1})) \tag{20}$$

$\mathrm{P}(I < I_{m-1})$ is estimated a similar manner; consequently: $\mathrm{P}(I_{m-1} < I < I_m) = \mathrm{P}(I < I_m) - \mathrm{P}(I < I_{m-1})$.

4.2.3 Step 3 : The $\mathrm{P}(S_i C_j A_k O_l | I_{m-1} < I < I_m)$ estimate

- The next step consists of searching for the $\mathrm{P}(S_i C_j A_k O_l | I_{m-1} < I < I_m)$ distribution (p distribution). These quantities correspond to the 190 probability distributions, with each distribution being considered separately. Each distribution will be the solution to a MinxEnt optimization. The previous large optimization is now broken down into 190 smaller optimizations.
- Another difference with the previous large optimization problem is the prior assumptions. Let's suppose that $\mathrm{P}(S_i C_j A_k O_l | I_{m-1} < I < I_m)$ has an a priori distribution, such as the $\mathrm{P}(S_i C_j A_k O_l)$ distribution rather than a uniform distribution, which implies supposing that $\mathrm{P}(S_i C_j A_k O_l)$ provides more information on $\mathrm{P}(S_i C_j A_k O_l | I_{m-1} < I < I_m)$ than Laplace's Principle of Insufficient Reason.

As demonstrated above, the most probable distribution is the solution to the minimum cross-entropy (MinxEnt) (subject to Stirling's Approximation). The MinxEnt formulation is as follows:

$$\min \; - \sum_{i,j,k,l} \mathrm{P}(S_i C_j A_k O_l | I_{m-1} < I < I_m) \ln \frac{\mathrm{P}(S_i C_j A_k O_l | I_{m-1} < I < I_m)}{\mathrm{P}(S_i C_j A_k O_l)} \tag{21}$$

Subject to constraints on deciles :

$$\begin{aligned}
\mathrm{P}(S_i | I_{m-1} < I < I_m) &= \mathrm{P}(I_{m-1} < I < I_m | S_i) \times \frac{\mathrm{P}(S_i)}{\mathrm{P}(I_{m-1} < I < I_m)} \\
\mathrm{P}(C_j | I_{m-1} < I < I_m) &= \mathrm{P}(I_{m-1} < I < I_m | C_j) \times \frac{\mathrm{P}(C_j)}{\mathrm{P}(I_{m-1} < I < I_m)} \\
\mathrm{P}(A_k | I_{m-1} < I < I_m) &= \mathrm{P}(I_{m-1} < I < I_m | A_k) \times \frac{\mathrm{P}(A_k)}{\mathrm{P}(I_{m-1} < I < I_m)} \\
\mathrm{P}(O_l | I_{m-1} < I < I_m) &= \mathrm{P}(I_{m-1} < I < I_m | O_l) \times \frac{\mathrm{P}(O_l)}{\mathrm{P}(I_{m-1} < I < I_m)}
\end{aligned} \tag{22}$$

$\mathrm{P}(S_i)$, $\mathrm{P}(C_j)$, $\mathrm{P}(A_k)$ and $\mathrm{P}(O_l)$ are all obtained from the synthetic population.

$\mathrm{P}(I_{m-1} < I < I_m | S_i)$, $\mathrm{P}(I_{m-1} < I < I_m | C_j)$, $\mathrm{P}(I_{m-1} < I < I_m | A_k)$ and $\mathrm{P}(I_{m-1} < I < I_m | O_l)$ have also been estimated by linear extrapolation for each modality in the same way as in Step 2 for the entire population. Ultimately, 19 constraints and 187 variables of interest are obtained.

The consistency of each optimization is verified by means of Linear Programming in using the same method as described in the previous approach. The rank deficiency of the constraints matrix is also treated just like in the previous large optimization by using QR decomposition.

An existing implementation of the MinxEnt R package was adapted to our specific problem (package "minxent" by Senay Asma, available at the Comprehensive R Archive Network (CRAN)).More specifically, we used a computer of 2 x 2.60GHz CPU cores and 16 GB RAM; the computation time for these 190 optimizations was less than 1.5 seconds.

5 Results

In the proposed heuristic, we have successively estimated $\mathrm{P}(S_i C_j A_k O_l)$ **(Step 1)**, $\mathrm{P}(I_{m-1} < I < I_m)$ **(Step 2)** and $\mathrm{P}(S_i C_j A_k O_l | I_{m-1} < I < I_m)$ **(Step 3)**. We are now able to set the $\mathrm{P}(I_{m-1} < I < I_m | S_i C_j A_k O_l)$ distribution as formulated in Equation 2. The total number of estimated probabilities equals 35 530 (i.e 190 modalities for the income $\times$ 187 possible combinations). After a detailed analysis of these probabilities, some incorrect values were identified:

- $\mathrm{P}(I < I_m | S_i C_j A_k O_l) > 1$
- $\mathrm{P}(I_{m-1} < I < I_m | S_i C_j A_k O_l) < 0$

These invalid probabilities are due to both the relatively large number of parameters to be estimated and the very small differences existing between two successive deciles. For example, in Table 2 between deciles D8 for complex households (27 843 euros) and D7 for age modality 40-49 (27 842 euros), the difference amounts to just one euro. To correct this error, we performed a post-processing step in order to filter these invalid probabilities. The post-processing procedure is described below :

- For each $\mathrm{P}(I < I_m | S_i C_j A_k O_l) > 1$, we set $\mathrm{P}(I_{m-1} < I | S_i C_j A_k O_l) = 0$; this renormalization of probability distribution $\mathrm{P}(I_{m-1} < I < I_m | S_i C_j A_k O_l)$ is intended to avoid sensitivity to maximum Income I_{max}, which has been arbitrarily fixed at 1.5 times the 9^{th} decile ($1.5 \times D9$).

- For each $\mathrm{P}(I_{m-1} < I < I_m | S_i C_j A_k O_l) < 0$, the interval is enlarged until $\mathrm{P}(I_{m-1} < I < I_{m+k} | S_i C_j A_k O_l) \geq 0$ with $k \geq 1$, which means that components m through m+k-1 are removed from income column vector $\mathbf{I}$.

This post-processing step has resulted in larger income brackets while also ensuring data consistency. The remaining probabilities after this step are those of income ranges according to household characteristics. A specific income is then to be allocated to each household of the synthetic population. Through the synthetic population, the number of households for each cross-modality is known, as is the income distribution for each cross-modality.

- We begin by compiling the number of households for each income range according to the income distribution. For example, the total number of $S_1 C_1 A_1 O_2$ households (size=1, single woman, under 30, tenant) equals 14 613. From the $\mathrm{P}(I_{m-1} < I < I_m | S_1 C_1 A_1 O_2)$ distribution, income groups are therefore assigned to households in this cross-modality.
- Next, a specific income is randomly assigned according to a continuous uniform distribution; this income lies between the lower and upper bounds of each income range.

Upon completion of of these two operations, each household in the synthetic population is allocated an income. The next section will verify the accuracy of our results.

5.1 Validation of results

To assess the performance of the proposed heuristic, let's compare our outputs (i.e. the simulated income for each household) with empirical observations (deciles from FiLoSoFi). New deciles from the specific simulated household incomes were first recalculated (Table 3) before applying two validation methods.

Table 3: Distribution of simulated deciles

Modalities	Deciles (euros)								
	D1	D2	D3	D4	D5	D6	D7	D8	D9
Entire population	9 993	13 333	16 053	18 632	21 249	24 152	27 684	32 500	41 549
1 person	9 699	12 664	14 972	17 160	19 361	21 842	24 969	29 408	37 239
2 persons	11 086	15 692	18 803	21 545	24 237	24 411	31 302	36 758	48 706
3 persons	9 746	13 397	16 649	19 713	22 840	25 931	29 685	35 263	47 193
4 persons	10 441	14 085	17 833	21 019	24 052	27 275	30 603	35 619	46 409
5 persons/+	8 867	11 120	12 758	16 048	20 769	24 991	29 652	35 879	49 401
Single woman	10 293	13 106	15 295	17 376	19 556	22 016	25 177	29 475	36 872
Single man	9 093	12 340	14 528	16 902	19 147	21 612	24 721	29 356	37 835
Couple without children	14 381	18 363	21 276	23 837	26 554	29 777	33 981	39 419	54 456
Couple with children	10 837	14 475	18 359	21 648	24 663	27 889	31 737	37 760	51 965
Single-parent family	8 703	10 617	12 029	13 782	15 818	18 248	21 331	25 680	32 993
Complex households	9 085	13 338	15 809	17 848	20 301	22 927	25 887	30 109	35 901
Under 30	8 463	11 465	13 820	16 049	18 104	20 246	22 700	26 016	31 803
30-39	10 355	13 206	16 002	18 701	21 425	24 464	27 706	31 898	39 302
40-49	10 049	13 002	15 594	18 577	21 637	24 943	28 894	33 657	44 341
50-59	10 719	13 849	17 042	20 082	23 333	26 831	30 972	37 151	50 267
60-74	12 564	15 818	19 060	22 002	25 108	28 716	32 658	39 107	54 787
75/+	14 055	16 638	18 967	21 299	23 833	26 691	30 533	35 770	46 406
Owner	17 025	20 499	23 125	25 598	28 221	31 306	35 196	41 078	53 325
Tenant	8 529	11 692	13 724	15 528	17 466	19 676	22 377	26 311	32 716

Note: The first decile (D1) is the income below which 10% of incomes are situated; The ninth decile (D9) is the income below which 90% of incomes are situated.

According to the first method, a graphical representation of the simulated and observed deciles was developed through quantile-quantile plots (Q-Q plots); these representations are a graphical evaluation of the fit between simulated and real data and moreover easily detect possible outliers. Figure 1 compares the distribution of simulated deciles to the distribution of real (observed) deciles for each modality as well as for the entire population. In these Q-Q plots, each point represents a simulated decile, and the reference line plotted is the first bisector. If the two distributions are identical, the points on the graph follow the line exactly.

The obtained Q-Q plots indicate a perfect fit between the simulated deciles and observed deciles from FiLoSoFi for the entire population. For all other plotted modalities, the majority of points (over 80%) fit along the first bisector, thus showing that the simulated deciles fit very well with the observed deciles, which demonstrates that the synthetic population has a simulated income on the whole consistent with the FiLoSoFi data.

We can also state that the fit of modalities "complex households", "5 persons or more" and "single-parent family" is not as good. It can also be noted that in general, deciles D8 and especially D9 deviate the most from the reference line.

Next, the second validation method estimates the accuracy of the proposed heuristic with two metrics typically used in a microsimulation context:

- Absolute error $|I_m - I_n|$
- Relative error $|\frac{I_m - I_n}{I_m}|$

where I_m and I_n are the observed and simulated deciles for each modality. Tables 4 and 5 respectively present the absolute and relative errors that exist between the two distributions. For the entire population deciles, the

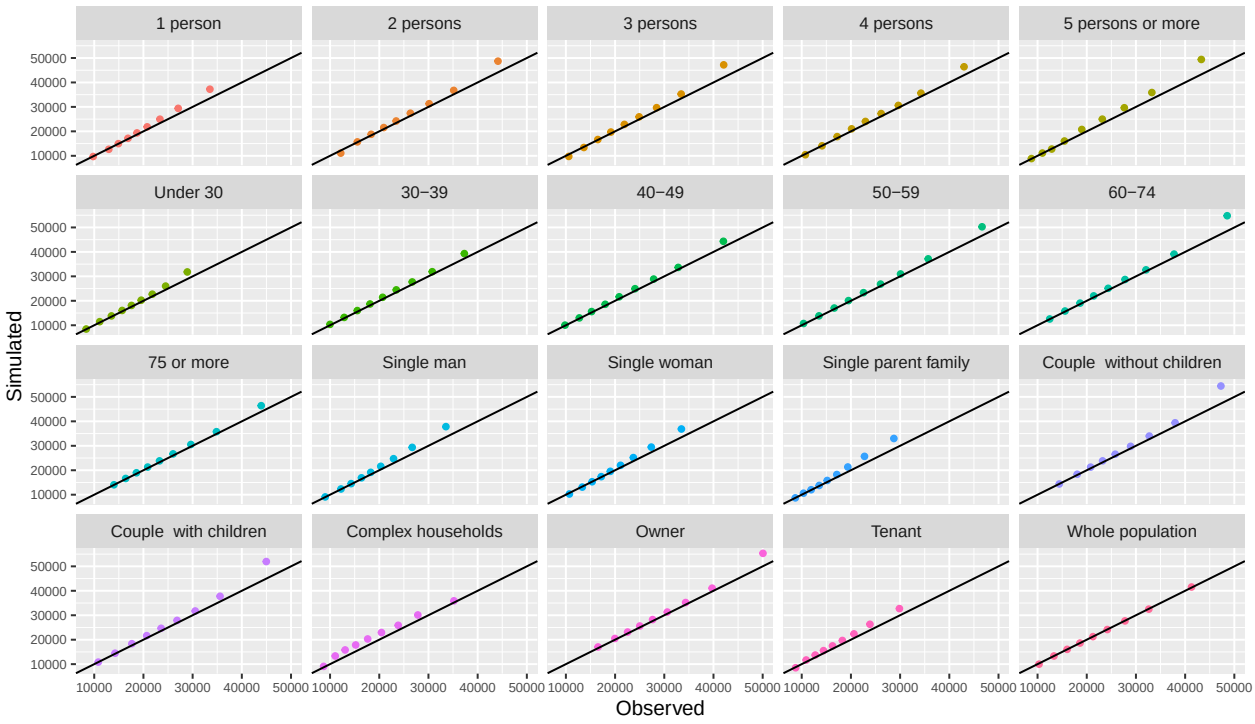

Figure 1: Quantile-Quantile plots between simulated and observed deciles

differences between simulated and observed deciles are very small (around 3% for the first decile D1 and less than 0.6% for the other deciles), with absolute differences of 3 euros and 1 euro respectively for deciles D2 and D4 (Table 4).

For the other modalities, small differences are also observed: 88% of the simulated deciles exhibit differences of less than 10% with the real deciles and about 69% of the simulated deciles have differences of less than 5%. The smallest relative errors are observed for the variables "age of the reference person" and "number of persons in the tax household", with differences of less than 1% for decile D1 of modalities "75/+", "60-74" and "1 person". Morever, our results show that the simulated incomes are more consistent with the FiLoSoFi data for deciles D1 through D7. More than 80% of these deciles have a relative difference of less than 5% with the real data.

In contrast, only 71% of the simulated D8 and D9 deciles have relative differences of less than 10% with the real deciles. As previously noted in Figure 1, larger differences are observed for the household family composition variable. For the "complex household" and "single parent family" modalities in particular, the respective relative differences amount to over 15% (for deciles D2 to D4) and over 10% (for deciles D7 to D9). Possible reasons for such discrepancies will be discussed in the next section.

Table 4: Absolute errors between observed and simulated deciles

Modalities	Deciles (euros)								
	D1	D2	D3	D4	D5	D6	D7	D8	D9
Entire population	310	3	29	1	14	36	90	120	241
1 person	95	297	58	295	674	1 079	1 612	2 339	3 725
2 persons	1 090	139	447	626	802	1 080	1 162	1 622	4 572
3 persons	838	259	160	568	947	1 040	1 245	1 831	5 114
4 persons	299	45	626	881	1 097	1 127	959	1 381	3 411
5 persons/+	109	130	121	581	1 778	1 827	2 014	2 641	6 109
Single woman	421	228	37	190	525	905	1 462	2 115	3 392
Single man	77	116	240	514	879	1 307	1 813	2 660	4 284
Couple without children	36	297	485	612	769	866	1 263	1 458	7 183
Couple with children	15	237	713	983	1 067	1 052	1 209	2 187	6 988
Single-parent family	1	250	114	225	639	1 113	1 961	2 919	4 260
Complex households	393	2 286	2 746	2 641	2 653	2 475	2 034	2 266	722
Under 30	92	348	319	371	532	689	897	1 503	2 883
30-39	370	334	463	579	737	1 001	1 002	1 127	2 002
40-49	222	269	368	584	798	888	1 052	820	2 323
50-59	348	337	425	573	772	801	877	1 441	3 609
60-74	90	236	419	590	748	938	609	1 356	6 239
75/+	50	249	384	430	558	663	885	921	2 461
Owner	482	533	580	576	595	694	860	1 351	5 265
Tenant	235	780	976	1 095	1 201	1 410	1 762	2 441	2 856

Note: The first decile (D1) is the income below which 10% of incomes are situated; The ninth decile (D9) is the income below which 90% of incomes are situated.

Table 5: Relative errors between observed and simulated deciles (%)

Modalities	Deciles								
	D1	D2	D3	D4	D5	D6	D7	D8	D9
Entire population	3.01	0.02	0.18	0.01	0.07	0.15	0.32	0.37	0.58
1 person	0.97	2.29	0.39	1.75	3.61	5.20	6.90	8.64	11.11
2 persons	8.95	0.89	2.44	2.99	3.42	4.10	3.86	4.62	10.36
3 persons	7.92	1.90	0.97	2.97	4.33	4.18	4.38	5.48	12.15
4 persons	2.78	0.32	3.64	4.37	4.78	4.31	3.24	4.03	7.93
5 persons/+	1.24	1.18	0.94	3.76	9.36	7.89	7.29	7.95	14.11
Single woman	3.93	1.71	0.24	1.11	2.76	4.29	6.16	7.73	10.13
Single man	0.85	0.95	1.68	3.14	4.81	6.44	7.91	9.96	12.77
Couple without children	0.25	1.64	2.33	2.64	2.98	3.00	3.86	3.84	15.19
Couple with children	0.14	1.66	4.04	4.76	4.52	3.92	3.96	6.15	15.54
Single-parent family	0.01	2.41	0.96	1.66	4.21	6.50	10.12	12.82	14.83
Complex households	4.52	20.68	21.02	17.37	15.03	12.10	8.53	8.14	2.05
Under 30	1.10	3.13	2.36	2.37	3.03	3.52	4.11	6.13	9.97
30-39	3.71	2.59	2.98	3.20	3.56	4.27	3.75	3.66	5.37
40-49	2.26	2.11	2.42	3.25	3.83	3.69	3.78	2.50	5.53
50-59	3.36	2.49	2.56	2.94	3.42	3.08	2.91	4.04	7.74
60-74	0.72	1.51	2.25	2.76	3.07	3.38	1.90	3.59	12.85
75/+	0.36	1.52	2.07	2.06	2.40	2.55	2.99	2.64	5.60
Owner	2.91	2.67	2.57	2.30	2.15	2.27	2.50	3.40	10.52
Tenant	2.68	7.15	7.66	7.59	7.38	7.72	8.55	10.23	9.56

Note: The first decile (D1) is the income below which 10% of incomes are situated; The ninth decile (D9) is the income below which 90% of incomes are situated.

6 Discussion

The discussion herein comprises two parts: an analysis of case study results, and an assessment of method applicability.

6.1 Case study analysis

Our findings suggest that the proposed methodology yields results consistent with most of the observed aggregate income data despite the fact that a number of income data points were removed in the post-processing stage. However, some larger differences were found for specific household modalities as well as for the highest deciles (D8 and D9).

6.1.1 Larger differences for some household modalities

Differences between modalities are due to the units surveyed in the two data sources used as highlighted in the problem description section (paragraph 3). According to the FiLoSoFi protocol, the survey unit is the tax household (all individuals included on the same income tax declaration). A person can thus be affiliated with a household from a tax perspective without necessarily residing in that household. According to the census protocol, only the inhabitants of a given dwelling are taken into account. Consequently, the number of households

and the definition of certain modalities in FiLoSoFi are not always the same as in the population census data. This phenomenon is especially true for the household family composition variable, for which differences exist in the definition of modalities between the two sources. Some households may therefore belong to two different modalities. For example, from the standpoint of the population census, a single-parent household comprises a lone parent and one or more single children (who are not parents). On the other hand, in the FiLoSoFi database, a tax household can be considered as a single-parent household if it is composed of several persons with the lead tax registrant being single, divorced or widowed. Such a household would be considered as a single-parent household in FiLoSoFi and as a complex household in the census. These discrepancies are therefore beyond our control and cannot be corrected by the proposed heuristic.

6.1.2 Larger differences for the highest deciles (D8 and D9)

Since no information is available between any two income deciles, we have assumed a linear income distribution. For each modality, we have arbitrarily set a maximum decile D10= D9$\times$1.5 for purposes of the linear approximation. This assumption may however be incorrect for the highest deciles and thus lead to certain biases.

6.2 Applicability of this methodology

The databases and R-scripts used herein are freely available. We will raise the level of community appropriation of these tools by means of developing an R library.

In our case study, aggregate data take the form of deciles. Our heuristic has been designed to process this type of information. However, our general methodology is able to incorporate various types of aggregate data: average income, number of square meters per type of household, etc. As an example, we can use typical aggregate data such as the mobility data available in Switzerland [4] or the United States [5], to determine a transportation mode or mobility status of synthetic individuals.

Our methodology could be formally applied to sample-less synthetic population generation as well, in which case the cross-modality probabilities are computed from the synthetic population. We did not test our methodology for this specific case.

Numerical problems arise when carrying out the large optimization method for our case study. We are currently working on alternative, more numerically robust algorithms. In the meantime, we feel that this method can be applied to problems with fewer constraints and variables of interest, as suggested by some preliminary simulations we conducted on smaller-scale problems. On such problems, the optimized and heuristic solutions are close to one another. These points will be subsequently studied in greater depth.

Since no information was available between two successive income deciles, income linearity was hypothesized. Moreover, for each modality we set a maximum income I_{max} equal to 1.5 times the 9^{th} decile and a minimum income I_{min} to zero. These assumptions however may be incorrect and produce erroneous results. This point can be remedied by studying the literature in income economics to gain better a priori knowledge of income distribution and, particularly, for these extreme incomes.

The accuracy of Stirling's Approximation depends on the number of households for a cross-modality with incomes. If this number is greater than 20, the approximation underestimated by less than 10%. To numerically resolve the optimization problem, the number of households for a cross-modality with income is considered as a real variable, thus suggesting that this number must be greater than 20. Caution is therefore required when handling results relative to a small number of households, which is a quite common cautious rule in statistics.

This last remark provides an a posteriori justification of our post-processing treatment, which yielded larger income brackets. One method improvement could consist of finding the right income discretization in order to produce wider income ranges that lead to a more suitable number of households for a cross-modality with incomes. This change would result in fewer invalid probabilities, as suggested by our post-processing treatment.

[4] https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases/tables.assetdetail.7226558.html, consulted on January 20, 2020.

[5] https://www.census.gov/content/census/en/data/tables/2020/demo/geographic-mobility/cps-2020.html, consulted on January 20, 2020.

7 Conclusion

This paper has presented a methodology to tackle the problem of adding variables to a synthetic population from aggregate data; this problem has to date received scant attention in the literature. Such a methodology integrates three distinct stages, the first of which theoretically models the problem by means of a multinomial distribution. The issue herein is to identify the most probable conditional probability of the income in knowing the cross-modalities of the synthetic population. This conditional probability derives from an entropy maximization problem based on the variables available in both the synthetic population and the aggregate data. In our specific case study, solving this problem directly proves to be impossible due to the high number of constraints. The second stage then presented a heuristic offering a practical solution to the problem. This heuristic combined Bayes' theorem and the cross-entropy minimization algorithm. However, given the large number of parameters needed to be estimated by the proposed heuristic, some of the results obtained were invalid. To rectify this shortcoming, a post-processing method was applied during a third stage so as to ensure the consistency of our results.

An income was allocated to each of the 157 000 households in the city of Nantes (France) based on aggregate data from the FiloSoFi through use of this method. The results were easy to compute and wound up being consistent with most of the aggregate data.

Special attention was paid to the reproducibility of our results with the databases and R-scripts used, all of which are freely available. This method remains general and indeed applicable to other variables with available aggregate data. To the best of our knowledge, no clear methodology for adding variables to a synthetic population from aggregate data already exists. An R library based on the findings presented herein is currently under development.

8 Acknowledgments

Linear programming and MinxEnt analyses were performed using the "boot" and "minxent" packages developed in R language. For graphical representations, the "ggplot2" and "lattice" R packages were run.

References

Avram, S., Figari, F., Leventi, C., Levy, H., Navicke, J., Matsaganis, M., Militaru, E., Paulus, A., Rastringina, O. & Sutherland, H. (2013). The distributional effects of fiscal consolidation in nine eu countries. Tech. rep., Euromod working paper

Barthelemy, J. & Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science*, *47*(2), 266–279

Bösch, P. M., Müller, K. & Ciari, F. (2016). The ivt 2015 baseline scenario. In *16th Swiss Transport Research Conference (STRC 2016)*. 16th Swiss Transport Research Conference (STRC 2016)

D'Orazio, M., Di Zio, M. & Scanu, M. (2006). *Statistical matching: Theory and practice*. John Wiley & Sons

Edwards, K. L. & Clarke, G. (2013). Simobesity: combinatorial optimisation (deterministic) model. In *Spatial Microsimulation: A reference guide for users*, (pp. 69–85). Springer

Felbermair, S., Lammer, F., Trausinger-Binder, E. & Hebenstreit, C. (2020). Generating synthetic population with activity chains as agent-based model input using statistical raster census data. *Procedia Computer Science*, *170*, 273–280

Gargiulo, F., Ternes, S., Huet, S. & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. *PloS one*, *5*(1), e8828

Golub, G. H., Loan, C. F. V., Loan, C. F. V. & Golub (1996). *Matrix Computations*. JHU Press

Hackl, J. & Dubernet, T. (2019). Epidemic spreading in urban areas using agent-based transportation models. *Future Internet*, *11*(4), 92

He, B. Y., Zhou, J., Ma, Z., Chow, J. Y. & Ozbay, K. (2020). Evaluation of city-scale built environment policies in new york city with an emerging-mobility-accessible synthetic population. *Transportation Research Part A: Policy and Practice*, *141*, 444–467

Hörl, S. & Balac, M. (2020). Reproducible scenarios for agent-based transport simulation: A case study for paris and île-de-france

Huynh, N. N., Barthelemy, J. & Perez, P. (2016). A heuristic combinatorial optimisation approach to synthesising a population for agent-based modelling purposes. *Journal of Artificial Societies and Social Simulation*, *19*(4), 11

Ilahi, A. & Axhausen, K. W. (2019). Integrating bayesian network and generalized raking for population synthesis in greater jakarta. *Regional Studies, Regional Science*, *6*(1), 623–636

Kapur, J. N. & Kesavan, H. K. (1992). Entropy Optimization Principles and Their Applications. In V. P. Singh & M. Fiorentino (Eds.), *Entropy and Energy Dissipation in Water Resources*, Water Science and Technology Library, (pp. 3–20). Dordrecht: Springer Netherlands
URL https://doi.org/10.1007/978-94-011-2430-0$_1$

Lovelace, R. & Ballas, D. (2013). 'truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, *41*, 1–11

Mattos, R. & Veiga, A. (2004). Entropy optimization: Computer implementation of the maxent and minexent principles. Tech. rep., Working Paper. Universidade Federal de Juiz de Fora, Brazil

Müller, K. & Axhausen, K. W. (2012). Multi-level fitting algorithms for population synthesis. *Arbeitsberichte Verkehrs-und Raumplanung*, *821*

Murata, T., Sugiura, S. & Harada, T. (2017). Income allocation to each worker in synthetic populations using basic survey on wage structure. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, (pp. 1–6). IEEE

Niven, R. K. (2005). Combinatorial Information Theory: I. Philosophical Basis of Cross-Entropy and Entropy
URL https://arxiv.org/pdf/cond-mat/0512017.pdf

O'Sullivan, D. (2008). Geographical information science: agent-based models. *Progress in Human Geography*, *32*(4), 541–550

Rich, J. (2018). Large-scale spatial population synthesis for denmark. *European Transport Research Review*, *10*(2), 63

Saadi, I., Mustafa, A., Teller, J., Farooq, B. & Cools, M. (2016). Hidden markov model-based population synthesis. *Transportation Research Part B: Methodological*, *90*, 1–21

Sallard, A., Balać, M. & Hörl, S. (2020). A synthetic population for the greater são paulo metropolitan region. *Arbeitsberichte Verkehrs-und Raumplanung*, *1545*

Sutherland, H. & Figari, F. (2013). Euromod: the european union tax-benefit microsimulation model. *International Journal of Microsimulation*, *6*(1), 4–26

Tanton, R. & Edwards, K. L. (2013). Introduction to spatial microsimulation: history, methods and applications. In *Spatial Microsimulation: A reference guide for users*, (pp. 3–8). Springer

Thiriot, S. & Sevenet, M. (2020). Pairing for generation of synthetic populations: the direct probabilistic pairing method. *arXiv preprint arXiv:2002.03853*

Tomintz, M. N., Clarke, G. P. & Rigby, J. E. (2008). The geography of smoking in leeds: estimating individual smoking rates and the implications for the location of stop smoking services. *Area*, *40*(3), 341–353

Vosooghi, R., Puchinger, J., Jankovic, M. & Vouillon, A. (2019). Shared autonomous vehicle simulation and service design. *Transportation Research Part C: Emerging Technologies*, *107*, 15–33

Yaméogo, B. F., Gastineau, P., Hankach, P. & Vandanjon, P.-O. (2020). Comparing methods for generating a two-layered synthetic population. *Transportation Research Record*. doi:10.1177/0361198120964734
URL http://dx.doi.org/10.1177/0361198120964734

Zhang, D., Cao, J., Feygin, S., Tang, D., Shen, Z.-J. M. & Pozdnoukhov, A. (2019). Connected population synthesis for transportation simulation. *Transportation Research Part C: Emerging Technologies*, *103*, 1–16