Estimating Cross-Classified Population Counts of Multidimensional Tables

Suesse, Thomas; Namazi-Rad, Mohammad Reza; Mokhtarian, Payam; Barthélemy, Johan

Published in:
Journal of Official Statistics

DOI:
10.1515/jos-2017-0048

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 1021–1050, http://dx.doi.org/10.1515/JOS-2017-0048

Estimating Cross-Classified Population Counts of Multidimensional Tables: An Application to Regional Australia to Obtain Pseudo-Census Counts

Thomas Suesse[1], *Mohammad-Reza Namazi-Rad*[1], *Payam Mokhtarian*[3], *and Johan Barthélemy*[2]

Estimating population counts for multidimensional tables based on a representative sample subject to known marginal population counts is not only important in survey sampling but is also an integral part of standard methods for simulating area-specific synthetic populations. In this article several estimation methods are reviewed, with particular focus on the iterative proportional fitting procedure and the maximum likelihood method. The performance of these methods is investigated in a simulation study for multidimensional tables, as previous studies are limited to 2 by 2 tables. The data are generated under random sampling but also under misspecification models, for which sample and target populations differ systematically. The empirical results show that simple adjustments can lead to more efficient estimators, but generally, at the expense of increased bias. The adjustments also generally improve coverage of the confidence intervals. The methods discussed in this article along with standard error estimators, are made freely available in the R package `mipfp`. As an illustration, the methods are applied to the 2011 Australian census data available for the Illawarra Region in order to obtain estimates for the desired three-way table for age by sex by family type with known marginal tables for age by sex and for family type.

Key words: Census data; IPFP; Log-linear model; model-based inference; count estimation; synthetic population.

1. Introduction

In many countries, census data are still the major source for geographically detailed estimates of populations and economies. Statistical agencies often provide public-use microdata files based on their census or surveys. To preserve confidentiality, some variables might be suppressed, or alternatively only marginal totals, also known as aggregated data, are released instead of the joint totals. For example, joint tables on age by sex and by income might not be released for small areas, as this could lead to disclosing the income of some people with specific age by sex. Instead, only separate tables of marginal totals, for example for age by sex and for income, are released.

[1] National Institute for Applied Statistics Research Australia, School of Mathametics and Applied Statistics, University of Wollongong, NSW 2522, Australia. Email: tsuesse@uow.edu.au
[2] SMART Infrastructure Facility, University of Wollongong, NSW 2522, Australia
[3] Damian Group, Fairfax Media, Sydney 2009 NSW, Australia

Acknowledgments: The authors wish to gratefully acknowledge the help of Dr Madeleine Strong Cincotta in the final language editing of this article. We are grateful to the associate editor and the anonymous referees for their helpful comments that greatly improved the article.

The generation of an artificial (or synthetic) population that realistically matches the population of interest from such limited tables, or more generally aggregated data, has become an important research area (Arentze et al. 2007; Gargiulo et al. 2010; Harland et al. 2012; Barthélemy and Toint 2013; Lenormand and Deffuant 2013; Geard et al. 2013; Huynh et al. 2016).

The conventional generation process of such a synthetic population (SP) is a two step procedure that integrates data from a fully disaggregated sample (for example derived from a survey) with aggregated data from a census (Beckman et al. 1996). The first step estimates the contingency table of all the attributes for the area of interest. The second step then randomly draws synthetic individuals from the sample in proportions that match the estimated contingency table. Using this SP generation approach, the risk of identification of population units and/or their sensitive information in the generated synthetic data is greatly reduced (Rubin 1987).

This article focuses on the first step, that is, the estimation of population counts in multidimensional contingency tables when a random sample is available together with known marginal population tables of lower dimension. It is also important to investigate the multidimensional case with several variables in which each variable has possibly more than two categories, as existing simulation studies, for example by Little and Wu (1991), only considered the unrealistic scenario of a 2 by 2 table.

The iterative proportional fitting procedure (IPFP) originally proposed by Deming and Stephan (1940) and the maximum likelihood (ML) method (Smith 1947) are the traditional methods for estimating cross-classified population counts. IPFP is a general purpose method to match marginal information and is not limited to surveys. The method has also been applied in small area estimation (SAE) to a slightly different situation when the complete table is replaced by some other source of information, such as a complete table from a previous census together with marginal tables which are not necessarily known but are based on some survey estimates. In this context, the method is known as structure preserving estimation (SPREE) (Purcell and Kish 1980; Zhang and Chambers 2004), as it preserves part of the structure of the implied log-linear models in both tables. IPFP has the same structure preserving property and SPREE can be thought of as a special case of IPFP. For example, Purcell and Kish (1980) have considered six different data situations and only referred to one as IPFP; however, all six situations were indeed solved with IPFP.

Section 2 introduces the main estimation methods IPFP and ML, as well as two other estimation methods. Data adjustments are also introduced, which are applied before the estimators are calculated in order to improve statistical properties. In Section 3, misspecification models are considered, that is models for which sample and population information differ systematically, including ML estimators for each of these misspecification models. In Section 4, a simulation-based empirical study is presented to investigate the performance of the methods discussed in this article under simple random sampling and under the misspecification models. The methods are then employed for estimating cross-classified population counts and probabilities for the Illawarra region using available one- and two-dimensional 2011 Australian census tables. This article concludes with a discussion of the results.

2. Estimating Cross-Tabulated Population Counts

The main methods for estimating cross-tabulated population counts and probabilities subject to known marginal population tables of lower dimensions are discussed in this section.

2.1. Iterative Proportional Fitting Procedure (IPFP)

IPFP was originally proposed by Deming and Stephan (1940) as an algorithm attempting to minimize the Pearson chi-squared statistic. For the purpose of population reconstruction, IPFP is often used as an algorithm attempting to adjust census tables so that table cells add up to totals in all required dimensions (Fienberg 1970; Gargiulo et al. 2010; Farooq et al. 2013; Barthélemy and Toint 2013). This application of iterative proportional fitting (IPF) to contingency tables with known margins is called raking (Stephan 1942). Raking (also known as raking ratio estimation) is a procedure which applies a proportional adjustment to the sample weights in a survey so that the adjusted weights add up the known population total when only the marginal population totals are known (Deville et al. 1991; Lu and Gelman 2003). Although raking is not a maximum likelihood (ML) method under random sampling, the raking estimates are consistent and best asymptotically normal (Arentze et al. 2007).

Ireland and Kullback (1968) showed that the estimator produced by the IPFP method minimizes the discrimination information criterion (also known as the Kullback-Leibler divergence, or relative entropy). Mosteller (1968) pointed out that IPFP also preserves the interaction structure of the initial table as defined by the conditional odd ratios.

For illustration purposes, we restrict ourselves to three-way tables, but the methods can be applied in a straightforward manner to tables with more variables. For a three-way contingency table referring to three categorical variables X_1, X_2, and X_3 each with A, B, and C levels, respectively, the population counts are denoted by N_{abc} with population size $N = \sum_{a=1}^{A}\sum_{b=1}^{B}\sum_{c=1}^{C} N_{abc} = N_{\bullet\bullet\bullet}$, where the dot (i.e., $\bullet$) refers to summation over the corresponding variable. The one-way marginal cell counts $N_{a\bullet\bullet}, N_{\bullet b\bullet}$, and $N_{\bullet\bullet c}$ are defined accordingly, for example $N_{a\bullet\bullet} = \sum_{b=1}^{B}\sum_{c=1}^{C} N_{abc}$. The two-way marginal totals are denoted by $N_{ab\bullet}$, $N_{a\bullet c}$, and $N_{\bullet bc}$ and defined by summing the N_{abc} over the respective index.

The main objective is to estimate the cell probabilities $\pi_{abc} = P(X_1 = a, X_2 = b, X_3 = c)$, or equivalently N_{abc}. All joint probabilities π_{abc} and marginal probabilities, such as $\pi_{ab\bullet}$ and $\pi_{a\bullet\bullet}$, need to sum up to one, as marginal probabilities also characterize a valid discrete distribution. When dealing with sample data, sample counts are denoted by y_{abc} with $n = y_{\bullet\bullet\bullet}$ denoting the total sample size.

In the classical IPFP presented by Deming and Stephan (1940), the initial value for the cell probabilities are set as $\pi_{abc}^{(0)} = (ABC)^{-1}$, which corresponds to the case of having no sample data available. When using IPFP for population synthesis, the initial cell probabilities are based on representative survey data with counts y_{abc} often referred to as the *seed data*, that is $\pi_{abc}^{(0)} = y_{abc}/n$. Let us assume for illustration purposes that the three two-way marginal population counts $N_{ab\bullet}, N_{a\bullet c}$, and $N_{\bullet bc}$ are available. We aim at finding

π_{abc} so that the following population constraints hold

$$\pi_{ab\bullet} = \frac{N_{ab\bullet}}{N}, \pi_{a\bullet c} = \frac{N_{a\bullet c}}{N} \text{ and } \pi_{\bullet bc} = \frac{N_{\bullet bc}}{N}. \quad (1)$$

Then one iteration of the IPFP consisting of a three-step cycle has the form

$$\pi_{abc}^{(k+1)} = \frac{\pi_{abc}^{(k)}}{\sum_{a=1}^{A} \pi_{abc}^{(k)}} \times \pi_{\bullet bc}, \quad \pi_{abc}^{(k+2)} = \frac{\pi_{abc}^{(k+1)}}{\sum_{b=1}^{B} \pi_{abc}^{(k+1)}} \times \pi_{a\bullet c},$$

$$\pi_{abc}^{(k+3)} = \frac{\pi_{abc}^{(k+2)}}{\sum_{c=1}^{C} \pi_{abc}^{(k+2)}} \times \pi_{ab\bullet}.$$

The algorithm is continued by setting $k := k + 3$ until convergence to the desired accuracy is attained. Importantly, the obtained estimates $\hat{\pi}_{abc} = \pi_{abc}^{(k)}$ will satisfy (1). The algorithm will converge to a unique solution provided the seed data contain strictly positive entries and provided the marginal constraints do not contradict each other. For example, the constraints $N_{ab\bullet}$ and $N_{a\bullet c}$ need to result in the same $N_{a\bullet\bullet}$, that is $N_{a\bullet\bullet} = \sum_b N_{ab\bullet} = \sum_c N_{a\bullet c}$.

Setting positive starting values $\left(\pi_{abc}^{(0)} > 0\right)$ ensures that each cell has a non-zero probability estimate, that is $\hat{\pi}_{abc} > 0$ (Gange 1995). If some zero cell counts are observed, that is $y_{abc} = 0$, then adjustments can be made. For example adding the value of 0.5 to all cells, the standard procedure for 2 by 2 tables (Agresti 2002, 71). An alternative proposed by Lang (2004) is to add a tiny constant (e.g., 10^{-6}) to all the zero cells to ensure that the estimates are strictly positive, that is $\hat{\pi}_{abc} > 0$.

Let $\boldsymbol{\pi}$ denote the vector $\boldsymbol{\pi} = (\pi_{111}, \ldots, \pi_{11C}, \ldots, \pi_{AB1}, \ldots, \pi_{ABC})^T$ of length $K = ABC$ and let the $AB + CB + AC$ constraints $N_{ab\bullet}/N$, $N_{a\bullet c}/N$ and $N_{\bullet bc}/N$ be stored in vector $\mathbf{c}$ and let matrix $\mathbf{A}$ be the $(AB + CB + AC) \times K$ matrix such that $\mathbf{A}\boldsymbol{\pi} = \mathbf{c}$. Then, following Little and Wu (1991), a (co)variance estimator for $\hat{\boldsymbol{\pi}}$ is:

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\pi}}) = n^{-1}\mathbf{U}(\mathbf{U}^T\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{D}^{-1}(\mathbf{p})\mathbf{U})(\mathbf{U}^T\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})\mathbf{U})^{-1}\mathbf{U}^T, \quad (2)$$

where $\mathbf{D}(\mathbf{a})$ is the diagonal matrix having vector $\mathbf{a}$ on its diagonal, and $\mathbf{p}$ is the vector of sample proportions, that is $\mathbf{p} = (p_{111}, \ldots, p_{11C}, \ldots, p_{AB1}, \ldots, p_{ABC})^T$ with $p_{abc} = y_{abc}/n$. Matrix $\mathbf{U}$ is an orthogonal complement of $\mathbf{A}$, such that $\mathbf{A}^T\mathbf{U} = 0$ and $(\mathbf{A}, \mathbf{U})$ has full rank. To achieve the full rank matrix $(\mathbf{A}, \mathbf{U})$, matrix $\mathbf{A}$ also needs to be of full rank. This requires removing three elements in vector $\mathbf{c}$ (and the corresponding rows in $\mathbf{A}$), as the second order constraints are linearly dependent, for example $N_{AB\bullet} = N - \sum_{a=1}^{A-1}\sum_{b=1}^{B-1} N_{ab\bullet}$.

Even though IPFP is often used to obtain population estimates $\hat{N}_{abc}$ via the simple formula

$$\hat{N}_{abc} = N\hat{\pi}_{abc}, \quad (3)$$

the (co)variance formula, see (2), to obtain confidence intervals for these population estimates is often not discussed in the literature on SP generation and is worth highlighting, as it provides an uncertainty measure.

It should be noted that the raking estimates denoted by $\hat{\pi}^r_{abc}$ based on all three second order population constraints are of the following form (Little and Wu 1991)

$$\log\left(\frac{\hat{\pi}^r_{abc}}{p_{abc}}\right) = \hat{\theta}^r + \hat{\theta}^r_{1(a)} + \hat{\theta}^r_{2(b)} + \hat{\theta}^r_{3(c)} + \hat{\theta}^r_{12(ab)} + \hat{\theta}^r_{13(ac)} + \hat{\theta}^r_{23(bc)}, \tag{4}$$

where θ are suitable parameters.

2.2. *Maximum Likelihood Approach*

The maximum likelihood method under random sampling (MLRS) has been considered for 2 by 2 tables by Smith (1947) and Little and Wu (1991) but has not been particularly extensively studied when dealing with more than two variables. For a three-way contingency table, Equation (1) can be expressed as $\mathbf{A}\boldsymbol{\pi} = \mathbf{c}$ with linearly dependent constraints removed.

Let $\mu_{abc} = E(y_{abc})$ with the corresponding vector $\boldsymbol{\mu}$ defined in a similar fashion as $\boldsymbol{\pi}$ and define the function $\mathbf{h}(\boldsymbol{\mu}) = \mathbf{A}\boldsymbol{\pi} - \mathbf{c}$ with $\boldsymbol{\pi} = \boldsymbol{\mu}/n$. With this definition, $\mathbf{h}(\boldsymbol{\mu}) = \mathbf{0}$ when $\mathbf{A}\boldsymbol{\pi} = \mathbf{c}$ and $\mathbf{h}(\boldsymbol{\mu}) \neq 0$ otherwise. Lang and Agresti (1994) and Lang (1996, 2004, 2005) provide a model framework and base estimation on maximising the log-likelihood subject to some arbitrary constraints expressed by $\mathbf{h}(\boldsymbol{\mu}) = 0$. This is achieved by using the famous method of Lagrange multipliers, which maximizes the constrained log-likelihood L_c

$$L_c = \text{constant} + \sum_{a,b,c} y_{abc} \log \pi_{abc} + \boldsymbol{\lambda}^T \mathbf{h}(\boldsymbol{\mu}), \tag{5}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{AB-1}, \ldots, \lambda_{AB+BC+AC-3})^T$ is a vector of the so-called Lagrange multipliers.

Joseph B. Lang provides an `R` function (mph.fit) available on http://homepage.stat.uiowa.edu/~jblang/mph.fitting/ for ML estimation of multinomial-Poisson homogeneous (MPH) models for contingency tables. Bergsma et al. (2009) provide a more efficient algorithm (`R` package `cmm`) to fit such models. Apart from obtaining estimates $\hat{\boldsymbol{\mu}}$, that will satisfy the population constraints, the ML method also provides a (co)variance matrix for $\hat{\boldsymbol{\mu}}$ as follows:

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\mu}}) = \mathbf{D}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T/n - \mathbf{D}(\hat{\boldsymbol{\mu}})\mathbf{H}(\mathbf{H}^T\mathbf{D}(\hat{\boldsymbol{\mu}})\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}(\hat{\boldsymbol{\mu}}), \tag{6}$$

where $\mathbf{H}(\boldsymbol{\mu}) = \frac{\partial \mathbf{h}^T(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ (Lang 2004).

Compared to log-linear models of the form $\log(\boldsymbol{\mu}) = \mathbf{X}\beta$ with design matrix $\mathbf{X}$ and the vector of fixed effects parameters $\boldsymbol{\beta}$, see, for example Agresti (2002), Formula (6) shows an additional term (the last term). This additional term reduces the variance imposed by the restrictions or constraints compared to the unconstrained model. Little and Wu (1991) proposed a different (co)variance formula for the ML method based on the delta method similar to (2) and is given by:

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\pi}}) = n^{-1}\mathbf{U}(\mathbf{U}^T\mathbf{D}(\hat{\boldsymbol{\pi}}^2/\mathbf{p})^{-1}\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{D}(\hat{\boldsymbol{\pi}}^2/\mathbf{p})^{-1}\mathbf{U})(\mathbf{U}^T\mathbf{D}(\hat{\boldsymbol{\pi}}^2/\mathbf{p})^{-1}\mathbf{U})^{-1}\mathbf{U}^T. \tag{7}$$

To obtain model-based population estimates $\hat{N}_{abc}$, Formula (3) is applied. Finally, the estimated (co)variance of the estimated population counts contained in the vector $\hat{\mathbf{N}} = \frac{N}{n}\hat{\boldsymbol{\mu}}$ is:

$$\widehat{\text{Cov}}(\hat{\mathbf{N}}) = N^2\widehat{\text{Cov}}(\hat{\boldsymbol{\pi}}). \tag{8}$$

It should be noted that the ML estimates $\hat{\pi}_{abc}^{ML}$ (based on second order population constraints) are of the form (see Appendix A)

$$\left(\frac{\hat{\pi}_{abc}^{ML}}{p_{abc}}\right)^{-1} = \hat{\theta}^{ML} + \hat{\theta}_{1(a)}^{ML} + \hat{\theta}_{2(b)}^{ML} + \hat{\theta}_{3(c)}^{ML} + \hat{\theta}_{12(ab)}^{ML} + \hat{\theta}_{13(ac)}^{ML} + \hat{\theta}_{23(bc)}^{ML}. \tag{9}$$

This ML method can also be used to fit standard log-linear models by including the model in the constraint function $\mathbf{h}(\boldsymbol{\mu})$ by setting $\mathbf{h}(\boldsymbol{\mu}) = \tilde{\mathbf{U}}^T \log \boldsymbol{\mu} = 0$, where $\tilde{\mathbf{U}}$ is a full column rank orthogonal complement of the design matrix $\mathbf{X}$. Lang (1996, 2004, 2005) extended this methodology to generalized log-linear models and homogeneous linear predictor models.

2.3. Other Estimation Methods

Two other popular estimation methods are the least squares method (LSQ) and the minimum chi-squared method (CHI2), see Little and Wu (1991). The LSQ estimates are obtained by minimizing the following criteria

$$\sum_{abc} \frac{(y_{abc} - n\pi_{abc})^2}{y_{abc}} + \boldsymbol{\lambda}^T \mathbf{h}(\boldsymbol{\mu}), \tag{10}$$

and CHI2 estimates by minimizing

$$\sum_{abc} \frac{(y_{abc} - n\pi_{abc})^2}{\pi_{abc}} + \boldsymbol{\lambda}^T \mathbf{h}(\boldsymbol{\mu}). \tag{11}$$

Similar to the ML method, solutions to (10) and (11) can be obtained by applying the Lagrange multiplier method. Under simple random sampling these methods are not recommended, because ML and IPFP provide more efficient estimates (Little and Wu 1991).

2.4. Adjusted Estimators

We also consider adjusted estimation methods "$+\alpha$", where the value of α is added to all counts y_{abc} before the estimators are calculated. Two such typical adjustments for contingency tables, namely $\alpha = 0.5$ and α being a tiny constant (the former to all counts and the latter only to zero counts), are mentioned in Subsection 2.1. Such adjustments of the data could lead to more efficient estimators. To further justify these adjustments, consider that the standard ML estimator for the multinomial distribution with K (here $K = ABC$) probabilities π_i and size n is $p_i = y_i/n$ for $i = 1, \ldots, K$ (the sample proportions). Under the Bayesian approach, the so-called Dirichlet distribution with parameters α_i is the conjugate family of priors for the multinomial distribution. The posterior distribution is also a Dirichlet distribution with parameters $y_i + \alpha_i$. The Bayesian estimate for π_i (the posterior mean) is $\frac{y_i+\alpha_i}{n+M}$ with $M = \sum_i \alpha_i$. Let $\gamma_i = \alpha_i/M$, then the Bayesian estimator (posterior mean) equals the weighted average

$$\frac{n}{n+M} \times p_i + \frac{M}{n+M} \times \gamma_i$$

with weights summing to one, that is $\frac{n}{n+M} + \frac{M}{n+M} = 1$ (Agresti and Hitchcock 2005).

A standard non-informative prior, the uniform prior, is obtained by setting $\alpha_i = 1$ (Jeffreys 1998; Gelman et al. 2003). In the binomial case ($K = 2$), this corresponds to assuming a uniform prior for π_i and leads to the Bayesian estimator $\frac{y_i+1}{n+2}$. In the multinomial case, this leads to $\frac{y_i+1}{n+K}$. In general, when adding α_i to cell y_i yielding new counts $\tilde{y}_i = y_i + \alpha_i$ and then applying the ML estimator $\tilde{y}_i/\tilde{n}$ with new sample size $\tilde{n} = n + M$ yields the Bayesian estimator, because $\frac{\tilde{y}_i}{\tilde{n}} \equiv \frac{y_i+\alpha_i}{n+M}$.

The case $\alpha_i = 1/2$ leads to the popular Jeffrey's prior. Often any prior with $\alpha_i = \alpha$ may be considered as non-informative (De Campos and Benavoli 2011). In this sense, "+0.5" and "+1" are special cases of non-informative priors. However, the concept of non-informative priors is debated in the Bayesian community, for example the uniform prior can be considered as highly informative and the Jeffrey's prior as non-informative, or vice versa. Agresti and Hitchcock (2005) noted the "lack of consensus about what 'non-informative' means". In this article, we consider all choices of a constant $\alpha_i = \alpha$ as non-informative following De Campos and Benavoli (2011).

The classical ML and Bayesian estimators apply to the unrestricted case (without imposing marginal constraints). When the margins/constraints are met by these classical estimators, however, they coincide with the restricted estimators, as $\boldsymbol{\lambda} = 0$ in this case, see Equation (5). If the prior distribution reflects the true sampling mechanism, then the Bayesian estimator has the highest efficiency by construction, as it is widely known that the (posterior) mean minimizes the mean squared error. Assuming the π_{abc} are not known and can be of any size, we anticipate that a choice of a "non-informative prior" ($\alpha > 0$) could also lead to improved efficiency in the restricted case.

Adjusting the observations and then applying standard Wald type confidence intervals (CI) has been applied by Agresti and Coull (1998), where the number of failures and successes was increased by two before the Wald type CI was applied (method "+2"). This method – the so-called Agresti-Coull-Interval – yielded better coverage than the standard Wald-type CI without "data adjustment" and even better coverage than the "exact" CI, in the sense that better means closer to the nominal 95% level, as the "exact" method is highly conservative. Based on these different choices for "+α" in the literature we also consider the methods "+2" and "+10" to investigate the effect of choosing a different value of $\alpha_i = \alpha$.

3. Misspecification Models

In theory, a probability sample is taken from a population, implying that both sample and population have the same characteristics. In practice, however, samples can differ systematically from the target population, due to, for example, omission of units or errors in the sampling frame, or very commonly due to the nonresponse of some selected units.

Let us now assume that the sample was obtained from a population, now referred to as the non-target population, which is not the same as the target population, the population of interest. We denote the probabilities referring to the non-target population by τ_{abc} and those to the target population by π_{abc}. Following Little and Wu (1991), we consider the following models relating π_{abc} and τ_{abc}

$$\left(\frac{\pi_{abc}}{\tau_{abc}}\right)^{\kappa} = \theta_\kappa + \theta_{1(a)\kappa} + \theta_{2(b)\kappa} + \theta_{3(c)\kappa} + \theta_{12(ab)\kappa} + \theta_{13(ac)\kappa} + \theta_{23(bc)\kappa}, \qquad (12)$$

where $\kappa = -1, 1, 2$ and $\kappa \to 0$ (in the following denoted by $\kappa = 0$) refers to the log-function, that is $\log\left(\frac{\pi_{abc}}{\tau_{abc}}\right)$. The specification of the θ parameters in (12) implies that second order population margins are provided. These four models specified by the value of κ provide flexible adjustments when sample and target population characteristics do not agree. Following similar arguments for the two-dimensional case as in Little and Wu (1991), we can show that the ML estimates for the model $\kappa = 0$ are provided by IPFP, see Appendix B. Similarly, it can be shown that the ML estimators for $\kappa = 1, -1, -2$ are of specific form. To summarize, the following results hold (Little and Wu 1991):

- For $\kappa = 1$: ML estimates are provided by LSQ
- For $\kappa = 0$: ML estimates are provided by IPFP
- For $\kappa = -1$: ML estimates are provided by ML(RS)
- For $\kappa = -2$: ML estimates are provided by CHI2.

The proofs for $\kappa = 1, -1, -2$ in the three-dimensional case are not shown to preserve space, but follow similar arguments as for $\kappa = 0$ and the two dimensional case.

Little and Wu (1991) compared all four estimation methods (IPFP, ML, CHI2, LSQ) in a simulation study while simulating data using random sampling and under each of the four misspecification models. The averaged results over a wide range of settings (Little and Wu 1991, see Table 1) show that under all five situations (random sampling and the four misspecification models), either IPFP or MLRS are the best performing methods. MLRS is best under random sampling and for the models $\kappa = -1, 2$, whereas IPFP is best under the models $\kappa = 0$ and $\kappa = 1$. To be more precise, MLRS is best for $\kappa = -1$ and IPFP is best for $\kappa = 0$, as expected but LSQ and IPFP are well performing methods for $\kappa = 1$ with IPFP being slightly better. Similarly, MLRS and CHI2 perform best when $\kappa = -2$, with a slightly better performance of MLRS than CHI2.

Even though these results are averaged over all simulations and limited to 2 by 2 tables, they still show that the commonly used IPFP and MLRS methods generally perform well, but their results refer to a single 2 by 2 table, an unrealistic situation for often sparse multidimensional tables. The next section considers a simulation study specially designed for multidimensional tables with a large number of cells.

4. Simulation Study

4.1. Setup

Little and Wu (1991) and Causey (1983) conducted empirical simulation studies based on 2 by 2 tables with constraints referring to the two (marginal) variables. It is not clear how these results can be extended to multidimensional tables with a large number of cells and more than two sets of population constraints.

When obtaining a sample (table) with small n from a population (table) with many cells, the sample table is often sparse. The simulation study considers table cells with low to relatively large probabilities by setting $A = 5$, $B = 4$ and $C = 2$ and where the $K = ABC = 40$ probabilities $\boldsymbol{\pi} = (\tilde{\pi}_{111}, \ldots, \tilde{\pi}_{11c}, \ldots, \tilde{\pi}_{AB1}, \ldots, \tilde{\pi}_{ABC})^T = (\tilde{\pi}_1, \tilde{\pi}_2 \ldots, \tilde{\pi}_K)^T$ are monotone increasing and the kth probability is $\tilde{\pi}_k \propto \exp([5(k-1)+1]/40)$ (proportional to exponential function and then normalized to sum to one), yielding

$\tilde{\pi}_{111} = \tilde{\pi}_1 = 0.0009 < \ldots < \tilde{\pi}_{ABC} = \tilde{\pi}_{40} = 0.1183$. We consider simple random sampling (RND) and the misspecification models in Section 3.

For each of these models, we sample randomly 10,000 population tables, where each table contains randomly generated counts denoted by y_{abc}^{pop} from a multinomial distribution with parameters $\tilde{\boldsymbol{\pi}}$ and N. The aim is to estimatethe population cell probabilities denoted by $\pi_{abc} = y_{abc}^{pop}/N$. For small near zero $\tilde{\pi}_{abc}$, the obtained y_{abc}^{pop} are often zero. This is a realistic scenario for multidimensional tables, as in practice some population counts will indeed be small and often be zero.

In Section 5, we use individual level sample data from a larger area (n large) to estimate population totals of a smaller area (N), such as $n > N$. This scenario does not warrant random sampling without replacement (for which $n < N$) and requires the consideration of misspecification models.

For simplicity, the misspecification models specified by (12) only include main effects $\theta_{1(a)\kappa}, \theta_{2(b)\kappa}, \theta_{3(c)\kappa}$, which are generated under a $N(\mu = 0, \sigma^2 = 1)$ distribution for $\kappa = 0$ and from a lognormal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$ for $\kappa = 1, -1, -2$ to have strictly positive parameters in the latter case. Rearranging Equation (12) in terms of τ_{abc} for $\kappa = 1, -1, -2$ gives

$$\tau_{abc} = \pi_{abc} \times (\eta_{abc})^{-\frac{1}{\kappa}}$$

and for $\kappa = 0$

$$\tau_{abc} = \pi_{abc} \times \exp(-\eta_{abc}),$$

where $\eta_{abc} = \theta_\kappa + \theta_{1(a)\kappa} + \theta_{2(b)\kappa} + \theta_{3(c)\kappa}$. The constant θ_κ is chosen such that all τ_{abc} sum to one. Then based on these τ_{abc}, a sample of size n can be obtained by random sampling to estimate the π_{abc}.

We investigate the performance of the estimators IPFP, MLRS (abbreviated here ML), LSQ, and CHI2 and their (co)variance estimators, and their adjusted versions "+0", "+0.5", "+1", "+2", and "+10" and any combination thereof, for example IPFP + 0 and CHI2 + 10.

To assess the efficiency we calculate the mean squared error (MSE) of IPFP + 0 and the relative MSE (RMSE) of all other methods relative to IPFP + 0. As each table has many cells, the MSE for a cell is defined as $E(\hat{\pi}_{abc} - \pi_{abc})^2$. The RMSE is always relative to IPFP + 0. A value greater than one indicates a larger MSE than the MSE of IPFP + 0 and a value less than one indicates a more efficient estimator. The bias is also assessed by calculating the relative bias, here defined as $E(\hat{\pi}_{abc} - \pi_{abc})/E(\pi_{abc})$.

For the confidence intervals (CI) and the ML method we consider Lang's formula and the delta method, see Formulae (6) and (7), however, due to similar performances and a generally slightly better performance of Lang's formula, we only show results based on Lang's formula.

One of the main questions is which (model-based) estimation method is best. If a true random sample is obtained, then the ML method is the appropriate choice. The package `mipfp` provides results of several goodness of fit (GOF) tests, such as Pearson's score test X^2, the Likelihood-Ratio statistic G^2 and the Wald statistic W^2, developed by Lang (2004). They test essentially whether the sample agrees with the population; in formula $H_0 : \mathbf{h}(\boldsymbol{\mu}) = 0$ versus $H_1 : \mathbf{h}(\boldsymbol{\mu}) \neq 0$. If H_0 is rejected, then there is a strong indication that the 'sample' is not a real random sample from the target population and one of the

misspecification models should be considered. To assess whether these GOF tests are useful for determining whether a true sample is provided or rather a misspecification model applies, we also recorded the rejection rate of the GOF tests. For zero cell counts or zero estimates these might not always be calculable and adjusted versions X^2_{adj}, G^2_{adj} and W^2_{adj} are considered. Similar to Lang's mph.fit implementation, X^2_{adj} is calculated over those cells for which $\hat{\pi}_{abc} > 0$, G^2_{adj} over those for which $\hat{\pi}_{abc} > 0$ and $y_{abc} > 0$, and W^2_{adj} over those for which $y_{abc} > 0$. Without these adjustments, the statistics would be undefined for many data sets, as they would contain zero valued denominators.

To illustrate the methods on higher dimensions, we also consider the five dimensional case where each dimension has three categories leading to $K = 3^5 = 243$ probabilities with $\tilde{\pi}_k \propto \exp(k/243)$.

4.2. Results

Table 1 shows the results of the MSE/RMSE for the IPFP and ML methods, similarly Table 2 shows the relative bias and Table 3 the coverage for these two methods along with the adjusted versions. Similar tables for LSQ and CHI2 are shown in Appendix D. The tables show n, N, the model (either RND or $\kappa = 0, 1, -1, 2$), and the expected probability $E(\pi) = E(\pi_{abc}) = \tilde{\pi}_{abc}$, calculated over all tables. The values of $E(\pi)$ are chosen such that the impact of relatively small, medium-sized and large probabilities (and likewise counts) can be observed, as different methods are expected to perform differently for cells of different sizes. The values of $E(\pi)$ are ordered, such that the first row under each configuration represents the smallest $E(\pi)$ (e.g., 0.09%), the second is the first tertile, a tertile is defined as the first three-quantile, (e.g., 0.40%), the third is the median (e.g., 0.97%) and the fourth is the largest of the $E(\pi)$ (e.g., 11.83%).

For RND, we observe that all methods (IPFP, ML, CHI2 and LSQ) perform similarly but ML is still generally the best in terms of efficiency and the smaller the n is, the larger is the performance improvement. The efficiency generally improves when $\alpha > 0$. For $N = 10{,}000$ and $n = 600$, "+10" appears best, whereas for $N = 600$ and $n = 200$ the size of α that has highest efficiency depends on $E(\pi)$. For small $E(\pi)$, "+0.5" appears best, for medium and large $E(\pi)$ it is "+2" and "+10". As can be seen in Table 2, the drawback of larger α is that the bias generally deteriorates. Table 3 shows that the coverage of the unadjusted "+0" method is often too low and improved by the adjusted methods $\alpha > 0$, an optimum appears around "+0.5" and "+1". An exception appears for the five-dimensional case and large $E(\pi)$, for which $\alpha = 0.5$ appears to lower the coverage slightly, but it still increases coverage significantly for small and medium sized $E(\pi)$.

Let us now focus on the misspecification models. When evaluating the unadjusted versions in terms of efficiency, for large $E(\pi)$ the ML method under the respective misspecification model is best. For example, CHI2 is best under $\kappa = -2$ when $E(\pi) = 11.83\%$. The results of Little and Wu (1991) are confirmed in the sense that under models $\kappa = 0$ and $\kappa = 1$ the LSQ and IPFP methods perform well, and under $\kappa = -1$ and $\kappa = -2$ the ML and CHI2 methods perform well.

It also apparent, however, that for large $E(\pi)$, ML is not efficient for $\kappa = 0, 1$, whereas IPFP is a well performing method for all κ. Overall, it appears that IPFP is the best method, as it performs well regardless of the misspecification model and the size of $E(\pi)$. In terms

Table 1. $E(\pi)$ in percentages, $10^5 \times$ actual MSE for IPFP + 0 (highlighted in bold) and RMSE for other methods relative to MSE of IPFP + 0, RMSE $<$ 1 indicates better and RMSE $>$ 1 worse, all based on 10,000 simulated data sets.

	IPFP					ML				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	**0.159**	0.362	0.237	0.172	0.120	1.027	0.360	0.240	0.186	0.200
0.40	**0.566**	0.666	0.485	0.302	0.085	1.003	0.663	0.480	0.294	0.076
0.97	**1.303**	0.855	0.742	0.580	0.198	0.996	0.851	0.739	0.577	0.197
11.8	**9.297**	0.974	0.963	0.955	0.937	0.996	0.970	0.960	0.957	1.157
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	**0.363**	0.290	0.302	0.333	0.368	1.092	0.292	0.311	0.355	0.431
0.40	**1.443**	0.380	0.293	0.276	0.345	1.005	0.375	0.288	0.273	0.352
0.97	**2.924**	0.654	0.500	0.378	0.352	0.993	0.647	0.494	0.373	0.368
11.8	**19.73**	0.932	0.903	0.863	0.633	0.991	0.925	0.903	0.897	1.127
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	**0.098**	0.673	0.742	0.940	1.506	1.022	0.641	0.690	0.864	1.497
0.40	**0.413**	0.759	0.668	0.642	1.048	0.982	0.722	0.630	0.592	0.945
0.97	**1.029**	0.877	0.790	0.687	0.702	0.935	0.810	0.730	0.632	0.616
11.8	**8.618**	0.978	0.971	0.976	1.141	0.732	0.717	0.714	0.718	0.937
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	**0.302**	0.559	0.597	0.634	0.643	1.308	0.530	0.463	0.445	0.525
0.40	**1.908**	0.407	0.404	0.425	0.458	1.553	0.769	0.594	0.451	0.309
0.97	**5.730**	0.460	0.428	0.425	0.432	2.416	1.563	1.287	0.989	0.446
11.8	**51.19**	0.576	0.604	0.689	0.994	4.655	3.740	3.653	3.505	2.764
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	**0.094**	0.688	0.771	0.986	1.586	1.090	0.699	0.728	0.900	1.581
0.40	**0.418**	0.755	0.671	0.652	1.037	1.181	0.845	0.722	0.646	0.934
0.97	**1.074**	0.888	0.815	0.733	0.776	1.315	1.153	1.037	0.885	0.708
11.8	**8.726**	0.989	0.992	1.012	1.229	1.724	1.688	1.663	1.629	1.620
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	**0.080**	0.709	0.778	1.016	1.795	1.024	0.703	0.768	1.004	1.883
0.40	**0.356**	0.802	0.715	0.678	1.090	0.988	0.789	0.700	0.660	1.077
0.97	**0.886**	0.908	0.839	0.749	0.740	0.983	0.892	0.824	0.734	0.716
11.8	**6.518**	0.985	0.981	0.989	1.148	0.925	0.910	0.906	0.915	1.198
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	**0.021**	0.171	0.142	0.122	0.079	0.999	0.206	0.204	0.223	0.255
0.07	**0.116**	0.174	0.132	0.113	0.083	0.996	0.176	0.151	0.153	0.170
0.17	**0.299**	0.484	0.298	0.162	0.062	0.999	0.483	0.299	0.165	0.073
2.05	**3.056**	0.977	0.922	0.803	0.313	1.000	0.990	0.956	0.877	0.456

Table 2. $E(\pi)$ in percentages, the relative bias of IPFP and ML relative to $E(\pi)$ in percentages based on 10,000 data sets.

	IPFP					ML				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	1.5	22.8	30.8	37.0	36.0	1.8	23.4	32.5	41.3	53.9
0.40	0.6	0.3	0.1	−0.1	0.4	0.7	0.2	−0.0	−0.3	−0.8
0.97	−0.4	0.1	0.6	1.3	3.8	−0.4	0.1	0.6	1.5	5.2
11.8	0.0	0.7	1.2	1.9	4.2	0.0	0.8	1.3	2.1	5.8
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	−2.2	31.6	35.6	35.6	22.6	0.8	34.4	42.1	48.0	54.5
0.40	0.9	−0.2	−0.3	−0.2	0.8	0.7	−0.3	−0.6	−0.8	−1.1
0.97	−0.3	1.1	2.1	3.3	4.4	−0.5	1.1	2.3	3.9	7.7
11.8	0.0	1.6	2.5	3.5	4.6	0.0	1.7	2.8	4.4	9.4
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	1.4	17.3	24.5	31.4	36.1	1.7	17.1	24.7	33.2	49.0
0.40	−0.1	−0.3	−0.5	−0.6	−0.3	−0.3	−0.5	−0.6	−0.7	−1.1
0.97	0.3	0.5	0.8	1.3	3.2	−0.0	0.3	0.6	1.1	4.0
11.8	0.0	0.5	0.9	1.5	3.6	−0.0	0.5	0.9	1.5	4.5
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	−0.7	25.9	28.2	28.9	23.6	1.4	20.7	26.7	33.5	47.4
0.40	1.1	−0.0	−0.1	−0.2	0.1	2.0	0.7	0.3	−0.0	−0.5
0.97	0.4	1.6	2.2	2.8	3.7	0.3	2.1	2.4	2.9	5.0
11.8	−0.0	1.6	2.1	2.6	3.3	0.1	2.1	2.4	3.1	5.7
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	−1.2	15.8	23.4	30.6	36.0	−1.1	15.5	23.3	32.2	49.4
0.40	−1.4	−1.4	−1.4	−1.3	−0.4	−1.7	−1.7	−1.6	−1.46	−1.1
0.97	0.2	0.6	0.9	1.4	3.5	0.2	0.6	0.9	1.5	4.3
11.8	−0.1	0.4	0.8	1.4	3.5	−0.1	0.4	0.8	1.4	4.3
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	1.1	16.2	23.8	31.7	39.0	1.5	16.3	24.4	33.7	51.4
0.40	−0.4	−0.6	−0.7	−0.8	−0.5	−0.4	−0.6	−0.7	−0.9	−1.2
0.97	−0.4	−0.1	0.2	0.7	3.0	−0.4	−0.1	0.2	0.8	3.7
11.8	−0.0	0.4	0.8	1.3	3.4	−0.0	0.4	0.8	1.4	4.3
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	−5.5	85.1	88.9	81.4	44.2	−5.7	100.5	119.1	131.6	144.1
0.07	−4.3	−32.0	−34.0	−33.7	−23.1	−4.3	−35.8	−41.5	−45.2	−49.0
0.17	5.1	6.0	6.7	7.3	5.6	5.1	6.2	7.4	9.0	12.2
2.05	−0.7	2.2	3.04	3.3	−0.1	−0.7	2.7	4.32	6.0	8.3

of adjusted versions, overall to find a good compromise between bias and efficiency, the methods "+0.5" and "+1" appear good methods and in particular IPFP + 1.

In terms of coverage, the unadjusted version "+0" often suffers from undercoverage. In general, we would expect that adding a constant α to each cell would lead to a decrease in coverage because, due to the artificially increased sample size the standard errors are smaller and the CIs are smaller. The results show, however, that the adjusted versions

Table 3. $E(\pi)$ in percentages, coverage of IPFP and ML methods and their adjusted versions in percentages based on 10,000 simulated data sets.

	IPFP					ML				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	41.5	100.0	100.0	100.0	100.0	41.5	100.0	100.0	100.0	100.0
0.40	99.1	99.7	99.9	100.0	100.0	99.1	99.7	99.9	100.0	100.0
0.97	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
11.8	94.9	95.1	95.0	94.5	89.8	95.0	95.2	95.1	94.5	85.9
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	15.8	42.5	42.6	42.5	40.3	14.5	42.6	42.6	42.6	41.5
0.40	80.1	90.8	90.9	91.0	90.7	80.0	90.8	90.9	91.0	90.6
0.97	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7
11.8	97.7	97.7	97.6	96.7	91.9	97.8	97.8	97.5	96.6	80.9
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	29.9	33.2	34.5	34.4	30.5	29.7	33.2	34.8	35.2	32.7
0.40	82.6	84.7	85.5	85.9	84.6	82.7	85.0	85.6	86.1	85.1
0.97	99.3	99.3	99.3	99.4	99.4	99.3	99.3	99.3	99.3	99.3
11.8	89.0	89.6	89.4	88.8	80.1	93.7	93.7	93.5	93.1	84.3
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	19.6	28.7	28.7	27.9	23.8	16.9	28.0	30.5	31.8	31.4
0.40	64.3	80.3	81.2	80.8	78.2	56.9	70.7	74.2	77.7	81.6
0.97	92.9	97.2	97.5	97.4	96.6	86.8	90.0	91.1	92.3	96.1
11.8	63.8	68.5	66.1	62.5	50.9	32.1	34.1	34.5	35.0	35.6
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	29.4	32.9	34.1	34.2	30.2	29.1	32.6	34.4	34.8	32.2
0.40	82.7	84.9	85.6	86.1	84.7	82.0	84.5	85.2	85.8	85.0
0.97	99.2	99.3	99.3	99.3	99.3	98.7	98.9	99.1	99.1	99.3
11.8	89.2	89.0	88.6	87.6	79.7	81.2	81.3	81.3	80.9	74.7
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	30.1	32.2	33.8	34.0	30.2	30.1	32.3	34.0	34.2	31.3
0.40	84.0	85.4	86.0	86.4	85.2	84.0	85.4	86.0	86.4	85.3
0.97	99.2	99.2	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3
11.8	93.1	93.3	93.0	92.7	85.8	94.2	94.1	94.1	93.8	84.7
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	7.6	76.0	76.0	76.0	75.9	7.6	76.0	76.0	76.0	76.0
0.07	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
0.17	65.2	98.4	99.8	100.0	99.9	65.3	98.5	99.9	100.0	99.9
2.05	93.2	92.5	91.8	90.7	89.2	93.1	92.7	91.8	90.1	84.7

"+0.5" and "+1" increase the coverage (improving undercoverage), and only for $\alpha = 2, 10$ the coverage appears to decrease compared to $\alpha = 0.5, 1$. Hence the adjusted versions "+0.5" and "+1" appear to be best, when aiming for the coverage to be near or above 95%. The results are similar to Agresti and Coull (1998) as they show that adding pseudo-observations improves upon coverage, but the results are also dissimilar, because our results rather suggest "+0.5" or "+1" and not "+2".

Figures 1 (three dimensions) and 2 (five dimensions) show boxplots of 10,000 estimates for the smallest and largest cells for ML, ML + 0.5, ML + 1, IPFP, IPFP + 0.5, and IPFP + 1. It shows the effect of reducing the MSE when adjusting the data.

The rejection rate of the GOF tests based on a five percent significance level is represented in Table 4 for the unadjusted data ("+0"), because from the results not presented here, it is clear that adjusting the data ("+α" with $\alpha > 0$) leads to too large type I error under H_0. For IPFP, the adjusted GOF versions G^2_{adj}, W^2_{adj} X^2_{adj} are recommended over the unadjusted versions G^2, W^2 X^2, and for ML and CHI2 all

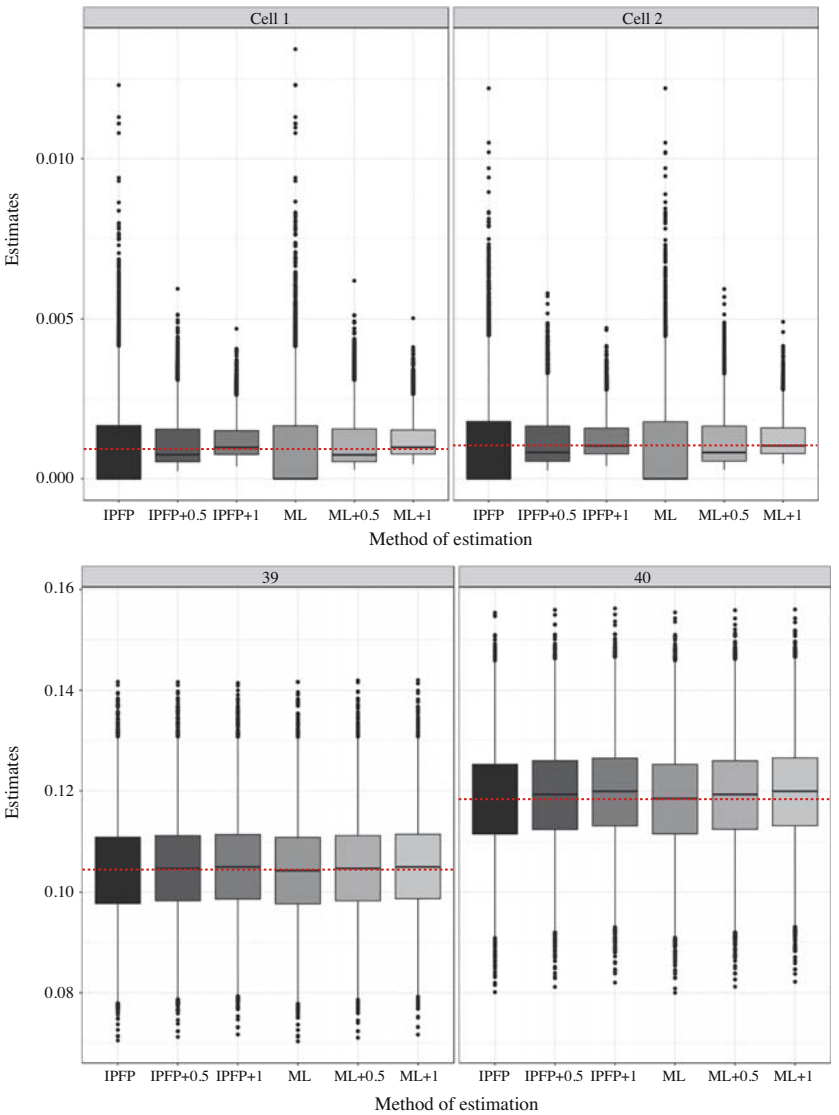

Fig. 1. Boxplots of 10,000 estimates of the methods IPFP, IPFP + 0.5, IPFP + 1, ML, ML + 0.5, and ML + 1 for the two smallest (top) and the two largest (bottom) out of 40 = 5 × 4 × 2 cells in the three dimensional case under random sampling with N = 600 and n = 100 compared with average population proportions (dotted line).

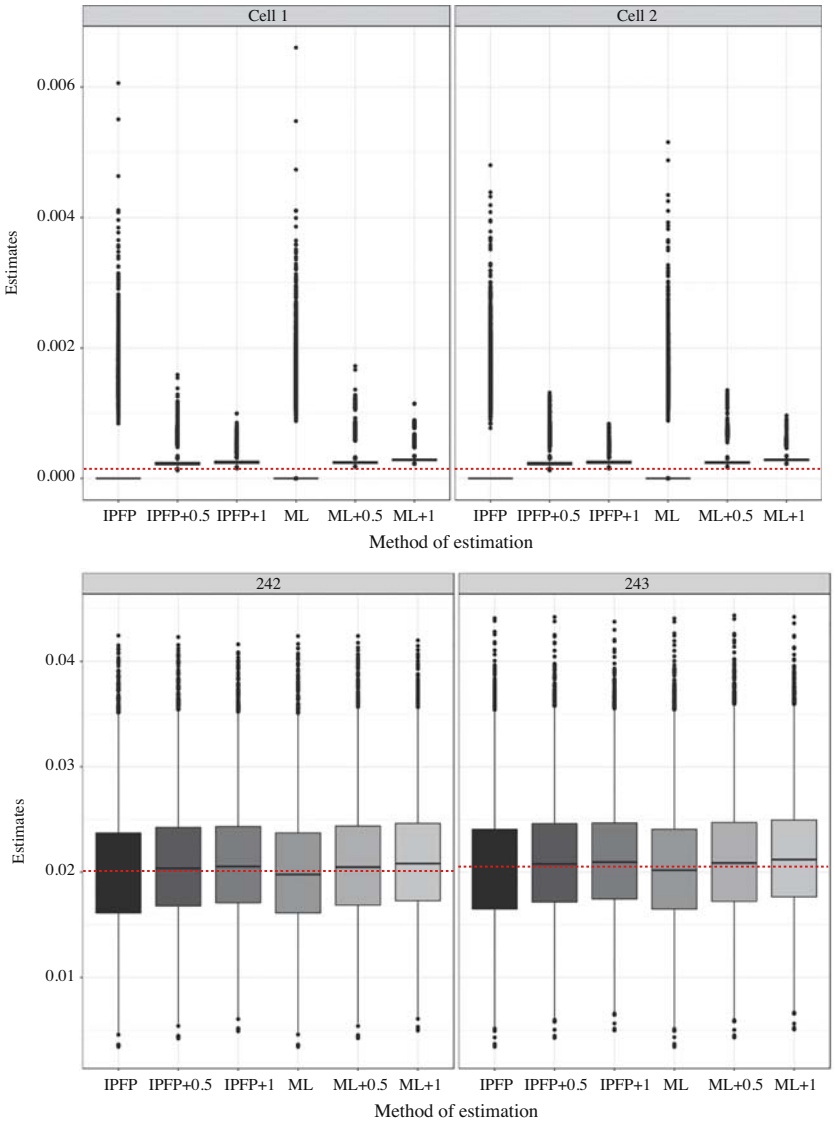

Fig. 2. Boxplots of 10,000 estimates for IPFP, IPFP + 0.5, IPFP + 1, ML, ML + 0.5, and ML + 1 for the two smallest (top) and the two largest (bottom) out of 243 = 3^5 cells in the five dimensional case with three categories in each dimension and under randomsampling with N = 600 and n = 100 compared with average population proportions (dotted line).

GOF versions maintain approximately the type I error, whereas for LSQ the type I error appears too large.

5. Estimating Multidimensional Population Counts for the Illawarra Region

The study area in this article is the Illawarra region in New South Wales, Australia, with a total population of 365,338 in 2011. The Illawarra is the coastal region situated immediately south of Sydney and north of the Shoalhaven or South Coast region

Table 4. Rejection rate of the GOF tests G^2, X^2, W^2 and their adjusted versions (adj) based on $\hat{\pi}$ of the four estimation methods IPFP, ML, CHI2, and LSQ.

	G^2	G^2_{adj}	W^2	W^2_{adj}	X^2	X^2_{adj}
			Dimension = 3, RND, $N = 10{,}000$, $n = 600$			
IPFP	0.143	0.043	0.069	0.068	0.286	0.044
ML	0.041	0.041	0.070	0.068	0.040	0.040
CHI2	0.043	0.043	0.069	0.068	0.038	0.038
LSQ	0.048	0.048	0.070	0.069	0.053	0.053
			Dimension = 3, RND, $N = 600$, $n = 200$			
IPFP	–	0.004	0.019	0.018	–	0.005
ML	0.004	0.004	0.032	0.029	0.003	0.003
CHI2	0.004	0.004	0.019	0.018	0.003	0.003
LSQ	0.095	0.095	0.109	0.091	0.099	0.099
			Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$			
IPFP	1.000	0.993	0.993	0.993	1.000	0.994
ML	0.993	0.993	0.993	0.993	0.993	0.993
CHI2	0.993	0.993	0.992	0.992	0.993	0.993
LSQ	0.994	0.994	0.993	0.993	0.994	0.994
			Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$			
IPFP	–	1.000	1.000	1.000	–	1.000
ML	1.000	1.000	1.000	1.000	1.000	1.000
CHI2	1.000	1.000	1.000	1.000	1.000	1.000
LSQ	1.000	1.000	1.000	1.000	1.000	1.000
			Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$			
IPFP	1.000	0.992	0.992	0.992	1.000	0.992
ML	0.992	0.992	0.992	0.992	0.992	0.992
CHI2	0.992	0.992	0.992	0.992	0.992	0.992
LSQ	0.992	0.992	0.992	0.992	0.993	0.992
			Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$			
IPFP	0.500	0.869	0.873	0.873	0.500	0.871
ML	0.868	0.868	0.873	0.873	0.868	0.868
CHI2	0.869	0.869	0.873	0.872	0.866	0.866
LSQ	0.873	0.873	0.873	0.873	0.876	0.876
			Dimension = 5, RND, $N = 10{,}000$, $n = 600$			
IPFP	–	0.042	0.054	0.050	–	0.039
ML	0.039	0.039	0.054	0.050	0.037	0.037
CHI2	0.042	0.042	0.054	0.050	0.035	0.035
LSQ	0.048	0.048	0.054	0.050	0.050	0.050

(see Figure 3). The smallest geographic area defined in the Australian Statistical Geography Standard (ASGS) is the Statistical Level 1 (SA1), indicated by index j, for which the data are available to our study. The number of males and females living within the study area and three major subregions is presented in Table 5.

The Australian census tables released by the Australian Bureau of Statistics (ABS) available through the ABS Table Builder Pro were used for this study. SA1-specific

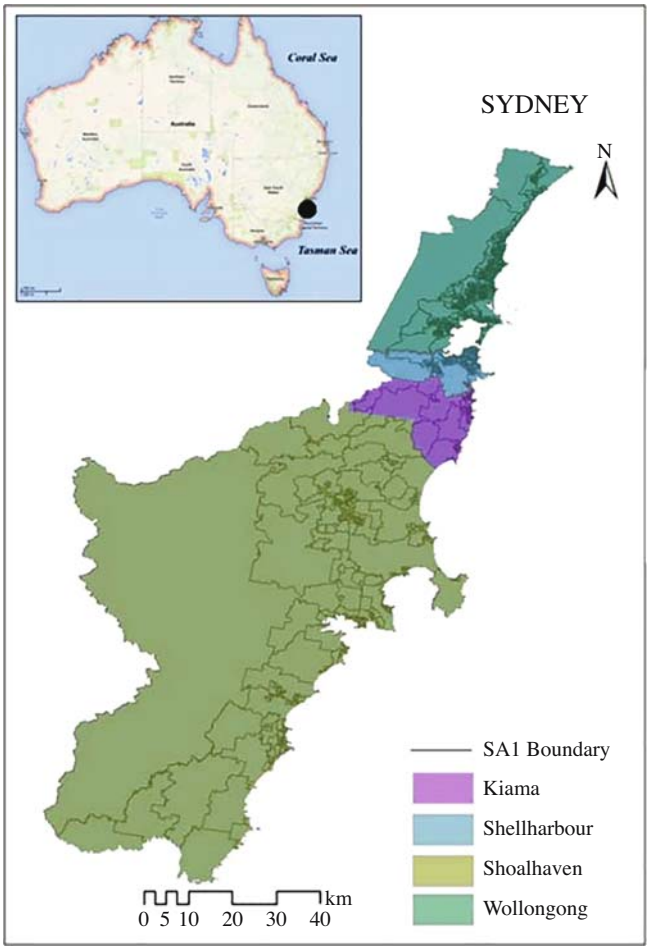

Fig. 3. Map of study area (Illawarra Region).

marginal population counts for age by gender and for family type are available from the census. These tables contain 18 age categories ($0-4, 5-9, \ldots, 80-84, >84$), two genders and four family categories (couple with no children, couple with children, one parent family, other family). Our aim is to find pseudo-census tables for age by sex by family type for each of the SA1 of the Illawarra region.

Table 5. Living population in the study area for the Illawarra and three greater subregions based on 2011 Australian Census Data.

Area	Males	Females	Total
Kiama and Shellharbour	40,160	42,184	82,344
Wollongong	94,986	97,079	192,065
Shoalhaven	44,667	46,262	90,929
Total	179,813	185,252	365,338

There are $144 = 18 \times 2 \times 4$ cells and corresponding probabilities $\pi_{abc}^{(j)}$ for each SA1 $j = 1, \ldots, 61$, and six of these $\pi_{abc}^{(j)}$ are set to zero because 'family couples without children' should have no family member in the age groups $0-4$, $5-9$, $10-14$ for both genders. This leaves 138 cells for each SA1 j.

A 1% Basic Census Sample File (CSF) with $n = 2{,}902$ housing units was available to this study through the Confidentialised Unit Record File (CURF) microdata system. As there is no geographical information (as SA1) attached to the 1% CSF, this sample is used for all of the 61 SA1 study areas with population sizes of $6 \leq N \leq 1060$. As $n > N$, the 1% Basic CSF can only be thought of as a pseudo-sample and not a random sample without replacement from the target-population. The `R` package `mipfp` (Barthélemy and Suesse 2016) is used to produce the raking (IPFP), ML(RS), CHI2, and LSQ estimates.

Figure 4 shows the results when using only the 1% CSF without imposing population constraints. The results do not vary across SA1s, as we have only one sample – the 1% CSF – containing people from the whole Illawarra region, ignoring the available known marginal totals for each SA1 j. Our approach of using this pseudo-sample might seem questionable, as sample and target populations are not the same, but as mentioned in Section 3, IPFP and ML also provide ML estimates under the misspecification models where $\kappa = 0, -1$. Here based on the known marginal totals, these models are of the specific form

$$\left(\frac{\pi_{abc}^{(j)}}{\tau_{abc}}\right)^{\kappa} = \theta^{(j)} + \theta_{1(a)}^{(j)} + \theta_{2(b)}^{(j)} + \theta_{3(c)}^{(j)} + \theta_{12(ab)}^{(j)}; \quad \kappa = -1, 0, 1, 2, \tag{13}$$

where the first variable is age, the second is gender and the third is family type. The specific SA1 is indicated by index j, however it should be noted that $\pi_{abc}^{(j)}$ contains superscript j whereas τ_{abc} does not have superscript j, because the same data set is used as a (pseudo-) sample from a population that is characterized by τ_{abc}. In contrast, each SA1 j has its specific

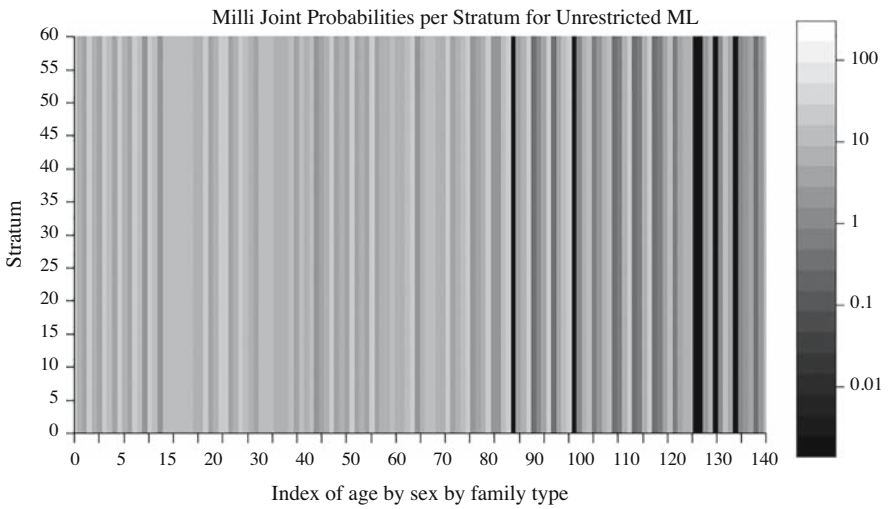

Fig. 4. Unrestricted ML estimator for 138 probabilities (columns) for each stratum (rows) based on 1% CSF file.

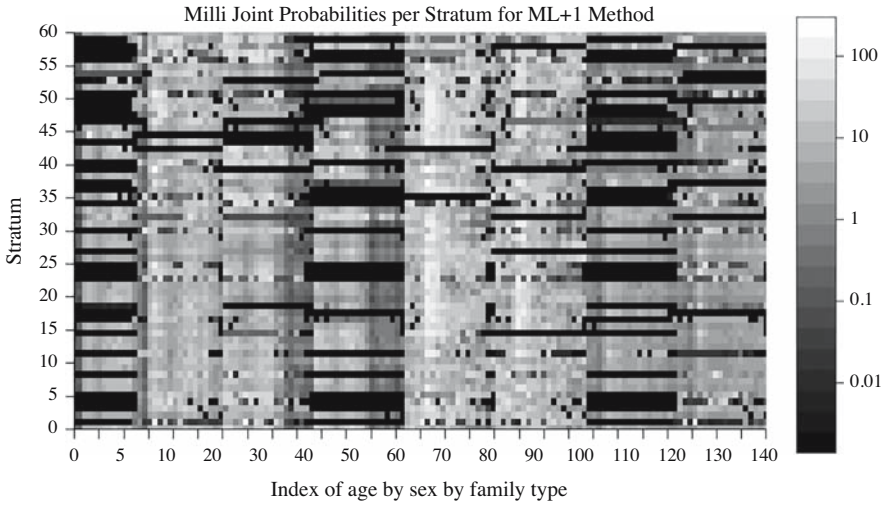

Fig. 5. ML + 1 estimates for 138 probabilities (columns) for each stratum (rows) based on 1% CSF file and known marginal population counts.

population distribution, denoted by $\pi_{abc}^{(j)}$, and its estimate will be different for each j, due to the availability of known marginal population counts that are specific for SA1 j.

The SA1 could be considered as another variable and the joint distribution containing $61 \times 138 = 8{,}418$ cells could be estimated at once, however this would yield the same results as when estimating 138 cells for each SA1 j separately and would also increase the number of constraints by a factor of 61. Usually the larger the number of cells and the number of constraints become, the more unstable becomes the optimisation algorithm due to the curse of dimensionality. Joint estimation would also complicate the specification of cells and the margins, increasing the chance or errors by the user. Generally, it is not

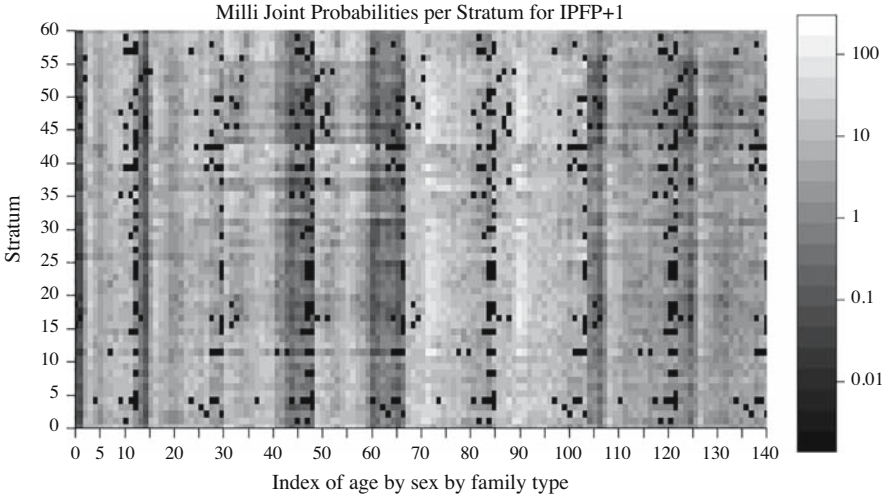

Fig. 6. IPFP + 1 estimates for 138 probabilities (columns) for each stratum (rows) based on 1% CSF file and known marginal population counts.

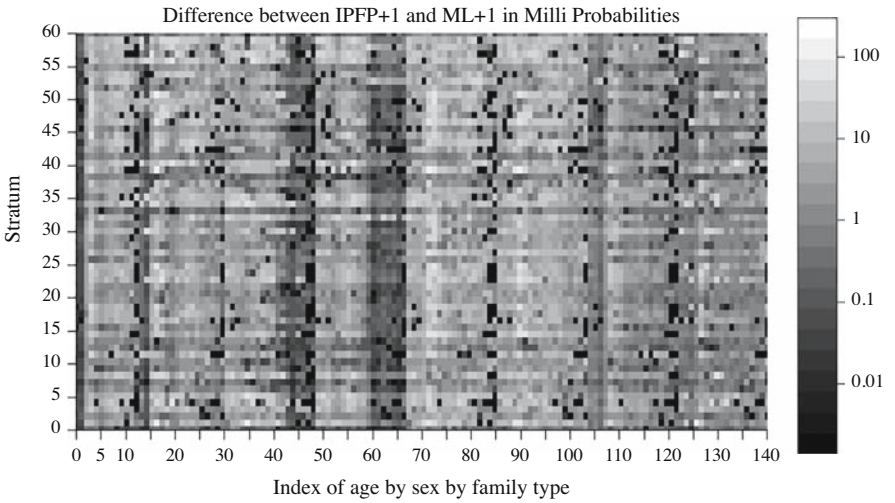

Fig. 7. Absolute differences between ML + 1 and IPFP + 1 estimates for 138 probabilities (columns) for each stratum (rows) based on 1% CSF file and known marginal population counts.

advisable to increase the dimensionality artificially, if not needed. We rather recommend separate estimation for each SA1 j, using for example the `R` package `mipfp`.

Figure 5 shows the results of the ML + 1 method and Figure 6 shows the results of IPFP + 1 for each geographical area (stratum) referring to 138 probabilities. The results differ, as can be seen in Figure 7. Based on the results of the simulation study and the fact that no true random sample is available, but only a pseudo-sample, IPFP + 1 is preferred to improve efficiency and to improve coverage. If bias is to be avoided, then IPFP + 0 is recommended for the point estimates.

All methods exactly match the population constraints. For example, for SA1 with ID 1114961 the population marginal proportions (relative to SA1 size) for family type are 0.191 (couple with no children), 0.292 (couple with children), 0.410 (one parent family) and 0.107 (other family). Results for IPFP + 0, ML + 0, and the CHI2 and LSQ methods are not shown to preserve space.

The age by sex by family type tables based on a pseudo-sample and two known marginal tables serve as pseudo-census counts/tables, as the true census counts are not released due to confidentiality restrictions. The results are also valuable for population reconstruction, as they form the basis for the simulation of area-specific SPs.

6. Discussion

The main objective of this article is to compare several estimation methods for obtaining population count estimates $\hat{N}_{abc} = N\hat{\pi}_{abc}$ or equivalently, estimates of the joint probabilities $\hat{\pi}_{abc}$, when a sample is available and when marginal population counts (subtables) are known. IPFP, also known as raking, is the standard method to deal with this problem (Ballas et al. 2005; Smith et al. 2005), due primarily to the popularity of IPFP and widely available software. IPFP can also be applied if the seed (sample) is only partially available, for example due to confidentiality restriction.

The ML method is not very popular because of several limitations: The availability of a representative sample is not always warranted, this sample needs to be a true random sample, and there are not many statistical packages that have implemented this approach. On the other hand, IPFP requires non-zero cells of the sample to converge and to provide a unique solution to the underlying optimisation problem. Nonetheless, it must be noted that even with zero cell counts, IPFP often still converges. Another problem with IPFP is that some of the estimated cell counts are zero, when some cells in the marginal population tables are zero, even if a solution exists with positive estimates. This means some combinations of attributes in the simulation process of synthetic populations will be impossible to sample due to the zero estimates. The ML method also has convergence problems when zero cell counts are present.

As an alternative we proposed the data adjustment methods of the form "$+\alpha$" with $\alpha = 0.5, 1, 2, 10$ and the simulation study showed that these methods generally improve efficiency at the costs of increased bias. They also generally improve coverage. Based on all the results, the adjustments "$+0.5$" and "$+1$" appear best. Overall, the simulation study suggests that under the mis-specification models that IPFP $+ 1$ appears to be a reasonable choice in terms of improved coverage, and in terms of efficiency at the costs of increased bias. Under random sampling, instead we recommend ML $+ 1$. Even if biased yet more efficient point estimates might not be desirable for some practitioners, the improved coverage of the confidence intervals is still worth highlighting.

It is sometimes difficult to determine whether random sampling or one of the misspecification models applies and this means that it is difficult to make a decision whether IPFP or ML should be applied. The GOF tests enable testing whether a random sample can be assumed. If the tests are not rejected, then we recommend using ML $+ 1$, and if rejected then IPFP $+ 1$. In some cases, as in our example, it is apparent, for example when $n > N$, that the sample cannot be a true random sample from the target population and the IPFP $+ 1$ method is preferred. The "$+1$" adjustment also solves the possible convergence problem, as now none of the cells is zero. If bias is a major concern, then either the IPFP $+ 0$ or ML $+ 0$ methods should be applied, depending on the results ofthe GOF tests.

IPFP is overall the preferred method under the misspecification models, however estimates of some cells of IPFP might be zero due to zero counts in the marginal tables. If this is undesirable, for example when the synthetic simulation process does not aim at excluding particular combinations of attributes, then the ML method is the preferred alternative (and its adjusted versions), if indeed the ML estimates are non-zero for all cells.

In the SP literature, the presented (co)variance estimators are often unknown and are worth highlighting, as they form the basis of Wald-type confidence intervals, the measure of uncertainty and precision.

Any data set that possesses features that are otherwise not available from the target population is recommended over using artificial data as the seed, as illustrated in Section 5. For example generally, age (by sex) is related to family type, as each family type usually has a particular age (by sex) distribution. While the population tables on age by sex and family type provide information about the marginal distributions, they do not provide information on how age by sex and family type are related. Because no sample was available, Barthélemy and Toint (2013) used equal weights as the seed, however, this will imply incorrectly that there is independence between age by sex and family type, see

Appendix C. In contrast, using some available related samples (even if not a real random sample in the classical sense) that have typical dependence between age by sex and family type present will preserve the relationship between age by sex and family type in the final estimates. This preservation of higher order terms in the respective log-linear model (Mosteller 1968), when IPFP is applied, is clearly advantageous. This is also advantageous in the classical sense. The seed is one data set and the marginal population tables form another data set. Using the two real data sets jointly will provide more information than a single real data set alone.

In this article, we only considered four misspecification models. In practice, however, this class of models might be too narrow to obtain accurate estimates in all cases. Developing a wider class of misspecification models and the investigation of the best performing method under this extended class will be the subject of future research.

Appendix A. Form of Maximum Likelihood Estimates

Let us write the constrained log-likelihood L_c, see (5), with second order population constraints as

$$L_c = \text{constant} + \sum_{a,b,c} y_{abc} \log \pi_{abc} + \sum_a \sum_b \lambda_{a,b}(\pi_{ab\bullet} - (N_{ab\bullet}/N))$$

$$+ \sum_a \sum_c \lambda_{a,c}(\pi_{a\bullet c} - (N_{a\bullet c}/N)) + \sum_b \sum_c \lambda_{b,c}(\pi_{\bullet bc} - (N_{\bullet bc}/N)).$$

Now let us take first derivatives with respect to π_{abc}

$$\frac{\partial L_c}{\partial \pi_{abc}} = \frac{y_{abc}}{\pi_{abc}} - \lambda_{a,b} - \lambda_{a,c} - \lambda_{b,c},$$

where $\lambda_{a,b}$, $\lambda_{a,c}$ and $\lambda_{b,c}$ are Lagrange multiplier determined by the ML algorithm.

Setting derivatives to zero $\frac{\partial L_c}{\partial \pi_{abc}} = 0$ and imposing typical constraints such as second order parameters sum to zero, estimates have the form

$$\left(\frac{\hat{\pi}^{ML}_{abc}}{p_{abc}}\right)^{-1} = \hat{\theta}^{ML} + \hat{\theta}^{ML}_{1(a)} + \hat{\theta}^{ML}_{2(b)} + \hat{\theta}^{ML}_{3(c)} + \hat{\theta}^{ML}_{12(ab)} + \hat{\theta}^{ML}_{13(ac)} + \hat{\theta}^{ML}_{23(bc)}. \tag{A.1}$$

It also shows that if second order population constraints are included, then the form of the estimates include second order terms. If for example first order population constraints are included, then the right hand side of (A.1) will only contain main effects (first order terms).

Appendix B. Showing that IPFP Estimates are ML Estimates under Model (12) with $\kappa \to 0$

Suppose sampling fractions are small, then y_{abc} are approximately multinomially distributed and the sample proportions p_{abc} are ML estimates of τ_{abc}. By Model (12) with $\kappa \to 0$, the population probabilities π_{abc} are given by

$$\pi_{abc} = \tau_{abc} \exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)}),$$

and ML estimates of the θ's are obtained by solving

$$\pi_{ab\bullet} = \sum_c \tau_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

$$\pi_{a\bullet c} = \sum_b \tau_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

$$\pi_{\bullet bc} = \sum_a \tau_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)}).$$

As the ML estimates of functions of τ_{abc} are the functions evaluated at $\hat{\tau}_{abc} = p_{abc}$, the ML estimates of π_{abc} are of the form

$$\hat{\pi}_{abc} = p_{abc}\exp(\hat{\theta} + \hat{\theta}_{1(a)} + \hat{\theta}_{2(b)} + \hat{\theta}_{3(c)} + \hat{\theta}_{12(ab)} + \hat{\theta}_{13(ac)} + \hat{\theta}_{23(bc)}),$$

where the $\hat{\theta}$ estimates are obtained by solving

$$\pi_{ab\bullet} = \sum_c p_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

$$\pi_{a\bullet c} = \sum_b p_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

$$\pi_{\bullet bc} = \sum_a p_{abc}\exp(\theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)})$$

These equations are solved by the raking estimates, see Equation (4).

Similar arguments can be shown to show that MLRS provides ML estimates for Model (12) with $\kappa = -1$, LSQ provides ML estimates for Model (12) with $\kappa = 1$ and CHI2 provides ML estimates for Model (12) with $\kappa = 2$.

Appendix C. Independence with Equal Weights

Suppose we have three variables and suppose equal initial weights as the seed, i.e., $\pi^{(0)}_{abc} \propto 1$, which implies that $\pi^{(0)}_{abc} = \frac{1}{\sum_{abc} 1} = \frac{1}{K}$ $(K = ABC)$.

$$\pi^{(0)}_{ab\bullet} = \sum_c \pi^{(0)}_{abc} = \frac{C}{K} = \frac{1}{AB}$$

$$\pi^{(0)}_{a\bullet c} = \sum_b \pi^{(0)}_{abc} = \frac{B}{K} = \frac{1}{AC}$$

$$\pi^{(0)}_{\bullet bc} = \sum_a \pi^{(0)}_{abc} = \frac{A}{K} = \frac{1}{BC}$$

$$\pi^{(0)}_{a\bullet\bullet} = \sum_{b,c} \pi^{(0)}_{abc} = \frac{BC}{K} = \frac{1}{A}$$

$$\pi^{(0)}_{\bullet b\bullet} = \sum_{a,c} \pi^{(0)}_{abc} = \frac{AC}{K} = \frac{1}{B}$$

$$\pi^{(0)}_{\bullet\bullet c} = \sum_{a,b} \pi^{(0)}_{abc} = \frac{AB}{K} = \frac{1}{C}$$

Now from these equations, it is apparent that the three categorical variables are independent when $\pi^{(0)}_{abc}$ would be the final estimates (zero iterations of IPFP).

Assuming that population counts are available for each of the three variables, it should be noted that the raking/IPFP estimates denoted by $\hat{\pi}^r_{abc}$ are of the following form (Little and Wu 1991)

$$\log\left(\frac{\hat{\pi}^r_{abc}}{p_{abc}}\right) = \hat{\theta}^r + \hat{\theta}^r_{1(a)} + \hat{\theta}^r_{2(b)} + \hat{\theta}^r_{3(c)},$$

similar to Equation (4), where $p_{abc} = \frac{1}{K}\left(= \pi^{(0)}_{abc}\right)$, because the "sample" consists of equal weights (pseudo-data), as no real data set/sample is available. Hence

$$\log\left(\hat{\pi}^r_{abc}\right) = \text{const} + \hat{\theta}^r_{1(a)} + \hat{\theta}^r_{2(b)} + \hat{\theta}^r_{3(c)}$$

and it follows that

$$\hat{\pi}^r_{abc} = \frac{1}{g_{\bullet\bullet\bullet}}\exp\left(\hat{\theta}^r_{1(a)}\right) \times \exp\left(\hat{\theta}^r_{2(b)}\right) \times \exp\left(\hat{\theta}^r_{3(c)}\right) = \frac{1}{g_{\bullet\bullet\bullet}}\alpha_a \times \alpha_b \times \alpha_c,$$

where $g_{\bullet\bullet\bullet} = \sum_{a,b,c}\alpha_a\alpha_b\alpha_c = \left[\sum_a \alpha_a\right] \times \left[\sum_b \alpha_b\right] \times \left[\sum_c \alpha_c\right]$. From this we obtain the estimated marginal probabilities

$$\hat{\pi}^r_{a\bullet\bullet} = \frac{1}{g_{\bullet\bullet\bullet}}\sum_{b,c}\alpha_a \times \alpha_b \times \alpha_c = \frac{\sum_{b,c}\alpha_a \times \alpha_b \times \alpha_c}{\sum_{a,b,c}\alpha_a \times \alpha_b \times \alpha_c} = \frac{\alpha_a}{\sum_a \alpha_a}$$

$$\hat{\pi}^r_{\bullet b\bullet} = \frac{1}{g_{\bullet\bullet\bullet}}\sum_{a,c}\alpha_a \times \alpha_b \times \alpha_c = \frac{\sum_{a,c}\alpha_a \times \alpha_b \times \alpha_c}{\sum_{a,b,c}\alpha_a \times \alpha_b \times \alpha_c} = \frac{\alpha_b}{\sum_b \alpha_b}$$

$$\hat{\pi}^r_{\bullet\bullet c} = \frac{1}{g_{\bullet\bullet\bullet}}\sum_{a,b}\alpha_a \times \alpha_b \times \alpha_c = \frac{\sum_{a,b}\alpha_a \times \alpha_b \times \alpha_c}{\sum_{a,b,c}\alpha_a \times \alpha_b \times \alpha_c} = \frac{\alpha_c}{\sum_c \alpha_c}$$

and therefore we conclude independence, because, for example, the following equation holds

$$\hat{\pi}^r_{abc} = \hat{\pi}^r_{a\bullet\bullet} \times \hat{\pi}^r_{\bullet b\bullet} \times \hat{\pi}^r_{\bullet\bullet c}.$$

When, for example, X_1 and X_2 are age and sex and X_3 is family type and the known population margins are provided for age by sex (i.e., (X_1, X_2) known) and for family type (X_3 known), then similarly final estimates will imply that still (X_1, X_2) and X_3 are independent.

Appendix D. Simulation Results of CHI2 and LSQ

Table D.6. $E(\pi)$ in percentages, RMSE for methods CHI2 and LSQ relative to IPFP + 0, RMSE $<$ 1 indicates better and RMSE $>$ 1 worse, all based on 10,000 simulated data sets.

	CHI2					LSQ				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	1.038	0.362	0.244	0.193	0.213	0.997	0.368	0.244	0.178	0.388
0.40	1.010	0.662	0.479	0.293	0.076	1.000	0.673	0.497	0.325	0.175
0.97	0.993	0.850	0.738	0.577	0.203	1.006	0.860	0.748	0.587	0.219
11.8	0.995	0.969	0.959	0.964	1.380	1.117	0.980	0.970	0.959	0.785
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	1.099	0.294	0.316	0.361	0.439	0.941	0.314	0.347	0.417	0.620
0.40	1.021	0.375	0.288	0.274	0.354	1.034	0.397	0.320	0.318	0.427
0.97	0.999	0.645	0.493	0.374	0.379	1.390	0.664	0.513	0.396	0.376
11.8	0.987	0.923	0.911	0.941	1.673	6.633	0.948	0.917	0.858	0.593
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	1.105	0.653	0.688	0.861	1.514	1.073	0.866	1.039	1.441	2.360
0.40	1.068	0.757	0.651	0.610	0.962	1.182	0.975	0.915	0.983	2.071
0.97	1.049	0.873	0.775	0.661	0.652	1.363	1.200	1.101	0.991	1.213
11.8	0.882	0.873	0.875	0.894	1.333	2.985	1.809	1.784	1.792	2.196
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	1.441	0.572	0.485	0.454	0.535	1.287	0.998	1.229	1.458	1.442
0.40	1.786	0.912	0.702	0.524	0.335	1.472	0.798	0.819	0.974	1.559
0.97	2.918	1.926	1.613	1.247	0.515	1.753	1.022	0.867	0.808	1.060
11.8	5.900	4.884	4.823	4.705	4.003	6.221	3.474	3.104	2.715	2.477
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	1.222	0.754	0.734	0.901	1.606	1.007	0.795	1.031	1.579	2.898
0.40	1.471	1.032	0.832	0.693	0.952	1.026	0.806	0.761	0.862	2.122
0.97	1.928	1.628	1.428	1.145	0.750	1.067	0.886	0.817	0.757	1.126
11.8	2.553	2.501	2.472	2.431	2.466	1.893	0.850	0.869	0.956	1.782
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	1.072	0.707	0.767	1.006	1.911	1.007	0.752	0.857	1.185	2.315
0.40	0.999	0.788	0.700	0.662	1.090	1.036	0.845	0.770	0.763	1.431
0.97	0.988	0.895	0.826	0.736	0.733	1.091	0.958	0.890	0.802	0.864
11.8	0.916	0.901	0.899	0.916	1.401	1.855	1.131	1.128	1.146	1.357
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	0.997	0.221	0.224	0.246	0.282	1.007	0.116	0.091	0.147	0.151
0.07	0.994	0.178	0.156	0.160	0.178	0.992	0.183	0.118	0.110	0.683
0.17	0.998	0.484	0.302	0.168	0.074	0.999	0.484	0.297	0.162	0.064
2.05	1.000	1.000	0.983	0.933	0.595	1.001	0.960	0.880	0.732	0.488

Table D.7. $E(\pi)$ in percentages, the relative bias of CHI2 and LSQ relative to $E(\pi)$ in percentages based on 10,000 data sets.

	CHI2					LSQ				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	1.7	23.5	33.2	42.7	56.2	2.4	22.0	27.4	24.8	−69.9
0.40	0.7	0.2	−0.1	−0.4	−1.2	0.6	0.3	0.2	0.3	3.9
0.97	−0.4	0.1	0.6	1.6	5.9	−0.4	0.1	0.5	1.2	2.3
11.8	0.0	0.7	1.3	2.3	7.0	0.1	0.7	1.2	1.7	1.5
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	0.2	35.5	43.9	50.2	57.1	2.0	24.7	12.9	−20.9	−98.9
0.40	0.7	−0.3	−0.7	−1.1	−1.8	−0.2	0.1	0.4	1.6	6.2
0.97	−0.4	1.1	2.5	4.4	8.7	0.3	1.0	1.8	2.6	0.8
11.8	0.0	1.8	3.1	5.0	12.5	−2.0	1.5	2.0	2.1	−2.7
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	2.0	17.0	24.8	34.0	51.3	2.4	17.6	24.5	28.5	−9.4
0.40	−0.1	−0.1	−0.6	−0.8	−1.4	0.0	−0.0	−0.2	−0.2	1.5
0.97	−0.1	0.2	0.4	1.1	4.5	0.6	0.8	1.1	1.5	2.4
11.8	0.1	0.5	0.8	1.5	4.9	−0.0	0.5	0.9	1.4	2.2
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	2.4	19.4	26.1	34.2	50.5	4.9	32.9	34.4	28.8	−16.1
0.40	2.0	0.71	0.2	−0.1	−0.2	0.6	−1.4	−1.3	−0.6	3.9
0.97	0.4	2.2	2.5	2.9	5.2	−0.3	1.3	1.9	2.5	2.1
11.8	0.3	2.4	2.7	3.2	6.2	1.9	1.8	2.2	2.4	0.6
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	−1.1	15.4	23.2	32.9	51.8	−0.7	15.8	23.0	27.5	−5.4
0.40	−1.9	−1.8	−1.7	−1.6	−1.4	−1.3	−1.3	−1.3	−1.1	1.5
0.97	0.4	0.6	0.9	1.4	4.7	0.4	0.7	1.0	1.4	2.6
11.8	0.1	0.5	0.8	1.4	4.7	−0.1	0.5	0.9	1.4	2.1
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	1.7	16.3	24.7	34.5	53.7	1.5	16.1	23.0	27.4	−13.8
0.40	−0.3	−0.6	−0.8	−0.9	−1.4	−0.5	−0.6	−0.7	−0.8	1.0
0.97	−0.3	−0.1	0.2	0.8	4.2	−0.4	−0.1	0.2	0.6	2.2
11.8	−0.0	0.4	0.8	1.4	4.9	−0.083	0.4	0.7	1.2	2.0
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	−5.9	106.2	126.8	140.4	153.5	−4.8	43.0	−25.8	−94.8	−97.2
0.07	−4.4	−37.2	−43.1	−46.9	−50.5	−4.6	−22.2	−9.3	20.1	115.6
0.17	5.1	6.5	7.9	9.7	12.4	5.1	5.7	5.8	5.7	2.7
2.05	−0.71	3.1	5.2	7.7	12.3	−0.68	1.4	1.0	−1.2	−13.2

Table D.8. $E(\pi)$ in percentages, coverage of CHI2 and LSQ methods and their adjusted versions in percentages based on 10,000 simulated data sets.

	CHI2					LSQ				
$E(\pi)$	+0	+0.5	+1	+2	+10	+0	+0.5	+1	+2	+10
Dimension = 3, RND, $N = 10{,}000$, $n = 600$										
0.09	41.3	99.9	99.9	99.9	99.9	41.5	99.9	99.9	99.8	42.6
0.40	99.0	99.6	99.9	100.0	100.0	99.0	99.6	99.8	100.0	100.0
0.97	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
11.8	95.0	95.2	95.2	94.5	81.2	94.7	94.9	94.8	94.2	92.0
Dimension = 3, RND, $N = 600$, $n = 200$										
0.09	14.3	42.5	42.6	42.6	41.5	17.2	41.6	39.1	30.2	0.9
0.40	79.9	90.7	90.9	90.9	90.6	79.4	90.7	90.9	90.9	90.4
0.97	99.6	99.6	99.6	99.6	99.6	98.1	99.6	99.6	99.6	99.6
11.8	97.8	97.8	97.5	96.3	68.3	90.4	97.6	97.4	96.6	91.7
Dimension = 3, $\kappa = -1$, $N = 500$, $n = 1{,}000$										
0.09	29.5	33.0	34.7	35.1	32.7	29.7	31.8	31.4	28.8	16.0
0.40	82.4	84.7	85.6	85.9	84.9	81.4	83.0	83.8	83.7	77.6
0.97	99.0	99.2	99.3	99.3	99.3	98.8	99.1	99.1	99.2	98.7
11.8	90.6	90.7	90.5	89.8	78.2	79.1	79.8	79.5	78.5	66.7
Dimension = 3, $\kappa = 0$, $N = 500$, $n = 1{,}000$										
0.09	16.1	26.5	29.4	31.4	31.3	15.3	20.3	17.5	13.6	6.9
0.40	53.5	66.6	70.7	75.1	80.8	51.8	63.7	64.1	61.0	44.8
0.97	83.3	87.2	88.3	90.2	95.0	80.5	89.1	91.2	92.1	85.9
11.8	25.7	27.9	28.4	28.7	29.5	28.9	31.3	31.6	31.3	23.3
Dimension = 3, $\kappa = 1$, $N = 500$, $n = 1{,}000$										
0.09	28.6	32.1	34.1	34.7	32.3	29.5	32.1	31.6	28.6	14.8
0.40	81.0	83.7	84.7	85.5	84.8	82.4	84.6	85.2	85.0	77.7
0.97	97.9	98.2	98.4	98.8	99.2	98.9	99.2	99.2	99.2	99.0
11.8	73.5	73.9	74.0	74.0	68.7	90.2	90.9	90.4	88.1	70.8
Dimension = 3, $\kappa = -2$, $N = 500$, $n = 1{,}000$										
0.09	29.9	32.1	33.9	34.2	31.3	30.1	32.0	33.2	31.9	17.2
0.40	84.0	85.4	85.9	86.4	85.2	83.7	85.1	85.5	86.0	83.3
0.97	99.2	99.2	99.2	99.2	99.2	99.1	99.2	99.2	99.2	99.2
11.8	94.1	94.2	94.1	93.6	81.7	90.8	91.4	91.0	90.4	82.5
Dimension = 5, RND, $N = 10{,}000$, $n = 600$										
0.01	7.7	76.0	76.0	76.0	76.0	7.7	75.9	65.3	4.8	0.0
0.07	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
0.17	65.3	98.5	100.0	100.0	99.9	65.2	98.3	99.8	100.0	99.8
2.05	93.1	92.9	91.8	89.8	79.2	93.3	92.2	91.6	89.8	72.0

7. References

Agresti, A. 2002. *Categorical Data Analysis*. New York: Wiley.

Agresti, A. and B.A. Coull. 1998. "Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions." *The American Statistician* 52: 119–126. Doi: http://dx.doi.org/10.1080/00031305.1998.10480550.

Agresti, A. and D.B. Hitchcock. 2005. "Bayesian Inference for Categorical Data Analysis." *Statistical Methods and Applications* 14: 297–330. Doi: http://dx.doi.org/10.1007/s10260-005-0121-y.

Arentze, T., H. Timmermans, and F. Hofman. 2007. "Creating Synthetic Household Populations: Problems and Approach." *Journal of the Transportation Research Board, 2014*, 85–91. Doi: http://dx.doi.org/10.3141/2014-11.

Ballas, D., G. Clarke, D. Dorling, H. Eyre, B. Thomas, and D. Rossiter. 2005. "Simbritain: A Spatial Microsimulation Approach to Population Dynamics." *Population, Space and Place* 11: 13–34. Doi: http://dx.doi.org/10.1002/psp. 351.

Barthélemy, J. and T. Suesse. 2016. "mipfp: Multidimensional Iterative Proportional Fitting and Alternative Models. R package version 3.1." Available from: http://CRAN.R-project.org/package=mipfp.

Barthélemy, J. and P.L. Toint. 2013. "Synthetic Population Generation without a Sample." *Transportation Science* 47: 266–279. Doi: http://dx.doi.org/10.1287/trsc.1120.0408.

Beckman, R., K. Baggerly, and M. McKay. 1996. "Creating Synthetic Baseline Populations." *Transportation Research Part A: Policy and Practice* 30: 415–429. Doi: http://dx.doi.org/10.1016/0965-8564(96)00004-3.

Bergsma, W., M. Croon, and J. Hagenaars. 2009. *Marginal Models for Dependent, Clustered and Longitudinal Categorical Data*. New York: Springer.

Causey, B.D. 1983. "Estimation of Proportions for Multinomial Contingency Tables Subject to Marginal Constraints." *Communications in Statistics-Theory and Methods* 12: 2581–2587. Doi: http://dx.doi.org/10.1080/03610928308828624.

De Campos, C.P. and A. Benavoli. 2011. "Inference with Multinomial Data: Why to Weaken the Prior Strength." In IJCAI Proceedings-International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain July 16–22, 2011, Volume 22, pp. 2107. Available at: http://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/view/3292.

Deming, W. and F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known." *Annals of Mathematical Statistics* 11: 367–484. Available at: http://www.jstor.org/stable/2235722.

Deville, J., C. Särndal, and O. Sautory. 1991. "Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 86: 87–95.

Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flotterod. 2013. "Simulation Based Population Synthesis." *Transportation Research Part B: Methodological* 58: 243–263. Doi: http://dx.doi.org/10.1016/j.trb.2013.09.012.

Fienberg, S. 1970. "An Iterative Procedure for Estimation in Contingency Tables." *Annals of Mathematical Statistics* 41: 907–917. Available at: http://www.jstor.org/stable/2239244.

Gange, S.J. 1995. "Generating Multivariate Categorical Variates Using the Iterative Proportional Fitting Algorithm." *American Statistician* 49: 134–138. Available at: http://www.tandfonline.com/doi/abs/10.1080/00031305.1995.10476130.

Gargiulo, F., S. Ternes, S. Huet, and G. Deffuant. 2010. "An Iterative Approach for Generating Statistically Realistic Populations of Households." *PLOS ONE* 5(1), e8828. Doi: http://dx.doi.org/10.1371/journal.pone.0008828.

Geard, N., J. McCaw, A. Dorin, K. Korb, and J. McVernon. 2013. "Synthetic Population Dynamics: A Model of Household Demography." *Journal of Artificial Societies and Social Simulation* 16(1): 1–23. Doi: http://dx.doi.org/10.18564/jasss.2098.

Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2003. *Bayesian Data Analysis* (2nd ed.). New York: Chapman and Hall/CRC Press.

Harland, K., A. Heppenstall, D. Smith, and M. Birkin. 2012. "Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques." *Journal of Artificial Societies and Social Simulation* 15: 1–24. Doi: http://dx.doi.org/10.18564/jasss.1909.

Huynh, N., J. Barthelemy, and P. Perez. 2016. "A Heuristic Combinatorial Optimisation Approach to Synthesising a Population for Agent Based Modelling Purposes." *Journal of Artificial Societies and Social Simulation* 19: 11. Doi: http://dx.doi.org/10.18564/jasss.3198.

Ireland, C. and S. Kullback. 1968. "Contingency Tables with Given Marginals." *Biometrika* 55: 179–199. Doi: https://doi.org/10.1093/biomet/55.1.179.

Jeffeys, H. 1998. *The Theory of Probability*. Oxford: Oxford University Press.

Lang, J. 1996. "Maximum Likelihood Methods for a Generalized Class of Loglinear Models." *Annals of Statistics* 24: 726–752.

Lang, J. 2004. "Multinomial-Poisson Homogeneous Models for Contingency Tables." *Annals of Statistics* 32: 340–383.

Lang, J. 2005. "Homogeneous Linear Predictor Models for Contingency Tables." *Journal of the American Statistical Association* 100: 121–134. Doi: http://dx.doi.org/10.1198/016214504000001042.

Lang, J. and A. Agresti. 1994. "Simultaneously Modelling Joint and Marginal Distributions of Multivariate Categorical Responses." *Journal of the American Statistical Association* 89: 625–632.

Lenormand, M. and G. Deffuant. 2013. "Generating a Synthetic Population of Individuals in Households: Sample-Free vs Sample-Based Methods." *Journal of Artificial Societies and Social Simulation* 16: 1–16. Doi: http://dx.doi.org/10.18564/jasss.2319.

Little, J. and M. Wu. 1991. "Models for Contingency Tables with Known Margins when Target and Sampled Population Differ." *Journal of the American Statistical Association* 86: 87–95.

Lu, H. and A. Gelman. 2003. "A Method for Estimating Design-Based Sampling Variances for Surveys with Weighting, Poststratification, and Raking." *Journal of Official Statistics* 19: 133–151.

Mosteller, F. 1968. "Association and Estimation in Contingency Tables." *Journal of the American Statistical Association* 63: 1–28. Doi: http://dx.doi.org/10.2307/2283825.

Purcell, N. and L. Kish. 1980. "Postcensal Estimates for Local Areas (or Domains)." *International Statistical Review* 43: 3–18.

Rubin, D. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Smith, D., G. Clarke, and K. Harland. 2005. "Improving the Synthetic Data Generation Process in Spatial Microsimulation Models." *Environment and Planning* A 41: 1251–1268. Doi: https://doi.org/10.1068/a4147.

Smith, J. 1947. "Estimation of Linear Functions of Cell Proportions." *Annals of Mathematical Statistics* 18: 231–254.

Stephan, F. 1942. "Iterative Method of Adjusting Frequency Tables when Expected Margins are Known." *Annals of Mathematical Statistics* 13(2): 166–178.

Zhang, L. and R. Chambers. 2004. "Small Area Estimates for Cross-Classifications." *Journal of the Royal Statistical Society:* Series B 66: 479–496. Doi: http://dx.doi.org/10.1111/j.1369-7412.2004.05266.x.

Received March 2015
Revised February 2017
Accepted February 2017