Multi-level fitting algorithms for population synthesis

Kirill Müller, Kay W. Axhausen
IVT
ETH Zurich, Zurich
Email: kirill.mueller@ivt.baug.ethz.ch

1 Introduction

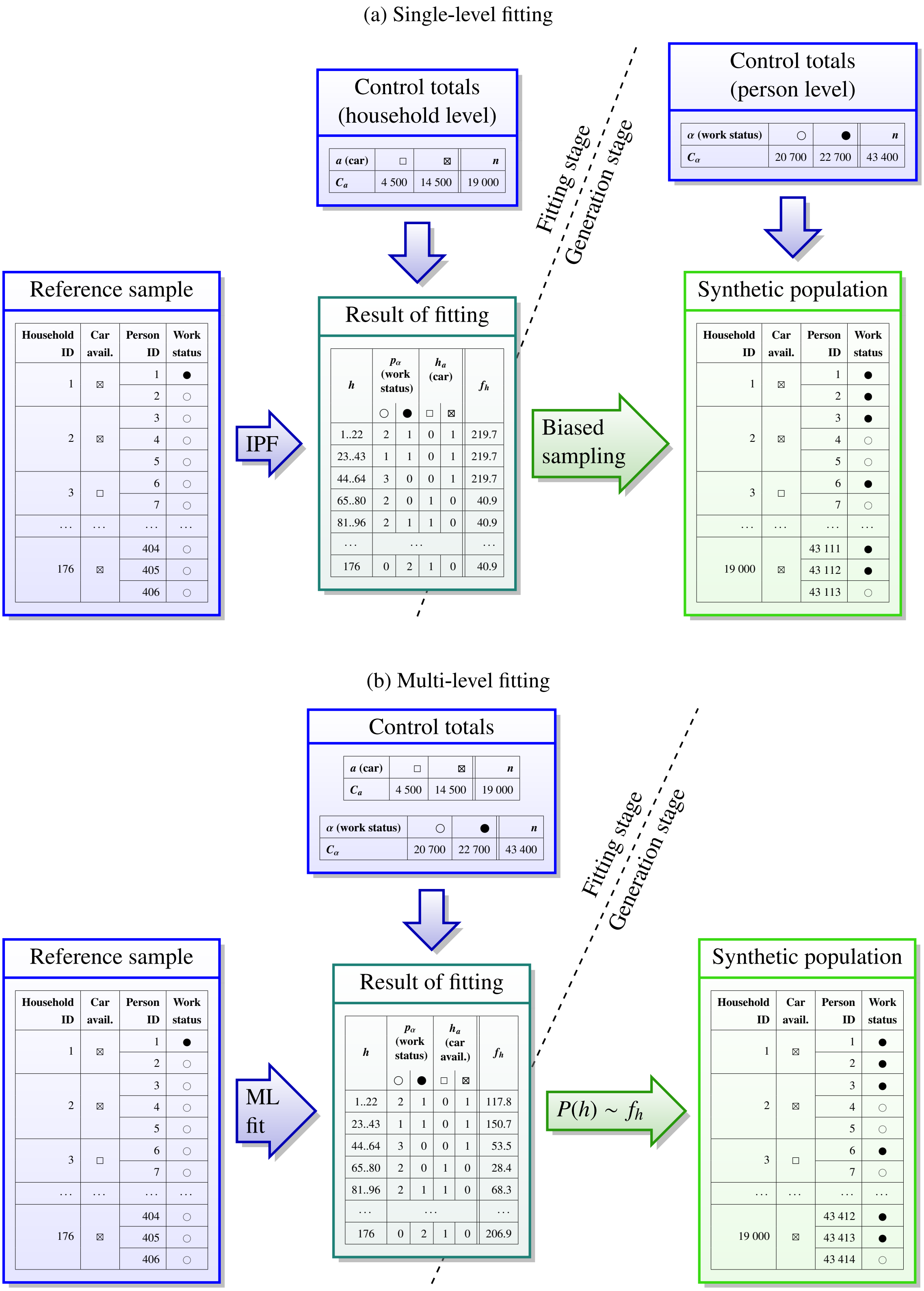
Agent-based microsimulation model systems for land use and transportation planning have come into widespread use. They simulate decisions of agents within an urban area, allowing for more detailed and accurate simulation and prediction of land pricing and travel demand than traditional aggregate models. Often, the *agents* represent the individual people living in the study area, grouped into households. This paper focuses on such person/household populations.

When implementing such a model system, the initial step is the definition of agents and their relationships; this process is called *population synthesis*. The main idea is to combine census microdata (the *reference sample*) with aggregate data at various levels in order 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 aggregate data. The *Synthetic Reconstruction* (SR) method [1] generates the synthetic population by drawing from a reweighted reference sample; see [2] for a literature review over SR techniques.

In reality, other household members may affect personal decisions [3]. Thus, properly replicating the household structure is necessary to be able to simulate these interactions. In this paper we assume that a reference sample of households that contains detailed data for all persons is provided. For the SR method, two options are available: (a) the weights obey only household-level constraints, person-level constraints are considered when selecting households (*single-level fitting*, cf. Fig. 1(a)), or (b) the weights obey constraints at both person and household levels, household selection is unconstrained (*multi-level fitting*, cf. Fig. 1(b)). The multi-level strategy greatly simplifies the construction of the final synthetic population using more complicated reweighting algorithms that have become available only recently [4, 5, 6, 7], and hence can be considered superior to the single-level strategy.

The main contribution of this paper is twofold. First, we propose an algorithmic framework for three multi-level fitting algorithms in which the implementation of each algorithm consists only of subtle changes. This suggests that the three algorithms are inherently similar, and that the choice of the algorithm should be driven by practical factors. Second, we demonstrate formal equivalence of one of the multi-level fitting algorithms to a special case of *generalized raking*, a procedure that

Figure 1: Illustration of single- and multi-level fitting algorithms



(a) Single-level fitting

a (car)	□	☒	n
C_a	4 500	14 500	19 000

α (work status)	○	●	n
C_α	20 700	22 700	43 400

Household ID	Car avail.	Person ID	Work status
1	☒	1	●
		2	○
2	☒	3	○
		4	○
		5	○
3	□	6	○
		7	○
…	…	…	…
176	☒	404	○
		405	○
		406	○

h	p_α (work status)		h_a (car)		f_h
	○	●	□	☒	
1..22	2	1	0	1	219.7
23..43	1	1	0	1	219.7
44..64	3	0	0	1	219.7
65..80	2	0	1	0	40.9
81..96	2	1	1	0	40.9
…	…				…
176	0	2	1	0	40.9

Household ID	Car avail.	Person ID	Work status
1	☒	1	●
		2	●
2	☒	3	●
		4	○
		5	○
3	□	6	●
		7	○
…	…	…	…
19 000	☒	43 111	●
		43 112	●
		43 113	○

(b) Multi-level fitting

a (car)	□	☒	n
C_a	4 500	14 500	19 000

α (work status)	○	●	n
C_α	20 700	22 700	43 400

Household ID	Car avail.	Person ID	Work status
1	☒	1	●
		2	○
2	☒	3	○
		4	○
		5	○
3	□	6	○
		7	○
…	…	…	…
176	☒	404	○
		405	○
		406	○

h	p_α (work status)		h_a (car avail.)		f_h
	○	●	□	☒	
1..22	2	1	0	1	117.8
23..43	1	1	0	1	150.7
44..64	3	0	0	1	53.5
65..80	2	0	1	0	28.4
81..96	2	1	1	0	68.3
…	…				…
176	0	2	1	0	206.9

Household ID	Car avail.	Person ID	Work status
1	☒	1	●
		2	●
2	☒	3	●
		4	○
		5	○
3	□	6	●
		7	○
…	…	…	…
19 000	☒	43 412	●
		43 413	●
		43 414	○

has been long known and used in the field of survey statistics but largely ignored by transportation planners. This allows for the first time to benefit from an enormous amount of theoretical and practical results from a field that focuses primarily on analyzing data from different sources.

2 Multi-level fitting

All three algorithms for multi-level fitting operate on (a) a representative reference sample that contains the characteristics of sampled households and all constituent persons, and (b) control totals for selected attributes on both household and person levels. The objective is to estimate a positive weight (or *expansion factor*) for each household so that all control totals are satisfied.

Figure 2 shows a framework of routines common to all three algorithms. The procedure **ML-FIT** in Figure 2(a) is the main routine. The household weights are initialized to unity. Then, the control totals are processed iteratively until convergence. For each category of each control total, household weights are adjusted to match (cf. procedures **H-FIT** and **P-FIT** in Figs. 2(b) and 2(c)). In turn, these procedures call **H-ADJUST** and **P-ADJUST** for adjustment of each controlled attribute (cf. Figs. 2(d) and 2(e)). The algorithms differ only in how the adjustment is carried out for control totals at person level; hence, only a skeleton is provided here for procedure **P-ADJUST**.

Ultimately, it is sufficient to implement the procedure **P-ADJUST**, and in one case to modify the procedure **P-FIT**, to obtain implementations of the three different multi-level fitting methods. The large amount of shared code suggests an inherent similarity between these algorithms.

3 Calibration

Most literature on generating synthetic population in the field of transport planning refers to the seminal paper by Beckman *et al.* [1]. In turn, this paper refers to works driven by the need to adjust (or *calibrate*) survey data to known marginal totals [8, 9, 10], a frequent task in the field of survey statistics. Calibration is used to correct nonresponse and selection bias before performing statistical analyses on response variables.

In 1992, Deville and Särndal [11] have proposed a common framework for weighting systems of which both *generalized regression (GREG) estimation* and IPF are special cases. The subsequent paper by Deville *et al.* [12] elaborates on this idea by focusing more on the reproduction of known marginal counts (*generalized raking*), and presents CALMAR, a software implementation for the statistical system SAS. A recent implementation for the R statistical software package [13], the `survey` package [14], has been provided by Lumley [15]. In the case of household surveys, generalized raking supports simultaneous calibration against household-level and person-level control totals.

The advances in the field of survey statistics do not seem to be widely applied within the transportation planning community, and in particular not within the population synthesis community.

Figure 2: A generic framework for multi-level fitting algorithms

(a) Procedure **ML-FIT**$(H, P, \mathbf{C}_{\mathbf{x}_1}, \ldots, \mathbf{C}_{\boldsymbol{\xi}_1}, \ldots)$ – the main routine of the generic framework

```
**Require:** Reference sample $(H, P)$
**Require:** Household-level control totals $\mathbf{C}_{\mathbf{x}_1}, \mathbf{C}_{\mathbf{x}_2}, \ldots$
**Require:** Person-level control totals $\mathbf{C}_{\boldsymbol{\xi}_1}, \mathbf{C}_{\boldsymbol{\xi}_2}, \ldots$
**Ensure:** Expansion factors $f_h$ obeying all control totals
  $f_h \leftarrow 1$ **for all** $h \in H$
  **while** convergence not reached **do**
      $f_h \leftarrow$ **H-FIT**$(H, f_h, \mathbf{C}_{\mathbf{x}_1}, \ldots)$
      $f_h \leftarrow$ **P-FIT**$(P, f_h, \mathbf{C}_{\boldsymbol{\xi}_1}, \ldots)$
  **return** $f_h$
```

(b) Procedure **H-FIT**$(H, f_h, \mathbf{C}_{\mathbf{x}_1}, \ldots)$ – subroutine for household-level fitting

```
**Require:** Households $H$ from reference sample
**Require:** Current expansion factors $f_h$
**Require:** Household-level control totals $\mathbf{C}_{\mathbf{x}_1}, \mathbf{C}_{\mathbf{x}_2}, \ldots$
**Ensure:** Improved expansion factors $f_h$
  **for all** household-level controls $\mathbf{C}_{\mathbf{x}}$ **do**
      **for all** attributes $x \in \mathbf{x}$ **do**
          $f_h \leftarrow$ **H-ADJUST**$(H, f_h, x, \mathbf{C}_{\mathbf{x}}(x))$
  **return** $f_h$
```

(c) Procedure **P-FIT**$(P, f_h, \mathbf{C}_{\boldsymbol{\xi}_1}, \ldots)$ – subroutine for person-level fitting

```
**Require:** Persons $P$ from reference sample
**Require:** Current expansion factors $f_h$
**Require:** Person-level control totals $\mathbf{C}_{\boldsymbol{\xi}_1}, \mathbf{C}_{\boldsymbol{\xi}_2}, \ldots$
**Ensure:** Improved expansion factors $f_h$
  **for all** person-level controls $\mathbf{C}_{\boldsymbol{\xi}}$ **do**
      **for all** attributes $\xi \in \boldsymbol{\xi}$ **do**
          $f_h \leftarrow$ **P-ADJUST**$(P, f_h, \xi, \mathbf{C}_{\boldsymbol{\xi}}(\xi))$
  **return** $f_h$
```

(d) Procedure **H-ADJUST**(H, f_h, x, C_x) – subroutine for fitting one category at household level

```
**Require:** Current household-level category $x$
**Require:** Value $C_x$ of the control total for category $x$
**Ensure:** Expansion factors $f_h$ that match control $C_x$
  $r \leftarrow C_x \div \sum_{h \in H_x} f_h$
  $f_h \leftarrow f_h \cdot r$ **for all** $h \in H_x$
  **return** $f_h$
```

(e) Procedure **P-ADJUST**(P, f_h, ξ, C_ξ) – subroutine for fitting one category at person level

```
**Require:** Current person-level category $\xi$
**Require:** Value $C_\xi$ of the control total for category $\xi$
**Ensure:** Expansion factors $f_h$ that match control $C_\xi$
  (implementation differs for each algorithm)
```

In fact, the *raking ratio* method presented in [12] is mathematically equivalent to the multi-level fitting algorithm introduced independently in [5, 6]. Also, generalized raking operates directly on the unrolled survey data instead of creating a crosstabulation – this has been suggested independently in [16]. Reasons for this might include the vast differences in terminology, notation, and perhaps application. This paper bridges the gap by formally demonstrating equivalence of the methods used in both fields, thus justifying the usage of theoretical results, algorithms and software implementations from survey statistics for the problem at hand.

4 Acknowledgements

The authors gratefully acknowledge the financial support of the Swiss National Science Foundation for the project “A generalized approach to population synthesis” (grant 138270).

References

[1] Beckman, R. J., K. A. Baggerly and M. D. McKay (1996) Creating synthetic baseline populations, *Transportation Research Part A: Policy and Practice*, **30** (6) 415–429.

[2] 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., January 2011.

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

[4] 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.

[5] 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.

[6] Lee, D.-H. and Y. Fu (2011) Cross-entropy optimization model for population synthesis in activity-based microsimulation models, *Transportation Research Record*, **2255**, 20–27.

[7] Müller, K. and K. W. Axhausen (2011) Hierarchical IPF: Generating a synthetic population for Switzerland, paper presented at the *51st Congress of the European Regional Science Association*, Barcelona, September 2011.

[8] Deming, W. E. and F. F. Stephan (1940) On the least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics*, **11** (4) 427–444.

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

[10] Little, R. J. A. and M.-M. Wu (1991) Models for contingency tables with known margins when target and sampled populations differ, *Journal of the American Statistical Association*, **86** (413) 87–95.

[11] Deville, J.-C. and C.-E. Särndal (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87** (418) 376–382.

[12] Deville, J.-C., C.-E. Särndal and O. Sautory (1993) Generalized raking procedures in survey sampling, *Journal of the American Statistical Association*, **88** (423) 1013–1020.

[13] R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, `http://www.r-project.org`.

[14] Lumley, T. (2012) Survey analysis in R, webpage, `http://faculty.washington.edu/tlumley/survey`. Accessed on 29/05/2012.

[15] Lumley, T. (2010) *Complex surveys: A guide to analysis using R*, vol. 565, John Wiley & Sons, Hoboken.

[16] Pritchard, D. R. and E. J. Miller (2012) Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously, *Transportation*, **39** (3) 685–704, May 2012.