Hierarchical IPF: Generating a synthetic population for Switzerland

Kirill Müller

Kay W. Axhausen

ERSA 2011 **June 2011**

IVT Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ERSA 2011

Hierarchical IPF: Generating a synthetic population for Switzerland

Kirill Müller, Kay W. Axhausen
IVT
ETH Zürich
8093 Zürich
phone: +41-44-633 33 17
fax: +41-44-633 10 57
{mueller,axhausen}@ivt.baug.ethz.ch

June 2011

Abstract

Agent-based microsimulation models for land use or transportation simulate the behavior of agents over time, although at different time scales and with different goals. For both kinds of models, the initial step is the definition of agents and their relationships. Synthesizing the population of agents often is the only solution, due to privacy and cost constraints. In this paper, we assume that the model simulates persons grouped into households, and a person/household population needs to be synthesized. However, the methodology presented here can be applied to other kinds of agent relationships as well, *e.g.*, persons and jobs/workplaces or persons and activity chains.

Generating a synthetic population requires (a) reweighting of an initial population, taken from census or other survey data, with respect to current constraints, and (b) choosing the households that belong to the generated population. We propose an algorithm that estimates household-level weights that fit the control totals at both person and household levels. This eliminates the need to account for person-level control during the generation of synthetic households. The algorithm essentially performs a proportional fitting in the domains of both households and persons, and introduces an entropy-optimizing fitting step to switch between these two domains. We evaluate the performance of our algorithm by generating a synthetic population for Switzerland and checking it against the complete Swiss census. The validation is performed using information-based and distance-based statistics, and proves that the new algorithm is highly competitive to the approaches presented by Ye *et al.* (2009) and Bar-Gera *et al.* (2009).

Keywords

Population synthesis, Microsimulation, Households, Disaggregation, IPF, Iterative Proportional Fitting, Hierarchical, Simultaneous Control, Multi-Level, Multi-Domain, Relative Entropy

1 Introduction

Agent-based microsimulation model systems for land use and transportation planning have come into widespread use (UrbanSim, 2011; MATSim-T, 2011; Bradley *et al.*, 2010; Beckx *et al.*, 2009; Roorda *et al.*, 2008; Bhat *et al.*, 2004; Ben-Akiva *et al.*, 2002; Bowman and Ben-Akiva, 2001; de Palma and Marchal, 2002; Mahmassani *et al.*, 1995). They simulate decisions of agents within an urban area, allowing for more detailed and accurate simulation and prediction of, *e.g.*, land pricing and travel demand than traditional aggregate models. Often, the *agents* represent the individual people living in the study area, grouped by households. Other kinds of agents and relationships are of interest as well, such as employees/firms or dwellings/buildings (Ryan *et al.*, 2009). In this paper we focus on person/household populations.

When implementing such a model system, the initial step is the definition of agents and their relationships. For many countries, one suitable data source is the national census that is collected on a regular basis. Census data must be prepared in order to be useful as input for microsimulation. First, complete census data is rarely available: Often, only a small subsample, the *public-use sample*, can be accessed. Information in that sample may be randomly rounded, aggregated, or removed altogether. Second, the census is collected rather infrequently: As much as 10 years can pass between two consecutive surveys.

The objective of *population synthesis* is to compensate for the difficulties above. The main idea is to combine census microdata with readily available up-to-date aggregate data. Both data source are used to generate a set of agents for which (a) the distribution and correlation of the agents' attributes are similar to those in the census microsample, and (b) the number of agents within each category matches the current aggregate data. Two fundamentally different kinds of population synthesis procedures are distinguished:

Synthetic reconstruction methods (SR) generate the synthetic population by combining joint distributions over different attribute sets using IPF, and then drawing from the reference sample using this fitted joint distribution. Recent contributions in the literature include (Auld and Mohammadian, 2010; Pritchard and Miller, 2009; Srinivasan and Ma, 2009; Ye *et al.*, 2009; Bar-Gera *et al.*, 2009; Guo and Bhat, 2007); see also (Müller and Axhausen, 2011) for a literature review over SR techniques.

Combinatorial optimization techniques (CO) estimate integer weights for the reference sample that minimize a suitable objective function. Ryan *et al.* (2009) have evaluated this approach for synthesizing a population of firms.

In reality, personal decisions are affected not only by personal attributes, but also by the individual family situation – *i.e.*, whether a partner, children, or other persons live in the same household (Jones *et al.*, 1983). Therefore, replicating the proper household structure is a major requirement for the agent population in order to be able to simulate these interactions. In this paper, we present a novel algorithm to fit a household sample with person information to given control totals, and compare two similar algorithms (Ye *et al.*, 2009; Bar-Gera *et al.*, 2009) to ours. We use all three algorithms to synthesize a population for Switzerland based on a 5 % sample of the Swiss census.

The remainder of this paper is structured as follows. In the next section we show the evolution of our algorithm from a variant of IPF and compare its results to those of two similar algorithms for a toy example. The subsequent section present the analysis of a synthetic population for Switzerland. We conclude with a summary and an outlook.

2 Algorithm

Generating synthetic populations using the Synthetic Reconstruction (SR) method consists of two principal stages: *fitting* and *generation*. The purpose of the fitting stage is to reweigh a disaggregate sample of agents (called *reference sample*), representing the full population of the study area, so that the reweighted sample matches aggregate constraints (referred to as *control totals* or *controls*). The fractional *expansion factor* is then used to construct a disaggregate set of persons and households with attributes required by the microsimulation model in the generation stage.

A frequently applied pattern is to estimate household-level expansion factors using IPF during the fitting stage, so that they match the control totals for the households, and then, using these expansion factors, generate a population of households that best fits the person-level control totals (Auld and Mohammadian, 2010; Srinivasan and Ma, 2009; Guo and Bhat, 2007). However, this complicates the generation stage and sometimes requires time-consuming computations not suitable for frequent repetition.

Recently, Ye *et al.* (2009) presented a technique called *Iterative Proportional Updating* (IPU) that estimates a fractional expansion factor for each household so that both household- and person-level controls are satisfied. Bar-Gera *et al.* (2009) use an entropy optimization approach to achieve the same goal, from hereon referred to as *Ent*. For all three procedures, construction of the final population is possible using fast, simple and well-understood probabilistic drawing with replacement, allowing to generate thousands of instances of similar yet different populations, *e.g.*, in the context of multiple imputation (Rubin, 1987). This paper presents *HIPF*, another algorithm for multi-level fitting serving the same purpose as IPU and Ent. The output of the three algorithms is compared in Section 3.

HIPF operates on the following input data:

- a representative reference sample that contains the characteristics of households and all persons belonging to the sampled households, and
- categorized control totals for selected attributes on both household and person levels.

Just like IPU and Ent, the algorithm is equivalent to IPF if control totals are provided for only the household or only the person level.

In the following, we describe the HIPF algorithm as an transition from IPF. We briefly summarize the IPF algorithm in its list-based version presented by Pritchard and Miller (2009) and introduce the missing pieces in order to finally present the HIPF algorithm.

2.1 Fitting Households

The list-based version of IPF estimates fractional expansion factors f_h^* for all households h of the reference sample so that the household-level control totals are satisfied. For the disaggregate reference sample given as a multiset of category tuples

$$H = \{(a_1, b_1, \dots), \dots, (a_h, b_h, \dots), \dots, (a_n, b_n, \dots)\}, \tag{1}$$

we define the set of households H_{ab} for a given combination of categories (a, b) as follows:

$$H_{ab} := \{h \in H : (a_h, b_h, \dots) = (a, b, \dots)\}. \tag{2}$$

In addition, we define the sum F_{ab} of all expansion factors for given categories a, b:

$$F_{ab} := \sum_{h \in H_{ab}} f_h. \tag{3}$$

For any category x, H_x and F_x are defined analogously. The control total for category a is denoted by C_a; we assume that only C_a and C_b are given. It is a trivial precondition for the convergence of IPF that the grand totals match for all attributes: $n = \sum_a C_a = \sum_b C_b$.

At $k = 0$, for the first iteration, the expansion factor is initialized to unity: $f_h^{(0)} := 1$. After that, the procedure FIT shown in Figure 1 is invoked repeatedly to compute a series $f_h^{(k)}$ of expansion factors.

We define f_h as the limit of $f_h^{(k)}$, if it exists: $f_h^* := \lim_{k\to\infty} f_h^{(k)}$. The convergence of IPF depends on the data. Two problems exist in the given context:

Missing observations If the control totals are nonzero for a combination of control categories without a corresponding reference sample, a division by zero occurs during execution of IPF. This is also referred to as the *zero-cell problem* – Figure 2(a) illustrates this.

Figure 1: Procedure FIT$(f_h, C_a, C_b, \ldots)$ – the Fitting Step in List-Based IPF

```
Require: Reference sample, expansion factors $f_h$, control totals $C_a$, $C_b$ and possibly others
Ensure: Expansion factors $f_h$ for the next iteration
  $f_h \leftarrow C_a \div \sum_i F_{a_h i}$ for all $h \in H$
  $f_h \leftarrow C_b \div \sum_i F_{i b_h}$ for all $h \in H$
  (accordingly for all other control totals)
  return $f_h$
```

Figure 2: Two Possible Reasons for Non-Convergence of IPF

(a) Missing observations ($C_a \neq 0$)

$$
\begin{array}{|c|c|c||c|}
\hline
 & 0 & & \\
 & \vdots & & \\
\ddots & 0 & \ddots & \vdots \\
 & \vdots & & \\
 & 0 & & \\
\hline
\hline
\cdots & C_a & \cdots & n \\
\hline
\end{array}
$$

(b) Conflicting control totals ($C_a \neq C_b$)

$$
\begin{array}{|c|c|c||c|}
\hline
 & 0 & & \\
\ddots & \vdots & \ddots & \vdots \\
 & 0 & & \\
\hline
0 \ \cdots \ 0 & 1 & 0 \ \cdots \ 0 & C_b \\
\hline
 & 0 & & \\
\ddots & \vdots & \ddots & \vdots \\
 & 0 & & \\
\hline
\hline
\cdots & C_a & \cdots & n \\
\hline
\end{array}
$$

Various remedies have been suggested in the literature, such as introducing the missing observations with arbitrarily small weights, combining rare categories, or borrowing from other regions; *cf.* (Müller and Axhausen, 2011) for an overview.

Conflicting controls If an observation is unique within two or more combinations of control categories, and the control totals for these categories differ, this observation's expansion factor will oscillate between the two control values. Figure 2(b) shows an example for the above case; however, more complicated settings exist where IPF does not converge. Pukelsheim and Simeone (2009) have proven necessary and sufficient conditions for the convergence of IPF, derivable directly from the input data.

These problems need to be taken care of when implementing IPF or a variant thereof. In the following, we assume that the reference sample and the control totals are well-conditioned so that no convergence problems occur; see also Section 3.1.

The procedure LIPF in Figure 3 implements the algorithm. It can be easily extended to fitting against more than two control variables, or fitting against cross-classified control variables.

Figure 3: Procedure LIPF$(C_a, C_b, \ldots)$ – the List-Based IPF Algorithm

```
Require: Reference sample, control totals $C_a$, $C_b$ and possibly others
Ensure: Expansion factors $f_h$ obeying all control totals
  $k \leftarrow 0$
  $f_h^{(0)} \leftarrow 1$ for all $h \in H$
  repeat
    $f_h^{(k+1)} \leftarrow \mathrm{FIT}(f_h^{(k)}, C_a, C_b, \ldots)$
    $k \leftarrow k + 1$
  until convergence
  return $f_h^{(k)}$
```

2.2 Fitting Persons

Recall that our reference sample features not only the households, but also the persons in these households. We denote this by another multiset of attribute tuples:

$$P = \{(\alpha_1, \beta_1, \ldots, h_1), \ldots, (\alpha_p, \beta_p, \ldots, h_p), \ldots, (\alpha_m, \beta_m, \ldots, h_m)\}. \tag{4}$$

For each person p, the attribute h_p specifies the household she belongs to. This also introduces a new implicit attribute p_h for each household h – the number of constituent members, defined as $p_h := |P_h|$.

In analogy to Section 2.1, we denote by P_α the set of persons that fall within category α:

$$P_\alpha := \{p \in P : (\alpha_p, \ldots) = (\alpha, \ldots)\}. \tag{5}$$

We also define the control totals C_α and the grand total $\nu = \sum_\alpha C_\alpha$.

At first glance, the household attribute h_p could be used as a control variable just like any other person-level attribute, using the household-level expansion factors f_h estimated with a household-level IPF run as control totals: $C_h := f_h \cdot p_h$. Two problems occur with this approach:

- The total number of persons implied by the new household-level control totals does not necessarily match the grand total required by the other controls: $\sum_h C_h \neq \nu$. As mentioned above, IPF does not converge in this case.
- Even in the case of convergence, the expansion factors are not necessarily equal for two persons within the same household. Consequently, this procedure does not yield the required household-level expansion factors straight away.

In what follows, we develop solutions for these problems.

2.3 Adjusting the Persons-per-Household Ratio

For fitting at the household level, specification of the desired number of persons by using the number of persons in the household as control variable introduces unjustified bias. In what follows, we present an alternative fitting step that goes well with IPF by adhering to the Principle of Minimum Discrimination Information (Kullback and Leibler, 1951; Ireland and Kullback, 1968).

For given expansion factors f_h, the objective is to estimate new expansion factors f'_h subject to the following restrictions:

$$\sum_h f'_h = n \tag{6}$$

$$\sum_h p_h \cdot f'_h = \nu. \tag{7}$$

From the infinite set of feasible solutions, we choose the one that minimizes the relative entropy from f_h to f'_h, defined as follows:

$$D(f'_h || f_h) = \sum_h f'_h \ln \frac{f'_h}{f_h}. \tag{8}$$

By this, we introduce the least possible amount of new information into our distribution. Nevertheless, the fit at the household level is potentially distorted.

Appendix A presents a detailed analysis of the underlying optimization problem.

2.4 Switching Between Domains

After adjusting the average number of persons per household, another IPF run can be carried out at the person level. The person-level expansion factors are copied from those at the household level: $f_p := f_{h_p}$. By explicitly controlling the household attribute h_p, the distribution of the household variables remains unchanged. However, as mentioned above, this procedure yields person-level expansion factors unsuitable for our purpose: These need to be converted into household-level factors.

A naïve approach to estimating household-level expansion factors is averaging:

$$f_h := \frac{1}{p_h} \sum_{p \in P_h} f_p. \tag{9}$$

Figure 4: Procedure HIPF($C_a, C_b, C_\alpha, C_\beta, \ldots$) – the Hierarchical IPF Algorithm

```
Require: Reference sample, control totals $C_a$, $C_b$, $C_\alpha$, $C_\beta$ and possibly others
Ensure: Expansion factors $f_h$ obeying all control totals
  $k \leftarrow 0$
  $f_h^{(0)} \leftarrow 1$ for all $h \in H$
  repeat
    $f_h^{(k+1)} \leftarrow \mathrm{FIT}(f_h^{(k)}, C_a, C_b, \ldots)$ for all $h \in H$
    $f_p^{(k+2)} \leftarrow f_{h_p}^{(k+1)}$ for all $p \in P$
    $f_p^{(k+3)} \leftarrow \mathrm{FIT}(f_p^{(k+2)}, C_\alpha, C_\beta, \ldots)$ for all $p \in P$
    $f_h^{(k+4)} \leftarrow p_h^{-1} \cdot \sum_{p \in P_h} f_p^{(k+3)}$ for all $h \in H$
    estimate $f_h^{(k+5)}$ from $f_h^{(k+4)}$ by adjusting the persons-per-household ratio (cf. Section 2.3)
    $k \leftarrow k + 5$
  until convergence
  return $f_h^{(k)}$
```

The bad news is that this completely undoes the efforts of person-level IPF, as the sum of the expansion factors of all household members – and hence also the average – is fixed by using household as control variable. As it turns out, the idea to explicitly control households during person-level fit must be completely abandoned to achieve a simultaneous fit at both levels. Nevertheless, the ingredients presented so far can be combined into an algorithm that solves the initial problem.

2.5 The Big Picture

In order to simultaneously control both levels of aggregation, we suggest fitting at household and person levels alternately. For this, it is necessary to convert household-level expansion factors into person-level ones, and vice versa – these conversions can be carried out as outlined above. In addition, an adjustment of the persons-per-household ratio has to be performed repeatedly. Figure 4 shows a pseudocode for the algorithm; its stop criterion can be specified using the relative change between iterations or the absolute difference to the control totals.

Note that, if fitting only at the household level, the algorithm is equivalent to IPF. While this is also true for the IPU and Ent, our approach follows the Principle of Minimum Discrimination Information more closely when fitting at two levels of hierarchy compared to IPU.

We repeat the toy example used for the presentation of IPU in (Müller and Axhausen, 2011) with our algorithm in Figure 5. In this example, we consider a sample of households and

Figure 5: Numeric Example for the Hierarchical IPF algorithm

h	a_h	p_h	p	α_p	k	0	1	2	3	4	5	10		∞	IPU	Ent
1..22	⊞	3	1	●	$f^{(k)}$	1.00	1.37	1.37	2.19	1.36	1.33	1.28	⋮	1.18	1.60	1.05
			2	○				1.37	0.94							
			3	○				1.37	0.94							
23..43	⊞	2	1	●			1.37	1.37	2.19	1.57	1.61	1.61		1.50	1.60	1.51
			2	○				1.37	0.94							
44..64	⊞	3	1	○			1.37	1.37	0.94	0.94	0.92	0.75		0.54	0.33	0.45
			2	○				1.37	0.94							
			3	○				1.37	0.94							
65..80	⊟	2	1	○			0.64	0.64	0.44	0.44	0.45	0.38		0.28	0.19	0.36
			2	○				0.64	0.44							
81..96	⊟	3	1	●			0.64	0.64	1.03	0.64	0.62	0.66		0.68	0.95	0.58
			2	○				0.64	0.44							
			3	○				0.64	0.44							
97..108	⊟	1	1	○			0.64	0.64	0.44	0.44	0.48	0.38		0.26	0.19	0.52
109..119	⊞	2	1	○			1.37	1.37	0.94	0.94	0.97	0.75		0.49	0.33	0.65
			2	○				1.37	0.94							
120..128	⊞	1	1	○			1.37	1.37	0.94	0.94	1.01	0.75		0.45	0.33	0.94
129..136	⊟	3	1	●			0.64	0.64	1.03	0.83	0.82	1.00		1.30	0.95	1.34
			2	●				0.64	1.03							
			3	○				0.64	0.44							
137..144	⊞	3	1	●			1.37	1.37	2.19	1.77	1.73	1.95		2.24	1.60	2.44
			2	●				1.37	2.19							
			3	○				1.37	0.94							
145..151	⊟	2	1	●			0.64	0.64	1.03	0.74	0.75	0.82		0.87	0.95	0.83
			2	○				0.64	0.44							
152..158	⊟	3	1	○			0.64	0.64	0.44	0.44	0.43	0.38		0.31	0.19	0.25
			2	○				0.64	0.44							
			3	○				0.64	0.44							
159..164	⊞	1	1	●			1.37	1.37	2.19	2.19	2.35	2.76		3.27	3.51	2.17
165..170	⊞	2	1	●			1.37	1.37	2.19	2.19	2.25	2.75		3.58	3.51	3.51
			2	●				1.37	2.19							
171..173	⊟	1	1	●			0.64	0.64	1.03	1.03	1.11	1.41		1.89	2.80	1.20
174..175	⊞	3	1	●			1.37	1.37	2.19	2.19	2.14	2.74		3.92	3.51	5.66
			2	●				1.37	2.19							
			3	●				1.37	2.19							
176	⊟	2	1	●			0.64	0.64	1.03	1.03	1.06	1.40		2.07	2.80	1.93
			2	●				0.64	1.03							
			n	145	Δ_h	39.00	0.00	?	?	-1.25	-2.82	-1.75	⋮	0.00	0.00	0.00
			$C_{\boxplus}$	190	Δ_a	14.00	0.00			2.10	0.00	0.00		0.00	0.00	0.00
			ν	227	Δ_p	100.00	85.18	85.18	0.00	49.46	48.63	28.71		0.00	0.00	0.00
			$C_{\bullet}$	434	Δ_α	28.00	-8.27	-8.27	0.00	0.00	0.00	0.00		0.00	0.00	0.00

their constituent members. The sample contains information whether the household owns a car (⊞/⊟), and whether each member of the household is working (●/○). The control totals postulate a somewhat larger population, a considerable shift towards occupation, and a slight preference for car possession. In the left part of the table, the reference sample is listed, with corresponding control totals below. The right part of the table shows the actual fitting procedure: For each iteration listed, the estimated expansion factors are shown. Below that, the Δ values denote the difference between desired and estimated counts. When fitting at person level, *e.g.*, in iteration 2, computing fitness values at household level is not meaningful – hence the question marks in the corresponding cells. The table is shown in condensed form: Households that are identical within the given attributes are listed only once.

The two rightmost columns list the results obtained with the IPU and the Ent algorithms for comparison. Note that the expansion factors obtained by the three techniques do not seem to exhibit a common pattern. Perhaps the most notable differences between HIPF and IPU can be observed for households 81..96 and 129..136: For both household structures, IPU estimates the same expansion factor of 0.95, while the households 81..96 occur almost twice as seldom in the HIPF solution as 129..136. The reason is that IPU treats both household types the same way when fitting for the person-level attribute: Both have at least one worker and at least one non-worker. In contrast, HIPF assigns a smaller expansion factor to 81..96, as this household type contradicts the increase of overall occupation specified by the person-level control totals. When comparing HIPF and Ent, in general, the expansion factors for one-person households (*e.g.*, 120..128) tend to be further away from 1, and those for three-person households (*e.g.*, 174..175) tend to be closer to 1 for HIPF. This is due to the fact that a change in the expansion factor has the same effect on the entropy regardless of household size, and a very large or very small expansion factor “pays off” for a large household if it helps the adjustment to the person-level control totals.

3 Experimental results

As a practical test for our algorithm, we generated a synthetic population for Switzerland. The Swiss census (Swiss Federal Statistical Office, 2000a) has been used as input for the reference sample. We computed household-level expansion factors for a random sample of the census using control totals derived from the census itself, using HIPF, IPU (Ye *et al.*, 2009), and the entropy optimization approach presented by Bar-Gera *et al.* (2009). The generated population has been validated, also against the census. While this is not a particularly realistic scenario, this setting allows us to test the quality of the algorithms under “perfect” conditions. In the following, we present the setup and results of the validation.

Table 1: Household Structure of the Base Sample

p	$\lvert H_p\rvert$	p	$\lvert H_p\rvert$	p	$\lvert H_p\rvert$	p	$\lvert H_p\rvert$	p	$\lvert H_p\rvert$	p	$\lvert H_p\rvert$
1	1 120 878	2	985 971	3	403 032	4	410 000	5	143 736	6	38 183
7	9 061	8	2 746	9	1 004	10	400	11	175	12	101
13	49	14	41	15	16	16	4	17	2		
$\lvert H\rvert$	3 115 399	$\lvert P\rvert$	6 992 811		$\lvert P\rvert \div \lvert H\rvert$	2.245					

Table 2: Cross Tabulations Available from the Swiss Federal Statistical Office

Level	Tabulation	Relevant	Chosen
H	Municipality × Household size	✓	✓
	Municipality × Household type × Children	✓	✓
P	Canton × Age × Sex × Foreigner	✓	✓
	Municipality × Age × Sex	✓	(✓)
	Municipality × Sex × Foreigner × Martial status	✓	✓
	Municipality × Religion		
	Municipality × Foreigner × Language		
	Municipality × Sex × Foreigner × Occupation type	✓	
	Municipality × Sex × Foreigner × Economic sector	✓	
	Municipality × Sex × Foreigner × Education	✓	✓
	Municipality × Sex × Foreigner × Professional activity	✓	
	Municipality × Sex × Foreigner × Place of birth		
	Municipality × Sex × Foreigner × Former residence	✓	

3.1 Description of the Data

True Population The Swiss census contains person and household attributes in a total of 7 452 000 records. For each person, the household ID and all attributes of that household are listed. Each place of residence of a person is represented by a distinct record; hence, we excluded 164 000 person records that correspond to secondary residences. In addition, we deleted 295 000 records that correspond to persons in collective households or group quarters. The true population, detailed in Table 1, consists of 6 993 000 persons and 3 115 000 households; the average household size is 2.245.

Control Totals Cross tabulations over attributes where aggregate data is available from the Swiss Federal Statistical Office are used as control totals. Table 2 lists the cross tabulations that are related to the census, highlighting those relevant for transportation planning and those finally selected for the synthesis procedure. Using the Municipality attribute for the

Table 3: Description of the Analyzed Attributes

Controlled	Level	Attribute	Scale	# Classes	Recoding
grouped by	H	Canton	Nominal	26 → 23	OW → NW, UR → GL, AI → AR
✓	H	Household size	Ratio	17 → 7	7, 8, 9, ... → 7
		Household type	Nominal	20 → 7	Coarser categories
		Children	Ratio	13 → 4	3, 4, 5, ... → 3
	P	Age	Ratio	→ 16	5-year steps; 75, 80, 85 ... → 75
		Sex	Nominal	2	
		Foreigner	Nominal	2	
		Martial status	Nominal	4 → 3	Widowed → Divorced
		Education	Ordinal	13 → 6	Coarser categories
	H	Municipality	Nominal	2 896	
		Age of head	Ratio		
		Age of oldest child	Ratio		
		Age of youngest child	Ratio		
	P	Workplace location	Ordinal	7	
		Commute mode	Nominal	143	

control totals would lead to many missing observations; to avoid this problem, the Canton attribute has been used instead. Table 3 lists the properties of the control variables and the uncontrolled attributes used in the validation described below.

Reference Sample We used a random sample of 5 % of the households in our true population as reference sample, as this is the sampling rate provided by the Swiss PUS (Swiss Federal Statistical Office, 2000b).

Missing Observations As dealing with missing observations is beyond the focus of this paper, a rather simple method has been chosen to avoid this problem. In the true population, all households that correspond to missing observations in the sample have been removed, and the control totals have been recomputed afterwards.

Conflicting Control Totals In a first run, control totals at the finest level available have been used. However, apart from a significant number of missing observations, control totals conflicted in various ways. To reduce the impact of this problem, rare categories have been merged manually, as suggested by Guo and Bhat (2007) and Auld and Mohammadian (2010). Table 3 also shows the recodings that have been applied to the data.

Grouping As the canton occurs in each control variable and also in the reference sample, synthesizing the population of each canton separately is advisable over synthesizing the whole population of Switzerland at once. In fact, none of the three algorithms exhibited satisfactory convergence behavior in the latter case.

3.2 Validation

All three algorithms (HIPF, IPU, and Ent) have been run with the prepared data, aborting after an absolute difference of less than 10^{-3} persons/households has been reached for each control total. Except for IPU, an optimal fit has been computed for each canton; IPU has converged to misfits of up to 40 persons/households for five of the smaller cantons. Apart from that, convergence rates are comparable.

In the next step, integer weights were computed using probabilistic drawing with replacement, or Monte-Carlo sampling, from each set of fractional household weights estimated by the three algorithms. The probability to draw a household has been set proportional to the household weight estimated by each algorithm. After that, for each algorithm, a set of fractional weights (at convergence) and integer weights (after sampling) is available for the validation; in each case, the total number of households in the sample equals that of the true population. Using these weights, three-dimensional joint distributions for each combination of the attributes listed in Table 3 are computed – separately for each canton, at the person level, using the values of the variables before recoding. The estimated joint distributions are then compared to the corresponding joint distribution of the true population using the information-based G^2 and the distance-based SRMSE metrics, as suggested by Pritchard and Miller (2009) and earlier by Knudsen and Fotheringham (1986). Denoting by N_{abc} the number of persons with attributes a, b, c in the true population, the metrics are defined as follows:

$$G^2 = 2\sum_{abc} N_{abc} \cdot \frac{F_{abc}}{N_{abc}} \qquad \text{SRMSE} = \frac{\sqrt{|A||B||C| \sum_{abc}(F_{abc} - N_{abc})^2}}{\sum_{abc} N_{abc}}. \tag{10}$$

For both statistics, a smaller value indicates a better fit.

The comparison results in twelve statistics (two for each HIPF, IPU, Ent, each before and after sampling) for each of the $\binom{14}{3} = 364$ joint distributions for each canton. Table 4 presents the results of the analysis. Each row shows data for one canton; as described in Section 3.1, some cantons were merged when preparing the data, and the analysis reflects this. The second column denotes whether IPU converged for the canton in question. Parentheses around a check mark mean that convergence has been achieved within a precision above 10^{-3}. The right half of the table presents, for each of the twelve statistics, the number of joint distributions where this statistics was the lowest among the three algorithms tested. For each canton, the algorithms performing best or second-best within the respective scenario are typeset in bold or italics. In most of the cases, HIPF seems to outperform both IPU and Ent in terms of the tested metrics. This is most notable for the statistics of the joint distributions after fitting and before the sampling. Also, note that for the cantons where IPU did not converge, all joint distributions estimated by IPU performed worse compared to the other algorithms w.r.t. the test statistics.

Table 4: Comparison of the goodness-of-fit of HIPF, IPU, and Ent for the synthetic population for Switzerland

Canton	IPU → 0	**Persons**	**Households**	**P/H**	**Number of joint distributions closest to the true population**											
					G^2						**SRMSE**					
					After fitting			**After sampling**			**After fitting**			**After sampling**		
					HIPF	**IPU**	**Ent**	**HIPF**	**IPU**	**Ent**	**HIPF**	**IPU**	**Ent**	**HIPF**	**IPU**	**Ent**
Zurich	✓	1 201 877	567 573	2.12	**201**	36	*127*	**141**	**141**	82	*92*	53	**219**	*127*	86	**151**
Bern	✓	916 135	415 901	2.20	**157**	90	*117*	**143**	*127*	94	**189**	34	*141*	*130*	60	**174**
Lucerne	✓	337 166	140 594	2.40	**217**	23	*124*	**293**	34	*37*	**239**	20	*105*	**201**	70	*93*
Glarus + Uri	✓	69 969	28 661	2.44	**298**	26	*40*	**190**	40	*134*	**253**	54	*57*	**183**	*92*	89
Schwyz	✓	124 140	50 017	2.48	**270**	30	*64*	**209**	76	*79*	**285**	*47*	32	**219**	*76*	69
Nidwalden + Obwalden		67 164	27 598	2.43	**270**	0	*94*	**291**	0	*73*	**291**	0	*73*	**272**	0	*92*
Zug		92 591	40 819	2.27	**335**	0	*29*	**262**	0	*102*	**303**	0	*61*	**255**	0	*109*
Fribourg	✓	230 908	94 093	2.45	**314**	19	*31*	**263**	36	*65*	**278**	15	*71*	**245**	38	*81*
Solothurn	✓	237 832	102 584	2.32	**229**	43	*92*	**197**	50	*117*	**209**	61	*94*	**181**	67	*116*
Basel-Stadt	(✓)	179 128	95 999	1.87	**316**	15	*33*	**287**	22	*55*	**257**	34	*73*	**273**	31	*60*
Basel-Landschaft	✓	252 761	111 675	2.26	**283**	9	*72*	**249**	19	*96*	**260**	24	*80*	**154**	65	*145*
Schaffhausen	✓	70 330	31 427	2.24	**248**	13	*103*	**204**	*121*	39	**253**	47	*64*	**162**	*104*	98
Both Appenzell	✓	65 265	26 862	2.43	**333**	1	*30*	**294**	3	*67*	**293**	14	*57*	**198**	9	*157*
St. Gallen	✓	437 732	183 750	2.38	**244**	*63*	57	**186**	*122*	56	**188**	77	*99*	*136*	**137**	91
Graubünden		174 137	77 781	2.24	**304**	0	*60*	**269**	0	*95*	**254**	0	*110*	**245**	0	*119*
Aargau	✓	532 903	224 128	2.38	**293**	35	*36*	**187**	49	*128*	**168**	*99*	97	*142*	**143**	79
Thurgau	(✓)	221 526	91 537	2.42	**183**	8	*173*	*144*	6	**214**	**197**	4	*163*	**201**	2	*161*
Ticino	✓	298 664	134 916	2.21	**205**	*118*	41	*131*	61	**172**	**188**	*93*	83	92	**144**	*128*
Vaud	✓	612 626	278 752	2.20	**278**	15	*71*	**234**	12	*118*	*118*	55	**191**	**168**	39	*157*
Valais	(✓)	258 435	107 378	2.41	**278**	0	*86*	**302**	1	*61*	**311**	0	*53*	**195**	0	*169*
Neuchâtel		161 223	74 049	2.18	**290**	0	*74*	*115*	0	**249**	**219**	0	*145*	*159*	0	**205**
Geneva	✓	383 841	181 611	2.11	**297**	26	*41*	**184**	*120*	60	**196**	50	*118*	*126*	**132**	106
Jura		65 958	27 471	2.40	**307**	0	*57*	**277**	0	*87*	**282**	0	*82*	**242**	0	*122*

3.3 Discussion

Only HIPF considers expansion factors at the person level; both IPU and Ent operate on household-level expansion factors only. The joint distributions computed for the analysis are person-based; it would be worthwhile to repeat the analysis with household-based joint distributions by additionally weighting each person record with the inverse of the household size.

The dissatisfactory results of IPU for the cantons where a perfect fit could not be estimated indicate the importance of a perfect fit in the fitting stage. Given that a fit exists (*cf.* Section 2.1), the fitting algorithm should be able to compute it.

While the probabilistic sampling procedure controls the total number of households, the person count may diverge after sampling. Various techniques are possible to control both household and person count during the sampling procedure: (a) repeated sampling, (b) probabilistic substitution of households, or (c) further combinatorial optimization. The same applies to retaining the controlled distributions. Given the availability of a true population, it will be possible to derive sampling strategies to generate accurate synthetic populations.

4 Conclusion

This paper presents HIPF, a novel algorithm for estimating household-level expansion factors for a sample consisting of persons grouped to households and control totals at both levels. The algorithm constantly switches between the household and the person domain, employing an entropy-optimizing adjustment step. Adaption to other kinds of populations is possible. Run time and convergence are comparable to other existing algorithms (IPU and Ent) when applied to the generation of a synthetic population for Switzerland from a 5 % sample.

A detailed analysis of the goodness-of-fit of the synthetic to the true population using the G^2 and SRMSE metrics shows that HIPF often outperforms IPU and Ent. This holds for the fractional expansion factors computed by the algorithms as well as for the final population derived from the expansion factors using probabilistic sampling. Future challenges are validation against other metrics, development of an optimal algorithm for the generation stage, and improvement of the algorithm's convergence speed.

5 References

Auld, J. and A. K. Mohammadian (2010) Efficient methodology for generating synthetic populations with multiple control levels, *Transportation Research Record*, **2183**, 19–28.

Bar-Gera, H., K. Konduri, B. Sana, X. Ye and R. M. Pendyala (2009) Estimating survey weights with multiple constraints using entropy optimization methods, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.

Beckx, C., T. A. Arentze, L. Int Panis, D. Janssens, J. Vankerkorn and G. Wets (2009) An integrated activity-based modelling framework to assess vehicle emissions: Approach and application, *Environment and Planning B*, **36** (6) 1086–1102.

Ben-Akiva, M. E., M. Bierlaire, H. Koutsopoulos and R. Mishalani (2002) Real time simulation of traffic demand-supply interactions with DynaMIT, in M. Gendreau and P. Marcotte (eds.) *Transportation and Network Analysis: Current Trends*, 19–36, Kluwer, Dordrecht.

Bhat, C. R., J. Y. Guo, S. Srinivasan and A. Sivakumar (2004) A comprehensive econometric microsimulator for daily activity-travel patterns, *Transportation Research Record*, **1894**, 57–66.

Bowman, J. L. and M. E. Ben-Akiva (2001) Activity-based disaggregate travel demand model system with activity schedules, *Transportation Research Part A: Policy and Practice*, **35** (1) 1–28.

Bradley, M. A., J. L. Bowman and B. Griesenbeck (2010) SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution, *Journal of Choice Modelling*, **3** (1) 5–31.

de Palma, A. and F. Marchal (2002) Real cases applications of the fully dynamic METROPOLIS tool-box: An advocacy for large-scale mesoscopic transportation systems, *Networks and Spatial Economics*, **2** (4) 347–369.

Guo, J. Y. and C. R. Bhat (2007) Population synthesis for microsimulating travel behavior, *Transportation Research Record*, **2014** (12) 92–101.

Ireland, C. T. and S. Kullback (1968) Contingency tables with given marginals, *Biometrika*, **55**, 179–188.

Jones, P. M., M. C. Dix, M. I. Clarke and I. G. Heggie (1983) *Understanding Travel Behaviour*, Gower, Aldershot.

Knudsen, D. C. and A. S. Fotheringham (1986) Matrix comparison, goodness-of-fit, and spatial interaction modelling, *International Regional Science Review*, **10** (2) 127–147.

Kullback, S. and R. A. Leibler (1951) On information and sufficiency, *Annals of Mathematical Statistics*, **22** (1) 79–86.

Mahmassani, H. S., T. Hu and R. Jayakrishnan (1995) Dynamic traffic assignment and simulation for advanced network informatics (DYNASMART), in N. H. Gartner and G. Improta (eds.) *Urban traffic networks: dynamic flow modeling and control*, Springer, Berlin.

MATSim-T (2011) Multi Agent Transportation Simulation Toolkit, webpage, http://www.matsim.org. Accessed on 03/04/2011.

Müller, K. and K. W. Axhausen (2011) Population synthesis for microsimulation: State of the art, paper presented at the *90th Annual Meeting of the Transportation Research Board*, Washington, D.C.

Pritchard, D. R. and E. J. Miller (2009) Advances in agent population synthesis and application in an integrated land use and transportation model, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.

Pukelsheim, F. and B. Simeone (2009) On the iterative proportional fitting procedure: Structure of accumulation points and L_1-error analysis, http://opus.bibliothek.uni-augsburg.de/volltexte/2009/1368, accessed on 29/07/2010.

Roorda, M. J., E. J. Miller and K. M. N. Habib (2008) Validation of TASHA: A 24-h activity scheduling microsimulation model, *Transportation Research Part A: Policy and Practice*, **42** (2) 360–375.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Ryan, J., H. Maoh and P. S. Kanaroglou (2009) Population synthesis: Comparing the major techniques using a small, complete population of firms, *Geographical Analysis*, **41** (2) 181–203.

Srinivasan, S. and L. Ma (2009) Synthetic population generation: A heuristic data-fitting approach and validations, paper presented at the *12th International Conference on Travel Behaviour Research (IATBR)*, Jaipur, December 2009.

Swiss Federal Statistical Office (2000a) Eidgenössische Volkszählung 2000, http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/blank/blank/vz/uebersicht.html.

Swiss Federal Statistical Office (2000b) Public use samples (PUS): Excerpts for general use from the Swiss federal population censuses 1970-2000, http://www.portal-stat.

admin.ch/pus/files/index_e.html.

UrbanSim (2011) Open Platform for Urban Simulation, webpage, http://www.urbansim.org.

Ye, X., K. Konduri, R. M. Pendyala, B. Sana and P. A. Waddell (2009) A methodology to match distributions of both household and person attributes in the generation of synthetic populations, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.

A Analysis of the Persons-per-Household Ratio Adjustment

In this section, we analyze the optimization problem that arises when adjusting the persons-per-household ratio respecting the Principle of Minimum Discrimination Information. The problem is defined as follows: Minimize

$$\sum_h f'_h \ln \frac{f'_h}{f_h} \tag{11}$$

subject to the constraints

$$\sum_h f'_h = n \tag{12}$$

$$\sum_h p_h \cdot f'_h = \nu. \tag{13}$$

Applying the method of Lagrange multipliers yields the following Lagrange function:

$$\Lambda = \sum_h f'_h \ln \frac{f'_h}{f_h} + \lambda_1 \left(\sum_h f'_h - n \right) + \lambda_2 \left(\sum_h p_h \cdot f'_h - \nu \right). \tag{14}$$

The necessary condition $\nabla \Lambda = 0$ for the optimum of the original problem leads to the following precondition, valid for all households h:

$$\begin{aligned} \frac{\partial}{\partial f'_h} \Lambda &= \ln \frac{f'_h}{f_h} + f'_h \frac{1}{f'_h} + \lambda_1 + \lambda_2 \cdot p_h \\ &= \ln \frac{f'_h}{f_h} + (1 + \lambda_1 + \lambda_2 \cdot p_h) \qquad\qquad = 0 \\ \frac{f'_h}{f_h} &= \mathrm{e}^{-1-\lambda_1} \cdot \left(\mathrm{e}^{-\lambda_2}\right)^{p_h} \\ &= c \cdot d^{p_h} \text{ (with } c := \mathrm{e}^{-1-\lambda_1} \text{ and } d := \mathrm{e}^{-\lambda_2}\text{).} \end{aligned} \tag{15}$$

This means that the ratio of new vs. old expansion factors is determined by the household size only, and that it follows a geometric progression with respect to the household size. Thus, we can rewrite the constraints:

$$(12) \iff \sum_p \sum_{h\colon p_h = p} f'_h = n \tag{16}$$

$$\iff \sum_p \sum_{h\colon p_h = p} f_h \cdot c \cdot d^{p_h} = n \tag{17}$$

$$\iff \sum_p \left(c \cdot d^p \cdot \sum_{h:\ p_h = p} f_h \right) = n \tag{18}$$

$$\iff c \cdot \sum_p F_p \cdot d^p = n \tag{19}$$

$$(13) \iff c \cdot \sum_p p \cdot F_p \cdot d^p = \nu. \tag{20}$$

Cancelling c yields a necessary and sufficient condition for d:

$$(19) \wedge (20) \iff \frac{1}{n} \cdot \sum_p F_p \cdot d^p = \frac{1}{\nu} \cdot \sum_p p \cdot F_p \cdot d^p \tag{21}$$

$$\iff \sum_p F_p \cdot d^p = \sum_p \big(n \cdot \nu^{-1} \cdot p\big) \cdot (F_p \cdot d^p) \tag{22}$$

$$\iff 0 = \sum_p \big(n \cdot \nu^{-1} \cdot p - 1\big) \cdot (F_p \cdot d^p) \tag{23}$$

$$\iff 0 = \sum_p \big(\big(n \cdot \nu^{-1} \cdot p - 1\big) \cdot F_p\big) \cdot d^p. \tag{24}$$

Except for ill-formed cases ($n > \nu$ or $n \cdot p \leq \nu$ for all p), for $d > 0$, this polynomial and its first $\lfloor \nu \cdot n^{-1} \rfloor$ derivatives have exactly one real-valued root, while the $\lfloor \nu \cdot n^{-1} \rfloor + 1$-th derivative is strictly greater than zero. Thus, if a solution exists, it is unique and can be found by solving the above polynomial for d. This in turn means that the f'_h derived from this solution constitute the only critical point for the Lagrange function (14). Due to the nature of the objective function this can only be a global minimum.