NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

10-15

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

Thomas Suesse, Mohammad-Reza Namazi-Rad, Payam Mokhtarian and Johan Barthelemy

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: anica@uow.edu.au

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

Thomas Suesse[a] *, Mohammad-Reza Namazi-Rad[a,b]
Payam Mokhtarian[a] and Johan Barthélemy[b]
[a] *National Institute for Applied Statistics Research Australia*
University of Wollongong, New South Wales 2522, Australia

[b] *SMART Infrastructure Facility, University of Wollongong*
New South Wales 2522, Australia

Abstract

Estimating population counts for multidimensional tables based on a representative sample data 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 (SPs). In order to generate a reliable SP, tabulating multidimensional tables of agents' socio-demographics is needed. In this paper we review the iterative proportional fitting procedure (IPFP) and the maximum likelihood (ML) method for estimating the cell counts in multidimensional tables subject to known population sub-tables. We also review two standard error estimators for ML and IPFP and investigate their performance in a simulation study, in which we consider mis-specification models, for which sample and target populations differ systematically. The empirical results show that a simple adjustment can lead to more efficient estimates when table probabilities are low. The methods discussed in this paper along with standard error estimators, one of which is relatively new, 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 Australia to obtain cell counts estimates for the desired three-way table for age by sex by family type subject to marginal tables for age by sex and family type.

Keywords: Census Data, IPFP, Log-linear Model, Model-Based Inference, Count Estimation, Synthetic Population.

*corresponding author, email: tsuesse@uow.edu.au

1 Introduction

While in many countries, census data are still the major source for geographically detailed estimates of populations and economies, such data are being released at higher levels of aggregation in shape of contingency tables, as releasing the fully informative disaggregated data while preserving the confidentiality is very challenging. One way to overcome such a challenge to generate an artificial population built from pseudo-census information obtained from anonymous survey and census data at the individual level in a way that the resulting artificial population realistically matches the observed population in a geographical zone for a given set of criteria (Beckman et al. 1996). Using this approach, the identification of population units or/and their sensitive information in the generated area-specific synthetic data will be difficult (Rubin 1987).

The purpose of generating a reliable synthetic population (SP) is to create a valid representation of the spatially distributed population units (e.g. individuals and households). In the literature on simulation-based population synthesis, approaches for generating an area-specific population of individuals and households aim at matching the aggregate-level information from the census (Arentze 2007, Gargiulo et al. 2010, Harland et al. 2012, Barthlémy and Toint 2013, Lenormand and Deffuant 2013, Geard et al. 2013, Namazi-Rad et al. 2014). In order to generate a reliable SP, tabulating multi-dimensional tables of agents' socio-demographics is needed.

This article focuses on the estimation of population counts in multidimensional contingency tables when a random sample is available together with known marginal population tables of lower dimensions. The commonly known approaches for estimating these multidimensional table counts - the iterative proportional fitting procedure (IPFP) originally described by Dem-

ing and Stephan (1940) and the the maximum likelihood (ML) method - are discussed in Section 2, along with some (co)variance estimators.

IPFP has also been applied in small area estimation to a slightly different situation when the complete table is replaced by some other source of information, for example a complete table from a previous census and the marginal tables 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 to both tables. As an illustration, Purcell and Kish (1980) used the table with 8 cells for the 3 dichotomous variables employment status, county and race. A complete table was available from a previous census and a current survey only provided the marginal tables for employment status and race.

IPFP has the same structure preserving property, as outlined in Section 2, and SPREE can be thought of as a special case of IPFP. Purcell and Kish (1980) have considered six different data situations and only referred to one as IPFP. However it is important to note that all six situations were solved with a method that is identical to IPFP. IPFP is a general purpose method to match marginal information and is not limited to surveys.

In Section 3, mis-specification models are considered, i.e. models for which sample and target population information differ systematically. In Section 4, we conduct a small simulation study to investigate the performance of the methods discussed in this paper under simple random sampling and under the mis-specification models. In particular, we focus on cells with small counts, as these occur frequently in multi-dimensional tables due to the large number of cells. As a case study, estimating the cell counts/proportions within multi-dimensional tables (required for generating

an area-specific synthetic population) for the Illawarra Region in Australia is considered in Section 5. Finally this paper concludes with a discussion.

2 Estimating Population Counts for Contingency Tables

In this Section we review common estimation methods for multi-dimensional tables (cross-calsssified by several variables of interest) based on a sample with known marginal tables, including various (co)variance estimators of these methods.

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. Here, the key objective of applying IPFP is to use the sample and census data for estimating population counts that are cross-classified by two or more characteristics of interest. These estimates can then be used in population reconstruction (Fienberg 1970, Gargiulo et al. 2010, Farooq et al. 2013, Barthelemy and Toint 2013). This application of iterative proportional fitting (IPF) to contingency tables with known margins is called *raking* discussed by Stephan (1942). Raking (also called raking ratio estimation) is a post-stratification procedure which applies a proportional adjustment to the sample weights in a survey so that the adjusted weights add up to known population totals for the post-stratified classifications when only the marginal population totals are known (Deville et al. 1991, Lu and Gelman 2003). While raking is not a maximum likelihood (ML) method under random sampling, yet the raking estimates are consistent and best asymptotically normal (Ireland and

Kullback 1968).

In operations research and econometrics, a tabulation of multi-dimensional variables is formulated by a bi-proportional matrix as presented by Stone (1962). IPFP (also known as the RAS method in economics) is also presented in the literature as an iterative scaling method whereby a multi-dimensional non-negative matrix is adjusted until its marginal sums equal certain values (Bacharach 1965, Schneider and Zenios 1990, Lahr and de Mesnard 2004, Onuki 2013). One alternative method to the IPFP is linear programming. However, as Lee (1993) pointed out, the final solution using this approach would likely contain some zero cell probabilities and the statistical properties of this method are not very well understood.

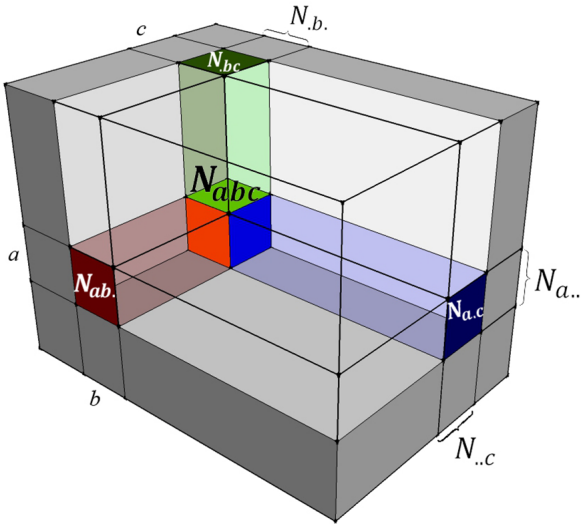
Ireland and Kullback (1968) showed that the estimator produced by the IPFP method minimizes the discrimination information criterion (also known as the Kullback-Leibler divergence, or relative entropy). Moreover Mosteller (1968) pointed out that the procedure preserves the interaction structure of the 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 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 ($\bullet$) 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 as:

$$N_{a\bullet\bullet} = \sum_{b=1}^{B}\sum_{c=1}^{C} N_{abc} = \pi_{a\bullet\bullet}N \ , \quad N_{\bullet b\bullet} = \sum_{a=1}^{A}\sum_{c=1}^{C} N_{abc} = \pi_{\bullet b\bullet}N$$
$$N_{\bullet\bullet c} = \sum_{a=1}^{A}\sum_{b=1}^{B} N_{abc} = \pi_{\bullet\bullet c}N \ . \tag{1}$$

The system of notations used for the cell frequencies and marginal totals used in this paper is presented in Figure 1.

Figure 1: The system of notation for the frequencies and marginal totals in a two-dimensional table



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}. The two-way marginal totals are denoted by $N_{ab\bullet}$, $N_{a\bullet c}$ and $N_{\bullet bc}$, and are given by:

$$N_{ab\bullet} = \sum_{c=1}^{C} N_{abc} = \pi_{ab\bullet} N \ , \ N_{a\bullet c} = \sum_{b=1}^{B} N_{abc} = \pi_{a\bullet c} N \ , \\ N_{\bullet bc} = \sum_{a=1}^{A} N_{abc} = \pi_{\bullet bc} N \ . \tag{2}$$

All joint probabilities π_{abc} and marginal probabilities, such as $\pi_{ab\bullet}$ and $\pi_{a\bullet\bullet}$,

need to sum up to 1.

$$\begin{aligned}
&\sum_{c=1}^{C} = \pi_{\bullet\bullet c} \sum_{b=1}^{B} \pi_{\bullet b\bullet} = \sum_{a=1}^{A} \pi_{a\bullet\bullet} = 1 \ , \\
&\sum_{a=1}^{A}\sum_{b=1}^{B} \pi_{ab\bullet} = \sum_{a=1}^{A}\sum_{c=1}^{C} \pi_{a\bullet c} = \sum_{b=1}^{B}\sum_{c=1}^{C} \pi_{a\bullet c} = 1 \ , \\
&\sum_{a=1}^{A}\sum_{b=1}^{B}\sum_{c=1}^{C} \pi_{abc} = 1 \ .
\end{aligned} \tag{3}$$

When dealing with sample data, sample counts are denoted by n_{abc} with $n = n_{\bullet\bullet\bullet}$ 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 $\pi^{(0)}_{abc} = (ABC)^{-1}$, which corresponds to the case of having no sample data available. When using IPFP for population synthesis, the initial cell probabilities are calculated using a representative survey data referred to as the *seed data*, i.e. $\pi^{(0)}_{abc} = n_{abc}/n$. Let us assume 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} such that the following population constrains hold

$$\pi_{ab\bullet} = \frac{N_{ab\bullet}}{N}, \pi_{a\bullet c} = \frac{N_{a\bullet c}}{N} \text{ and } \pi_{\bullet bc} = \frac{N_{\bullet bc}}{N} \ , \tag{4}$$

other margins could also be available, such as $N_{a\bullet\bullet}$ and $N_{\bullet bc}$ or $N_{a\bullet\bullet}$, $N_{\bullet b\bullet}$ and $N_{\bullet\bullet c}$. Then one iteration of the IPFP consisting of a three-step cycle

has the form

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

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

The algorithm is continued by setting $k := k + 3$ until convergence to the desired accuracy is attained at iteration k. Importantly, the obtained estimates $\hat{\pi}_{abc} = \pi_{abc}^{(k)}$ will satisfy (4). The algorithm will converge to a unique solution provided the seed data contain strictly positive entries and provided the marginal constrains do not contradict each other, for example the constrains $N_{ab\bullet}$ and $N_{a\bullet c}$ need to result in the same $N_{a\bullet\bullet}$, i.e. $N_{a\bullet\bullet} = \sum_b N_{ab\bullet} = \sum_c N_{a\bullet c}$.

Setting positive starting values for cell probabilities ($\pi_{abc}^{(0)} > 0$) ensures that each cell has a non-zero probability, i.e. $\pi_{abc} > 0$ (Gange 1995). If we observe some $n_{abc} = 0$, then we might apply some corrections, for example we can apply $\pi_{abc}^{(0)} = (n_{abc} + 0.5)/n$ to all cells. This is the standard procedure for two-by-two tables (Agresti 2002, p. 71). An alternative proposed by Lang (2004) is to add a tiny constant (e.g. 10^{-6}) to all the cells to ensure that the estimates are strictly positive, i.e. $\pi_{abc} > 0$.

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

estimator for $\hat{\boldsymbol{\pi}}$ is:

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\pi}}) = n^{-1}\mathbf{U}(\mathbf{U}^T\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{D}^{-1}(\mathbf{p})\mathbf{U})(\mathbf{U}^T\mathbf{D}^{-1}(\hat{\boldsymbol{\pi}})\mathbf{U})^{-1}\mathbf{U}^T \ , \tag{5}$$

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

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

$$\hat{N}_{abc} = N\hat{\pi}_{abc}, \tag{6}$$

the (co)variance formula as in (5) to obtain confidence intervals for these population estimates is often not discussed in the literature on SP generation and is worth high-lighting, as this provides an uncertainty measure.

2.2 Log-linear Models

Some studies such as Gonzalez and Hoza (1978) suggested synthetic estimation methods that effectively multiply total population estimates for the targeted small areas by national level estimates of population proportions in each cell in the cross-classification. The main issue with such synthetic estimation methods is to justify the major underlying assumption that such small area cross-classifications are just scaled-down versions of the corresponding national level cross-classification. Purcell and Kish (1980) outlined a generalization of synthetic estimation called structure preserving

estimation (SPREE) which attempts to update the contingency table from a previous census using direct estimators. SPREE, as discussed by Purcell and Kish (1980), used the IPFP (Deming and Stephan 1940) to update the cells in a contingency table based on a primary data source such that they sum to the margins of a secondary data source. Since SPREE uses IPFP, SPREE inherits its structure preserving property from IPFP. IPFP and its structure preserving property referring to parameters of a log-linear model will be discussed next.

A fully-saturated log-linear model (Bishop et al. 1975, Agresti 2002) for a three-way table has the form:

$$\log(\pi_{abc}) = \theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)} + \theta_{123(abc)} \; ; \\ a \in \{1, \ldots, A\}, b \in \{1, \ldots, B\}, c \in \{1, \ldots, C\}, \tag{7}$$

The subscripts of the parameters in model (7) show the variables for which we seek an estimation and levels are shown in brackets. For example, for $\theta_{13(ac)}$, the subscript '13' refers to variables X_1 and X_3 while a is the level of X_1 and c the level of X_3. Here, θ is the intercept, $\theta_{1(a)}$, $\theta_{2(b)}$ and $\theta_{3(c)}$ are the main effects, $\theta_{12(ab)}$, $\theta_{13(ac)}$, and $\theta_{23(bc)}$ are the two-way interaction effects and $\theta_{123(abc)}$ are the three-way interaction effects. The θ parameters are set to satisfy the restrictions:

$$\begin{aligned} &\textstyle\sum_a \theta_{1(a)} = \sum_b \theta_{2(b)} = \sum_c \theta_{3(c)} = 0, \\ &\textstyle\sum_a \theta_{12(ab)} = \sum_a \theta_{13(ac)} = \sum_b \theta_{12(ab)} = \sum_b \theta_{23(bc)} = \sum_c \theta_{13(ac)} = \sum_b \theta_{23(ab)} = 0, \\ &\textstyle\sum_a \theta_{123(abc)} = \sum_b \theta_{123(abc)} = \sum_c \theta_{123(abc)} = 0 \; . \end{aligned} \tag{8}$$

Here $\mu_{abc} = n \times \pi_{abc}$ is the mean for the 'abc'.

Alternatives to the fully-saturated models, as in (7), are log-linear models

with some of the higher order terms removed. For example, a model without three-way interaction terms is:

$$\log(\pi_{abc}) = \theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{12(ab)} + \theta_{13(ac)} + \theta_{23(bc)} \ . \qquad (9)$$

When population counts are available from census data, i.e. N_{abc}, or by a representative sample from the population, i.e. n_{abc}, any log-linear model can be fitted using the IPF algorithm (as discussed by Smith (1947) and Bishop et al. (1975) to estimate the θ parameters via the ML approach.

Any log-linear model, as in (7) and (9), can be expressed as follows:

$$\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}, \qquad (10)$$

where $\boldsymbol{\mu} = E(\mathbf{y}) = n\boldsymbol{\pi}$ and $\mathbf{y} = (y_{111}, \ldots, y_{11C}, \ldots, y_{AB1}, \cdots, y_{ABC})^T$. The design matrix $\mathbf{X}$ is of dimension $(ABC) \times p$, where p is the number of model parameters contained in $\boldsymbol{\beta}$. Then, the ML estimates $\hat{\pi}_{abc}$ are obtained by maximizing the log-likelihood function based on a multinomial distribution (for a random sample without replacement from a population $\mathbf{y}$ is approximately multinomially distributed for $n << N$) given by

$$L = \text{constant} + \sum_{a,b,c} y_{abc} \log \pi_{abc}. \qquad (11)$$

Iterative approaches such as Fisher-Scoring or Newton-Raphson algorithms are discussed in the literature (e.g. Agresti 2002, Singh 2009) for obtaining ML estimates for log-linear models (7) and (9). The estimated asymptotic (co)variance matrix of the estimated $\boldsymbol{\beta}$ coefficients (Agresti 2002, p. 340) is

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \left\{ \mathbf{X}^T \left[\mathbf{D}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T/n \right] \mathbf{X} \right\}^{-1}. \qquad (12)$$

When the aim is to estimate $\boldsymbol{\mu}$ directly (and not $\boldsymbol{\beta}$) and we imply a saturated model, as in (7), the asymptotic estimated (co)variance matrix of $\hat{\boldsymbol{\mu}} = n\boldsymbol{\pi}$ is:

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\mu}}) = \mathbf{D}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T/n. \tag{13}$$

Suppose the two-way (second order) population margins $N_{ab\bullet}$, $N_{a\bullet c}$ and $N_{\bullet bc}$ are known, the raking estimates $\hat{\pi}^r_{abc}$ obtained with IPFP will be of the form of a log-linear model of the form:

$$\log(\hat{\pi}^r_{abc}) = \hat{\theta}^r + \hat{\theta}^r_{1(a)} + \hat{\theta}^r_{2(b)} + \hat{\theta}^r_{3(c)} + \hat{\theta}^r_{12(ab)} + \hat{\theta}^r_{13(ac)} + \hat{\theta}^r_{23(bc)} + \hat{\theta}^r_{123(abc)}. \tag{14}$$

All parameters of up to second order are different from the estimates obtained for (7) without population constrains, except the three-way interaction terms $\hat{\theta}^r_{abc(123)}$, which are preserved by the raking procedure, i.e. $\hat{\theta}^r_{123(abc)}$ are identical to the ML estimates $\hat{\theta}^{ML}_{123(abc)}$ without population constrains. In our case the available margins are $N_{ab\bullet}$ and $N_{\bullet\bullet c}$, and the two-way interaction terms $\hat{\theta}^r_{23(bc)}$ and $\hat{\theta}_{13(ac)}$ in addition to $\hat{\theta}_{123(abc)}$ are also preserved (see Brick et al. (2003) for more discussion). When using the log-linear models for obtaining raking estimates, only the lower order parameters up to the order of the population margins are adjusted, while higher order parameters are preserved. This preservation of higher order parameters lead to the naming of the SPREE method.

In contrast to IPFP, the ML method under random sampling has not been widely discussed in the literature, particularly when dealing with more than two variables. For a three-way contingency table, equation (4) can be expressed as $\mathbf{A}\boldsymbol{\pi} = \mathbf{c}$ (with linearly dependent constrains removed). Let us define the function $\mathbf{h}(\boldsymbol{\mu}) = \mathbf{A}(\boldsymbol{\mu}/n) - \mathbf{c}$. With this definition, $\mathbf{h}(\boldsymbol{\mu}) = \mathbf{0}$ when $\mathbf{A}\boldsymbol{\pi} = \mathbf{c}$. Lang and Agresti (1994) and Lang (1996, 2004, 2005) provide a

model framework to achieve maximising the log-likelihood subject to some arbitrary constrains expressed by $\mathbf{h}(\boldsymbol{\mu}) = \mathbf{0}$ by maximising the constrained likelihood

$$L_c = \text{constant} + \sum_{a,b,c} y_{abc} \log \pi_{abc} + \boldsymbol{\lambda}^T \mathbf{h}(\boldsymbol{\mu}), \tag{15}$$

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, applying the famous method of Lagrange multipliers.

Joseph Lang provides an R function (mph.fit) for maximum likelihood fitting 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 constrains, the method also provides a (co)variance matrix for $\hat{\boldsymbol{\mu}}$ as follows:

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\mu}}) = \mathbf{D}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T/n - \mathbf{D}(\hat{\boldsymbol{\mu}})\mathbf{H}(\mathbf{H}^T\mathbf{D}(\hat{\boldsymbol{\mu}})\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}(\hat{\boldsymbol{\mu}}), \tag{16}$$

where $\mathbf{H}(\boldsymbol{\mu}) = \frac{\partial \mathbf{h}^T(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$. This formula shows an additional term (the last term) compared to formula (13). The additional term reduces the variance imposed by the restrictions or constrains compared to the unconstrained model. Little and Wu (1991) proposed a different formula based on the delta method, similar to (5), given by:

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\pi}}) = n^{-1}\mathbf{U}(\mathbf{U}^T\mathbf{D}(\hat{\boldsymbol{\pi}}^2/\mathbf{p})^{-1}\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{D}(\hat{\boldsymbol{\pi}}^2/\mathbf{p})^{-1}\mathbf{U})(\mathbf{U}^T\mathbf{D}(\hat{\boldsymbol{\pi}}^2/\mathbf{p})^{-1}\mathbf{U})^{-1}\mathbf{U}^T\,. \tag{17}$$

To obtain model-based population counts, formula (6) 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:

$$\widehat{\mathrm{Cov}}(\hat{\mathbf{N}}) = \frac{N^2}{n^2}\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\mu}}) = N^2\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\pi}}) \ . \tag{18}$$

As an interesting feature, the ML estimates $\hat{\pi}^{ML}_{abc}$ (based on second order population constrains) will be of the form (see Appendix A)

$$\left(\frac{\hat{\pi}^{ML}_{abc}}{p_{abc}}\right)^{-1} = \hat{\theta}^{ML} + \hat{\theta}^{ML}_{1(a)} + \hat{\theta}^{ML}_{2(b)} + \hat{\theta}^{ML}_{3(c)} + \hat{\theta}^{ML}_{ab(12)} + \hat{\theta}^{ML}_{ac(13)} + \hat{\theta}^{ML}_{bc(23)} \ . \tag{19}$$

The proposed ML method can also be used to fit standard log-linear modelsby including the model in the constrain function $\mathbf{h}(\boldsymbol{\mu})$ by setting $\mathbf{h}(\boldsymbol{\mu}) = \mathbf{U}^T \log \boldsymbol{\mu} = 0$, where $\mathbf{U}$ is a full column rank orthogonal complement of $\mathbf{X}$, see model (10). Lang (1996, 2004, 2005) extended the methodology to generalized log-linear models and homogeneous linear predictor models.

3 Estimation Methods under Model Miss-specification

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

Before we proceed, let us note that raking estimates $\hat{\pi}^{r}_{abc}$ using all three second order population constrains will be of the following form

$$\log\left(\frac{\hat{\pi}^{r}_{abc}}{p_{abc}}\right) = \hat{\theta} + \hat{\theta}^{r}_{1(a)} + \hat{\theta}^{r}_{2(b)} + \hat{\theta}^{r}_{3(c)} + \hat{\theta}^{r}_{12(ab)} + \hat{\theta}^{r}_{13(ac)} + \hat{\theta}^{r}_{23(bc)}, \tag{20}$$

leading to the previous equation (14) with $\hat{\theta}^{r}_{123(abc)} = \log(p_{abc})$, see Little and Wu (1991) for a two-way contingency table with first order constrains.

Let us now assume that the unknown sample cell probabilities are denoted by τ_{abc} and those of the population by π_{abc}. Also suppose again that second order population margins are provided. Following Little and Wu (1991), we consider the following models relating π_{abc} and τ_{abc}

$$\left(\frac{\pi_{abc}}{\tau_{abc}}\right)^{\kappa} = \theta + \theta_{1(a)} + \theta_{2(b)} + \theta_{3(c)} + \theta_{ab(12)} + \theta_{ac(13)} + \theta_{bc(23)}, \qquad (21)$$

where $\kappa = -1, 1, 2$ and $\kappa \to 0$ refers to the log-function, i.e. $\log(\frac{\pi_{abc}}{\tau_{abc}})$. These four models provide flexible adjustments when sample and target population characteristics do not agree.

Following similar arguments as in Little and Wu (1991), we can show that the ML estimates for the model $\kappa \to 0$ are provided by IPFP , see Appendix B. The ML method for the model with $\kappa = -1$ is identical to the MLRS method introduced in Section 2.3.

Little and Wu (1991) also considered the cases $\kappa = 1, 2$. The ML estimates for $\kappa = 1$ are provided by the least squares method (LSQ) and ML estimates for $\kappa = 2$ are provided by minimum chi-squared method (MCSQ) (Little and Wu 1991).

Little and Wu (1991) compared all four methods in a simulation study while simulating under random sampling and under each of the four misspecification models. The averaged results over a wide range of settings (see Table 1 in Little and Wu 1991) clearly show that under all five models, either IPFP or MLRS are the best. MLRS is best under random sampling and $\lambda = -1, 2$, whereas IPFP is best under $\lambda \to 0$ and $\lambda = 1$. Even though these results are averaged over all simulation and limited to two-by-two tables, they still show that the commonly used IPFP and MLRS methods are generally best. However their results refer to a single two-by-

two table, an unrealistic situation for often highly sparse multi-dimensional tables. The next section considers a simulation study specially designed for multidimensional tables.

4 Simulation Study

Little and Wu (1991) and Causey (1983) conducted empirical simulation studies based on two-by-two tables with constrains referring to the two (marginal) variables. It is not clear how these results extend to tables with many cells and more than two sets of population constrains.

When obtaining a sample for a table with many cells, the sample table is often sparse. The simulation study considers cells with low to relatively large probabilities by setting $A = 5$, $B = 4$ and $C = 2$ and where the $ABC = 40$ probabilities $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_{111}, \ldots, \tilde{\pi}_{11c}, \ldots, \tilde{\pi}_{AB1}, \ldots, \tilde{\pi}_{ABC})^T$ are monotone increasing and the kth probability is $\tilde{\pi}_k \propto \exp(k/40)$, 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 mis-specification models in Section 3 with $\kappa = 0, 1, -1, -2$, where $\kappa = 0$ stands for $\kappa \to 0$ (log-function). For each of those models we sample randomly a table of population counts y_{abc}^{pop} from a multinomial distribution with parameters $\tilde{\boldsymbol{\pi}}$ and N. Now the population proportions are $\pi_{abc} = y_{abc}^{pop}/N$ which we aim to estimate. This means that for small $\tilde{\pi}_{abc}$ we will obtain often $\pi_{abc} = 0$ as the expected counts $\tilde{\pi}_{abc}N$ are small. This is a realistic scenario for multidimensional tables, as some population counts will be indeed be small and often be zero.

Under simple random sampling $n < N$, and under the mis-specification models we also consider $n > N$, as motivated by the example in Section 5. For simplicity, the mis-specification models (21) only include main effects $\theta_{1(a)}, \theta_{1(b)}, \theta_{1(c)}$, which were all generated under $N(0, 0.2^2)$ for each simu-

lated data set. This ensures that sample and population are systematically different. According to the mis-specification models, see (21), the τ_{abc} are obtained from the π_{abc}.

We investigate the performance of the estimators IPFP, MLRS (abbreviated here ML), LSQ and MCSQ and their (co)variance estimators by calculating the emprircal relative bias $E(\hat{\pi}_{abc} - \pi_{abc})/\pi_{abc}$, the relative mean squared error (MSE) relative to IPFP, the coverage and the length of the confidence intervals, all calculated over $10,000$ simulated data sets. The MSE for a cell is defined as $E(\hat{\pi}_{abc} - \pi_{abc})^2$. The relative MSE is defined as the MSE of a particular method divided by the MSE of IPFP. IPFP serves as a benchmark and its relative MSE is set to 1.000.

The results of LSQ and MCSQ are omitted, as their performance is dominated by either IPFP or ML, which is similar to Little and Wu (1991)'s findings. We also consider adjusted versions called ML+1 and IPFP+1, where a one was added to all sample cell counts. This was done in the anticipation of obtaining better estimates for cells with low probabilities. For two-by-two tables it is common to add 0.5 to each cell (Agresti 2002). Here we add a one to each cell, motivated by the Bayesian approach. For the binomial distribution assuming the uniform distribution as a non-informative prior for the unknown success probability, the posterior mean estimate is $(y + 1)/(n + 2)$ when y are the number of successes and n the number of trials. In contrast, the well-known ML estimator is y/n. It is known that the Bayesian estimator $(y + 1)/(n + 2)$ performs better compared to y/n when using a squared error loss function (Agresti 2002). Similar results hold for the multinomial distribution and led to considering ML+1 and IPFP+1.

For the confidence intervals (CI) and the ML method we consider the delta method (D) and Lang's formula (L), see formulae (17) and (16). Table

4 shows the coverage and length of 95% confidence intervals and Table 4 shows the relative MSE and the relative bias for the π_{abc}. The tables show n, N, the model (either RND or $\kappa = 0, 1, -1, 2$), and an indicator for whether the cells to be estimated are small (S), medium (M) and large (L), followed by the expected cell count in the sample, i.e. $E(n_{abc}) = n \times \pi_{abc}$.

In terms of coverage, the method ML+1-L performs generally best across scenarios S and M and all models. Only for scenario L and $\kappa = 0, 1$, IPFP or IPFP+1 are better. The last column of Table 4 shows which method is best in terms of relative MSE and bias. In terms of relative MSE, ML+1 and IPFP+1 are best, whereas in terms of bias ML and IPFP are best. Across scenarios IPFP+1 seems to perform slightly better than IPFP+1 in terms of relative MSE. In terms of bias, ML seems to perform slightly better than IPFP across all scenarios.

The results for RND and $n = 100$, $N = 600$ are graphically shown in Figure 3 for the methods ML and ML+1. They show the methods are virtually idential for larger cells but for smaller cells, in particular cells 1-8, the method ML+1 provides more efficient estimates.

Scenario	ML-L	ML-D	ML+1-L	ML+1-D	IPFP-D	IPFP+1-D
$N = 600$, $n = 100$						
RND - M - 0.28	23.0 (0.60)	23.4 (0.83)	81.0 (1.25)	80.4 (0.92)	22.4 (0.55)	79.9 (0.89)
RND - M - 0.52	38.1 (1.03)	38.5 (1.24)	95.1 (1.42)	91.7 (1.06)	37.3 (0.97)	91.8 (1.10)
RND - S - 0.10	7.16 (0.15)	7.42 (0.43)	45.5 (0.70)	42.1 (0.38)	6.04 (0.086)	41.0 (0.36)
RND -L - 10.5	95.0 (8.77)	95.5 (8.95)	95.5 (7.59)	97.1 (8.36)	95.2 (8.97)	97.1 (8.42)
$N = 10,000$, $n = 600$						
RND - M - 1.68	80.6 (0.61)	80.6 (0.62)	99.5 (0.70)	98.4 (0.63)	80.6 (0.62)	98.2 (0.63)
RND - M - 3.55	89.6 (1.02)	90.1 (1.03)	95.4 (1.02)	94.0 (0.93)	90.1 (1.03)	94.0 (0.94)
RND - S - 0.62	44.7 (0.25)	45.3 (0.25)	100.0 (0.38)	99.9 (0.30)	45.3 (0.25)	99.8 (0.30)
RND -L - 63.0	95.2 (3.69)	95.2 (3.70)	95.3 (3.58)	95.6 (3.65)	95.2 (3.71)	95.6 (3.65)
$N = 800$, $n = 1,000$						
$\kappa = 0$ - M - 2.80	73.9 (0.48)	74.2 (0.49)	86.3 (0.55)	85.5 (0.51)	74.4 (0.49)	85.5 (0.52)
$\kappa = 0$ - M - 5.92	88.0 (0.79)	88.8 (0.81)	93.7 (0.80)	92.9 (0.76)	89.0 (0.81)	93.2 (0.77)
$\kappa = 0$ - S - 1.03	41.5 (0.20)	41.8 (0.20)	54.7 (0.30)	53.0 (0.26)	41.8 (0.20)	52.5 (0.26)
$\kappa = 0$ - L - 105	92.6 (2.85)	93.1 (2.91)	92.7 (2.80)	93.5 (2.88)	95.0 (2.95)	95.2 (2.92)
$N = 800$, $n = 1,000$						
$\kappa = 1$ - M -2.80	72.3 (0.47)	72.3 (0.49)	85.3 (0.55)	84.2 (0.52)	73.4 (0.49)	84.7 (0.52)
$\kappa = 1$ - M - 5.92	85.5 (0.78)	86.3 (0.81)	92.1 (0.79)	91.1 (0.76)	87.9 (0.82)	92.1 (0.77)
$\kappa = 1$ - S - 1.03	40.4 (0.20)	40.5 (0.20)	54.4 (0.30)	52.5 (0.26)	40.9 (0.20)	52.0 (0.26)
$\kappa = 1$ - L - 105	83.8 (2.84)	84.8 (2.92)	84.0 (2.79)	85.1 (2.89)	91.2 (2.98)	91.4 (2.95)
$N = 800$, $n = 1,000$						
$\kappa = -1$ - M -2.80	73.7 (0.48)	73.9 (0.49)	86.2 (0.55)	85.6 (0.52)	73.7 (0.49)	85.0 (0.52)
$\kappa = -1$ - M - 5.92	88.0 (0.79)	89.1 (0.81)	94.1 (0.80)	93.4 (0.76)	88.6 (0.82)	92.6 (0.77)
$\kappa = -1$ - S - 1.03	41.3 (0.20)	41.5 (0.20)	54.3 (0.30)	52.5 (0.26)	41.5 (0.20)	51.8 (0.26)
$\kappa = -1$ - L - 105	94.1 (2.85)	94.7 (2.92)	94.1 (2.80)	94.9 (2.89)	91.8 (2.97)	92.3 (2.93)
$N = 800$, $n = 1,000$						
$\kappa = -2$ - M - 2.80	75.4 (0.48)	75.3 (0.49)	86.5 (0.55)	85.6 (0.51)	75.2 (0.49)	85.5 (0.51)
$\kappa = -2$ - M - 5.92	88.8 (0.79)	89.3 (0.80)	94.5 (0.79)	93.5 (0.75)	89.2 (0.80)	93.3 (0.76)
$\kappa = -2$ - S - 1.03	42.6 (0.20)	42.7 (0.20)	54.6 (0.30)	53.0 (0.25)	42.7 (0.20)	52.7 (0.25)
$\kappa = -2$ - L - 105	94.3 (2.85)	94.6 (2.87)	94.3 (2.80)	94.7 (2.85)	93.9 (2.89)	94.1 (2.86)

Table 1: Coverage and in brackets average length both in percentages for the ML methods and the IPFP using the Delta method (D) and Lang's standard errors (L); "+1" indicates adding ones to all cell counts

Scenario	ML	ML+1	IPFP	IPFP+1	best method mse/bias
$N = 600$, $n = 100$					
RND - M - 0.28	1.016 (0.220)	0.161 (27.70)	1.000 (-3.188)	0.159 (22.01)	IPFP+1/ML
RND - M - 0.52	1.012 (0.037)	0.157 (-15.21)	1.000 (-2.410)	0.159 (-12.31)	ML+1/ML
RND - S - 0.10	1.513 (14.35)	0.243 (35.90)	1.000 (-25.73)	0.240 (28.10)	IPFP+1/ML
RND -L - 10.5	0.966 (0.024)	0.726 (2.810)	1.000 (0.371)	0.743 (2.323)	ML+1/ML
$N = 10,000$, $n = 600$					
RND - M - 1.68	1.000 (0.156)	0.491 (11.52)	1.000 (0.173)	0.490 (11.02)	IPFP+1/ML
RND - M - 3.55	0.998 (-0.426)	0.592 (-3.571)	1.000 (-0.427)	0.594 (-3.420)	ML+1/ML
RND - S -0.62	1.000 (-0.122)	0.2163 (22.32)	1.000 (-0.088)	0.217 (21.19)	ML+1/ML
RND -L - 63.0	0.996 (-0.011)	0.943 (0.812)	1.000 (-0.011)	0.947 (0.774)	ML+1/ML
$N = 800$, $n = 1,000$					
$\kappa = 0$ - M - 2.80	1.020 (0.176)	0.659 (8.158)	1.000 (0.265)	0.666 (8.094)	ML+1/ML
$\kappa = 0$ - M - 5.92	1.035 (0.278)	0.746 (-2.124)	1.000 (0.268)	0.736 (-2.117)	IPFP+1/IPFP
$\kappa = 0$ - S -1.03	1.024 (-1.112)	0.563 (16.35)	1.000 (-1.190)	0.577 (15.98)	ML+1/ML
$\kappa = 0$ - L - 105	1.123 (-0.031)	1.082 (0.527)	1.000 (-0.017)	0.967 (0.541)	IPFP+1/IPFP
$N = 800$, $n = 1,000$					
$\kappa = 1$ -M - 2.80	1.127 (-0.388)	0.703 (7.668)	1.000 (-0.355)	0.668 (7.666)	IPFP+1/IPFP
$\kappa = 1$ - M - 5.92	1.155 (0.036)	0.812 (-2.311)	1.000 (0.036)	0.731 (-2.337)	IPFP+1/IPFP
$\kappa = 1$ - S -1.03	1.070 (-0.669)	0.547 (16.90)	1.000 (-0.757)	0.557 (16.65)	ML+1/ML
$\kappa = 1$ - L - 105	1.707 (0.054)	1.651 (0.618)	1.000 (-0.033)	0.990 (0.553)	IPFP+1/IPFP
$N = 800$, $n = 1,000$					
$\kappa = -1$ - M - 2.80	0.993 (0.179)	0.634 (8.244)	1.000 (0.286)	0.6644 (8.250)	ML+1/ML
$\kappa = -1$ - M - 5.92	0.961 (0.279)	0.682 (-2.148)	1.000 (0.164)	0.726 (-2.189)	ML+1/IPFP
$\kappa = -1$ - S -1.03	1.029 (0.620)	0.529 (17.31)	1.000 (0.466)	0.556 (17.02)	ML+1/IPFP
$\kappa = -1$ - L - 105	0.718 (0.063)	0.692 (0.623)	1.000 (0.107)	0.956 (0.665)	ML+1/ML
$N = 800$, $n = 1,000$					
$\kappa = -2$ - M - 2.80	0.989 (-0.253)	0.677 (7.504)	1.000 (-0.188)	0.685 (7.324)	ML+1/IPFP
$\kappa = -2$ - M - 5.92	0.991 (0.332)	0.733 (-1.971)	1.000 (0.301)	0.743 (-1.923	ML+1/IPFP
$\kappa = -2$ - S -1.03	1.003 (-0.379)	0.580 (16.80)	1.000 (-0.414)	0.588 (16.29)	ML+1/ML
$\kappa = -2$ - L - 105	0.930 (-0.017)	0.902 (0.533)	1.000 (-0.012)	0.971 (0.523)	ML+1/IPFP

Table 2: Relative MSE and in brackets relative bias (relative to true parameters) for the ML methods and the IPFP methods

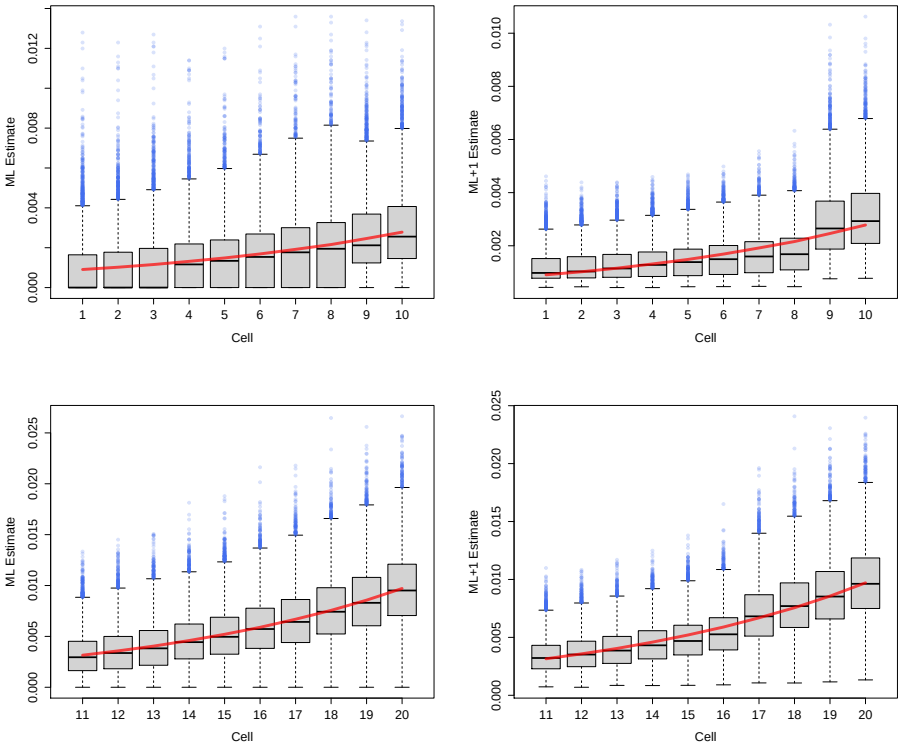


Figure 2: Boxplots of ML (left) and ML+1 (right) estimates for the first 20 (small to medium) out of 40 cells comparing with average population proportions (red).

5 Estimating Multi-Dimensional Population Counts for the Illawarra Region

The study area in this paper is the Illawarra region in New South Wales, a state in Australia, with the total population of 365,388 individuals in 2011. Illawarra is the coastal region situated immediately south of Sydney and north of the Shoalhaven or South Coast region (see Figure 2). The smallest geographic area defined in the Australian Statistical Geography Standard (ASGS) is the Statistical Level 1 (SA1) for which the data are available to our study. Numbers of males and females living within the study area are

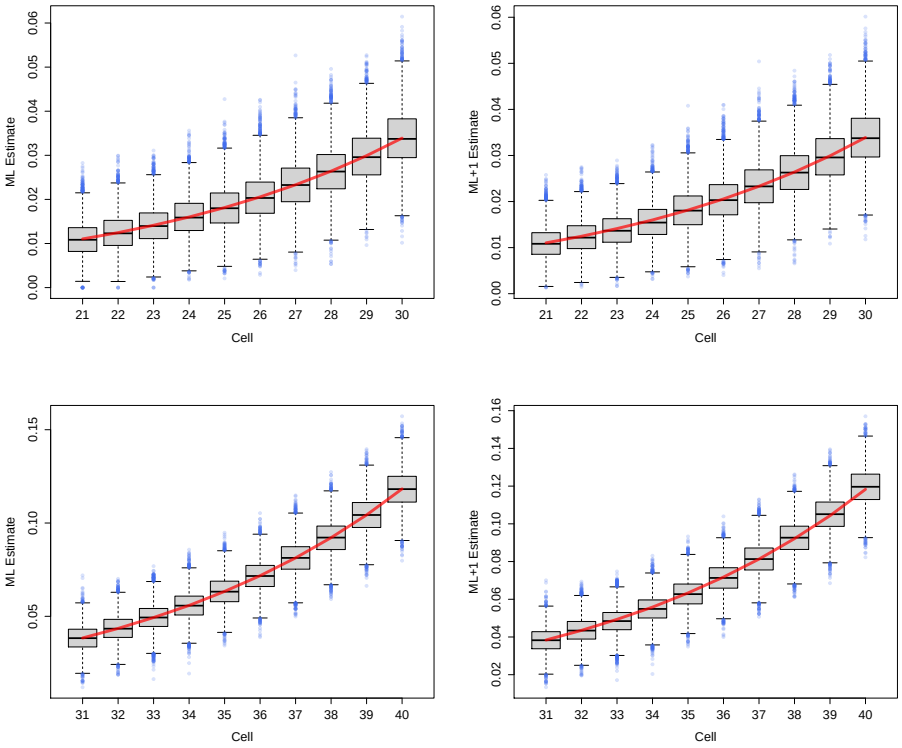


Figure 3: Boxplots of ML (left) and ML+1 (right) estimates for the last 20 out of 40 (medium to large) cells comparing with average population proportions (red).

presented in Table 1.

Table 3: Living population in the study area

Area	Males	Females	Total
Kiama-Shellharbour	40,160	42,184	82,344
Wollongong	59,982	60,968	120,950
Dapto-Port Kembla	35,004	36,111	71,115
Shoalhaven	44,667	46,262	90,929
Total	179,813	185,252	365,338

The sets of Australian census tables released by the Australian Bureau of Statistics (ABS) are available to this study by including individual-related

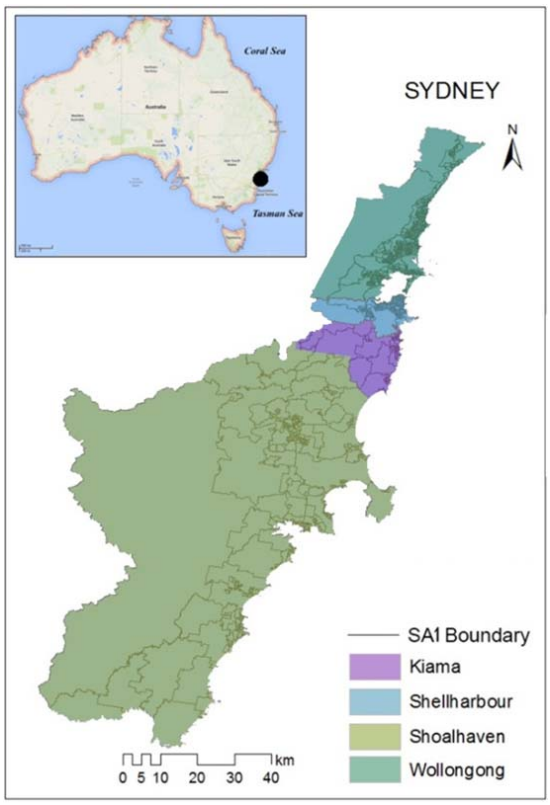


Figure 4: Map of study area (Illawarra Region)

tables (e.g. distribution of age by gender, and relationship in household by age by gender) as well as household related tables (e.g. age by sex tables; family composition tables; and family composition by gender and age). There are 18 age categories (0-4,5-9,. . . ,80-84,>84), two genders and 4 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.

A 1% Basic Census Sample File (CSF) is available to this study through the Confidentialised Unit Record File (CURF) microdata system. The data are used as a pseudo sample for the tree-way table of socio-characteristics (e.g. age, sex and family type) required for the estimation and for simulating

area-specific synthetic populations. SA1-specific marginal population counts for age by sex and for family type are also available from the census. The `R` package `mipfp` is used to generate the raking (IPFP) and the ML estimates (Barthelemy and Suesse 2014).

Figure 5 shows the results when using only the CSF without imposing population constrains. The results do not vary across SA1's (these geographical areas also called strata in this context), as we have only one sample - the CSF - available across the whole Illawarra which is used as a sample for each SA1, as we do not have any geographical information (as SA1) attached to the CSF. This approach might seem questionable, as sample and target populations do not agree, however as we noted in Section 3, IPFP and ML provide also ML estimates under the mis-specification models with $\kappa = 0, -1$ and either method is best under each of the five models (RND, $\kappa = 0, 1, -1, -2$) considered in Section 3.

Figure 6 shows the results of the ML+1 method and Figure 7 shows the results for IPFP+1. The results differ, as is seen in Figure 8. The ML method seems to yield smoother results than the IPFP, smoother in the sense that probabilities do not seem to vary as much as for IPFP.

Both methods exactly match the population constrains. As an example for SA1 with ID1114961 the population margins 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), expressed as proportions relative to SA1 size.

Results for IPFP and ML are not shown to preserve space, but the results of those are similar to ML+1 and IPFP+1 except that small (and in particular zero) cell counts have estimates that are further away from zero.

These age by sex by family type tables based on a pseudo sample and

two know 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 SR, as they form the basis for the simulation of area-specific synthetic populations.

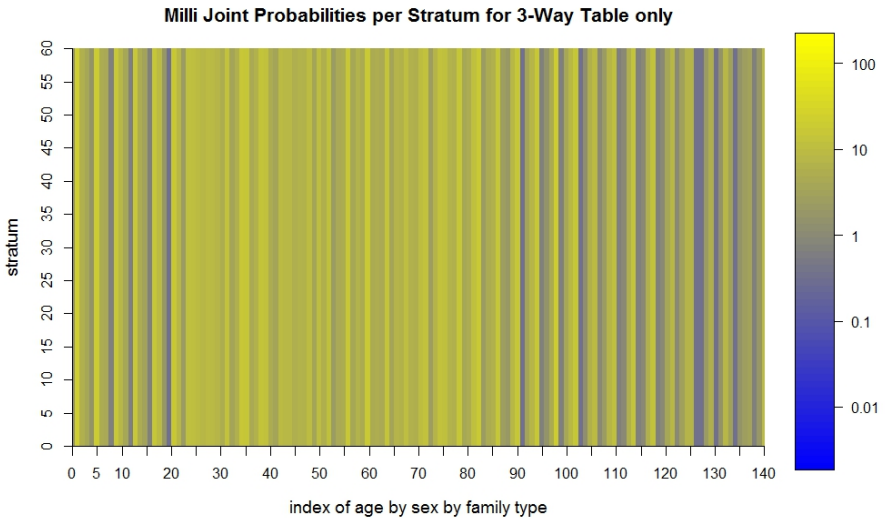


Figure 5: Estimated π_{abc} per stratum based on 1%CSF file without marginal population constrains

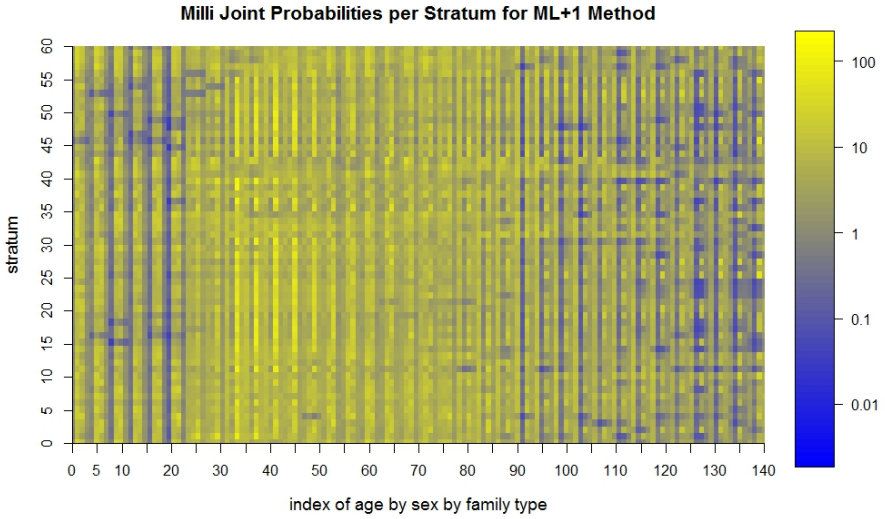


Figure 6: Estimated π_{abc} per stratum (area) using ML+1 method based on 1% CSF file and known marginal population constrains

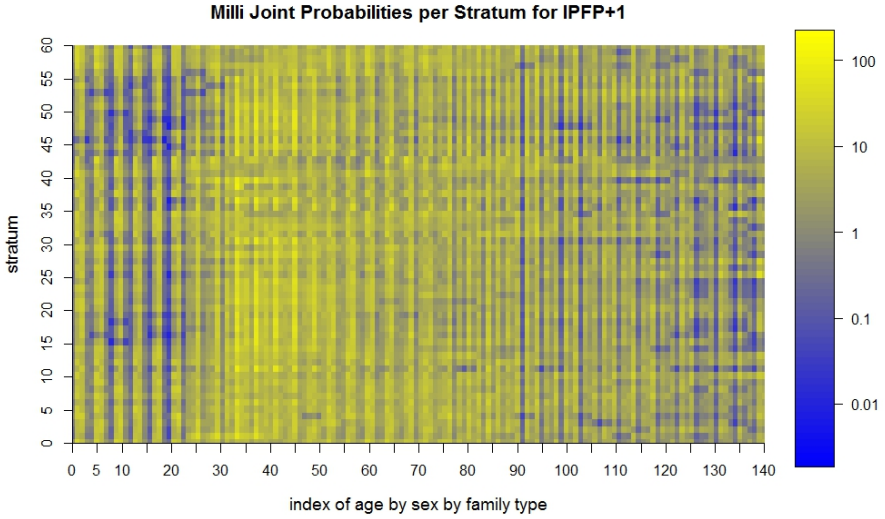


Figure 7: Estimated π_{abc} per stratum using IPFP+1 based on 1% CSF file and known marginal population constrains

6 Discussion

In this paper, our focus is on two methods (IPFP and ML) to obtain population count estimates $\hat{N}_{abc} = N\hat{\pi}_{abc}$ or equivalently to obtain estimates of

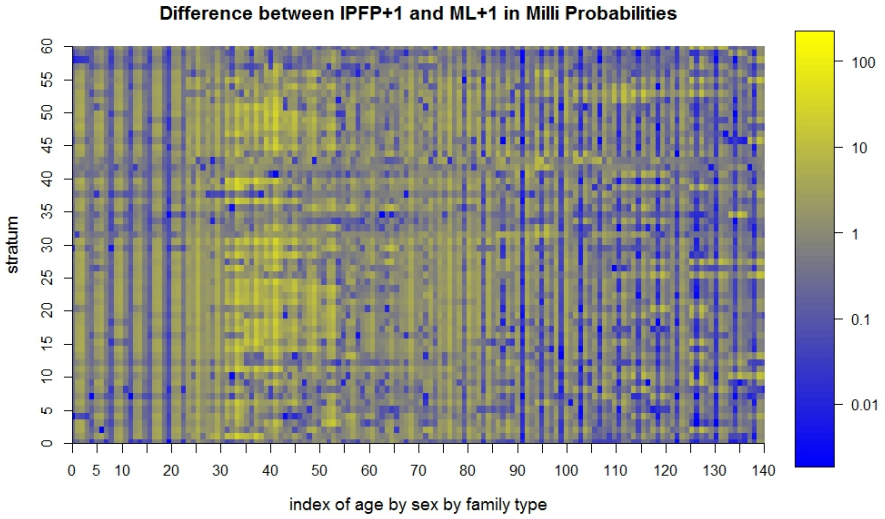


Figure 8: Absolute differences between ML+1 method and IPFP+1

the joint probabilities $\hat{\pi}_{abc}$ when a sample is available and when marginal population counts (sub-tables) are known. IPFP, also known as raking, is the standard method for this problem (Ballas et al. 2005, Smith et al. 2009), supposedly mainly due to the popularity of IPFP and widely available software. IPFP can also be applied if the seed (sample) is partially observed, e.g. due to confidentiality restriction, for example when a marginal table of the sample is available. In this survey context IPFP is often known as SPREE.

The ML method is relatively unknown supposedly due to various factors: only applicable when a sample is available, software unavailability and unfamiliarity of the underlying constrained maximum likelihood approach in applied sciences. IPFP requires non-zero cells of the sample to converge and to provide a unique solution of the underlying optimization problem. Even with zero cell counts, IPFP often still converges and provides estimates. The ML method has convergence problems when zero cell counts are

present. As an alternative we presented the methods ML+1 and IPFP+1 (adding ones to all cells) and the simulation study showed that these methods work well for small expected cell counts. In a simulation study, the Method ML+1 in combination with Lang's covariance estimator (ML+1-L) performed best in terms of coverage and average length under all models (including mis-specification models for which sample and population can differ systematically), when expected cells were of small to medium size. ML+1-L can be generally recommended for multi-dimensional contingency tables, as those will always contain many small to medium 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.

To make all these methods freely available we provided an R package `mipfp` (Barthelemy and Suesse 2014) that provides the IPFP, ML, MCSQ and LSQ methods. The ML method is important as it serves as an important alternative to IPFP. For issues with IPFP, see Brick et al. (2003). In particular when true random sampling can be assumed, ML (and ML+1) are worthy alternatives to IPFP (and IPFP). The methods ML+1 and IPFP+1 are not directly implemented, as it only requires adding 1 to each cell of the sample before the IPFP and ML methods are implied.

Generally speaking when efficiency is the main aim, then ML+1 and IPFP+1 are generally recommended with a slight preference for IPFP+1. When obtaining unbiased estimates is the main aim, then the methods ML and IPFP are recommended, with a slight preference for ML over IPFP.

In this paper, our main focus was on three-way contingency tables being driven mainly by our data example, extending the two-way table situation considered by Little and Wu (1991). The methods can be easily extended to

more than 3 variables in a straightforward manner and similar results can be expected.

From some of the results and models, it is apparent that the more population information is available, the more accurate will the estimators be under mis-specification. For example when two-way population margins are given then the mis-specification model (21) contains two-way interaction (2nd order) parameters. If only one-way population margins are provided, then the model will only contain main effects (1st order parameters), meaning that final estimates would be less reliable compared to the situation when two-way population margins are provided, as two-way interaction terms provide higher flexibility compared to only having main effects.

The main message is that the more population information in form of marginal totals are available the better are the final estimates and the better are the estimates against any form of mis-specification. This is not surprising and expected, but still worth highlighting, as in the general context statistical accuracy is often only increased by increasing sample size or reducing non-response. Here variability can be reduced by a combination of factors: increasing sample size n but also by providing more marginal population counts, as the latter makes estimation more robust against mis-specification.

Bibliography

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

Arentze, T., Timmermans, H.J.P., and Hofman, F. (2007). Creating Synthetic Household Populations: Problems and Approach. *Journal of the Transportation Research Board* 2014: 85-91.

Bacharach, M. (1965). Estimating Nonnegative Matrices from Marginal Data. *International Economic Review* 6(3): 294-310.

Barthélemy, J., and Toint, P.L. (2013). Synthetic Population Generation without a Sample. *Transportation Science* 47(2)': 266-279.

Barthélemy, J., and Suesse, T. (2014). *mipfp: Multidimensional Iterative Proportional Fitting*. R package version 1.0. `http://CRAN.R-project.org/package=mipfp`

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

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

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariable Analysis: Theory and Practice*. Cambridge: MIT Press.

Bricks, A. (2003). Identifying Problems with Raking Estimators. In *Proceedings of the Joint Statistical Meeting*, Section Survey Research Methods. San Francisco: 710-717.

Causey, B.D. (1983). Estimation of Proportions for Multinomial Contingency Tables Subject to Known Marginal Constraints. *Communications in Statistics-Theory and Methods* 12: 2581-2587.

Deming, W.E., and Stephan, F.F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Annals of Mathematical Statistics* 11(4): 367-484.

Deville, J., Sarndal, C., and Sautory, O. (1991). Raking Procedures in Survey Sampling. *Journal of the American Statistical Association* 86(413): 87-95.

Farooq, B., Bierlaire, M., Hurtubia, R., and Flettered, G. (2013). *Simulation Based Population Synthesis Transportation Research Part B: Methodological* 58: 243-263.

Fienberg, S.E. (1970). An Iterative Procedure for Estimation in Contingency Tables. *Annals of Mathematical Statistics* 41(3): 907-917.

Gange, S.J. (1995). Generating Multivariate Categorical Variates Using the Iterative Proportional Fitting Algorithm. *The American Statistician* 49(2): 134-138.

Geard, N., McCaw, J.M., Dorin, A., Korb, K.B., and McVernon, J. (2013). Synthetic Population Dynamics: A Model of Household Demography. *Journal of Artificial Societies and Social Simulation* 16(1): 8.

Gargiulo, F., Ternes, S., Huet, S., and Deffuant, G. (2010). An Iterative Approach for Generating Statistically Realistic Populations of Households. *PLOS ONE* 5(1): e8828.

Gonzalez, M. E., and Hoza, C. (1978). Small-area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association* 73(361): 7-15.

Harland, K., Heppenstall, A., Smith, D., and Birkin, M. (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.

Ireland, C.T., and Kullback, S. (1968). Contingency Tables with Given Marginals. *Biometrika* 55(1): 179-199.

Lahr, M.L., and de Mesnard, L. (2004). Biproportional Techniques in Input-Output Analysis: Table Updating and Structural Analysis. *Economic Systems Research* 16(2): 115-134.

Lang, J.B. (1996). Maximum Likelihood Methods for a Generalized Class of Log-Linear Models. *Annals of Statistics* 24(2): 726-752.

Lang, J.B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *Annals of Statistics* 32(1): 340-383.

Lang, J.B. (2005). Homogeneous Linear Predictor Models for Contingency Tables. *Journal of the American Statistical Association* 100(469): 121-134.

Lang, J.B., and Agresti, A. (1994). Simultaneously Modelling Joint and Marginal Distributions of Multivariate Categorical Responses. *Journal of the American Statistical Association* 89(426): 625-632.

Lee, A.J. (1993). Generating Random Binary Deviates Having Fixed Marginal Distributions and Specified Degrees of Association. *The American Statistician* 47(3): 209-215.

Little, J.A., and Wu, M.M. (1991). Models for Contingency Tables with Known Margins when Target and Sampled Population Differ. *Journal of the American Statistical Association* 86(413): 87-95.

Lu, H., and Gelman, A. (2003). A Method for Estimating Design-based Sampling Variances for Surveys with Weighting, Poststratification, and Raking. *Journal of Official Statistics* 19(2): 133-151.

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association* 63: 1-28.

Namazi-Rad, M., Mokhtarian, P., and Perez, P. (2014). Generating a Dynamic Synthetic Population - Using an Age-Structured Two-Sex Model for Household Dynamics. *PLOS ONE* 9(4): e94761.

Onuki, Y. (2013). Extension of the Iterative Proportional Fitting Preocedure and Its Evaluation Using Agent-Based Models. In Murata T, Terano, T, Shingo, S (eds.) *Agent-Based Approaches in Economic and Social Complex Systems VII: Post-proceedings of The AESCS International Workshop 2012*. Springer.

Purcell, N.J., and Kish, L. (1980). Postcensal Estimates for Local Areas (or Domains). *International Statistical Review* 48, 3-18.

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

Smith, J.H. (1947). Estimation of Linear Functions of Cell Proportions. *Annals of Mathematical Statistics* 18: 231-254.

Schneider, M.H., and Zenios, S.A. (1990). A Comparative Study of Algorithms for Matrix Balancing. *Operations Research* 38(3): 439-455.

Singh, A.C. (2009). Categorical Data Analysis for Simple and Complex Surveys. In Pfeffermann, D., and Rao, C.R. (eds.) *Sample Surveys: Design, Methods and Applications, Handbook of Statistics 29* Vol. 29B, 329-370. Noth-Holland: Elsevier.

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

Stone, R. (1942). Multiple Classifications in Social Accounting. *Bulletin de l'Institut International de Statistique* 39: 215-233.

Zhang, L.C., and Chambers, R.L. (2004). Small Area Estimates for Cross-Classifications. *Journal of the Royal Statistical Society: Series B* 66(2): 479-496.

Appendix

A Form of Maximum Likelihood Estimates

Let us write the constrained log-likelihood L_c, see (15), with second order population constrains 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 constrains such as second order parameters sum to zero, estimates have the form

$$\left(\frac{\hat{\pi}^{ML}_{abc}}{p_{abc}}\right)^{-1} = \hat{\theta}^{ML} + \hat{\theta}^{ML}_{1(a)} + \hat{\theta}^{ML}_{2(b)} + \hat{\theta}^{ML}_{3(c)} + \hat{\theta}^{ML}_{ab(12)} + \hat{\theta}^{ML}_{ac(13)} + \hat{\theta}^{ML}_{bc(23)}. \quad (22)$$

It also shows that if second order population constrains are included, then the form of estimates include second order terms. If e.g. first order population constrains are included, then the right hand side of (22) will only contain main effects (first order terms).

B Showing that IPFP Estimates are ML Estimates under Model (21) with $\lambda \to 0$

Assume sampling fractions are small, then n_{abc} are multinomially distributed and the sample proportions p_{abc} are ML estimates of τ_{abc}. By model (21) 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

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

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

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

Now we note that these equations are solved by the raking estimates, see equation (20) in combination with (4).

Similar arguments can be shown to show that MLRS provides ML estimates for model (21) with $\kappa = -1$, LSQ provides ML estimates for model (21) with $\kappa = 1$ and MCSQ provides ML estimates for model (21) with $\kappa = 2$.