UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND
MATHEMATICAL SCIENCES

Social Statistics and Demography

Multivariate Structure Preserving Estimation for Population Compositions

by

Ángela Luna Hernández

Thesis for the Degree of Doctor of Philosophy

October 2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Social Statistics and Demography

Thesis for the degree of Doctor of Philosophy

MULTIVARIATE STRUCTURE PRESERVING ESTIMATION FOR POPULATION COMPOSITIONS

by Ángela Luna Hernández

This document introduces a new Structure Preserving Estimator for Small Area compositions, using data from a proxy and a sample compositions. The proposed estimator, the Multivariate Structure Preserving Estimator (MSPREE), extends the two main SPREE-type estimators: the SPREE and the GSPREE. The additional flexibility of the MSPREE may lead to estimates with less MSE than its predecessors. An extension of the MSPREE including cell specific random effects, the Mixed MSPREE (MMSPREE), is also presented, in an attempt to further reduce the size of the bias when the associated sample size allows for it. In order to estimate the variance components governing the variance structure of the random effects in the MMSPREE, an unbiased moment-type estimator is proposed. Furthermore, an estimator for the variance of the MSPREE, as well as methodologies to evaluate the unconditional and finite population MSE of both MSPREE and MMSPREE, are developed. The behaviour of the proposed estimators is illustrated in a controlled setting via a simulation exercise, and in a real data application.

Contents

List of Tables

List of Figures

Declaration Of Authorship

I, Ángela Luna Hernández, declare that the thesis entitled Multivariate Structure Preserving Estimation for Population Compositions and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Signed: ....

Date ....

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Nikos Tzavidis and Dr. Li-Chun Zhang, for the generosity with which they shared their time and knowledge with me, for their support and the infinite patience they have shown me through this difficult but awesome journey. Their curiosity, openness and hard work have been very encouraging and inspiring.

I am also indebted to my family and friends who have been unconditionally supportive of my decisions and have blessed me with their affection and understanding. I am deeply grateful to my mom, my brother and my nephew, for enduring my absence and always finding a way to bring happiness and hope to my life. To Olivier, who despite having experienced first hand the difficulties of completing a PhD, stayed by my side until the end. Thanks for your understanding, love and dedication.

All my gratitude and affection to Kristine, the wonderful woman I have learnt from and grown with for the past 4 years and whose friendship has been my strength in my weakest hours. To Dilek and Sarah - thanks for your care, affection and consistent support. To all my colleagues and friends from Building 58: thanks for all the great moments together; for sharing your energy, ideas and uncertainties.

I am grateful to the Economic and Social Research Council and the Office of National Statistics for providing the funding that allowed this project to be completed. Finally, I would also like to express my gratitude to Amos, Pauline and Glenn on behalf of the Southampton Doctoral Training Centre, for the invaluable logistic and financial support you gave me during the PhD.

List of Abbreviations

APS	Annual Population Survey
BLP	Best Linear Predictor
BLUE	Best Linear Unbiased Estimator
BLUP	Best Linear Unbiased Predictor
BP	Best Predictor
DEFT	Design factor
EBLUP	Empirical Best Linear Unbiased Predictor
GLMM	Generalized Linear Mixed Model
GLSM	Generalized Linear Structural Model
GLSMM	Generalized Linear Structural Mixed Model
GSPREE	Generalized SPREE-type estimator of Zhang and Chambers (2004), or its corresponding estimate
IPF	Iterative Proportional Fitting
IWLS	Iterative Weighted Least Squares
LA	Local Authority
LFS	Labour Force Survey
ML	Maximum Likelihood
MSE	Mean Square Error
MSPREE	Multivariate SPREE-type estimator defined in chapter 2, or its corresponding estimate
MMSPREE	Multivariate Mixed SPREE-type estimator defined in chapter 3, or its corresponding estimate
MSPREE(J)	Special case of the MSPREE with only J free parameters
NSI	National Statistical Institute
ONS	Office for National Statistics
PQL	Partial Quasi-Likelihood
REML	Restricted (residual) Maximum Likelihood
SA	Small Area
SAE	Small Area Estimation, or Small Area Estimate

SPREE	Structure Preserving, or, SPREE-type estimator of Purcell and Kish (1980), or its corresponding estimate
VML	Virtual Microdata Laboratory

Nomenclature

$\mathbf{Y}$	Composition of interest, with components Y_{aj} for $a = 1\ldots, A$ and $j = 1, \ldots, J$
$\mathbf{Y}_{a+}$	Column vector containing the row margins of $\mathbf{Y}$
$\mathbf{Y}_{+j}$	Row vector containing the column margins of $\mathbf{Y}$
θ^{Y}	Composition of interest (within-area proportions)
$\mathbf{X}$	Proxy composition
θ^{X}	Proxy composition (within-area proportions)
$\alpha_0^Y, \alpha_a^Y, \alpha_j^Y,$ α_{aj}^Y	Global, row, column and interaction terms in the representation of $\mathbf{Y}$ as a saturated log-linear model, analogously for $\mathbf{X}$
$\boldsymbol{\alpha}_a^Y$	Column vector containing the interaction terms of $\mathbf{Y}$ for area a, analogously for $\mathbf{X}$
$\mathbf{I}_{(M)}$	Identity matrix of dimension $M \times M$ (the dimension can be missing if it is obvious from the context)
$\mathbf{1}_{(M \times K)}$	Matrix of dimension $M \times K$ with all elements equal to 1; the case $K = M$ will be denoted by $\mathbf{1}_{(M)}$
$\mathfrak{y}$	Sample composition
$\hat{\mathbf{Y}}$	Composition of direct estimates of $\mathbf{Y}$
$\hat{\mathbf{Y}}^{S}$	SPREE of $\mathbf{Y}$
$\hat{\mathbf{Y}}^{G}$	GSPREE of $\mathbf{Y}$
$\hat{\theta}^{Y,G}$	GSPREE of θ^{Y}
β	Parameter of the GSPREE
$\hat{\mathbf{Y}}^{M}$	MSPREE of $\mathbf{Y}$
$\boldsymbol{\beta}$	Matrix $J \times J$ containing the parameters of the MSPREE
$\otimes$	Kronecker product
$\mathbb{1}^{j}_{(1 \times J)}$	Row vector of dimension (J) with value of 1 on the j-th component and zero everywhere else
$\mathbb{1}^{jj}_{(J)}$	Square $\mathbf{0}$ matrix of dimension (J) with an entry 1 on component (j, j)

$\mathrm{vec}\left(\cdot\right)$ Vector operator. Transforms a matrix into a column vector by stacking the columns of the matrix.

$\mathrm{diag}\left(\cdots\right)$ Diagonal or block-diagonal matrix with the components in parenthesis on the main diagonal.

Introduction

This thesis addresses the problem of obtaining Small Area (SA) estimates for the cell counts or proportions of a population *composition*. That term would be used through this document to denote a set of vectors with positive components, arranged as rows in a two way table. Each vector may contain the frequencies of a categorical variable in a given area, or the corresponding within-area proportions. For instance, a labour force composition in England may contain the frequencies of individuals belonging to the categories "employed", "unemployed" and "inactive" disaggregated by Local Authority (LA). Furthermore, it will be assumed that, given some covariates, the areas that constitute the rows of the composition are exchangeable. Hence the methods hereby presented seem more intuitive when the *areas* correspond to some geographical classification than when they correspond to *domains* in a more general sense.

The estimators that are proposed in this thesis presuppose the availability of a *proxy* and *sample* estimate of the target composition, as well as its true margins. In practice, a proxy of the composition of interest can be obtained from a population census or an administrative source, referring to the same set of areas in a previous time period, maybe under a slightly different definition or covering only partially the population of interest. A sample estimate of the target composition can often be obtained from the surveys routinely carried out by most National Statistical Institutes (NSIs). However, because such estimate is usually not accurate enough for the inner cells of the composition, the estimation of the composition as a whole is still considered a small area problem.

Information regarding the row and column margins of the composition of interest can be found in hard sources, particularly administrative data. In the absence of this type of information, sample estimates at very aggregated levels can usually be considered accurate enough as for the estimation error to be disregarded. For instance, in the previous example, the row margin of the target composition is given by the size of the labour force population by LA which may be obtained via demographic methods. The column margin is the total number of employed, unemployed and inactive people at the country level and can be accurately estimated using survey data.

Furthermore, in order to propose the estimators that are the core of this document, it will be assumed that all cells in the target and proxy compositions have strictly positive values. In that situation, *Structure Preserving* estimators, hereby called SPREE-type[1] estimators, can be used to produce estimates of the composition of interest as explained below.

SPREE-type estimators

Let us denote the composition of interest by $\mathbf{Y}$. In the conditions above referred, it is possible to induce an additive decomposition where the natural logarithm of the quantity in each cell is expressed as the summation of four components: a global effect, row and column specific effects and an *interaction* term (see for instance Agresti, 2013, section 9.1.3). The first three components conform to what has been called by Purcell and Kish (1980) the *allocation structure* whereas the interaction terms constitute the *association structure* of $\mathbf{Y}$.

SPREE-type estimators provide an estimate of $\mathbf{Y}$ that *preserves* the association structure observed in $\mathbf{X}$ while keeping the allocation structure implied by the known margins. The exact meaning of the word *preserves* is given by a functional relationship assumed between the association structures of $\mathbf{X}$ and $\mathbf{Y}$, that we will call the *structural assumption*. This assumption is used to derive an estimate of the association structure of $\mathbf{Y}$ which in turn, leads to an estimate of $\mathbf{Y}$ via imposition of the known margins, typically using a multiplicative raking algorithm such as Iterative Proportional Fitting (IPF, Deming and Stephan, 1940). Notice that in this approach, any benchmarking to the known row and column margins is incorporated as part of the modelling, rather than as an ex-post adjustment.

The first SPREE-type estimator, called simply SPREE, was proposed in Purcell and Kish (1980). The structural assumption of this estimator is that $\mathbf{X}$ and $\mathbf{Y}$

[1] The acronym SPREE will be used in this document both as a abbreviation for *Structure Preserving* and to identify the SPREE-type estimator in Purcell and Kish (1980). The meaning of the acronym should be clear according to the context.

have the same association structure. This assumption is practically convenient because no sample data is required to derive an estimate of $\mathbf{Y}$. However, it is difficult to justify why the association structure does not change, especially when there is a big time gap between the reference periods of $\mathbf{X}$ and $\mathbf{Y}$, or when the definitions of the target populations are considerably different. Departures from this assumption lead to a biased estimator.

Provided that a direct estimate of the target composition, $\hat{\mathbf{Y}}$, is also available, Zhang and Chambers (2004) relaxed the assumption of equality, allowing both association structures to be proportional. Their proposed estimator will hereby be called Generalized SPREE (GSPREE). Even though this estimator is considerably more flexible and contains the SPREE of Purcell and Kish (1980) as a particular case, the use of one common proportionality constant governing the relationship between the proxy and target composition results insufficient in many practical situations, in which the GSPREE will be biased. Zhang and Chambers (2004) also proposed a version of the GSPREE including cell-specific random effects as a way to minimise the risk of bias. Unfortunately, if the sample sizes are small, the inclusion of random effects may introduce a considerable amount of variance that does not always compensate for the bias reduction.

Problem statement

As with many other areas of SAE, there is growing need for estimation of SA compositions. Particularly, it can be foreseen that future requirements of users will involve domains defined by very detailed geographic classifications. Assuming that sample sizes will not increase accordingly, this would indicate a future decrease in the size of the, already small, areas for which estimates are of interest, and a likely increase in the number of areas without sample data. In this setting, fixed effects estimators for SA compositions under more flexible structural assumptions acquire relevance with respect to estimators that rely on random effects as the main way to reduce the potential bias, because the prediction of random effects requires area-specific information: little gain could be expected in cases where a considerable number of areas are out of sample. Moreover, even if the sampling design ensures that all areas are in-sample, the trade-off between bias and variance mentioned above seems unfavourable for this type of estimator in a context with progressively smaller sample sizes.

As aforesaid, a main drawback of the GSPREE estimator is the assumption of the same proportionality constant holding for all the columns in the composition. In this thesis, we propose a new SPREE-type estimator called Multivariate Structure Preserving Estimator (MSPREE), which generalises the proportionality assumption underpinning the GSPREE without incurring in additional data requirements. Because the estimator does not rely in random effects, bias reduction respect to the existing SPREE-type estimators can be obtained even in the cases with very small sample sizes.

Furthermore, an extension of the MSPREE including cell specific random effects, called Mixed MSPREE (MMSPREE), as well as an estimator for the variance components, are proposed. Because the random effects are included directly in the linear predictor, the association structure under the mixed model remains well defined after the inclusion of the random effects. Moreover, neither the estimation of the variance components, nor the calculation of the MMSPREE, require of computationally intensive methods.

Outline of the document

The document is divided in six chapters as follows. Chapter 1 introduces some preliminary concepts related to the SAE problem and presents existing estimators for population compositions, with emphasis on SPREE-type estimators. Chapter 2 and 3 are devoted to the development of the proposed estimators, MSPREE and MMSPREE. Chapter 4 illustrates the use of those estimators in a simulation exercise. Chapter 5 presents the results of an application of the proposed estimators, using data from the 2011 Census in England and the Annual Population Survey conducted by the Office for National Statistics (ONS). Finally, Chapter 6 presents the conclusions of this document and suggests lines for future research.

Chapter 1

Preliminaries

1.1 The Small Area Estimation (SAE) problem

The interest for obtaining estimates of population characteristics has been traditionally satisfied by the use of existent sources of data, such as administrative databases, or by collecting specific data, either via total enumeration when feasible, or via survey samples. Use of existent sources of information has the advantage that no additional cost is associated with the collection process. However, unless existent sources collect exactly the data required to produce the indicators of interest, this approach would only provide an approximation to the target phenomena under study. Censuses, on the other hand, have been traditionally the gold standard because, ignoring considerations of coverage and measurement error, inference from them has no statistical uncertainty involved. However, the costs associated with such an operation and the level of burden that they imply in the population, make them an exception more than a rule in terms of collecting data for statistical purposes.

Well-developed sample surveys have proved their ability to produce valid inferences. Their broad use in almost all aspects of social research is a clear indication of this. Even though classic books of survey sampling, such as Cochran (1977) or Kish (1965) are still in common use, much research has been devoted in the past 50 years to the development of sampling strategies that are efficient in terms of bias and variance (regarding to the so-called design-based approach of inference, as it will be explained next), while keeping costs under control.

Survey sampling differs from most areas of statistics in regard to the main approach of inference in use. Predominantly, survey sampling practitioners rely on design-based inference, which uses the randomization induced by the sampling design while assuming the values of the variables of interest are fixed. Estimates are calculated using sampling weights that can be obtained from the sampling design or calculated a posteriori using auxiliary information (Särndal et al., 1992). Desirable estimators in this context are typically design-consistent, i.e., their Mean Square Error (MSE) goes to zero as the sample size increases, with the expectation calculated with respect to the distribution induced by the sampling design. Between the two terms that compose the MSE, the variance is commonly the dominant term.

However, in recent decades, interest in estimates for subgroups of the population, i.e., *domains*, has increased. Under a design-based approach, if such domains could be identified in advance, a sample strategy involving stratification and sample allocation could, in principle, provide reliable estimates using only domain-specific and perhaps auxiliary data. We will refer to those estimates as *direct*. Unfortunately, in practice it is impossible to identify in advance all domains that might be of interest and even in such a case, the sum of all the required sample sizes could be extremely costly. In the case of unplanned domains, direct estimation is unfeasible for out-of sample domains and may be of limited use for in-sample domains due to big variances, especially if they correspond to a small fraction of the population.

When a domain-specific sample is too small for reliable direct estimates to be produced, that domain is considered *small*. The term *Small Area Estimation* refers to the problem of producing reliable estimates for such domains. Notice that the names *area* and *domain* are used interchangeably in this document. Unlike direct estimators, most estimators used in SAE are *indirect* or model-based (Rao and Molina, 2015, Chapter 3), i.e., they are built under the assumption that a given statistical model holds for the population of interest, respect to which the statistical inference is performed. Moreover, because different domains have common attributes according to the assumed model, data from other domains can be used to predict the value of a given domain, hence *borrowing strength* by the increase of effective sample size available for domain-specific estimation. Model-based estimators are not new to survey sampling (see for instance Valliant et al., 2000), but have acquired a predominant role in the case of small domains.

The model underpinning a given indirect estimator may be implicitly or explicitly stated. A further distinction is made between estimators supported by explicit models that explain between-domain heterogeneity only via covariates, e.g., marginal models, and explicit models that include additional elements to take into account extra heterogeneity, e.g., mixed effect models with area-specific random effects or M-Quantile models (Chambers and Tzavidis, 2006). In the former case, the estimators are called *synthetic* and can be calculated for both in-sample and out-of-sample areas. In the latter, only in-sample areas can be estimated. A compromise approach, common in practice for cases where there is sample in many but not all areas of interest, is to work with a mixed effects model, such as the General Linear Mixed Model (GLMM) which will be introduced in section 1.2, and use the synthetic predictor obtained from the fixed part of the model for out-of sample areas.

Because model-based predictors may perform poorly in the face of model misspecification, model selection and diagnostics are at the core of the production of SA estimates. Moreover, even though the inference is still model-based, the practice of testing the validity of SA predictors using design-based simulation exercises has gained popularity. For a more detailed discussion on these topics see ESSNet (2012) and Tzavidis et al. (2016).

This thesis addresses the issue of obtaining SA estimates for *compositions*. The term has been previously used by Aitchison (2003, p. 26) to indicate a vector with positive components whose sum is 1. The distribution of a categorical variable in a given domain constitutes an example of a composition according to this definition. As previously mentioned, in a SA setting, we are interested in obtaining estimates for several domains, therefore, we will hereby extend the term to refer not to an individual vector but to a set of them, arranged as rows in a two way table. Such a table may contain the within-domain distribution or the raw frequencies of a categorical variable.

In the next sections, a revision of the existent literature in SAE that we consider relevant for the estimation of compositions is provided. Readers interested in SAE methods in general may find a very comprehensive account in Rao and Molina (2015), as well as a review of the most important developments of the last decade in Pfeffermann (2013). The remaining of this chapter is organised as follows. Section 1.2 introduces the GLMM, according to Chapter 5 of Rao and Molina (2015). The GLMM covers many of the most commonly used models in SAE. The notation in this section has been kept in accordance with the one in Rao and Molina (2015), and may hence differ from the one used in the rest of the document. Section 1.3 introduces the Structure Preserving (SPREE)-type estimators, that are the main focus of this thesis, and discusses the issues that motivate the development of the estimators proposed in this document. Finally, Section 1.4 discusses other existing estimators for small area compositions.

1.2 General Linear Mixed Model (GLMM)

Denote by $\mathbf{y}$ a vector of sample observations of dimension $n \times 1$. Let $\mathbf{X}$ and $\mathbf{Z}$ be known matrices of dimension $n \times p$ and $n \times h$ and $\mathbf{u}$ and $\mathbf{e}$ be random vectors independently distributed with zero mean and variance covariance matrices $\mathbf{G}$ and $\mathbf{R}$ respectively, governed by a set of parameters $\boldsymbol{\delta} = (\delta_1, \ldots \delta_q)^T$ called *variance components*. It is assumed that the sample data follows the GLMM

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu} + \mathbf{e}, \tag{1.1}$$

where the vector $\boldsymbol{\nu}$ represents a set of random effects and $\mathbf{e}$ denotes the error terms of the model. The unconditional variance of $\mathbf{y}$ under the model is hence $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T$.

BLUP and EBLUP

The aim is to obtain a predictor for the linear combination

$$\mu = \mathbf{l}^T\boldsymbol{\beta} + \mathbf{m}^T\boldsymbol{\nu}, \tag{1.2}$$

with $\mathbf{l}$ and $\mathbf{m}$ known vectors of constants. We use the term *predictor* instead of *estimator* to acknowledge that $\boldsymbol{\nu}$ in the equation above is a realization of a random entity. Assuming the vector of variance components $\boldsymbol{\delta}$ is known, Henderson (1950) showed that the Best Linear Unbiased Predictor (BLUP) of μ under model (1.1) is

$$\tilde{\mu}^H(\boldsymbol{\delta}) = \mathbf{l}^T\tilde{\boldsymbol{\beta}} + \mathbf{m}^T\tilde{\boldsymbol{\nu}} \tag{1.3}$$

with $\tilde{\boldsymbol{\beta}}$ the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$,

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{1.4}$$

and

$$\tilde{\boldsymbol{\nu}} = \mathbf{G}\mathbf{Z}^{\mathsf{T}}\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right). \tag{1.5}$$

This result extends to the simultaneous estimation of $r \geqslant 2$ linear combinations, $\boldsymbol{\mu} = \mathbf{L}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{\nu}$, for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_r)$ and $\mathbf{L}$ and $\boldsymbol{M}$ matrices of constants. Moreover, assuming $\boldsymbol{\beta}$ known and $\boldsymbol{\nu}$ and $\boldsymbol{e}$ normally distributed, (1.5) with $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ is the best predictor (BP) of $\boldsymbol{\nu}$. Without the normality assumption but with $\boldsymbol{\beta}$ known, it remains the best among the class of linear predictors of $\boldsymbol{\nu}$.

In practice, the vector $\boldsymbol{\delta}$ containing the variance components is unknown, therefore the matrices $\mathbf{G}$, $\mathbf{R}$ and $\mathbf{V}$ required for the calculation of the BLUP are not available. A two stage estimator for μ is developed by obtaining an estimate $\hat{\boldsymbol{\delta}}(\mathbf{y})$ and using it as a substitute for $\boldsymbol{\delta}$ in the definition of the above mentioned matrices. The resulting estimator $\tilde{\mu}^H(\hat{\boldsymbol{\delta}})$ is called the empirical best linear unbiased predictor (EBLUP). Kackar and Harville (1981) showed that the EBLUP remains unbiased for μ as long as three conditions are satisfied: i) $E\left[\tilde{\mu}^H(\hat{\boldsymbol{\delta}})\right]$ is finite; ii) the distributions of $\mathbf{u}$ and $\boldsymbol{e}$ are symmetric around $\mathbf{0}$; and iii) the variance components estimators are even and translation-invariant i.e., $\hat{\boldsymbol{\delta}}(-\mathbf{y}) = \hat{\boldsymbol{\delta}}(\mathbf{y})$ and $\hat{\boldsymbol{\delta}}(\mathbf{y} - \mathbf{X}\mathbf{b}) = \hat{\boldsymbol{\delta}}(\mathbf{y})$ for all $\mathbf{y}$ and $\mathbf{b}$. Kackar and Harville (1981) also showed that some of the most well known procedures to obtain estimates of the variance components, such as Maximum Likelihood (ML), Restricted (or residual) Maximum Likelihood (REML) and the method of Fitting constants (Henderson, 1953), satisfy this property. For the sake of completeness, ML and REML estimators under the assumption of normality will be briefly introduced next.

ML and REML estimation

Under normality of $\boldsymbol{\nu}$ and $\boldsymbol{e}$, ML estimates of the variance components can be obtained iteratively using the Fisher-scoring algorithm. Denote by $\mathbf{V}_{(j)}$ the first derivative of $\mathbf{V}$ with respect to δ_j and by $\mathbf{V}^{(j)}$ the first derivative of $\mathbf{V}^{-1}$ with respect to δ_j, noticing that $\mathbf{V}^{(j)} = -\mathbf{V}^{-1}\mathbf{V}_{(j)}\mathbf{V}^{-1}$. The vector of first derivatives of the log-likelihood $l(\boldsymbol{\beta}, \boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$ has components $s_1, \ldots, s_q$ with:

$$s_j(\boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \delta_j} = -\frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{V}_{(j)}\right) - \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}}\mathbf{V}^{(j)}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right). \tag{1.6}$$

The Fisher Information matrix has elements (j, k) given by

$$\mathcal{I}_{jk}(\boldsymbol{\delta}) = -\frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{(j)}\mathbf{V}_{(k)}\right). \tag{1.7}$$

In iteration $i+1$, a new estimate of $\boldsymbol{\delta}$ is calculated as

$$\boldsymbol{\delta}^{(i+1)} = \boldsymbol{\delta}^{(i)} + \left[\mathfrak{I}\left(\boldsymbol{\delta}^{(i)}\right)\right]^{-1} \mathbf{s}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}^{(i)}), \boldsymbol{\delta}^{(i)}\right], \tag{1.8}$$

with $\tilde{\boldsymbol{\beta}}$ and $\mathbf{s}$ defined as in equations (1.4) and (1.6). ML estimators of $\boldsymbol{\delta}$ do not take into account the loss of degrees of freedom due to the estimation of $\boldsymbol{\beta}$ (Harville, 1977). In order to take them into consideration, a restricted log-likelihood function $l_r(\boldsymbol{\delta})$ (see for instance Pawitan, 2013, section 17) can be maximised to obtain REML estimates of $\boldsymbol{\delta}$ using a Fisher algorithm analogous to (1.8). The new matrix of first derivatives $\mathbf{s}_R$ has components $s_{R_1}, \dots s_{R_q}$ defined as

$$s_{R_j}(\boldsymbol{\delta}) = \frac{\partial l_R(\boldsymbol{\delta})}{\partial \delta_j} = -\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\mathbf{V}_{(j)}\right) + \frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{V}_{(j)}\mathbf{P}\mathbf{y}, \tag{1.9}$$

with $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$. The Fisher Information matrix of the restricted likelihood has elements (j,k) given by

$$\mathfrak{I}_{R,jk}(\boldsymbol{\delta}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{P}\mathbf{V}_{(j)}\mathbf{P}\mathbf{V}_{(k)}\right). \tag{1.10}$$

Unfortunately, both ML and REML methods can lead to solutions for which the variance matrices $\mathbf{G}$ and $\mathbf{R}$ are not positive definite. Solutions to this issue proposed for specific versions of the GLMM are discussed in Rao and Molina (2015), section 5.2.4. Asymptotically, ML and REML estimates are equally efficient with variance covariance matrices $V(\tilde{\boldsymbol{\beta}}) = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$ and $V(\hat{\boldsymbol{\delta}}) = [\mathfrak{I}(\boldsymbol{\delta})]^{-1}$.

Mean Square Error estimation

Because of their nature of *borrowing strength* from data in different domains and from auxiliary sources, indirect estimators usually exhibit less variance than direct estimators. However, because the models they rely cannot hold perfectly in practice, indirect estimators also suffer from bigger biases. Estimation of MSE is, therefore, a topic that has received appreciable attention in the literature of SAE.

Following Rao and Molina (2015), Sections 5.2.2-5.2.6, let μ denote the target of estimation, defined as in equation (1.2) and denote by $t(\boldsymbol{\delta})$ and $t(\hat{\boldsymbol{\delta}})$ the BLUP and EBLUP of μ, respectively. Consider the decomposition of the error of the

EBLUP

$$t(\hat{\delta}) - \mu = [t(\delta) - \mu] + \left[t(\hat{\delta}) - t(\delta)\right]. \quad (1.11)$$

Under normality assumptions and provided that $\hat{\delta}$ is translation invariant, squaring both sides and taking the expectation of (1.11) leads to

$$\text{MSE}\left[t(\hat{\delta})\right] = \text{MSE}\left[t(\delta)\right] + E\left[t(\hat{\delta}) - t(\delta)\right]^2. \quad (1.12)$$

The first term on the right hand side of (1.12), the MSE of the BLUP, can be further decomposed as the sum of two terms: $g_1(\delta)$, which corresponds to the variance of the BLUP with respect to the true μ when β is known, and $g_2(\delta)$, which carries the variability due to the estimation of β. For the GLMM defined in equation 1.1, the terms g_1 and g_2 are given by

$$g_1(\delta) = \mathbf{m}^T(\mathbf{G} - \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\mathbf{G})\mathbf{m}$$

and

$$g_2(\delta) = \mathbf{d}^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{d},$$

for $\mathbf{d}^T = \mathbf{l}^T - \mathbf{b}^T\mathbf{X}$ and $\mathbf{b}^T = \mathbf{m}^T\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}$.

The second term on the right hand side of (1.12) cannot be simplified in general. The use of a Taylor linearisation for $t(\hat{\delta}) - t(\delta)$ ignoring the terms corresponding to derivatives of the error of the BLUE of β, $(\tilde{\beta} - \beta)$, leads to the approximation $E\left[t(\hat{\delta}) - t(\delta)\right]^2 \approx g_3(\delta)$, for

$$g_3(\delta) = \text{tr}\left[\left(\frac{\partial \mathbf{b}^T}{\partial \delta}\right)\mathbf{V}\left(\frac{\partial \mathbf{b}^T}{\partial \delta}\right)^T \bar{\mathbf{V}}(\hat{\delta})\right]$$

where $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T$ and $\bar{\mathbf{V}}(\hat{\delta})$ is the asymptotic covariance matrix of $\hat{\delta}$. Finally, an approximation of the MSE of the EBLUP is given by

$$\text{MSE}\left[t(\hat{\delta})\right] \approx g_1(\delta) + g_2(\delta) + g_3(\delta). \quad (1.13)$$

Studying the nested-error regression model (Battese et al., 1988), the random regression coefficient model (Dempster et al., 1981) and the Fay-Herriot model (Fay and Herriot, 1979), Prasad and Rao (1990) proposed the estimator

$$\widehat{\text{MSE}}\left[t(\hat{\delta})\right] = g_1(\hat{\delta}) + g_2(\hat{\delta}) + 2g_3(\hat{\delta}) \quad (1.14)$$

which is second-order unbiased under normality assumptions, provided that $\hat{\delta}$ is translation-invariant. Subsequently, under the assumption of unbiasedness for the estimator of variance components, Harville and Jeske (1992) proposed the estimator (1.14) for the GLMM. Furthermore, Lahiri and Rao (1995) showed the robustness of this estimator to departures from normality of the random effects, for the Fay-Herriot model, as long as some moment conditions are satisfied.

1.3 SPREE-type estimators

The underpinning idea of SPREE-type estimators can be tracked back to Deming and Stephan (1940) but was first used in the context of SAE by Chambers and Fenney (1977); Bousfield (1977) (as referenced in Purcell and Kish (1980)) and Gonzalez and Hoza (1978). It is in Purcell and Kish (1980) where the concept is formalized with the introduction of the SPREE estimator.

Denote by $\mathbf{Y}$ the population composition of interest, constituted by the counts Y_{aj} for $a = 1, \dots A; j = 1 \dots, J$. $\mathbf{Y}$ can be represented in the form of a saturated log-linear model as

$$\log Y_{aj} = \alpha_0^Y + \alpha_a^Y + \alpha_j^Y + \alpha_{aj}^Y. \tag{1.15}$$

The quantities α_0^Y, α_a^Y, α_j^Y and α_{aj}^Y can be defined in several ways. Hereby we will use a *centred-constraints* parametrisation, given by

$$\begin{aligned}
\alpha_0^Y &= \frac{1}{AJ} \sum_a \sum_j \log Y_{aj}, \\
\alpha_a^Y &= \frac{1}{J} \sum_j \log Y_{aj} - \alpha_0^Y, \\
\alpha_j^Y &= \frac{1}{A} \sum_a \log Y_{aj} - \alpha_0^Y, \\
\alpha_{aj}^Y &= \log Y_{aj} - \alpha_a^Y - \alpha_j^Y - \alpha_0^Y,
\end{aligned} \tag{1.16}$$

satisfying the constraints $\sum_a \alpha_a^Y = \sum_j \alpha_j^Y = \sum_a \alpha_{aj}^Y = \sum_j \alpha_{aj}^Y = 0$. Expressed as in (1.15), it is possible to decompose $\mathbf{Y}$ in two structures (Purcell and Kish, 1980):

1. The *association structure:* corresponds to the set of interaction terms α_{aj}^Y, for $a = 1, \dots A; j = 1 \dots, J$. It determines the relationship between rows

and columns. In the theoretical case where both dimensions are independent, $\alpha^{Y}_{aj} = 0$ for all pairs (a, j).

2. The *allocation structure:* it is given by the sets of terms α^{Y}_{0}, α^{Y}_{a}, and α^{Y}_{j} for $a = 1, \dots A$; $j = 1 \dots, J$. It carries information about the scale of the composition and the disparities within the sets of rows and columns.

SPREE-type estimators make use of the fact that the true row and column margins of $\mathbf{Y}$ may be known or sufficiently accurately estimated to assume so. In such a case, only the association structure of the composition of interest needs to be estimated because the corresponding allocation structure is exclusively determined by the margins. SPREE-type estimators use a proxy composition $\mathbf{X}$ (and a composition of sample estimates $\hat{\mathbf{Y}}$ if available), to build an estimate of the association structure of $\mathbf{Y}$ and then impose the known margins. Ensuring that the margins of the final estimate coincide with the known margins is desirable, not only as a way to satisfy the necessity of arithmetic consistency between different sets of estimates (for example, two compositions addressing the same reference population) but also as protection against model misspecification, avoiding undesired departures of the model-based estimates from the corresponding direct estimates at levels at which the latter are considered reliable.

1.3.1 SPREE

Purcell and Kish (1980) addresses the problem of how to produce postcensal estimates of frequency characteristics for local areas or domains. Denote by $\mathbf{Y}$ the composition of interest, with cell counts Y_{aj} where a indexes the Local areas and j indexes the values of the categorical variable, for $a = 1, \dots A$, $j = 1 \dots J$. Purcell and Kish (1980) considers several scenarios of data availability. For illustration purposes, we will hereby simplify the problem by assuming that a proxy composition $\mathbf{X}$ of the same dimension as $\mathbf{Y}$ can be obtained, for instance, from a past census, and that sets of margins $\mathbf{Y}_{a+} = (Y_{1+}, \dots, Y_{A+})$ and $\mathbf{Y}_{+j} = (Y_{+1}, \dots, Y_{+J})$, where the symbol $+$ substitutes the index in the summation, can be estimated from *hard* data sources such as a census or administrative data, i.e., they are accurate enough for the estimation error to be disregarded. The aim is to produce an estimate $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ with the same association structure observed in $\mathbf{X}$, satisfying the allocation structure determined by the known sets of margins. Purcell and Kish (1980) propose to obtain such an estimate using an iterative procedure described as follows:

1. Rescale the rows of $\mathbf{X}$ as:

$$\hat{Y}^{(1)}_{aj} = X_{aj}\frac{\mathbf{Y}_{+j}}{\mathbf{X}_{+j}},$$

2. Rescale the columns of $\hat{\mathbf{Y}}^{(1)}$ as:

$$\hat{Y}^{(2)}_{aj} = \hat{Y}^{(1)}_{aj}\frac{\mathbf{Y}_{a+}}{\hat{Y}^{(1)}_{a+}},$$

3. Rescale the rows of $\hat{\mathbf{Y}}^{(2)}$ as:

$$\hat{Y}^{(3)}_{aj} = \hat{Y}^{(2)}_{aj}\frac{\mathbf{Y}_{+j}}{\hat{Y}^{(2)}_{+j}}.$$

Steps (2) and (3) are repeated alternately until convergence, using in each step the estimates obtained in the previous step. This algorithm, due to Deming and Stephan (1940), is known in the literature as Iterative Proportional Fitting (IPF) and can be found in most textbooks about categorical data (e.g. Agresti, 2013, section 9.7.2). Because the final estimate, $\hat{\mathbf{Y}}^{S}$, has the association structure observed in the proxy composition, the implicit structural assumption of the SPREE of Purcell and Kish (1980) is

$$\alpha^{Y}_{aj} = \alpha^{X}_{aj} \quad \text{for all } a, j, \tag{1.17}$$

with α^{Y}_{aj} and α^{X}_{aj} defined as in equation (1.16).

When only one set of margins is available no iteration is required. The resulting estimator minimises the χ-squared distance

$$D_{\chi^2} = \sum_a \sum_j \frac{(Y_{aj} - \hat{Y}_{aj})^2}{Y_{aj}},$$

as well as the Kullback-Leibler discrimination information measure

$$D_{KL} = \sum_a \sum_j Y_{aj} \log \frac{Y_{aj}}{\hat{Y}_{aj}}.$$

When more than one set of margins is available, iteration is required. Ireland and Kullback (1968) showed that if $Y_{aj} > 0$ for all a, j, the IPF algorithm converges to the optimal solution according to the discrimination information measure.

Noble et al. (2002) showed that generalized linear models (GLMs) can also be used to calculate the SPREE. In the first step, a saturated log-linear model is fitted to the proxy composition and the matrix $\boldsymbol{\alpha}^X$ containing the interaction terms α^X_{aj} for $a = 1, \dots, A$, $j = 1, \dots, J$ is calculated. Then, a new composition $\tilde{\mathbf{X}}$ is generated using the known sets of margins and assuming independence, i.e., $\tilde{X}_{aj} = (Y_{a+}Y_{+j})/Y_{++}$ for $a = 1, \dots, A$, $j = 1, \dots, J$. Finally, the SPREE is obtained by fitting a non-saturated log-linear model to $\tilde{\mathbf{X}}$, with offset $\boldsymbol{\alpha}^X$. Noble et al. (2002) argue in favour of using a Generalized Linear Model instead of the IPF algorithm as a way to allow for sources of auxiliary information besides the proxy composition, including discrete and continuous covariates.

1.3.2 GSPREE

Zhang and Chambers (2004) addresses the problem of obtaining an estimate for a population composition, θ^Y, that contains the within-area proportions, $\theta^Y_{aj} = Y_{aj}/Y_{a+}$ for $a = 1, \dots, A; j = 1 \dots, J$, using information of a proxy composition θ^X and a composition of direct estimates $\hat{\theta}^Y$. As a generalization of the SPREE of Purcell and Kish (1980), this estimator will be hereby called Generalized SPREE (GSPREE).

The GSPREE assumes that the target and proxy compositions relate via the Generalized Linear Structural Model (GLSM)

$$\tau^Y_{aj} = \lambda_j + \beta\tau^X_{aj}, \tag{1.18}$$

with $\sum_j \lambda_j = 0$, for $a = 1, \dots, A$; $j = 1, \dots, J$, where

$$\tau^Y_{aj} = \log\theta^Y_{aj} - \frac{1}{J}\sum_{l=1}^{J}\log\theta^Y_{al}, \tag{1.19}$$

and τ^X_{aj} is analogously defined. The model is fitted using the sample estimate $\hat{\theta}^Y$ in place of θ^Y. The GSPREE estimate of θ^Y, denoted $\hat{\theta}^{Y,G}$, is obtained applying the inverse function

$$\hat{\theta}^{Y,G}_{aj} = \frac{\exp\hat{\tau}^Y_{aj}}{\sum_{l=1}^{J}\exp\hat{\tau}^Y_{al}}$$

on the fitted values. Notice that equation (1.19) is generally not invertible. However, as shown in Lemma 1 in page 19, the set of constraints $\theta^Y_{a+} = 1$ for $a = 1, \dots, A$ ensures the existence of an inverse function in this particular case.

Because the estimation problem is formulated in the within-area proportion scale, information regarding the margins of the target composition is not necessary. If $\mathbf{Y}_{a+}$ or $\mathbf{Y}_{+j}$ were known, a GSPREE of the composition of counts, denoted by $\hat{\mathbf{Y}}^G$, could be obtained imposing the known margins on $\hat{\theta}^{Y,G}$ using IPF, without altering the fitted association structure.

The GLSM is reminiscent of a GLM but receives the name *structural* because equation (1.18) is supposed to hold for the finite-population composition instead of for its expected value, as in the GLM setting. Assuming that $\hat{\theta}^Y$ is an unbiased estimator of θ^Y with known block diagonal covariance matrix, Zhang and Chambers (2004) proposes to fit the GLSM using an Iterative Weighted Least Squares (IWLS) algorithm. Alternatively, if it is assumed that $\hat{\mathbf{Y}}$ is a realization of a random composition with product multinomial distribution, the GLSM can be considered a GLM and the IWLS leads to the ML estimate of the proportions that parameterize that distribution (see for instance Agresti, 2013, section 4.6).

Now we will turn to identify the *structural assumption* of the GSPREE. Notice that

$$\begin{aligned}\log \theta^Y_{aj} &= \log Y_{aj} - \log Y_{a+} \\ &= \alpha^Y_0 + (\alpha^Y_a - \log Y_{a+}) + \alpha^Y_j + \alpha^Y_{aj} \\ &:= \alpha^Y_0 + \tilde{\alpha}^Y_a + \alpha^Y_j + \alpha^Y_{aj}\end{aligned}$$

where the α^Y_0, α^Y_a, α^Y_j and α^Y_{aj} are the corresponding terms for the table of counts defined in (1.16). Also, note that $\sum_{l=1}^{J} \log \theta^Y_{al} = J(\alpha^Y_0 + \tilde{\alpha}^Y_a)$ given that in the centred-constraints parametrisation $\sum_l \alpha^Y_l = \sum_l \alpha^Y_{al} = 0$. Hence, the link function defined in (1.19) is equivalent to

$$\tau^Y_{aj} = \alpha^Y_j + \alpha^Y_{aj}, \tag{1.20}$$

i.e. τ^Y_{aj} *isolates* the terms corresponding to the column effects and the interactions of θ^Y. Note that because these two sets of terms are the same whether the composition is considered in the count or in the within-area proportions scale, the implied model for θ^Y also holds for $\mathbf{Y}$. Replacing (1.20) in the definition of the GLSM, equation (1.18), and adding over the areas, we obtain

$$\begin{aligned}\alpha_j^Y + \alpha_{aj}^Y &= \lambda_j + \beta(\alpha_j^X + \alpha_{aj}^X)\\ \Rightarrow \sum\nolimits_a(\alpha_j^Y + \alpha_{aj}^Y) &= A\lambda_j + \beta\sum\nolimits_a(\alpha_j^X + \alpha_{aj}^X)\\ \Rightarrow \alpha_j^Y &= \lambda_j + \beta\alpha_j^X. \end{aligned} \tag{1.21}$$

The last line given by the sum to zero constraints of the centred-constraints parameterization. Moreover, the constraint $\sum \lambda_j = 0$ stated in the definition of the model is implicit in the sum of (1.21). Substituting (1.21) and (1.20) in the definition of the GLSM, we obtain to the structural assumption of the GSPREE:

$$\begin{aligned}\tau_{aj}^Y = \lambda_j + \beta\tau_{aj}^X, \Rightarrow \alpha_j^Y + \alpha_{aj}^Y &= \lambda_j + \beta(\alpha_j^X + \alpha_{aj}^X)\\ \Rightarrow \lambda_j + \beta\alpha_j^X + \alpha_{aj}^Y &= \lambda_j + \beta(\alpha_j^X + \alpha_{aj}^X)\\ \Rightarrow \alpha_{aj}^Y &= \beta\alpha_{aj}^X. \end{aligned} \tag{1.22}$$

Zhang and Chambers (2004) indicates that the sets of equations (1.21) and (1.22) for $a = 1, \ldots, A$, $j = 1 \ldots, J$ are equivalent to equation (1.18) to define the GLSM. However, the only true assumption underpinning the GSPREE is (1.22) because given β it is always possible to set $\lambda_j = \alpha_j^Y - \beta\alpha_j^X$ to satisfy (1.21). In this sense, notice that the set of parameters $\lambda_1, \ldots, \lambda_J$ are nuisance parameters in the GLSM. The GSPREE assumes proportionality between the association structures of the target and proxy compositions.

1.3.2.1 Alternatives to the GLSM

Even if a product multinomial distribution is assumed for $\hat{\mathbf{Y}}$, bespoke software is necessary to estimate β using the procedure described in Zhang and Chambers (2004) because the link function used by the GLSM is not included in the available packages for the fitting of GLMs. Lemmas 2 and 3 in subsection 1.3.2.3 present two new equations that also induce the structural assumption (1.22) and can be used as alternatives to the GLSM to obtain the GSPREE using a product multinomial or a Poisson likelihood and standard software.

1.3.2.2 GSPREE with random effects

Zhang and Chambers (2004) proposed a mixed effects version of the GSPREE with the aim of reducing the risk of bias due to model misspecification with respect to the estimator with only fixed effects. However, because the predicted random effects can be unstable, particularly under small sample sizes, it is possible for this estimator to exhibit a higher MSE than its fixed-effects counterpart. The GSPREE with random effects is defined by the Generalized Linear Structural Mixed Model (GLSMM)

$$\tau_{aj}^{Y} = \lambda_j + \beta\tau_{aj}^{X} + \nu_{aj} \tag{1.23}$$

where $(\nu_{a2}, \dots, \nu_{aJ})^T \overset{\text{IID}}{\sim} N(0, \Sigma)$ and $\nu_{a1} = -\sum_{j\neq 1} \nu_{aj}$, for $a = 1, \dots, A$; and $j = 1 \dots, J$.

Following a process analogous to the one used with the GLSM, it is possible to show that the GLSMM is equivalent to the sets of equations:

$$\alpha_j^{Y} = \lambda_j + \beta\alpha_j^{X} + \bar{\nu}_{+j} \tag{1.24}$$

$$\alpha_{aj}^{Y} = \beta\alpha_{aj}^{X} + (\nu_{aj} - \bar{\nu}_{+j}) \tag{1.25}$$

with $\bar{\nu}_{+j} = (1/A)\nu_{+j}$. Note that the term ν_{a+} is zero by construction and therefore, also ν_{++}. On the other hand, nothing ensures that $\nu_{+j} = 0$. Asymptotically, as the number of areas increases, it is expected that the random effects affect only the association structure, leading to the structural assumption

$$\alpha_{aj}^{Y} = \beta\alpha_{aj}^{X} + \nu_{aj};$$

however, formally speaking, equations (1.24) and (1.25) imply that the random effects will affect not only the association but also the allocation structure. It would be desirable to impose the additional constraint $\nu_{+j} = 0$ but this would introduce correlation across areas and complicate substantially the estimation. Instead, Zhang and Chambers (2004) suggests to perform the estimation assuming that the ν_{aj} are independent between areas and standardize the estimates ex-post. Because such adjustment is of order $\mathcal{O}_p(A^{-1/2})$, the expected difference is small even for a moderate number of areas.

On the other hand, notice that equations (1.24) and (1.25) are sufficient to guarantee a set of predicted interactions that satisfy the sum-zero constraints assumed by the centred-constraints parameterization because:

$$\begin{aligned}\sum_a \alpha^Y_{aj} &= \beta \sum_a \alpha^X_{aj} + \sum_a (\nu_{aj} - \bar{\nu}_{+j}) \\ &= \nu_{+j} - A\bar{\nu}_{+j} \\ &= 0,\end{aligned}$$

given that $\alpha^X_{+j} = 0$. Analogously,

$$\begin{aligned}\sum_j \alpha^Y_{aj} &= \beta \sum_j \alpha^X_{aj} + \sum_j (\nu_{aj} - \bar{\nu}_{+j}) \\ &= \nu_{a+} - \frac{1}{A}\nu_{++} \\ &= 0,\end{aligned}$$

because $\alpha^X_{a+} = \nu_{a+} = \nu_{++} = 0$.

1.3.2.3 Complementary Material

In this section, we expand the available results for the GSPREE (Zhang and Chambers, 2004) in two directions. First, we proof that the link function thereby used in the formulation of the GLSM is invertible. Second, we propose two sets of equations that are equivalent to the GSPREE assumption of proportional interactions and therefore, can be used as an alternative to the GLSM for the estimation of the proportionality parameter.

Throughout this section, denote by θ^Y the target composition in the scale of within-area proportions, $\theta^Y_{aj} = Y_{aj}/Y_{a+}$ for $a = 1, \ldots, A$; $j = 1, \ldots, J$. Denote by $\alpha^Y_0, \alpha^Y_a, \alpha^Y_j$ and α^Y_{aj} the terms of its representation according to equation (1.15).

Invertibility of the link function of the GLSM

This lemma shows the invertibility of the link function used in the formulation of the GLSM. It is referenced in page 16.

Lemma 1. *Let* g *be a function defined in* $\mathbb{R}^{A \times J}$*, the space of the real-valued matrices of dimension* $A \times J$ *by* $g : \theta \mapsto \boldsymbol{\tau}$*, with* τ_{aj} *defined as in equation* (1.19)*:*

$$\tau_{aj} = \log \theta_{aj} - \frac{1}{J} \sum_{l=1}^{J} \log \theta_{al},$$

where the constraints $\theta_{a+} = 1$ *for* $a = 1, \ldots, A$ *are satisfied. Then,* $g(\theta) = g(\tilde{\theta})$ *if and only if* $\theta = \tilde{\theta}$.

Proof. The proof of $\theta = \tilde{\theta}$ implying $g(\theta) = g(\tilde{\theta})$ is trivial. To proof the remaining implication, define $\gamma_{aj} = \log \theta_{aj}$ and $\tilde{\gamma}_{aj} = \log \tilde{\theta}_{aj}$ and notice that

$$g(\theta) = g(\tilde{\theta}) \quad \Longleftrightarrow \quad \gamma_{aj} - \frac{1}{J}\gamma_{a+} = \tilde{\gamma}_{aj} - \frac{1}{J}\tilde{\gamma}_{a+},$$

therefore,

$$\gamma_{aj} = \tilde{\gamma}_{aj} + k_a, \tag{1.26}$$

for $k_a = \frac{1}{J}(\gamma_{a+} - \tilde{\gamma}_{a+})$. On the other hand, as $\theta_{aj} = \exp \gamma_{aj}$, the constraints $\theta_{a+} = 1$ for $a = 1, \ldots, A$ imply

$$\sum_j \exp \gamma_{aj} = 1. \tag{1.27}$$

Substituting equation (1.26) in equation (1.27) leads to

$$1 = \sum_j \exp \gamma_{aj} = \exp k_a \sum_j \exp \tilde{\gamma}_{aj} = \exp k_a, \tag{1.28}$$

because equation (1.27) also holds for the set of $\tilde{\gamma}_{aj}$. The proof is complete because, as $k_a = 0$ for $a = 1, \ldots, A$, then $\gamma_{aj} = \tilde{\gamma}_{aj}$ and hence $\theta_{aj} = \tilde{\theta}_{aj}$ for $a = 1, \ldots, A$ and $j = 1, \ldots, J$, because of the invertibility of the logarithm function. □

GSPREE using a logit link

This lemma proposes equations to obtain the GSPREE using a logit link. It is referenced in Section 1.3.2.1, page 17.

Lemma 2. *Define the logit function with reference category* r,

$$\rho^{Y,r}_{aj} = \log \theta^{Y}_{aj} - \log \theta^{Y}_{ar}.$$

Assume analogous definitions for the proxy composition θ^{X}. *The two following sets of equations are equivalent:*

$$\rho^{Y,r}_{aj} = \phi_j + \beta(\alpha^{X}_{aj} - \alpha^{X}_{ar}), \quad \textit{for } a = 1, \ldots, A;\ j = 1, \ldots, J;\ j \neq r \tag{1.29}$$

$$\alpha^{Y}_{aj} = \beta \alpha^{X}_{aj}, \quad \textit{for } a = 1, \ldots, A;\ j = 1, \ldots, J. \tag{1.30}$$

Proof. Notice that

$$\rho_{aj}^{Y,r} = (\alpha_j^Y - \alpha_r^Y) + (\alpha_{aj}^Y - \alpha_{ar}^Y), \tag{1.31}$$

for $a = 1, \ldots, A;\ j = 1, \ldots, J;\ j \neq r$. Substitute (1.31) into (1.29) to obtain,

$$(\alpha_j^Y - \alpha_r^Y) + (\alpha_{aj}^Y - \alpha_{ar}^Y) = \varphi_j + \beta(\alpha_{aj}^X - \alpha_{ar}^X). \tag{1.32}$$

Adding (1.32) over the areas:

$$\alpha_j^Y - \alpha_r^Y = \varphi_j, \tag{1.33}$$

because $\alpha_{a+}^Y = \alpha_{a+}^X = 0$ for $a = 1, \ldots, A$. Substituting (1.33) back in (1.29):

$$\alpha_{aj}^Y - \alpha_{ar}^Y = \beta(\alpha_{aj}^X - \alpha_{ar}^X), \tag{1.34}$$

To prove the theorem it is enough to prove the equivalence between (1.34) and (1.30) because for any value of β it is possible to set $\varphi_j = (\alpha_j^Y - \alpha_r^Y)$ so that (1.33) is satisfied.

Adding (1.34) over j:

$$\sum_{j \neq r} \alpha_{aj}^Y - (J-1)\alpha_{ar}^Y = \beta \left(\sum_{j \neq r} \alpha_{aj}^X - (J-1)\alpha_{ar}^X \right)$$
$$\alpha_{ar}^Y = \beta \alpha_{ar}^X$$

because $\alpha_{a+}^Y = \alpha_{a+}^X = 0$ for $a = 1, \ldots, A$. As the choice of the reference category is arbitrary, we conclude that (1.34) $\Longrightarrow$ (1.30). Finally, (1.30) $\Longrightarrow$ (1.34) can be easily seen by subtracting the equation for the reference category from all the other equations in (1.30). □

GSPREE using a log link
This lemma proposes equations to obtain the GSPREE using a log link. It is referenced in Section 1.3.2.1 in page 17.

Lemma 3. *The set of equations*

$$\zeta_{aj}^Y = \gamma_a + \tilde{\lambda}_j + \tilde{\beta}\alpha_{aj}^X, \quad \textit{for } a = 1, \ldots, A;\ j = 1, \ldots, J \tag{1.35}$$

where $\zeta_{aj}^Y = \log Y_{aj}$, *is also equivalent to* (1.30).

Proof. Remember that $\zeta^Y_{aj} = \alpha^Y_0 + \alpha^Y_a + \alpha^Y_j + \alpha^Y_{aj}$ defined as in equation (1.15). Substituting this in (1.35) and summing either over rows or columns it is possible to arrive to the sets of equations:

$$\alpha^Y_0 + \alpha^Y_a = \gamma_a$$
$$\alpha^Y_j = \tilde{\lambda}_j$$
$$\alpha^Y_{aj} = \beta\alpha^X_{aj}.$$

for $a = 1, \ldots, A;\ j = 1, \ldots, J.$ □

The extra set of nuisance parameters γ_a represents the row scale of the table, therefore they do not appear in the equations where the link function is standardized with respect to the row scale, such as the centred-log used by the GLSM or the logit link.

1.3.3 Inclusion of random effects on SPREE-type estimators

Notice that none of the previously presented SPREE-type mixed effect estimators consider the possibility of including area-specific or column-specific random effects. The reason behind this is that only cell-specific random effects would have an impact on the final estimates, given the raking procedure that constitutes the last step of estimation for this type of estimators. For illustration, consider the fixed effects GSPREE as starting point and assume that $\alpha^Y_{aj} = \beta\alpha^X_{aj} + \nu_a$, where $\nu_a, a = 1, \ldots A$ are independent area-specific random effects with $E(\nu_a) = 0$ and $V(\nu_a) = \sigma^2_\nu$. An estimate of the interaction in cell a, j would have the form:

$$\hat{\alpha}^Y_{aj} = \hat{\beta}\alpha^X_{aj} + \hat{\nu}_a.$$

Denote by $\hat{Y}^0_{aj} = \exp(\hat{\alpha}^Y_{aj}) = \exp(\hat{\beta}\alpha^X_{aj})\exp(\hat{\nu}_a)$. Hence,

$$\hat{Y}^0_{a+} = \sum_j \hat{Y}^0_{aj} = \exp(\hat{\nu}_a)\sum_j \exp(\hat{\beta}\alpha^X_{aj}).$$

According to section 1.3.1, the set of estimates that satisfy the row margin Y_{a+} is obtained from the equation $\hat{Y}^1_{aj} = Y_{a+}\dfrac{\hat{Y}^0_{aj}}{\hat{Y}^0_{a+}}$. In this case,

$$\hat{Y}^1_{aj} = Y_{a+}\frac{\exp(\hat{\beta}\alpha^X_{al})\exp(\hat{\nu}_a)}{\exp(\hat{\nu}_a)\sum_j \exp(\hat{\beta}\alpha^X_{al})} = Y_{a+}\frac{\exp(\hat{\beta}\alpha^X_{al})}{\sum_j \exp(\hat{\beta}\alpha^X_{al})},$$

i.e., an area-specific random effect ν_a would have no effect on the estimates once they have been benchmarked by row. Analogously, column-specific random effects would disappear when the benchmarking of the column margins.

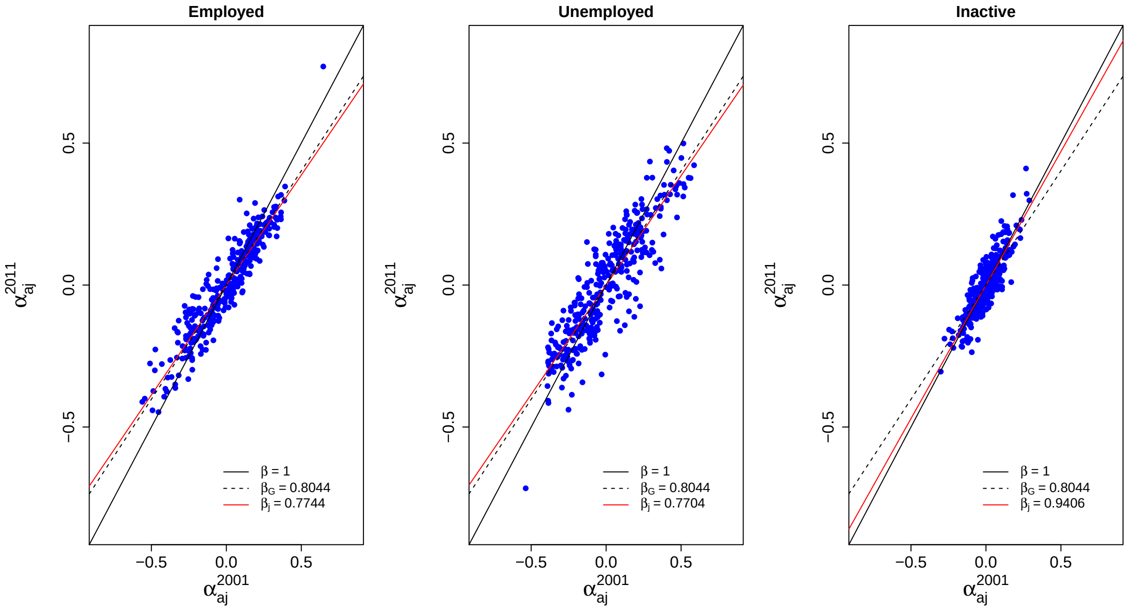
The inclusion of a set of cell-specific random effects at the interaction level is desirable but it poses the additional difficulty of imposing on the random effects the same sum-to-zero row and column constraints satisfied by the interaction terms under the centred constraints parametrisation, in order to obtain a valid structure for the table of interest. The estimation of the model parameters under such constraints can be computationally difficult, specially if a fully parametric approach is being used.

1.3.4 Limitations of the SPREE and GSPREE estimators

The previously discussed fixed effects SPREE estimators assume at maximum one parameter to control the relationship between the interaction structures of the proxy and target compositions. Further reductions in the bias of the estimator are obtained by the inclusion of random effects. However, the interactions terms corresponding to different categories of a given area are related to each other because they satisfy sum zero constraints, something that is ignored by the above mentioned SPREE-type estimators. Moreover, the same proportionality constant is assumed for all categories, which can lead to model misspecification if more than one constant of proportionality is required. Finally, notice that random effects cannot be accurately predicted if the sample sizes are too small.

We will illustrate these issues considering data from the Population Censuses 2001 and 2011 in England. For each census, a composition of the population counts disaggregated by LA and Economic Activity status (Employed, Unemployed, Inactive) was obtained. Such compositions where transformed into the log-scale and the interaction terms α_{aj} defined in equation (1.16) were calculated. Figure 1.1 presents scatter plots of the corresponding pairs of those interaction terms for each one of the categories of the labour force variable, being those of the 2001 composition in the X axis, and those of the 2011 composition in the Y axis. In each plot, the continuous black line represents the assumption made by the SPREE, i.e., that the interaction terms of both compositions are the same (Y=X). The black dashed line represents the assumption made by the GSPREE, i.e., that the interaction terms of the 2011 composition are proportional to those in the 2001 composition, with a proportionality constant that, for this data, is estimated as 0.8044. Finally, the continuous red line corresponds to an OLS fit independently for each category of the variable and corresponds to the best possible linear fit between the interaction terms for each category.

Figure 1.1: Interaction terms α_{aj} in the compositions of Local Authority by Economic Activity. Population censuses 2001 and 2011 in England.
Lines: Continuous black : SPREE. Dashed black: GSPREE. Continuous red: Category-specific OLS fit.



As the points in figure 1.1 do not accumulate above any of the black lines, it is possible to conclude that both SPREE and GSPREE estimators are biased. However, because the line corresponding to the GSPREE assumption is closer to the OLS fit for categories employed and unemployed, it is expected for the GSPREE to present less bias than the SPREE for those categories. Furthermore, as the slope of the OLS fit for category inactive (0.9406) is considerably different from the slopes for the other two categories (around 0.77), estimates for that category will much benefit from an estimator which allows for category-specific proportionality constants instead of imposing a common one.

Unfortunately, a natural extension based on an assumption such as $\alpha^{Y}_{aj} = \beta_j \alpha^{X}_{aj}$ would impose the additional constraint $\sum_j \beta_j \alpha^{X}_{aj} = 0$ for $a = 1, \ldots, A$, in order to ensure valid estimates of the interaction structure. Given that in practice

$A \geqslant J$, and because the column rank of the matrix that contains the interaction terms is at most $J-1$, such constraint cannot be satisfied in general.

1.4 Other estimators for population compositions

1.4.1 Multinomial models for the estimation of compositions

Multinomial GLMs have also been proposed for the estimation of population compositions. Molina et al. (2007) addresses the issue of estimating the proportions of people in three categories of labour force: *unemployed*; *employed* and *not in the labour force*, in a composition disaggregated by unitary and local authorities in the UK. Indicator variables of region, and age by sex groups, were used as covariates, as well as the proportion of unemployment according to administrative data which was included in the logarithmic scale. Molina et al. (2007) uses the logistic mixed model:

$$\log\left(\frac{\theta_{ijk}}{\theta_{ij3}}\right) = \mathbf{x}_{ij}^{T}\boldsymbol{\beta}_{k} + \vartheta_{i} \tag{1.36}$$

where k indexes the categories, i indexes the areas and $j = 1\ldots,6$ indexes the gender by age groups. The random effects ϑ_i are assumed independent and normally distributed with mean zero and variance σ^2_{ϑ}.

The estimation of the parameters of the model is performed using a combination of penalized quasi-likelihood (PQL, Breslow and Clayton (1993)) for the fixed terms of the model and the random effects, and REML for the variance components. Mean square error estimates of the small area estimates are obtained via parametric bootstrap.

Scealy (2010) extends the model proposed in Molina et al. (2007) to include category-specific random effects. In this case, the interest is in obtaining estimates for population counts according to the three categories of labour force mentioned above, by local government areas in Australia. Scealy (2010) indicates that initial analysis carried out using separate logistic mixed models for the three categories of the labour force variable, showed evidence of different variances in the random effects structure, which motivated the extension of equation (1.36) to include category-specific random effects, ϑ_{ik}, following a bivariate normal distribution in each area.

The model includes indicator variables of area, age by sex and remoteness, as well as household type and socio-economic indexes at the area level, obtained either from the Labour Force Survey or from the Census of Population and Housing. Moreover, benefit payment variables obtained from administrative sources are also included in the model. An approach analogous to the one presented in Molina et al. (2007) is used to obtain estimates for the parameters in the model and the random effects, as well as mean square estimates of the small area counts. Furthermore, the properties of the estimation strategy are also evaluated using parametric bootstrap.

Once again in the context of estimation of SAE of labour characteristics in the UK, Saei and Taylor (2012) proposes another extension of the model considered in Molina et al. (2007). Using a multinomial model with category-specific random effects following a bivariate normal distribution, An EBLUP-type predictor is developed. Starting with Pseudo-Likelihood estimates of the components of the model, ML is used to obtain final estimates of the fixed terms and predicted random effects, whereas the estimation of the variance components is performed via REML. Saei and Taylor (2012) compare two models with category-specific random effects (uncorrelated or potentially correlated) and the model with area-specific random effects proposed in Molina et al. (2007). According to their analysis, the model with correlated category-specific effects performs considerably better, in particular for the category *unemployed*.

Model (1.36) considered at the area level has also been used to propose extensions such as the inclusion of independent category-specific or time-correlated random effects. See López-Vizcaíno et al. (2013) and López-Vizcaíno et al. (2015) for more details.

1.4.2 Berg and Fuller (2014)

Berg and Fuller (2014) develops a small area estimator for a two way table containing the distribution of occupied people among categories of occupation and provinces in Canada. Direct estimates of the within area proportions are available from the Canadian Labour Force Survey, but due to the small sizes in the cells, they have prohibitively large variances. Auxiliary information is available, in the form of the same table obtained from the last population census. Indexing the areas by $k = 1, \dots, K$ and the categories of Occupation by

$i = 1, \dots, R$, the model proposed is defined by the two equations:

$$\hat{P}_{ik} = P_{ik} + e_{ik} \tag{1.37}$$

$$P_{ik} = P_{T,ik} + u_{ik} \tag{1.38}$$

where P_{ik} is the within-province proportion of interest, $\hat{P}_{ik}$ denotes its direct estimator and $P_{T,ik}$ is a function of the covariates, $P_{T,ik} = g(\mathbf{x}'_{ik}\boldsymbol{\lambda})$ such that two conditions are satisfied: i) $0 \leqslant g(\mathbf{x}'_{ik}\boldsymbol{\lambda}) \leqslant 1$ and ii) $\sum_{i=1}^{R} g(\mathbf{x}'_{ik}\boldsymbol{\lambda}) = 1$. In particular, Berg and Fuller (2014) uses a logistic function,

$$P_{T,ik} = g(\mathbf{x}'_{ik}\boldsymbol{\lambda}) = \frac{e^{x'_{ik}\lambda}}{\left(1 + \sum_{i=2}^{R} e^{x'_{ik}\lambda}\right)}$$

for $i = 2, \dots, R$ and $P_{T,1k} = 1 - \sum_{i=2}^{R} P_{T,ik}$. The vector $\mathbf{x}_{ik} = (x_{ik,1}, \dots, x_{ik,R})$ of covariates is composed by $R - 1$ indicator variables of categories $i = 2, \dots, R$ and the interaction term corresponding to the same cell in the auxiliary table, α^{X}_{aj}.

The sampling error component, $\boldsymbol{e}_k = (e_{1k}, \dots, e_{Rk})$ has $E(\boldsymbol{e}_k|\mathbf{P}_k) = \mathbf{0}$ and $V(\boldsymbol{e}_k|\mathbf{P}_k) = \boldsymbol{\Sigma}_{ee,k,c}$, where $\mathbf{P}_k = (P_{1k}, \dots, P_{Rk})$. Finally, $\mathbf{u}_k = (u_{1k}, \dots, u_{Rk})$ is a random component with $E(\mathbf{u}_k) = \mathbf{0}$ and $V(\mathbf{u}_k) = \boldsymbol{\Sigma}_{uu,k}$. It is assumed that $\boldsymbol{\Sigma}_{uu,k} = \psi \boldsymbol{\Gamma}_{\mathbf{P}_{T,k}}$, where $\boldsymbol{\Gamma}_{\mathbf{P}_{T,k}}$ is the variance covariance matrix of a multinomial distribution parametrized by $\mathbf{P}_{T,k} = (P_{T,1k}, \dots, P_{T,Rk})$. Besides simplifying considerably the model, this specification has two additional advantages: i) it ensures that $V(P_{ik}) \to 0$ when $P_{T,ik} \to 0$ and ii) because $\boldsymbol{\Sigma}_{uu,k}$ is not full rank, $\sum_i u_{ik}$ is constrained and as consequence $\sum_i P_{ik} = 1$.

Use the symbol $^{(1)}$ to indicate that in a vector or matrix, the first component or the first row and column, respectively, has been removed. Assuming $\boldsymbol{\lambda}, \psi$ and the unconditional variance-covariance matrix of the sampling errors $\boldsymbol{\Sigma}_{ee,k}$ known, Berg and Fuller (2014) shows that the BLUP of $\mathbf{P}_k$ is given by

$$\hat{\mathbf{P}}^{(1)}_{k} = \mathbf{P}^{(1)}_{T,k} + \boldsymbol{\Sigma}^{(1)}_{uu,k}(\boldsymbol{\Sigma}^{(1)}_{uu,k} + \boldsymbol{\Sigma}^{(1)}_{ee,k})^{-1}(\hat{\mathbf{P}}^{(1)}_{k} - \mathbf{P}^{(1)}_{T,k}). \tag{1.39}$$

The predictor for the first category can be obtained by subtraction. Furthermore, the same BLUP estimates do not depend on the category being removed. Unfortunately, the predictor in equation (1.39) may not remain in the interval $(0, 1)$. For the cases where the predictor falls outside the parameter space, Berg and Fuller (2014) propose to obtain the predictor $\mathbf{P}_k^{*(1)}$ that minimises

$$(\hat{\mathbf{P}}_k^{(1)} - \mathbf{P}_k^{*(1)})^{\mathsf{T}} \left[\boldsymbol{\Sigma}_{uu,k}^{(1)} + \boldsymbol{\Sigma}_{ee,k}^{(1)}\right]^{-1} (\hat{\mathbf{P}}_k^{(1)} - \mathbf{P}_k^{*(1)}),$$

subject to $\delta_{ik} \leqslant P_{ik}^* \leqslant 1 - \delta_{ik}$, where δ_{ik} are specified positive constants.

Furthermore, assuming that $V(\boldsymbol{e}_k|\mathbf{u}_k) = n_k^{-1} d_k[\mathrm{diag}(\mathbf{P}_k) - \mathbf{P}_k\mathbf{P}_k^{\mathsf{T}}]$, where n_k is the total sample size in province k and d_k plays the role of an average design effect (Rao and Scott, 1981) for that province, we obtain

$$\boldsymbol{\Sigma}_{ee,k} = E\left[V(\boldsymbol{e}_k|\mathbf{u}_k)\right] = n_k^{-1} c_k[\mathrm{diag}(\mathbf{P}_{T,k}) - \mathbf{P}_{T,k}\mathbf{P}_{T,k}^{\mathsf{T}}],$$

where $c_k = d_k(1-\psi)$. Under this working model, the predictor in (1.39) simplifies to:

$$\tilde{P}_{ik} = P_{T,ik} + \gamma_k(\hat{P}_{ik} - P_{T,ik}), s$$

with $\gamma_k = \psi/(\psi + c_k n_k^{-1})$. Notice that for this working model, as γ_k does not depend on the category, the predicted $\tilde{P}_{ik}$ automatically belong to the interval $(0, 1)$.

The estimators of $\boldsymbol{\lambda}$, $P_{T,ik}$, c_k and ψ are obtained iteratively, using either Maximum Likelihood or Generalized Least Squares. The estimation of the MSE of the predictor without benchmark can be performed using a closed form approximation that is proposed in Berg and Fuller (2014). If benchmark has been used, an adaptation from the moment-matching bootstrap approach considering the first two moments is proposed.

Discussion

Berg and Fuller (2014) presents a model for proportions in a population composition that is simple but very flexible. In principle, no restrictions are imposed on the type of covariate information that is allowed in the model because the constrains satisfied by $g(\cdot)$ ensure that the synthetic estimator $P_{T,ik}$ belongs to the parameter space. In particular, the choice of $g(\cdot)$ and x_{ik} presented at the beginning of this section, induce on the $P_{T,ik}$ the structural assumption of the GSPREE.

The proposed predictors for the small area proportions are built over a very general form for the variance covariance matrix of the sampling errors and random effects. Particular working models suggested in Berg and Fuller (2014) lead to substantial simplifications without compromising the efficiency of the predictors. The proposed methods for estimating the parameters of the model are not computationally intensive.

The inclusion of random effects at the probability scale, as in equation (1.38), is convenient but may allow for within-province proportions of interest P_{ik} which are not necessarily between 0 and 1. The first predictor proposed, derived as the BLUE for the proportions of interest, may also provide estimates outside the parameter space. However, an alternative predictor with the desirable property of being close to the BLUE is proposed. Furthermore, this problem can avoided by using the working model proposed for $\boldsymbol{\Sigma}_{ee,k}$, as long as the data supports this alternative.

1.4.3 Dostál et al. (2016)

Another estimator for population compositions was proposed by Dostál et al. (2016) as a way to obtain coverage corrections by areas and subgroups of the population, for the German census 2011. The proposed estimator is an extension of the SPREE of Purcell and Kish (1980), which minimises the Chi-square distance to the proxy composition while satisfying constraints given by sets of estimated margins.

For a given composition $\mathbf{M}$ with cells m_{ij}, for $i = 1, \dots, I$; $j = 1, \dots, J$, the idea is to obtain an estimator $\mathbf{N}$ with cells n_{ij} such that the distance

$$d = \sum_i \sum_j \frac{\left(\frac{n_{ij}}{n_{++}} - \frac{m_{ij}}{m_{++}} \right)^2}{m_{ij}}$$

is mimimal, subject to constraints on the row and column margins n_{i+} and n_{+j}. When only one margin is fixed, the solution to this problem is given by the SPREE (see Purcell and Kish, 1980, Section 4) and can be calculated using IPF.

Using Lagrange multipliers, Dostál et al. (2016) show that when two margins are fixed, the solution to the minimisation problem above is given by

$$\hat{n}_{ij} = \left(\lambda_i + \mu_j \right) m_{ij}, \tag{1.40}$$

where the vector of parameters $\boldsymbol{\psi}^{\mathsf{T}} = (\lambda_1, \dots, \lambda_I, \mu_1, \dots, \mu_J)$ is estimated as the solution of the linear system $\mathbf{A}\boldsymbol{\psi} = \mathbf{b}$, for

$$\mathbf{A} = \left[\begin{array}{c:c} \text{diag}(m_{i+}) & \mathbf{M} \\ \hdashline \mathbf{M} & \text{diag}(m_{+j}) \end{array}\right]$$

and $\mathbf{b}^T = (n_{1+}, \ldots, n_{I+}, n_{+1}, \ldots, n_{+J})$. Because $\mathbf{A}$ is not full rank, a Moore-Penrose inverse of $\mathbf{A}$ is used to solve the linear system. All solutions are given by $\hat{\boldsymbol{\psi}} = \mathbf{A}^+\mathbf{b} + \left(\mathbf{I} - \mathbf{A}^+\mathbf{A}\right)\mathbf{Z}$, with $\mathbf{A}^+$ the Moore-Penrose inverse of $\mathbf{A}$ and $\mathbf{Z}$ a free vector. The λ_i and μ_j are not identifiable, but $(\lambda_i + \mu_j)$ is unique.

Discussion

The estimator proposed in Dostál et al. (2016) only assumes the existence of a proxy table $\mathbf{M}$ and the true sets of margins n_{i+} and n_{+j}. The methodology proposed is very convenient because the estimator has a closed form. Furthermore, as no particular treatment is given to the rows or columns of the composition, no assumption of exchangeability across rows is required, therefore it can be use in a wide range of situations.

In order to study the structural assumption underpinning the estimator, taking the logarithm on both sides of equation (1.40) notice that the counts of the proxy composition $\mathbf{M}$ act as an offset in determining the association structure of the estimated composition, as in the case of the SPREE. However, because the term $(\lambda_i + \mu_j)$ is cell-specific, it induces an *update* of the association structure which is governed by $I + J$ parameters. Dostál et al. (2016) report encouraging results for the proposed estimator in a simulation study, except for outliers respect to the association structure. However, considering that the only estimates for the cells that are assumed to be available correspond to the proxy composition, it seems somehow strange to use only marginal information to update the association structure rather than preserving the one observed in the proxy composition. Further research, for instance comparing this estimator with the estimators proposed in this document may be desirable.

Chapter 2

MSPREE

This chapter introduces the Multivariate SPREE (MSPREE). The chapter begins with the introduction of the structural assumption underpinning the estimator. Then, alternatives to estimate the matrix of parameters that governs the estimator are presented. Finally, an analytic estimator to approximate the variance of the estimator, and bootstrap procedures to estimate its MSE are proposed. In what follows, it will be assumed that the target composition $\mathbf{Y}$ and the proxy composition $\mathbf{X}$, both of dimension $A \times J$ with $A > J$, have no structural zeroes, i.e., it will be assumed that all cell counts (or their expectations, if $\mathbf{Y}$ is assumed random) are strictly positive. Moreover, it will be assumed that the two true sets of margins $\mathbf{Y}_{a+} = (Y_{1+}, \cdots, Y_{A+})$ and $\mathbf{Y}_{+j} = (Y_{+1}, \cdots, Y_{+J})$ are known. Furthermore, it will be assumed that sample data about $\mathbf{Y}$, in the form of a composition of direct estimates $\hat{\mathbf{Y}}$ or sample counts $\mathbf{y}$ or the corresponding within-area proportions, is available.

Throughout this chapter, whenever a particular lemma or theorem is required to support a given assertion, this is indicated including the corresponding theorem number and page. However, in order to avoid breaking the flow of the text, all lemmas and theorems and their respective proofs are presented as part of the Complementary Material, in subsection 2.4.2, at the end of this chapter. To ease the cross-referencing between lemmas and theorems and the results they are linked to, a paragraph has been added before the enunciate of each lemma or theorem, indicating the context and pages where they are being used.

2.1 Multivariate SPREE (MSPREE)

The Multivariate SPREE is built upon the assumption that, for each area a, the set of interactions of the target composition, $\boldsymbol{\alpha}_a^Y = (\alpha_{a1}^Y, \ldots, \alpha_{aJ}^Y)^T$, is a linear combination of the interactions of the same area in the auxiliary composition, $\boldsymbol{\alpha}_a^X = (\alpha_{a1}^X, \ldots, \alpha_{aJ}^X)^T$, with coefficients that may vary from one category to another but are common for all areas (between area exchangeability). The *structural assumption* of the MSPREE can hence be written as:

$$\begin{array}{ccccc} \begin{bmatrix} \alpha_{a1}^Y \\ \vdots \\ \alpha_{aJ}^Y \end{bmatrix} & = & \begin{bmatrix} \beta_{11} & \cdots & \beta_{1J} \\ \vdots & \ddots & \vdots \\ \beta_{J1} & \cdots & \beta_{JJ} \end{bmatrix} & \begin{bmatrix} \alpha_{a1}^X \\ \vdots \\ \alpha_{aJ}^X \end{bmatrix} \\ \boldsymbol{\alpha}_a^Y & = & \boldsymbol{\beta} & \boldsymbol{\alpha}_a^X, \end{array} \tag{2.1}$$

for $a = 1, \ldots, A$, or equivalently, as

$$\alpha_{aj}^Y = \sum_l \beta_{jl} \alpha_{al}^X \,, \tag{2.2}$$

for $a = 1, \ldots, A$ and $j, l = 1, \ldots, J$.

Lemma 4 in page 47 shows that under equation (2.2), the set of constraints $\beta_{+l} = \beta_{j+} = 0$ for $j, l = 1, \ldots, J$, are required in order to ensure a valid association structure for $\mathbf{Y}$, i.e., to ensure that the set of α_{aj}^Y above defined satisfies the sum-zero constraints of the centred-constraints parametrisation. Notice that despite $\boldsymbol{\beta}$ being a $J \times J$ matrix, such constraints imply that the relationship between the two association structures is characterized by $(J-1)^2$ free parameters. Moreover, because the interaction terms are centred around zero and the matrix $\boldsymbol{\alpha}^Z$ containing as rows the vectors $\boldsymbol{\alpha}_a^Z$ for $Z \in \{\mathbf{X}, \mathbf{Y}\}$ has a maximum rank of $J-1$, as long as exchangeability between areas is assumed, $(J-1)^2$ is the maximum number of parameters that can characterise a linear relationship between two compositions in the interaction scale.

The MSPREE of $\mathbf{Y}$ is obtained by imposing the allocation structure determined by the margins $\mathbf{Y}_{a+}$ and $\mathbf{Y}_{+j}$, on an estimate of the association structure obtained under the structural assumption (2.2), via IPF, i.e.:

$$\hat{\mathbf{Y}}^M = \text{IPF}\left(\exp \hat{\boldsymbol{\alpha}}^Y ; \mathbf{Y}_{a+} ; \mathbf{Y}_{+j}\right), \tag{2.3}$$

where for each area:

$$\hat{\boldsymbol{\alpha}}_a^Y = \hat{\boldsymbol{\beta}}\boldsymbol{\alpha}_a^X. \tag{2.4}$$

Several alternatives to obtain $\hat{\boldsymbol{\beta}}$ will be proposed in section 2.2.

The MSPREE has been called *Multivariate* in an intention to emphasize that, in each area, the quantity of interest is the *vector* of interactions of the target composition, in contrast with all existing estimators under the SPREE approach, that deal with the interaction terms in an univariate fashion. Approaching the problem in a multivariate way may lead to potential gains in efficiency derived from the association between the sample counts or direct estimates in each area, in an analogous way to the Multivariate Fay-Herriot model of Datta et al. (1991). Furthermore, a multivariate approach provides a mathematically convenient way of generalising the structural assumptions of SPREE and GSPREE. Those two estimators are, indeed, special cases of the proposed MSPREE. In order to justify this claim, it is enough to prove that the structural assumption stated in equation (2.2) covers the assumptions underpinning SPREE and MSPREE as special cases because, given an estimated association structure for the target composition, the procedure used to impose the known margins is the same for all estimators. Such a proof is presented in Lemma 6 in page 49 and its corollary.

The definition of the MSPREE given in section 2.1 is not complete until we have proposed a methodology to obtain an estimate of the matrix of parameters $\boldsymbol{\beta}$. In the next section, we will discuss the issues associated with an attempt to produce such estimates directly from equation (2.2) and, in consequence, propose a methodology based on alternative expressions of this assumption and the use of ML or IWLS.

2.2 Estimation of $\boldsymbol{\beta}$

The formulation of a statistical model to obtain estimates of the matrix $\boldsymbol{\beta}$ directly from the structural assumption of the MSPREE, i.e., in the interaction scale, would require a sample estimate of the matrix $\boldsymbol{\alpha}^Y$ containing the terms α_{aj}^Y for $a = 1, \dots, A$ and $j = 1, \dots, J$. An intuitive estimator is given by $\hat{\boldsymbol{\alpha}}^Y = \boldsymbol{\alpha}^{\hat{Y}}$, the association structure of the composition of direct estimates. However, given that the sample sizes in the SA are assumed to be small, it is not unlikely that some of the sample counts in the cells are zero despite the assumption of no zeroes in $\mathbf{Y}$, a situation in which the association structure of neither $\hat{\mathbf{Y}}$ nor $\mathbf{y}$ can be calculated. Moreover, even if $\mathbf{y}$ and hence $\hat{\mathbf{Y}}$ contain no zeroes, the logarithmic transformation required to obtain $\boldsymbol{\alpha}^{\hat{Y}}$ from $\hat{\mathbf{Y}}$ would imply bias for the former due to non-linearity. Therefore, a methodology of estimation of the β_{jl} using either $\hat{\mathbf{Y}}$ or $\mathbf{y}$ or the corresponding proportions would be preferred.

We present in this section two alternative equations for the MSPREE assumption, which can be used in conjunction with suitable distributional assumptions in order to obtain estimates of the β_{jl} via ML or IWLS. The validity of this approach relies on the equivalence between the parameters of the alternative equations and those of the structural assumption of the MSPREE, which is proved in Theorems 7 and 8 in pages 50 and 51, at the end of this chapter.

2.2.1 Alternative equations

Log link

$$\zeta^{Y}_{aj} = \log Y_{aj} = \gamma_a + \lambda_j + \sum_{l} \beta_{jl}\alpha^{X}_{al}\,, \tag{2.5}$$

for $a = 1,\ldots,A$ and $j,l = 1,\ldots,J$ with $\beta_{j+} = \beta_{+l} = \lambda_+ = 0$. Denoting by $\text{vec}(\cdot)$ the vector operator, which transforms a matrix into a column vector by stacking its columns, equation (2.5) can be written in matrix notation as

$$\boldsymbol{\zeta}^{Y} = \mathbf{Z}_{\log}\boldsymbol{\Psi}_{\log},$$

where $\boldsymbol{\zeta}^{Y} = \text{vec}(\log(\mathbf{Y}))$, $\mathbf{Z}_{\log}$ is a design matrix of dimension $AJ \times (A + J(J-1))$ given by:

$$\mathbf{Z}_{\log} = \left[\begin{array}{c:c:c} \mathbf{1}_{(J\times 1)} \otimes \mathbf{I}_{(A)} & \mathbf{T} \otimes \mathbf{1}_{(A\times 1)} & \mathbf{T} \otimes (\boldsymbol{\alpha}^{X}\mathbf{T}) \end{array}\right], \tag{2.6}$$

with $\boldsymbol{\alpha}^{X}$ the $A \times J$ matrix containing the interactions of the auxiliary composition, $\otimes$ denoting the Kronecker product and $\mathbf{T}$ defined as

$$\mathbf{T}_{(J\times(J-1))} = \left[\begin{array}{c} \mathbf{I}_{(J-1)} \\ \hdashline -\mathbf{1}_{1\times(J-1)} \end{array}\right]. \tag{2.7}$$

$\boldsymbol{\Psi}_{\log}$ is the column vector of parameters:

$$\boldsymbol{\Psi}_{\log} = \left[\begin{array}{ccccccccccc} \gamma_1 & \ldots & \gamma_A & \lambda_1 & \ldots & \lambda_{J-1} & \beta_{1,1} & \beta_{1,2} & \ldots & \beta_{J-1,1} & \beta_{J-1,J-1} \end{array}\right]^{T}.$$

The constraints $\beta_{j+} = \beta_{+l} = 0$ and $\lambda_+ = 0$ have been implicitly included in the design matrix, by setting $\beta_{jJ} = -\sum_{l<J} \beta_{jl}$, $\beta_{Jl} = -\sum_{j<J} \beta_{jl}$ for $j, l = 1, \ldots, J$ and $\lambda_J = -\sum_{j<J} \lambda_j$, respectively. Neither $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ or their estimates depend on which subset of $J-1$ categories has free parameters.

Logit link

$$\rho_{aj}^{Y,r} = \log\left(\frac{\theta_{aj}^{Y}}{\theta_{ar}^{Y}}\right) = \phi_j^r + \sum_{l \neq r} \tilde{\beta}_{jl}^r (\alpha_{al}^X - \alpha_{ar}^X), \tag{2.8}$$

with $\theta_{aj}^Y = Y_{aj}/Y_{a+}$ for $a = 1, \ldots, A$ and $j, l = 1, \ldots, J$, $j, l \neq r$. Using category J as reference and stacking $\mathbf{Y}$ by column, equation (2.8) can be written in matrix notation as:

$$\boldsymbol{\rho}^{Y,J} = \mathbf{Z}_{\text{logit}}^{J} \boldsymbol{\Psi}_{\text{logit}}^{J},$$

where

$$\boldsymbol{\rho}^{Y,J} = \left[\begin{array}{ccccccc} \log(\theta_{1,1}^Y/\theta_{1,J}^Y) & \ldots & \log(\theta_{A,1}^Y/\theta_{A,J}^Y) & \ldots & \log(\theta_{1,J-1}^Y/\theta_{1,J}^Y) & \ldots & \log(\theta_{A,J-1}^Y/\theta_{A,J}^Y) \end{array}\right]^T,$$

and $\mathbf{Z}_{\text{logit}}^{J}$ is a design matrix of dimension $A(J-1) \times (J(J-1))$ given by:

$$\mathbf{Z}_{\text{logit}}^{J} = \left[\begin{array}{c:c} \mathbf{I}_{(J-1)} \otimes \mathbf{1}_{(A\times 1)} & \mathbf{1}_{(A\times 1)} \otimes (\boldsymbol{\alpha}^X \mathbf{T}) \end{array}\right], \tag{2.9}$$

with $\boldsymbol{\alpha}^X$ and $\mathbf{T}$ defined as in (2.7). $\boldsymbol{\Psi}_{\text{logit}}^{J}$ is the column vector of parameters:

$$\boldsymbol{\Psi}_{\text{logit}}^{J} = \left[\begin{array}{cccccccc} \phi_1^r & \ldots & \phi_{(J-1)}^r & \tilde{\beta}_{1,1}^J & \tilde{\beta}_{1,2}^J & \ldots & \tilde{\beta}_{J-1,1}^J & \tilde{\beta}_{J-1,J-1}^J \end{array}\right]^T.$$

In general, the vectors of parameters corresponding to two different reference categories differ. However, there is a one-to-one relationship between the set of parameters $\tilde{\beta}_{jl}^r$ corresponding to any reference category r, and the set of β_{jl} used in the definition of the structural assumption of the MSPREE (equation 2.2). This equivalence, which is presented as part of the proof of Theorem 8 (equation (2.28), page 52), ensures that the matrix $\boldsymbol{\beta}$ in equation 2.1 is invariant to the reference category chosen.

Further remarks

It was mentioned at the beginning of this section that the MSPREE of $\mathbf{Y}$ could be obtained by imposing the known margins on an estimate of its association structure given by $\exp\hat{\boldsymbol{\alpha}}^Y = \left\{\exp\hat{\boldsymbol{\alpha}}_1^Y, \ldots, \exp\hat{\boldsymbol{\alpha}}_A^Y\right\}^T$ where for each area, $\hat{\boldsymbol{\alpha}}_a^Y = \hat{\boldsymbol{\beta}}\boldsymbol{\alpha}_a^X$, with $\hat{\boldsymbol{\beta}}$ obtained directly from equation (2.5), or from (2.8) using equation (2.28) (page 52). Notice that the same estimate would be obtained if such margins were imposed directly on the composition of fitted counts or within area proportions because all the parameters related to the allocation structure would be recalculated in the IPF process. In this sense, the additional parameters γ_a and λ_j, or ϕ_j^r, depending on which of the two alternative equations is used, can be considered nuisance parameters.

2.2.2 ML estimates

ML estimates of β_{jl} for $j, l = 1, \ldots, J$ can be obtained making suitable distributional assumptions for Y_{aj} and using the alternative equations to formulate an associated GLM, as follows:

- Assuming $Y_{aj}|\boldsymbol{\alpha}_a^X \overset{\text{ind}}{\sim} \text{Poisson}(\mu_{aj})$ for $a = 1, \ldots, A$, $j = 1, \ldots, J$. After substituting Y_{aj} by μ_{aj}, equation (2.5) defines a Poisson Regression Model.
- Assuming $\mathbf{Y}_a|\boldsymbol{\alpha}_a^X \overset{\text{ind}}{\sim} \text{Multinomial}\left(Y_{a+}, \boldsymbol{\pi}_a^Y\right)$ for $a = 1, \ldots, A$. Substituting θ_{aj} by π_{aj}, equation (2.8) defines a Multinomial Logistic Regression Model.

Both alternatives are equivalent, i.e., they lead to the same estimate of the association structure, because a term for the margin that is assumed fixed under the product multinomial distribution is included in (2.5) (see Birch, 1963).

The use of a Poisson or Multinomial likelihood to obtain MLE estimates of $\boldsymbol{\beta}$ is practically convenient because such models can be fitted using standard software. An illustration using the functions `glm` and `mlogit` (packages `stats` and `mlogit` in the software **R**) is included in section 2.4.3, at the end of this chapter. However, if the sample has been obtained using a complex sampling procedure, this approach can lead to model misspecification, as will be discussed next.

First, the use of a complex sampling procedure may result in *informativeness* of the sampling design (see for instance Pfeffermann 1993, Chambers and Clark 2012, Section 1.4). This term has been used to indicate the situation where the model which holds for the sample data is different from the model holding for the population from which such sample has been selected. In the context of model-based estimation, and specially in SAE, accounting for informativeness is of key importance in order to avoid bias in the SA predictors.

Perhaps the most common situation that induces informativeness is the use of sampling designs with unequal selection probabilities, whenever such probabilities are related to the study variable of interest. This can occur either at the area level, if not all areas are sampled, or at a lower level, when households or individuals are selected. For instance, sample boosts which target individuals in a given domain, in order to produce estimates of a minimum level of accuracy, may result in informative sampling designs.

Regarding the use of SPREE-type estimators, notice that as long as all the information concerning the design of the sample is available, it would be possible to account for informativeness below the area level, by using appropriate direct estimators for the cells in the composition, e.g, Horvitz-Thompson estimators which weight each observation by the inverse of its inclusion probability (Särndal et al., 1992, Section 2.8). However, it would be difficult in that case, to justify a fully distributional assumption for the estimators in the sample, and therefore an IWLS approach instead of ML would be recommended. That route of analysis will be discussed below, in section 2.2.3.

Furthermore, notice that because the row margins of the target composition are assumed known, the use of different sampling fractions across areas would not necessarily constitute informativeness when using SPREE-type estimators, as long as the within-area proportions observed in the sample remain unbiased for the corresponding population quantities. In such a case, a ML approach using the sample composition could be used to obtain the MSPREE. Unfortunately, due to the assumption of exchangeability across areas, neither the existing nor the new estimators proposed in this document could take into account informativeness at the area level.

Second, even if informativeness is not an issue, it is still possible that the variance structure assumed under a Poisson or Multinomial likelihood does not represent appropriately the variability observed in the sampling data, e.g., due to the selection of clusters instead of individual units, as in multi-stage sampling. Model misspecification in this sense would affect the efficiency of the estimators of β_{jl}, but would not compromise their asymptotic unbiasedness (see Liang and Zeger, 1986). Alternative methods have been proposed in the literature to deal with the modelling of count data with *over-dispersion*, for instance, the use of Quasi-Likelihood functions (Wedderburn 1974, McCullagh 1983) or negative binomial regression (see Agresti, 2013, section 14.4). Studying the efficiency of any of those approaches is considered beyond the scope of this document.

On a final note, in accordance with what was mentioned before about the difficulties of obtaining estimates of the β_{jl} directly from equations (2.1) or (2.2), notice that a MLE approach in the interaction scale would be difficult to justify given that there is not an intuitive candidate for the joint distribution of the α^{Y}_{aj}. On the other hand, MLE seems a natural alternative when considering the MSPREE assumption at the proportion or count scale via the alternative equations.

2.2.3 IWLS estimates

The discussion above illustrates why, in some cases, a fully distributional approach could be undesirable for the purposes of obtaining the estimates of the matrix of parameters $\boldsymbol{\beta}$. On the other hand, considering the target composition $\mathbf{Y}$ as fixed and assuming that the $\boldsymbol{\beta}$ that better represents the relationship between $\mathbf{Y}$ and $\mathbf{X}$ is also the one that better represents the relationship between the composition of direct estimates of the counts $\hat{\mathbf{Y}}$ or proportions $\hat{\boldsymbol{\theta}}^{Y}$ and $\mathbf{X}$, $\hat{\mathbf{Y}}$ and an estimate of its variance-covariance matrix are enough to obtain estimates of the parameters of interest via IWLS.

For the sake of completeness, we will briefly describe the procedure here. For a more detailed explanation, see for instance Jiang (2007), Section 1.4.3. For this description, let us assume that $\mathbf{Y}$ is a random vector with n components, such that $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $V(\mathbf{Y}) = \mathbf{V}$. We are interested in obtaining estimates for $\boldsymbol{\Psi}$, the vector of unknown parameters of the (possibly non-linear) model:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\Psi} \tag{2.10}$$

where $\boldsymbol{\eta}$ is a $n \times 1$ vector, g is an invertible link function and $\mathbf{Z}$ is a design matrix of dimension $n \times p$. A linear approximation of $g(\mathbf{Y})$ around $\boldsymbol{\mu}$ is given by the expression:

$$\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta} + g'(\boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu}) \tag{2.11}$$

where g' is the matrix of first derivatives of g with respect to $\mathbf{Y}$, i.e., $g'_{i,j} = \dfrac{\partial \eta_i}{\partial Y_j}$. Considering (2.10) and (2.11) together, it is clear than an estimate for $\boldsymbol{\Psi}$ can be obtained from the linear model:

$$\tilde{\boldsymbol{\eta}} = \mathbf{Z}\boldsymbol{\Psi} + \boldsymbol{e} \tag{2.12}$$

for $\boldsymbol{e} = g'(\boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})$. Notice that $\mathrm{E}(\boldsymbol{e}) = \mathbf{0}$ and:

$$\mathrm{V}(\boldsymbol{e}) = \left(g'(\boldsymbol{\mu})\right)\ \mathbf{V}\ \left(g'(\boldsymbol{\mu})\right)^{\mathsf{T}} := \boldsymbol{W}^{-1}.$$

In general, $\boldsymbol{W}^{-1}$ is not proportional to the identity matrix, therefore Weighted Least Squares (WLS) are necessary in order to fit (2.12). As both $\tilde{\boldsymbol{\eta}}$ and $\boldsymbol{W}$ depend on the unknown $\boldsymbol{\mu}$, an iterative algorithm such as the following is required:

1. Define an initial estimate of $\boldsymbol{\mu}$, denoted by $\boldsymbol{\mu}^*$

2. Calculate $\tilde{\boldsymbol{\eta}} = g(\boldsymbol{\mu}^*) + g'(\boldsymbol{\mu}^*)(\boldsymbol{y} - \boldsymbol{\mu}^*)$, where $\boldsymbol{y}$ is a sample observation of the random vector Y

3. Calculate $\boldsymbol{W}^* = \left[(g'(\boldsymbol{\mu}^*))\ \mathbf{V}\ (g'(\boldsymbol{\mu}^*))^{\mathsf{T}}\right]^{-1}$. If unknown, $\mathbf{V}$ can be replaced by a consistent estimate

4. Obtain an estimate of the vector of parameters, $\boldsymbol{\Psi}^* = \left(\mathbf{Z}^{\mathbf{T}}\boldsymbol{W}^*\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathbf{T}}\boldsymbol{W}^*\tilde{\boldsymbol{\eta}}$

5. Calculate a new estimate of $\mathrm{E}(\mathbf{Y})$, $\boldsymbol{\mu}^* = g^{-1}(\mathbf{Z}\boldsymbol{\Psi}^*)$

6. Repeat from 2 until convergence is achieved.

Under mild conditions, Jiang et al. (2007) showed that: i) the IWLS algorithm converges with probability tending to one as the sample size increases and ii) the limit estimator of $\boldsymbol{\Psi}$ is consistent and as efficient as the BLUE asymptotically. In this case "sample size" refers to the number of rows, A. Furthermore, the variance-covariance matrix of the estimated parameters can be estimated by $\hat{\mathbf{V}}_{\hat{\boldsymbol{\Psi}}} = (\mathbf{Z}^{\mathbf{T}}\hat{\boldsymbol{W}}^{-\mathbf{1}}\mathbf{Z})^{-1}$.

Turning back to the MSPREE, direct estimators and their observed values play the roles of the random vector $\mathbf{Y}$ and its sample observation $\boldsymbol{y}$, and the fixed composition acts as $\boldsymbol{\mu}$ in the above description. Notice that it is being implicitly assumed that the direct estimators are unbiased or approximately unbiased. An initial estimate of $\boldsymbol{\mu}$ can be obtained using the classic SPREE, i.e., imposing the known margins on the auxiliary composition via IPF.

The alternative equations presented in section 2.2.1 can be used to set up an IWLS algorithm to obtain estimates of the matrix $\boldsymbol{\beta}$ as follows:

Using the Log link
In this case $\boldsymbol{\eta} = g(\mathbf{Y}) = \log(\mathbf{Y})$ and $\partial\eta_{aj}/\partial Y^*_{sl} = (Y^*_{aj})^{-1}$ if $s = a; l = j$ and zero otherwise, i.e., the matrix of first derivatives is diagonal. $\mathbf{Y}^*$ is an initial estimate of $\mathbf{Y}$ in the first iteration and is obtained as $\mathbf{Y}^* = \exp(\mathbf{Z}_{\log}\hat{\boldsymbol{\Psi}}_{\log})$ in the following iterations.

Using the Logit link
Assuming category J as the reference and considering the target and proxy compositions stacked by column, $\boldsymbol{\eta} = g(\boldsymbol{\theta}^Y) = \boldsymbol{\rho}^{Y,J}$ as defined in section 2.2.1. The matrix of first derivatives has components:

$$\frac{\partial\eta_{aj}}{\partial\theta^*_{sl}} = \begin{cases} (\theta^*_{aj})^{-1} + (\theta^*_{aJ})^{-1} & \text{if} \quad s = a;\ l = j \\ (\theta^*_{aJ})^{-1} & \text{if} \quad s = a;\ l \neq j \\ 0 & \text{if} \quad s \neq a;\ l \neq j \end{cases}$$

for $a, s = 1, \ldots, A$, $j, l = 1, \ldots, J-1$. The matrix of first derivatives $g'(\boldsymbol{\theta}^*)$, of dimension $A(J-1) \times A(J-1)$, is formed by blocks of diagonal matrices of dimension $A \times A$. $\boldsymbol{\theta}^*$ has entries θ^*_{aj} that correspond to initial estimates of θ^Y_{aj} in the first iteration and are defined by:

$$\theta^*_{aj} = \begin{cases} \dfrac{\exp\eta^*_{aj}}{1 + \sum\limits_{j}^{J-1}\exp\eta^*_{aj}} & \text{for} \quad j \neq J \\ \dfrac{1}{1 + \sum\limits_{j}^{J-1}\exp\eta^*_{aj}} & \text{for} \quad j = J \end{cases}$$

in the following iterations. In each iteration, the matrix $\boldsymbol{\eta}^*$ is the estimated linear predictor $\mathbf{Z}^{\mathbf{J}}_{\text{logit}}\boldsymbol{\Psi}^{*\mathbf{J}}_{\text{logit}}$.

IWLS with V proportional to a Multinomial covariance matrix
Direct estimates of the variances and covariances of the direct estimators can be very unstable when the sample sizes in the different areas are small. Generalized Variance Functions (see Wolter, 2007, Chapter 7) or some type of smoothing of the variance estimates prior to the estimation of $\boldsymbol{\beta}$ may be nece-

ssary in order to address this issue.

Moreover, in many practical situations, an estimate of $\mathbf{V} = V(\hat{\mathbf{Y}})$, may not be available at all. Using the logit-link, a possible solution in that case is to ignore any correlation among estimates belonging to different areas and use the covariance matrix of a product multinomial distribution parametrized by $\hat{\theta}^Y$ and $\mathbf{Y}_{a+}$, multiplied by a scalar, in place of $\mathbf{V}$. This approach coincides with the idea of using the variance estimate corresponding to a simple random sample without replacement (SRSWOR) selected independently in each area, multiplied by a design effect, Δ. An estimate of the covariance matrix of $\hat{\theta}^Y$ in area a would hence have entries:

$$\hat{V}_{a,jl} = \begin{cases} \dfrac{\Delta\,\hat{\theta}_{aj}(1-\hat{\theta}_{aj})}{n_a} & \text{if} \quad j = l \\ \dfrac{-\Delta\,\hat{\theta}_{aj}\hat{\theta}_{al}}{n_a} & \text{if} \quad j \neq l \end{cases}$$

2.3 MSE estimation

Five alternatives will be proposed to estimate a measure of uncertainty of the MSPREE. In this section we will introduced three of them, assuming that equation (2.2) holds. The first proposal is an analytical approximation for the variance of the estimator, using a Taylor linearisation on the first cycle of the IPF. Such an approximation has already been used by Rao (1986) to propose a variance approximation for the SPREE, however, that result is not extensible to the MSPREE because only the uncertainty associated with the estimation of the margins is taking into consideration in that case. The second proposal uses the bootstrap to estimate

$$\text{MSE}(\hat{\mathbf{Y}}^M) = E(\hat{\mathbf{Y}}^M - \mathbf{Y})^2,$$

with the expectation calculated both over $\mathbf{Y}$ and the sampling mechanism. The third proposal, aims to estimate a Finite Population MSE defined as

$$\text{FP-MSE}(\hat{\mathbf{Y}}^M) = E(\hat{\mathbf{Y}}^M - \mathbf{Y}|\mathbf{Y}),$$

using the bootstrap as well. The last two proposals attempt to estimate the MSE and FP-MSE above mentioned under a more general structural equation than (2.2) and will be presented in the next chapter, in section 3.3.

2.3.1 Analytical approximation to $\mathbf{V}(\hat{\mathbf{Y}}^{M})$

This alternative approximates the variance of the MSPREE using an estimate of the variance covariance matrix of $\hat{\boldsymbol{\beta}}$ and a first order Taylor approximation to the first cycle of the IPF algorithm. Because the resulting estimator is analytic and closed-form, its calculation is simple and does not involve considerable computational efforts, on top of the estimation of $\mathbf{Y}$. However, any potential bias resulting from model misspecification in case equation (2.2) does not hold is ignored.

Let h_1 be a function defined in $\mathbb{R}^{A\times J}$, the space of real-valued matrices of dimension $A \times J$, and h_2 and h_3 be functions defined in $\{\mathbb{R}^{+}\}^{A\times J}$, the subspace of matrices $A \times J$ with only positive entries; h_1, h_2 and h_3 defined as follows:

- $h_1 : \boldsymbol{\alpha} \mapsto \boldsymbol{\upsilon}$ with $\upsilon_{aj} = \exp \alpha_{aj}$
- $h_2 : \boldsymbol{\upsilon} \mapsto \mathbf{Y}^{(1)}$; $\quad Y^{(1)}_{aj} = \left(\dfrac{\upsilon_{aj}}{\upsilon_{a+}}\right) Y_{a+}$
- $h_3 : \mathbf{Y}^{(1)} \mapsto \mathbf{Y}^{(2)}$; $\quad Y^{(2)}_{aj} = \left(\dfrac{Y^{(1)}_{aj}}{Y^{(1)}_{+j}}\right) Y_{+j}$

where $\mathbf{Y}_{\cdot +} = (Y_{1+}, \ldots, Y_{A+})$ and $\mathbf{Y}_{+ \cdot} = (Y_{+1}, \ldots, Y_{+J})$ are fixed positive vectors, $\upsilon_{a+} = \sum_j \upsilon_{aj}$ and $Y^{(1)}_{+j} = \sum_a Y^{(1)}_{aj}$.

It is possible to see the IPF algorithm presented in section 1.3.1 as a successive composition of functions h_2 and h_3 until convergence is achieved. Moreover, considering only the first iteration cycle of the IPF it is possible to approximate the MSPREE of composition $\mathbf{Y}$, previously defined in equation (2.3) as:

$$\hat{\mathbf{Y}}^{M} = \text{IPF}\left(\exp \hat{\boldsymbol{\alpha}}^{Y} ; \mathbf{Y}_{\cdot +} ; \mathbf{Y}_{+ \cdot}\right)$$

by $\tilde{h}\left(\hat{\boldsymbol{\alpha}}^{Y}\right) = h_3 \circ h_2 \circ h_1 \left(\hat{\boldsymbol{\alpha}}^{Y}\right)$.

A linear approximation to $\hat{\mathbf{Y}}^{M}$ can be used to obtain an approximation of the variance of the MSPREE, $V\left(\hat{\mathbf{Y}}^{M}\right)$. A first order Taylor approximation of $\tilde{h}(\cdot)$ around the true association structure $\boldsymbol{\alpha}^{Y}$, is given by:

$$\tilde{\mathbf{Y}}^{M} = \tilde{h}(\boldsymbol{\alpha}^{Y}) + \tilde{h}'(\boldsymbol{\alpha}^{Y})(\hat{\boldsymbol{\alpha}}^{Y} - \boldsymbol{\alpha}^{Y})$$

and hence, an analytic approximation for $V\left(\hat{\mathbf{Y}}^{M}\right)$ is given by:

$$\mathrm{AV}\left(\hat{\mathbf{Y}}^{M}\right) = V\left(\tilde{\mathbf{Y}}^{M}\right) = \left(\tilde{h}'(\boldsymbol{\alpha}^{Y})\right) \mathbf{V}_{\hat{\alpha}} \left(\tilde{h}'(\boldsymbol{\alpha}^{Y})\right)^{T} \tag{2.13}$$

where $\tilde{h}'$ is the matrix of first derivatives with terms $\partial \tilde{Y}^{M}_{aj}/\partial \hat{\alpha}^{Y}_{rk}$, evaluated in the set of true α^{Y}_{aj}.

The deduction of the matrix of first derivatives $\tilde{h}'$ is presented in Lemma 9 in page 53. Having stacked the compositions by column, $\tilde{h}'$ is composed by block matrices of dimension $A \times A$. The blocks in the main diagonal of $\tilde{h}'$ have diagonal elements:

$$\frac{\partial \hat{Y}^{M}_{aj}}{\partial \hat{\alpha}_{aj}} = \left[\frac{Y_{+j}}{Y^{(1)}_{+j}}\right] \left[1 - \frac{Y^{(1)}_{aj}}{Y^{(1)}_{+j}}\right] \left[1 - \frac{Y^{(1)}_{aj}}{Y_{a+}}\right] Y^{(1)}_{aj} \tag{2.14}$$

and off-diagonal ones:

$$\frac{\partial \hat{Y}^{M}_{aj}}{\partial \hat{\alpha}_{sj}} = \left[\frac{Y_{+j}}{Y^{(1)}_{+j}}\right] \left[\frac{Y^{(1)}_{aj}}{Y^{(1)}_{+j}}\right] \left[1 - \frac{Y^{(1)}_{sj}}{Y_{s+}}\right] \left[-Y^{(1)}_{sj}\right]. \tag{2.15}$$

Analogously, the blocks outside the main diagonal have diagonal elements:

$$\frac{\partial \hat{Y}^{M}_{aj}}{\partial \hat{\alpha}_{al}} = \left[\frac{Y_{+j}}{Y^{(1)}_{+j}}\right] \left[1 - \frac{Y^{(1)}_{aj}}{Y^{(1)}_{+j}}\right] \left[\frac{Y^{(1)}_{aj}}{Y_{a+}}\right] \left[-Y^{(1)}_{al}\right] \tag{2.16}$$

and off-diagonal ones:

$$\frac{\partial \hat{Y}^{M}_{aj}}{\partial \hat{\alpha}_{sl}} = \left[\frac{Y_{+j}}{Y^{(1)}_{+j}}\right] \left[\frac{Y^{(1)}_{aj}}{Y^{(1)}_{+j}}\right] \left[\frac{Y^{(1)}_{sj}}{Y_{s+}}\right] Y^{(1)}_{sl}. \tag{2.17}$$

An estimate for $V\left(\hat{\mathbf{Y}}^{M}\right)$ can be obtained, substituting the unknown $\boldsymbol{\alpha}^{Y}$ and $\mathbf{V}_{\hat{\alpha}}$ in equation (2.13) by their corresponding estimates.

The only remaining thing to be discussed is how to obtain an estimate of $\mathbf{V}_{\hat{\alpha}}$. Because for each area, $\hat{\boldsymbol{\alpha}}^{Y}_{a} = \hat{\boldsymbol{\beta}}\boldsymbol{\alpha}^{X}_{a}$, all the variability arises from the estimation of the matrix of coefficients $\boldsymbol{\beta}$. If one of the alternative equations provided in section 2.2 was used to obtain $\hat{\boldsymbol{\beta}}$, the corresponding block of $\hat{\mathbf{V}}_{\boldsymbol{\Psi}}$ is an estimate of their variance-covariance matrix. Such an estimate can be easily obtained as the the inverse of the Hessian matrix calculated at the MLE if the ML approach has been followed, or the inverse of $(\mathbf{Z}^{\mathbf{T}}W^{-1}\mathbf{Z})$ if IWLS was used. Denoting this estimate by $\hat{\mathbf{V}}_{\beta}$, $\mathbf{V}_{\hat{\alpha}}$ can be estimated by:

$$\hat{\mathbf{V}}_{\alpha} = \mathbf{Z}_{\alpha}\hat{\mathbf{V}}_{\beta}\mathbf{Z}_{\alpha}^{\mathbf{T}}, \tag{2.18}$$

where $\mathbf{Z}_{\alpha} = \mathbf{T} \otimes (\boldsymbol{\alpha}^{\mathbf{X}}\mathbf{T})$, for $\mathbf{T}$ defined as in equation (2.7) and $\boldsymbol{\alpha}^{\mathbf{X}}$ denoting the $A \times J$ matrix containing the interactions of the auxiliary composition. Remember that if the logit link function was used, the parameter estimates correspond to the set of $\tilde{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}$ and need to be multiplied on the left by the matrix $[\mathbf{I} - (1/J)\mathbf{1}]$ (see the equivalence between the two sets of parameters, as part of the proof of Theorem 8 in page 51). Hence, if the logit link was used, $\hat{\mathbf{V}}_{\beta} = [\mathbf{I} - (1/J)\mathbf{1}]\,\hat{\mathbf{V}}_{\tilde{\beta}}\,[\mathbf{I} - (1/J)\mathbf{1}]^{\mathsf{T}}$.

Remark
A requirement for the proposed approximation to work is that the initial estimate of $\mathbf{Y}$ that enters the IPF, i.e., $\boldsymbol{\upsilon}$ in this case, is not substantially far from the scale given by the known margins. Unfortunately, that is not the case of the MSPREE defined as above, because the initial $\boldsymbol{\upsilon} = \exp \boldsymbol{\alpha}^{\mathbf{Y}}$ ignores the terms corresponding to the allocation structure of $\mathbf{Y}$. For the purposes of variance estimation, we suggest creating an initial estimate $\tilde{\boldsymbol{\upsilon}}$ by exponentiating the sum of a sensible allocation structure and the estimated association structure. The matrix of first derivatives $\tilde{h}'(\cdot)$ defined by equations 2.14 - 2.17 can then be evaluated on $\tilde{\mathbf{Y}}^{(1)} = h_2(\tilde{\boldsymbol{\upsilon}})$. Notice that this procedure has no impact, either on the point estimate or on the variance of the MSPREE, because any initial allocation added to the estimated structure would lead to the same MSPREE after IPF. The only aim of this procedure is to improve the quality of the approximation of one-cycle IPF to the IPF after convergence. Adding an allocation structure can be done in several ways, for instance, by including in the initial estimate the nuisance parameters of the equations for the estimation of $\boldsymbol{\beta}$ presented in section 2.2.1, i.e., using the fitted values instead of only the interaction terms, or by making an initial IPF of the estimated association structure to the margins of $\mathbf{X}$.

2.3.2 Estimation of MSE$(\hat{\mathbf{Y}}^{M})$

The MSPREE of $\mathbf{Y}$ and one of the GLMs associated with the alternative equations proposed in section 2.2.1 can be used to set up a parametric bootstrap to estimate $\mathrm{MSE}(\hat{\mathbf{Y}}^M) = E(\hat{\mathbf{Y}}^M - \mathbf{Y})^2$. As in section 2.2.2, assume that $\mathbf{Y}_a|\boldsymbol{\alpha}_a^X \overset{\text{ind}}{\sim}$ Multinomial $(Y_{a+}, \boldsymbol{\pi}_a^Y)$. Furthermore, suppose that $\boldsymbol{\pi}_a = \hat{\boldsymbol{\theta}}_a^M$, the estimated vector of within-area proportions of the MSPREE for area a, $a = 1, \dots, A$. The bootstrap procedure consists in generating repeatedly population compositions from the above distribution, selecting a sample from each population and calculating the MSPREE based on the bootstrap sample. If B populations have been generated, $\hat{\mathbf{Y}}^{M,b}$ denotes the MSPREE for the composition of population b, $\mathbf{Y}^b$, and $\mathrm{vec}(\cdot)$ denotes the vector operator, the estimate of $\mathrm{MSE}(\hat{\mathbf{Y}}^M)$ is obtained using the Monte Carlo approximation:

$$\widehat{\mathrm{MSE}}(\hat{\mathbf{Y}}^M) := \frac{1}{B} \sum_b \mathrm{vec}\left(\hat{\mathbf{Y}}^{M,b} - \mathbf{Y}^b\right) \mathrm{vec}\left(\hat{\mathbf{Y}}^{M,b} - \mathbf{Y}^b\right)^T.$$

2.3.3 Estimation of FP-MSE($\hat{\mathbf{Y}}^M$)

Assume $\hat{\mathbf{Y}}^M$, the MSPREE calculated over the original sample, as the fixed parameter of interest. By sampling repeatedly from the population defined by that composition, it is possible to propose a bootstrap estimate of uncertainty that only considers the variability due to the sampling procedure, having an interpretation that is closer to the uncertainty in the design-based approach of inference. The quantity of interest, $\text{FP-MSE}(\hat{\mathbf{Y}}^M) = E(\hat{\mathbf{Y}}^M - \mathbf{Y}|\mathbf{Y})$, can be estimated using the Monte Carlo approximation:

$$\widehat{\text{FP-MSE}}(\hat{\mathbf{Y}}^M) := \frac{1}{B} \sum_b \mathrm{vec}\left(\hat{\mathbf{Y}}^{M,b} - \hat{\mathbf{Y}}^M\right) \mathrm{vec}\left(\hat{\mathbf{Y}}^{M,b} - \hat{\mathbf{Y}}^M\right)^T$$

with $\hat{\mathbf{Y}}^{M,b}$ the MSPREE of the fixed $\hat{\mathbf{Y}}^M$ calculated over sample b.

Notice that, because all populations for the first estimator, and all samples for the second, have been derived on base of $\hat{\mathbf{Y}}^M$, neither $\widehat{\mathrm{MSE}}(\hat{\mathbf{Y}}^M)$, nor $\widehat{\text{FP-MSE}}(\hat{\mathbf{Y}}^M)$ are able to take into account any potential misspecification in case equation (2.2) does not hold.

2.4 Complementary Material

2.4.1 Parameter interpretation

The MSPREE assumes that the interactions in the target composition are a linear combination of the interactions in the proxy composition, with coefficients given by a $J \times J$ matrix denoted by $\boldsymbol{\beta}$ with components β_{jl}. Because the scale $\boldsymbol{\beta}$ depends on the number of categories of the composition, an intuitive interpretation of the estimated parameters $\hat{\beta}_{jl}$ is difficult. To illustrate this, consider that according to the corollary of Lemma 6 in page 49, SPREE is a special case of MSPREE with $\boldsymbol{\beta} = \mathbf{C}$, for $\mathbf{C} = \mathbf{I} - J^{-1}\mathbf{1}$. This means that the matrix of coefficients $\boldsymbol{\beta}$ for SPREE has diagonal components $\beta_{jj} = (J-1)/J$ and off-diagonal ones $\beta_{jl} = -1/J$. For instance, if $J = 5$, the diagonal terms will be 0.8 but if $J = 3$, they would be approximately 0.66 even though in both cases the interactions are assumed to remain unaltered.

On the other hand, the SPREE and GSPREE have an intuitive interpretation in terms of the assumption of proportional interactions. Extending the GSPREE idea to J proportionality constants, it is possible to think of another special case of the MSPREE with only J free parameters and a matrix of coefficients $\boldsymbol{\beta} = \mathbf{CBC}$ where $\mathbf{B} = \text{diag}\{b_1, b_2, ...b_J\}$, that we will denote MSPREE(J). It is straightforward to show that the structural assumption of the MSPREE(J) is

$$\alpha^Y_{aj} = b_j \alpha^X_{aj} - \frac{1}{J}\sum_l b_l \alpha^X_{al}.$$

Because the second term on the right hand side is the same for all the interaction terms in a given area, its task is only to ensure that α^Y_{aj} satisfy the constraint $\sum_a \alpha^Y_{aj} = 0$. Hence, the coefficients $\{b_1, b_2, ...b_J\}$ can be considered as category-specific proportionality constants, with b_i only affecting the interaction terms of category i.

Paying attention to the expression $\boldsymbol{\beta} = \mathbf{CBC}$, notice that the multiplication on the left and on the right by $\mathbf{C}$ aims to ensure that the constraints $\beta_{j+} = \beta_{+l} = 0$ for $j, l = 1, \ldots, J$ are satisfied. Moreover, because $\mathbf{C}$ is idempotent, the SPREE and GSPREE can be written analogously in a very convenient way, with $\mathbf{B} = \mathbf{I}$ and $\mathbf{B} = \beta\mathbf{I}$ respectively. In all those cases, the structural assumptions of the estimator are clearly stated in the corresponding matrix $\mathbf{B}$. Unfortunately, there is an identifiability problem on the relationship between $\boldsymbol{\beta}$ and $\mathbf{B}$ because $\boldsymbol{\beta} = \mathbf{CBC} = \mathbf{C}(\mathbf{B} + k\mathbf{1}\mathbf{1}^T)\mathbf{C}$ for any scalar k. However, if $\mathbf{B}$ is forced to be diagonal in order to make an interpretation of the parameters in the spirit of proportional interactions, the identifiability problem ceases.

For the MSPREE in the general case, it is always possible to write $\boldsymbol{\beta} = \mathbf{CBC}$ but clearly a diagonal matrix would not be able to manage the $(J-1)^2$ free components of $\boldsymbol{\beta}$. For parameter interpretation purposes, we propose to rescale $\boldsymbol{\beta}$ in the general case to a matrix $\mathbf{B}^p$ with components b^p_{jl} such that the diagonal components are free but the off-diagonal ones are forced to sum to zero, by row and column, as is the case for the three estimators mentioned above. Such a matrix is given by

$$b^p_{jl} = \beta_{jl} + \frac{1}{J-2}\left(\beta_{jj} + \beta_{ll} - \frac{\mathrm{Tr}(\boldsymbol{\beta})}{J-1}\right). \tag{2.19}$$

The details of the derivation of (2.19) are given in Lemma 10, page 54. It is straightforward to show that $\mathbf{B}^p$ corresponding to estimators SPREE, GSPREE and MSPREE(J) are $\mathbf{I}$, $\beta\mathbf{I}$ and $\mathrm{diag}\{b_1, b_2, ...b_J\}$ respectively. Hence, the superscript p has been added to emphasize that the parameters on that scale relate to the proportional interactions assumption.

The suggested interpretation in terms of $\mathbf{B}^p$ is as follows. The diagonal terms b^p_{jj} are an indicator of how the interactions of a given category in the target composition shrink or expand with respect to the same terms in the proxy composition, as in the case of the GSPREE. On the other hand, the off-diagonal terms would indicate how the interactions of the remaining categories can be better arranged in order to compensate for the unequal proportionality constants among columns, in order to satisfy the constraints of the centred-constraints representation. An illustration of this interpretation using real data will be shown in chapter 4.

2.4.2 Proofs

Constraints imposed on $\boldsymbol{\beta}$

This lemma proves that the constraints $\beta_{+l} = \beta_{j+} = 0$ for $j, l = 1, \ldots, J$, together with the MSPREE structural assumption, ensure a well defined interaction structure for the target composition. It is referenced in the definition of the MSPREE in page 32.

Lemma 4. *Equation* (2.2) *with the constraints* $\beta_{+l} = \beta_{j+} = 0$ *for* $j, l = 1, \ldots, J$ *imply:*

1. $\alpha^Y_{a+} = \alpha^Y_{+j} = 0$

2. *If* $\alpha^Y_{aj} = \sum_l \beta_{jl}\alpha^X_{al} = \sum_l \tilde{\beta}_{jl}\alpha^X_{al}$, *then* $\beta_{jl} = \tilde{\beta}_{jl}$ *for* $j, l = 1, \ldots, J$.

Proof. We will start by proving item 1. In order to satisfy the constraint $\alpha^Y_{+j} = 0$ it is only necessary to ensure that the same $\boldsymbol{\beta}$ is used for all areas, given that

$$\alpha^Y_{+j} = \sum_a \sum_l \beta_{jl}\alpha^X_{al} = \sum_l \beta_{jl} \sum_a \alpha^X_{al} = 0,$$

since the α^X_{al} sum to zero by column. On the other hand,

$$\alpha^Y_{a+} = \sum_j \sum_l \beta_{jl}\alpha^X_{al} = \sum_l \alpha^X_{al} \sum_j \beta_{jl},$$

so the constraint $\alpha^Y_{a+} = 0$ is only satisfied if $\beta_{+l} = 0$ for $l = 1 \ldots, J$. The column constraints ensure that equation (2.2) is well defined in the interaction scale.

Regarding item 2, notice that

$$\sum_l \beta_{jl}\alpha^X_{al} = \sum_l \tilde{\beta}_{jl}\alpha^X_{al} \iff \sum_l \alpha^X_{al}(\beta_{jl} - \tilde{\beta}_{jl}) = 0.$$

for $j = 1, \ldots, J$. The last equality is satisfied as long as $\beta_{jl} - \tilde{\beta}_{jl} = k_j$, independent of l because:

$$k_j \sum_l \alpha^X_{al} = 0$$

given that the α^X_{aj} sum to zero by row. This evidences an identifiability issue in the sense that a particular β_{jl} and any $\tilde{\beta}_{jl} = \beta_{jl} - k_j$ would lead to the same association structure for the target composition. The constraints $\beta_{j+} = \tilde{\beta}_{j+} = 0$ for $j = 1, \ldots, J$ ensure identifiability because:

$$\sum_l \beta_{jl} = Jk_j + \sum_l \tilde{\beta}_{jl} = Jk_j = 0$$

implies $k_j = 0$ and hence $\beta_{jl} = \tilde{\beta}_{jl}$ for $j = 1, \ldots, J$. □

Matrix C

This lemma shows some properties of the matrix $\mathbf{C}$ that are used to proof several lemmas and theorems across the document. Is used in the proof of Lemma 6 (page 49), Theorem 8 (page 51), Lemma 10 (page 54), Lemma 11 (page 68), Lemma 13 (page 71), Lemma 14 (page 72) and Theorem 15 (page 73).

Lemma 5. *Define the $J \times J$ matrix $\mathbf{C} = \mathbf{I} - J^{-1}\mathbf{1}$, where $\mathbf{I}$ denotes the identity matrix and $\mathbf{1}$ is a squared matrix with all component equal to 1. Hence,*

1. *$\mathbf{C}$ is symmetric.*

2. *$\mathbf{C}$ is idempotent.*
3. *For any matrix $\mathbf{B}$ of dimension $J \times J$:*
   - *(a) The product $\mathbf{CB}$ sums to zero by column.*
   - *(b) The product $\mathbf{BC}$ sums to zero by row.*
   - *(c) The product $\mathbf{CBC}$ sums to zero by both row and column.*

Proof. The first two properties are easily derived from the definition of $\mathbf{C}$. First, notice that $\mathbf{C}^T = \mathbf{I}^T - J^{-1}\mathbf{1}^T = \mathbf{C}$ because both matrices are symmetric. Second, see that:

$$\mathbf{CC} = \mathbf{C}\left(\mathbf{I} - J^{-1}\mathbf{1}\right) = \mathbf{C} - J^{-1}\mathbf{C1} = \mathbf{C} - J^{-1}\left(\mathbf{I} - J^{-1}\mathbf{1}\right)\mathbf{1} = \mathbf{C} - J^{-1}\left(\mathbf{1} - J^{-1}\mathbf{11}\right).$$

As $\mathbf{11} = J\mathbf{1}$, we can conclude that $\mathbf{C}$ is idempotent. For a proof of item 3, notice that any matrix $\boldsymbol{A}$ sums to zero by column if and only if, $\mathbf{1}\boldsymbol{A} = \mathbf{0}$, the zero matrix of the corresponding dimension. Analogously, $\boldsymbol{A}$ sums to zero by row if and only if $\boldsymbol{A}\mathbf{1} = \mathbf{0}$. Consider the product:

$$\mathbf{1CB} = \mathbf{1}\left(\mathbf{I} - J^{-1}\mathbf{1}\right)\mathbf{B} = \mathbf{1B} - J^{-1}\mathbf{11B}.$$

As $\mathbf{11} = J\mathbf{1}$, we conclude that $\mathbf{1CB} = \mathbf{0}$ i.e., $\mathbf{CB}$ sums to zero by column. For a proof of (b), consider the product:

$$\mathbf{BC1} = \mathbf{B}\left(\mathbf{I} - J^{-1}\mathbf{1}\right)\mathbf{1} = \mathbf{B1} - J^{-1}\mathbf{B11},$$

using the same argument, we conclude that $\mathbf{BC}$ sums to zero by row. Finally, for (c), notice that $\mathbf{1CBC} = (\mathbf{1CB})\,\mathbf{C} = \mathbf{0}$, according to (a). Analogously, $\mathbf{CBC1} = \mathbf{C}\,(\mathbf{BC1}) = \mathbf{0}$ according to (b). Therefore, $\mathbf{CBC}$ sums to zero by both row and column. □

SPREE and GSPREE as special cases

This lemma and its corollary show that the SPREE and GSPREE are special cases of the MSPREE. They are referenced in pages 33 and 46.

Lemma 6. *Equation* (2.1) *with $\boldsymbol{\beta} = \beta\mathbf{C}$ is equivalent to the structural assumption of the GSPREE, $\alpha^Y_{aj} = \beta\alpha^X_{aj}$ for $a = 1, \ldots, A$ and $j = 1, \ldots, J$.*

Proof. $\mathbf{C}$ satisfies the constraints imposed on $\boldsymbol{\beta}$ in equation (2.1), $\beta_{j+} = \beta_{+l} = 0$ for $j, l = 1, \ldots, J$, given that:

$$\mathbf{1C} = \mathbf{1}\left(\mathbf{I} - J^{-1}\mathbf{1}\right) = \mathbf{0}$$

because $\mathbf{1}\mathbf{1} = J\mathbf{1}$. As well, $\mathbf{C}\mathbf{1} = \left(\mathbf{1}\mathbf{C}^T\right) = \mathbf{0}$ because $\mathbf{C}$ is symmetric. As β multiplies the entire matrix, $\beta\mathbf{C}$ also satisfies the constraints. Substituting $\beta\mathbf{C}$ in equation (2.1) leads to the set of equations:

$$\alpha^Y_{aj} = \beta\left(1 - \frac{1}{J}\right)\alpha^X_{aj} - \beta\left(\frac{1}{J}\right)\sum_{l \neq j}\alpha^X_{al} = \beta\alpha^X_{aj} - \beta\left(\frac{1}{J}\right)\sum_{l}\alpha^X_{al},$$

for $j = 1, \ldots, J$. However, the last term at the right hand side vanishes given that $\alpha^X_{a+} = 0$. □

Corollary 6.1. *The structural assumption of the SPREE, $\alpha^Y_{aj} = \beta\alpha^X_{aj}$ for $a = 1, \ldots, A$ and $j = 1, \ldots, J$, is obtained for $\beta = 1$.*

Equivalence between the alternative equations and the structural assumption of the MSPREE

The following two theorems show the equivalence between the two alternative equations and the structural assumption of the MMSPREE. They are referenced in pages 34, 35, 44 and 61.

Theorem 7. *Denote $\zeta^Y_{aj} = \log Y_{aj}$, for $a = 1, \ldots, A$, $j, l = 1, \ldots, J$. The set of equations* (2.5)*:*

$$\zeta^Y_{aj} = \gamma_a + \lambda_j + \sum_l \beta_{jl}\alpha^X_{al}$$

where $\beta_{j+} = \beta_{+l} = 0$ for $j, l = 1, \ldots, J$ and $\lambda_+ = 0$, is equivalent to the structural assumption of the MSPREE (equation 2.2):

$$\alpha^Y_{aj} = \sum_l \beta_{jl}\alpha^X_{al}\,,$$

for $a = 1, \ldots, A$, $j = 1, \ldots, J$.

Proof. We will start the proof by showing that equation (2.5) implies (2.2). As $\zeta^Y_{aj} = \log Y_{aj}$, using the centred-constraint parametrisation given in equation (1.16), it is possible to rewrite equation (2.5) as:

$$\zeta^Y_{aj} = \alpha^Y_0 + \alpha^Y_a + \alpha^Y_j + \alpha^Y_{aj} = \gamma_a + \lambda_j + \sum_l \tilde{\beta}_{jl}\alpha^X_{al}\,. \tag{2.20}$$

Given the constraints satisfied by the set of α^Y_{aj}, and the constraints $\tilde{\beta}_{+l} = 0$ for $l = 1, \ldots, J$ and $\lambda_+ = 0$, by summing across j in both sides of (2.20) we obtain:

$$\alpha^Y_0 + \alpha^Y_a = \gamma_a\,. \tag{2.21}$$

Substituting (2.21) in (2.20) and summing across a we obtain:

$$\alpha_j^Y = \lambda_j \,. \tag{2.22}$$

Substituting equations (2.21) and (2.22) back in equation (2.20), we obtain equation (2.2), completing the proof of the first implication. The equivalence is proved because, as there are no constraints regarding the possible values of γ_a and the only constraint for λ_j, i.e., $\lambda_+ = 0$ is automatically satisfied by $\lambda_j = \alpha_j^Y$, it is always possible to set γ_a and λ_j as in equations (2.21) and (2.22) to obtain (2.20). □

Theorem 8. *Denote by* $\rho_{aj}^{Y,r}$ *the logit between category* j *and category* r *in area* a, *i.e.,* $\rho_{aj}^{Y,r} = \log\left(Y_{aj}/Y_{ar}\right)$, *for* $a = 1, \ldots, A;\ \ r \in \{1, \ldots, J\};\ \ j, l = 1, \ldots, J;\ \ j, l \neq r$. *The set of equations* (2.8)*:*

$$\rho_{aj}^{Y,r} = \phi_j + \sum_{l \neq r} \tilde{\beta}_{jl}(\alpha_{al}^X - \alpha_{ar}^X)$$

with $\phi_+ = \sum_{j \neq r} \phi_j = 0$ *is equivalent to the structural assumption of the MSPREE (equation 2.2):*

$$\alpha_{aj}^Y = \sum_{l} \beta_{jl} \alpha_{al}^X \,.$$

with $\beta_{j+} = \beta_{+l} = 0$ *for* $a = 1, \ldots, A$, $j = 1, \ldots, J$.

Proof. We will prove first the implication (2.8) ⇒ (2.2). Notice that:

$$\begin{aligned}\rho_{aj}^{Y,r} &= \log Y_{aj} - \log Y_{ar} \\ &= (\alpha_j^Y - \alpha_r^Y) + (\alpha_{aj}^Y - \alpha_{ar}^Y)\end{aligned} \tag{2.23}$$

Substituting (2.23) in (2.8) and summing over a we obtain:

$$(\alpha_j^Y - \alpha_r^Y) = \phi_j \,, \tag{2.24}$$

given that $\alpha_{a+}^Y = 0$ under this parametrisation. Substituting (2.23) and (2.24) back in (2.8) leads to:

$$(\alpha_{aj}^Y - \alpha_{ar}^Y) = \sum_{l \neq r} \tilde{\beta}_{jl}(\alpha_{al}^X - \alpha_{ar}^X), \tag{2.25}$$

for $a = 1, \ldots, A$; $j = 1, \ldots, J$, $j \neq r$.

Consider now the sum of (2.25) over j:

$$\sum_{j\neq r}\left(\alpha^{Y}_{aj}-\alpha^{Y}_{ar}\right)=\sum_{j\neq r}\sum_{l\neq r}\tilde{\beta}_{jl}(\alpha^{X}_{al}-\alpha^{X}_{ar})$$

$$-J\alpha^{Y}_{ar}=\left(\sum_{l\neq r}\alpha^{X}_{al}\sum_{j\neq r}\tilde{\beta}_{jl}\right)-\left(\alpha^{X}_{ar}\sum_{l\neq r}\sum_{j\neq r}\tilde{\beta}_{jl}\right)$$

$$\alpha^{Y}_{ar}=\frac{1}{J}\left(\tilde{\beta}_{++}\alpha^{X}_{ar}-\sum_{l\neq r}\tilde{\beta}_{+l}\alpha^{X}_{al}\right). \tag{2.26}$$

Finally, substituting (2.26) in (2.25) we obtain:

$$\alpha^{Y}_{aj}=\sum_{l\neq r}\left(\tilde{\beta}_{jl}-\frac{1}{J}\tilde{\beta}_{+l}\right)\alpha^{X}_{al}+\left(\frac{1}{J}\tilde{\beta}_{++}-\tilde{\beta}_{j+}\right)\alpha^{X}_{ar}. \tag{2.27}$$

Equations (2.26) and (2.27) are equivalent to the structural assumption of the MSPREE (2.2), with:

$$\beta_{jl}=\begin{cases}\tilde{\beta}_{jl}-\frac{1}{J}\tilde{\beta}_{+l} & \text{for } j\neq r,\ l\neq r\\ \frac{1}{J}\tilde{\beta}_{++}-\tilde{\beta}_{j+} & \text{for } j\neq r,\ l=r\\ -\frac{1}{J}\tilde{\beta}_{+l} & \text{for } j=r, l\neq r\\ \frac{1}{J}\tilde{\beta}_{++} & \text{for } j=r, l=r\end{cases} \tag{2.28}$$

Equivalently, the square matrix containing the $(J-1)^2$ independent terms β_{jl} for $j, 1=1,\ldots,J$; $j, l\neq r$ can be obtained as:

$$[\mathbf{I}-(1/J)\mathbf{1}]\,\tilde{\boldsymbol{\beta}}. \tag{2.29}$$

The remaining $\beta_{j,r}, \beta_{r,j}$ for $j=1,\ldots,J$ and β_{rr} can be obtained by substraction, due to the constraints $\beta_{j+}=\beta_{+l}=0$. Notice that the matrix $[\mathbf{I}-(1/J)\mathbf{1}]$ above, corresponds to the first $(J-1)$ rows and columns of matrix $\mathbf{C}_J$ defined in Lemma 5.

To prove the implication (2.2) ⇒ (2.8) it is enough to prove (2.2) ⇒ (2.25) because ϕ_j can always be set as in (2.24) to satisfy (2.8). Subtracting equation (2.2) for the reference category from the equation corresponding to category j and using the definition of β_{jl} in (2.28) we obtain:

$$\begin{aligned}
\alpha^Y_{aj} - \alpha^Y_{ar} &= \sum_l (\beta_{jl} - \beta_{rl})\alpha^X_{al} \\
&= \sum_{l \neq r} \left(\tilde{\beta}_{jl} - \frac{1}{J} + \frac{1}{J}\right) \alpha^X_{al} + \left(\frac{1}{J}\tilde{\beta}_{++} - \tilde{\beta}_{j+} - \frac{1}{J}\tilde{\beta}_{++}\right) \alpha^X_{ar} \\
&= \sum_{l \neq r} \left(\tilde{\beta}_{jl}\alpha^X_{al} - \tilde{\beta}_{j+}\alpha^X_{ar}\right) \\
&= \sum_{l \neq r} \tilde{\beta}_{jl} \left(\alpha^X_{al} - \alpha^X_{ar}\right)
\end{aligned}$$

i.e., (2.2) ⇒ (2.25) completing the proof. □

Matrix of first derivatives of $\tilde{h}(\cdot)$
This lemma shows the first derivatives of $\tilde{h}(\cdot)$. It is referenced in page 43.

Let h_1 be a function defined in $\mathbb{R}^{A \times J}$, the space of the real-valued matrices of dimension $A \times J$ by $h_1 : \boldsymbol{\alpha} \mapsto \boldsymbol{\upsilon}$ with $\upsilon_{aj} = \exp \alpha_{aj}$. Let h_2, h_3 be functions in the subspace of $\mathbb{R}^{A \times J}$ of matrices with only positive entries, defined by:

$$h_2 : \boldsymbol{\upsilon} \mapsto \mathbf{Y}^{(1)}; \quad Y^{(1)}_{aj} = \left(\frac{\upsilon_{aj}}{\upsilon_{a+}}\right) Y_{a+}$$

$$h_3 : \mathbf{Y}^{(1)} \mapsto \mathbf{Y}^{(2)}; \quad Y^{(2)}_{aj} = \left(\frac{Y^{(1)}_{aj}}{Y^{(1)}_{+j}}\right) Y_{+j}$$

where $\mathbf{Y}_{.+} = (Y_{1+}, \ldots, Y_{A+})$ and $\mathbf{Y}_{+.} = (Y_{+1}, \ldots, Y_{+J})$ are fixed positive vectors, $\upsilon_{a+} = \sum_j \upsilon_{aj}$ and $Y^{(1)}_{+j} = \sum_a Y^{(1)}_{aj}$.

Lemma 9. *The composition of the functions defined above, $\tilde{h} = h_3 \circ h_2 \circ h_1$, has first derivatives:*

$$\tilde{h}' : \frac{\partial Y^{(2)}_{aj}}{\partial \alpha_{sl}} = \left(\frac{Y_{+j}}{Y^{(1)}_{+j}}\mathbb{1}_{[s=a]} - \frac{Y_{+j}}{Y^{(1)}_{+j}}\frac{Y^{(1)}_{aj}}{Y^{(1)}_{+j}}\right)\left(Y^{(1)}_{sl}\mathbb{1}_{[l=j]} - Y^{(1)}_{sl}\frac{Y^{(1)}_{sj}}{Y_{s+}}\right)$$

for $a, s = 1, \ldots, A$ and $j, l = 1, \ldots, J$. $\mathbb{1}_{[x=y]}$ denotes the indicator function of $x = y$.

Proof. The derivatives of each one of the functions h_1, h_2 and h_3 are:

$$h_1' : \frac{\partial \upsilon_{aj}}{\partial \alpha_{sl}} = \upsilon_{aj} \mathbb{1}_{[s=a;l=j]}, \tag{2.30}$$

$$\begin{aligned} h_2' : \frac{\partial Y^{(1)}_{aj}}{\partial \upsilon_{sl}} &= \frac{1}{\upsilon_{a+}} \left(Y_{a+} \mathbb{1}_{[l=j]} - Y_{a+} \frac{\upsilon_{aj}}{\upsilon_{a+}} \right) \mathbb{1}_{[s=a]} \\ &= \frac{1}{\upsilon_{a+}} \left(Y_{a+} \mathbb{1}_{[l=j]} - Y^{(1)}_{aj} \right) \mathbb{1}_{[s=a]} \end{aligned} \tag{2.31}$$

$$h_3' : \frac{\partial Y^{(2)}_{aj}}{\partial Y^{(1)}_{sl}} = \left(\frac{Y_{+j}}{Y^{(1)}_{+j}} \mathbb{1}_{[s=a]} - \frac{Y_{+j}}{Y^{(1)}_{+j}} \frac{Y^{(1)}_{aj}}{Y^{(1)}_{+j}} \right) \mathbb{1}_{[l=j]} \tag{2.32}$$

Using the chain rule,

$$\tilde{h}' : \frac{\partial Y^{(2)}_{aj}}{\partial \alpha_{sl}} = \frac{\partial Y^{(2)}_{aj}}{\partial Y^{(1)}_{sj}} \frac{\partial Y^{(1)}_{sj}}{\partial \upsilon_{sl}} \frac{\partial \upsilon_{sl}}{\partial \alpha_{sl}} \tag{2.33}$$

because all the other terms are zero. Substituting (2.30), (2.31) and (2.32) in (2.33), and using the fact that $\upsilon_{sl}/\upsilon_{s+} = Y^{(1)}_{sl}/Y_{s+}$, the theorem is proved. □

Matrix $\mathbf{B}^p$ for parameter interpretation

This lemma shows how the matrix **C** of Lemma 5 can be used to transform $\boldsymbol{\beta}$ into the alternative parameterization **B**. The new matrix **B** is used to provide an interpretation of the coefficients of the MSPREE in terms of the assumption of proportional interactions. It is referenced in page 47.

Lemma 10. *Let* $\boldsymbol{\beta}$ *and* $\mathbf{C}$ *be* $J \times J$ *matrices, with* $\mathbf{C} = \mathbf{I} - J^{-1}\mathbf{1}$ *and* $\beta_{j+} = \beta_{+l} = 0$ *for* $j, l = 1 \ldots, J$.

1. $\boldsymbol{\beta}$ *can be written in the form* $\boldsymbol{\beta} = \mathbf{CBC}$ *with* $\mathbf{B}$ *a matrix with components* b_{jl} *such that* $b_{j+} = b_{jj}$ *and* $b_{+l} = b_{ll}$ *for* $j, l = 1, \ldots, J$.

2. $\mathbf{B}$ *is unique.*

Proof. We will start by studying the product $\mathbf{CBC} := \tilde{\boldsymbol{\beta}}$. Given the definition of $\mathbf{C}$ we have

$$\tilde{\beta}_{jl} = b_{jl} - \bar{b}_{j+} - \bar{b}_{+l} + \bar{b}_{++}$$

where $\bar{b}_{j+} = (1/J) \sum_l b_{jl}$, $\bar{b}_{+l} = (1/J) \sum_j b_{jl}$, and $\bar{b}_{++} = (1/J^2) \sum_j \sum_l b_{jl}$. Given the constraints imposed on **B**, this equation can be rewritten as

$$\tilde{\beta}_{jl} = b_{jl} - \frac{1}{J} b_{jj} - \frac{1}{J} b_{ll} + \frac{1}{J^2} \mathrm{Tr}(\mathbf{B}). \tag{2.34}$$

Simplifying, the diagonal elements of $\tilde{\boldsymbol{\beta}}$ are

$$\tilde{\beta}_{jj} = \frac{1}{J^2}\left(J(J-2)b_{jj} + \mathrm{Tr}(\mathbf{B})\right). \tag{2.35}$$

Let us turn now to $\mathrm{Tr}(\tilde{\boldsymbol{\beta}})$. Substituting (2.35) in the definition of the trace we obtain $\mathrm{Tr}(\tilde{\boldsymbol{\beta}}) = ((J-1)/J)\mathrm{Tr}(\mathbf{B})$, from which $\mathrm{Tr}(\mathbf{B}) = (J/(J-1))\mathrm{Tr}(\tilde{\boldsymbol{\beta}})$. Substituting this in (2.35),

$$\tilde{\beta}_{jj} = \frac{(J-2)}{J}b_{jj} + \frac{1}{J(J-1)}\mathrm{Tr}(\tilde{\boldsymbol{\beta}}),$$

or equivalently,

$$b_{jj} = \frac{J}{(J-2)}\left[\tilde{\beta}_{jj} - \frac{1}{J(J-1)}\mathrm{Tr}(\tilde{\boldsymbol{\beta}})\right]. \tag{2.36}$$

Substituting (2.36) in (2.34), it can be seen that the off-diagonal elements of $\tilde{\boldsymbol{\beta}}$ are

$$\tilde{\beta}_{jl} = b_{jl} - \frac{1}{J-2}\left[\tilde{\beta}_{jj} + \tilde{\beta}_{ll} - \frac{\mathrm{Tr}(\tilde{\boldsymbol{\beta}})}{J-1}\right].$$

from where

$$b_{jl} = \tilde{\beta}_{jl} + \frac{1}{J-2}\left[\tilde{\beta}_{jj} + \tilde{\beta}_{ll} - \frac{\mathrm{Tr}(\tilde{\boldsymbol{\beta}})}{J-1}\right]. \tag{2.37}$$

Equation (2.37) is enough to define, for a given $\tilde{\boldsymbol{\beta}}$, the matrix $\mathbf{B}$ such that $\tilde{\boldsymbol{\beta}} = \mathbf{CBC}$ because for the diagonal elements (2.37) is equivalent to (2.36). Notice that the only constraints imposed so far on $\tilde{\boldsymbol{\beta}}$ are $\tilde{\beta}_{j+} = \tilde{\beta}_{+l} = 0$ for $j, l = 1, \ldots, J$, which is derived from $\tilde{\boldsymbol{\beta}} = \mathbf{CBC}$ and Lemma 5. Hence, by taking $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$, item 1 is proven. To prove item 2, it is enough to notice that all the multiple solutions of $\boldsymbol{\beta} = \mathbf{CBC}$ have the form $\mathbf{M} = \mathbf{B} + k\mathbf{1}\mathbf{1}^T$ for some scalar k. Because of the constraints $b^p_{j+} = b^p_{jj}$ and $b^p_{+l} = b^p_{ll}$ for $j, l = 1, \ldots, J$, $\mathbf{B}$ is unique. □

2.4.3 Illustration. MSPREE via ML using Poisson/Multinomial regression in R

```
A <- 6  #Number of rows (areas)
J <- 4  #Number of columns (categories)

###### Data

Xtable <- matrix(c(1238, 216, 1981, 1128, 2419, 62, 1105, 581, 908, 846,
2717, 1047, 2384, 217, 2121, 979, 1881, 258, 1561,
1142, 2215, 307, 1814, 1124),
nrow = A, ncol = J, byrow = T) #Auxiliary composition

ytable <- matrix(c(1828, 285, 1858, 946, 2265, 36, 569, 1632, 884, 748,
3603, 435, 2767, 151, 1787, 1642, 1415, 199, 2776,
905, 2659, 269, 2914, 1193),
nrow = A, ncol = J, byrow = T) #sample composition

Ya. <- c(4917, 4502, 5670, 6347, 5295, 7035) #True row margins
Y.j <- c(11818, 1688, 13507, 6753) #True column margins

###### Column id

col <- kronecker(matrix(seq(1:J),1,J),matrix(1,1,A))

###### Interaction structure of X

f.LLRep<-function(y){
A <- nrow(y); J <- ncol(y);
Z <- log(y)
alpha0 <- mean(Z)
alphaa <- rowSums(Z)/J-alpha0
alphaj <- colSums(Z)/A-alpha0
alphaaj <- Z- matrix(alphaa,nrow=A,ncol=J,byrow=F)-
matrix(alphaj,nrow=A,ncol=J,byrow=T)-
matrix(alpha0,nrow=A, ncol=J)
alphaaj
}
alpha.Xaj <- f.LLRep(Xtable)

###### Design matrices

zg <- kronecker(matrix(1,J,1),diag(A))  #area effects
zl <- kronecker(rbind(diag(J-1),rep(-1,J-1)),matrix(1,A,1)) #column effects
matrixT <- rbind(diag(J-1),matrix(-1,1,J-1))
#for the poisson, the beta coefficients for the last column are -sum(row)
#for the multinomial, the last column is the reference
talpha <- alpha.Xaj%*%matrixT  #auxiliary structure
```

```
z.pois <- cbind(zg,zl,kronecker(matrixT,talpha))
z.mult <- kronecker(matrix(1,J,1),talpha)

###### Poisson fitting

model.poisson <- glm(c(ytable) ~ -1+z.pois,family = "poisson")

###### Multinomial fitting

library(mlogit)

data1 <- as.data.frame(cbind(factor(c(col)),c(ytable),z.mult))
data2 <- mlogit.data(data1, choice = "V1", shape="wide", alt.levels = seq(1,J))
model.multi <- mlogit(V1~1|V3+V4+V5, weights=V2, data=data2, reflevel = J)

###### Comparison of the two fitted structures

Yhat.poisson <- matrix(fitted(model.poisson),A,J,byrow=F)
struc.poisson <- f.LLRep(Yhat.poisson)

Yhat.multi<- matrix(fitted(model.multi),A,J,byrow=F)
struc.multi <- f.LLRep(Yhat.multi)

max(abs(struc.poisson - struc.multi))

par(mfrow=c(2,2))
for (i in 1:ncol(struc.poisson)){
plot(struc.poisson[,i],struc.multi[,i])
abline(0,1)
}

###### Imposition of the known column and row margins

Yhat.mspree <- loglin(outer(Ya.,Y.j)/sum(Ya.),margin=list(1,2),
start=Yhat.poisson, fit=TRUE, eps=1.e-08, iter=100)$fit
```

Chapter 3

Mixed MSPREE

This chapter introduces the MMSPREE, a mixed effects version of the MSPREE presented in the previous chapter, in an attempt to reduce the bias in the cases when the sample size allows for it. The chapter is composed of four sections. Section 3.1 starts with a motivation and then defines the structural assumption underpinning the MMSPREE. Section 3.2, presents a proposal to obtain estimates of the variance components and predictors of the random effects. Section 3.3 contains the proposed methodology to estimate the MSE of the mixed effects estimator. Finally, section 3.4 contains all the proofs corresponding to this chapter.

3.1 Mixed MSPREE (MMSPREE)

The MSPREE introduced in the previous chapter is built on the structural assumption

$$\alpha^{Y}_{aj} = \sum_{l} \beta_{jl}\alpha^{X}_{al},$$

where α^{Y}_{aj} and α^{X}_{aj} are the interaction terms corresponding to the association structure of the target and proxy composition, respectively, for $a = 1, \dots, A$ and $j = 1, \dots, J$, and β_{jl} are unknown parameters that satisfy $\beta_{j+} = \beta_{+l} = 0$. However, in practice, it is difficult to expect the equality to hold. Notice though that all differences between the true association structure of $\mathbf{Y}$ and the one assumed by the MSPREE can be expressed in terms of a set of unknown fixed quantities

$$m_{aj} := \alpha^{Y}_{aj} - \sum_{l} \beta_{jl}\alpha^{X}_{al}. \tag{3.1}$$

for $a = 1, \dots, A$ and $j = 1, \dots, J$ such that, differently from the structural assumption of the MSPREE, the assumption

$$\alpha^Y_{aj} = \sum_l \beta_{jl} \alpha^X_{al} + m_{aj}$$

always holds. Summing both sides of (3.1) over a or over j, it is clear that the m_{aj} terms satisfy $m_{a+} = m_{+j} = 0$ for $a = 1, \dots, A$ and $j = 1, \dots, J$.

The Mixed MSPREE (MMSPREE) is an extension of the MSPREE with the aim of reducing the risk of bias due to misspecification in the sense of (3.1). Substituting the fixed m_{aj} in the above equation by cell-specific random effects u_{aj}, and acknowledging, in principle, that the set of parameters that better explain the relationship between $\boldsymbol{\alpha}^Y$ and $\boldsymbol{\alpha}^X$ may differ between structural assumptions with or without random effects, lead us to the *mixed* structural assumption:

$$\alpha^Y_{aj} = \sum_l \tilde{\beta}_{jl} \alpha^X_{al} + u_{aj} \,. \tag{3.2}$$

In order to ensure that the constraints $u_{a+} = 0$ and $u_{+j} = 0$ are, not only on expectation but always, satisfied, the random effects are defined as a linear transformation of a set of independent cell-specific random variables ϑ_{aj}, with $E(\vartheta_{aj}) = 0$ and $V(\vartheta_{aj}) = \sigma^2_j$, as

$$u_{aj} = \vartheta_{aj} - \frac{1}{A}\vartheta_{+j} - \frac{1}{J}\vartheta_{a+} + \frac{1}{AJ}\vartheta_{++}. \tag{3.3}$$

Arranging the set of ϑ_{aj} in an $(A \times J)$ matrix denoted by $\boldsymbol{\vartheta}$ and using vec$(\cdot)$ to denote the vector operator, it is possible to write equation (3.3) in matrix notation as

$$\mathbf{u} := \text{vec}(\mathbf{C}_{(A)} \boldsymbol{\vartheta} \mathbf{C}_{(J)}), \tag{3.4}$$

where $\mathbf{C}$ is the matrix defined in Lemma 5 as $\mathbf{C}_{(K)} = \mathbf{I}_k - K^{-1}\mathbf{1}_{(K)}$ (page 48).

As it is shown in Lemma 11 (page 68), equation (3.4) induces

$$\boldsymbol{\Sigma}_u := V(\mathbf{u}) = (\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)}) \boldsymbol{\Sigma}_\vartheta (\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)}) \tag{3.5}$$

with $\otimes$ denoting the Kronecker product, $\boldsymbol{\Sigma}_\vartheta = V(\text{vec}(\boldsymbol{\vartheta})) = \text{diag}(\boldsymbol{\sigma}^2) \otimes \mathbf{I}_A$ and $\boldsymbol{\sigma}^2 = [\sigma^2_1, \dots, \sigma^2_J]^T$. For computational purposes, notice that the product on the right hand side of equation (3.5) can also be written as $(\mathbf{C}_{(J)} \text{diag}(\boldsymbol{\sigma}^2) \mathbf{C}_{(J)}) \otimes \mathbf{C}_{(A)}$.

As in other SPREE-type estimators, the MMSPREE of $\mathbf{Y}$ is obtained by imposing, via IPF, the known row and column margins of $\mathbf{Y}$, $\mathbf{Y}_{a+}; Y_{+j}$ on an estimate of its association structure, i.e.,

$$\hat{\mathbf{Y}}^{MM} = \mathrm{IPF}\left(\exp\left\{\tilde{\boldsymbol{\alpha}}^{Y} + \hat{\mathbf{u}}\right\} ; \mathbf{Y}_{a+} ; \mathbf{Y}_{+j}\right). \tag{3.6}$$

The details behind the estimation of $\tilde{\boldsymbol{\alpha}}$ and the prediction of the random effects will be explained in detail in the next section.

3.2 Estimation of the association structure

In section 2.2.1 we proposed an equation based on a log-link that is equivalent to the structural assumption of the MSPREE and can be used to obtain estimates for the parameters of interest. In a similar way, it is possible to see that

$$\zeta^{Y}_{aj} = \log Y_{aj} = \tilde{\gamma}_a + \tilde{\lambda}_j + \sum_{l} \tilde{\beta}_{jl}\alpha^{X}_{al} + u_{aj}\,, \tag{3.7}$$

is equivalent to the mixed structural assumption (3.2) that is the base of the MMSPREE. A formal proof is analogous to the proof of Theorem 7 (page 50) because the set of terms u_{aj} satisfy $u_{a+} = u_{+j} = 0$ always (not only in expectation).

Hereafter, we will assume that the set of parameters $\tilde{\beta}_{jl}$ that determine the MMSPREE is the same as the MSPREE, and hence, the fixed part of the structural assumption is the same for both estimators. This is convenient because it allows us to see the MSPREE as the synthetic estimator derived from the MMSPREE when ignoring the random effects. Moreover, as will be shown next, the two sets of parameters are actually the same if normality is assumed for the random effects. To see that, exponentiate both sides of equation (3.7) to obtain

$$Y_{aj} = \exp\left\{\tilde{\gamma}_a + \tilde{\lambda}_j + \sum_{l} \tilde{\beta}_{jl}\alpha^{X}_{al}\right\} e^{u_{aj}}\,.$$

Conditioning on the proxy composition and the known margins and taking the expectation on both sides, we have

$$E[Y_{aj}|\mathbf{X}, \mathbf{Y}_{a+}, \mathbf{Y}_{+j}] = \exp\left\{\tilde{\gamma}_a + \tilde{\lambda}_j + \sum_{l} \tilde{\beta}_{jl}\alpha^{X}_{al}\right\} E[e^{u_{aj}}]\,. \tag{3.8}$$

The last term on the right hand side of equation (3.8) is the moment-generating function of u_{aj}, $M_{u_{aj}}(t) = E(e^{tu_{aj}})$ evaluated at $t = 1$. Under the assumption of normality for u_{aj}, $E[e^{u_{aj}}] = e^{\frac{1}{2}\sigma^2_{u_{aj}}}$. However, according to the corollary of Lemma 11 (page 68),

$$\sigma^2_{u_{aj}} = V(u_{aj}) = \frac{(A-1)}{AJ}\left[(J-2)\sigma^2_j + \bar{\sigma}^2\right],$$

with $\bar{\sigma}^2 = J^{-1}(\sigma^2_+)$, i.e., $\sigma^2_{u_{aj}}$ depends only on j. This means that under the normality assumption, the introduction of random effects has a multiplicative effect on the expected value of Y_{aj} that it is not cell-specific. Hence, in expectation, it may modify the remaining parameters of (3.7) but not the ones that control the relationship between the $\boldsymbol{\alpha}^Y$ and $\boldsymbol{\alpha}^X$, i.e., the matrix of parameters $\boldsymbol{\beta}$.

Therefore, hereby we propose to obtain estimates of the vector of variance components $\boldsymbol{\sigma}^2$, as well as predictions of the vector of random effects $\mathbf{u}$, using the equation

$$\eta_{aj} = \zeta^{Y,M}_{aj} - \hat{\alpha}^Y_{aj} = \tilde{\gamma}_a + \tilde{\lambda}_j + u_{aj} + e_{aj}, \tag{3.9}$$

where $\zeta^{Y,M}_{aj}$ is the first order Taylor approximation of $\log Y_{aj}$ around the corresponding MSPREE, $\hat{Y}^M_{aj}$, given by

$$\zeta^{Y,M}_{aj} := \log \hat{Y}^M_{aj} + \frac{1}{\hat{Y}^M_{aj}}(Y_{aj} - \hat{Y}^M_{aj}); \tag{3.10}$$

$\hat{\alpha}^Y_{aj}$ is the estimated interaction for cell (a, j) obtained from the MSPREE of the composition, $\hat{\mathbf{Y}}^M$; u_{aj} are the random effects defined in equation (3.3) and e_{aj} is an error term measured in the log-scale, with known variance-covariance matrix. Stacking all the relevant components by column and using matrix notation, equation (3.9) can be written as

$$\boldsymbol{\eta} = \boldsymbol{\zeta}^{Y,M} - \text{vec}(\hat{\boldsymbol{\alpha}}^Y) = \mathbf{Z}\tilde{\boldsymbol{\Psi}} + \mathbf{u} + \mathbf{e}, \tag{3.11}$$

where $\mathbf{Z}$ is the design matrix of dimension $(AJ \times (A + J - 1))$ defined as

$$\mathbf{Z} = \left[\; \mathbf{1}_{(J\times 1)} \otimes \mathbf{I}_{(A)} \;\vdots\; \mathbf{T} \otimes \mathbf{1}_{(A\times 1)} \;\right]$$

with $\mathbf{T}$ defined previously in section 2.2.1, i.e., $\mathbf{T}_{(J\times(J-1))} = \left[\; \mathbf{I} \;\vdots\; -\mathbf{1} \;\right]^T$, and

$\tilde{\boldsymbol{\Psi}} = \left[\begin{array}{cccccc} \tilde{\gamma}_1 & \dots & \tilde{\gamma}_A & \tilde{\lambda}_1 & \dots & \tilde{\lambda}_{J-1} \end{array} \right]^T$ contains the nuisance parameters due to the rows and columns, once the estimated interactions have been taken into account. Furthermore, it is assumed that $E[\boldsymbol{e}] = \mathbf{0}$ and $V(\boldsymbol{e}) = \boldsymbol{\Sigma}_e$ known, and that $Cov(\mathbf{u}, \boldsymbol{e}) = \mathbf{0}$.

Conditioning on $\hat{\mathbf{Y}}^M$ and substituting $\mathbf{Y}$ by the composition of direct estimates $\hat{\mathbf{Y}}$ in the definition of $\boldsymbol{\zeta}^{Y,M}$, equation (3.11) can be seen as a linearised version of a particular case of the GLMM defined in section 1.2. This equation defines an area-level model, which is different from the standard Fay-Harriot model (Fay and Herriot, 1979) and also different from its extension including correlated sampling errors (see Rao and Molina, 2015, section 8.2) because both matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ are allowed to contain non-zero correlations.

Notice that, in the definition of $\boldsymbol{\eta}$, a Taylor approximation to $\log(\cdot)$ was preferred over a direct transformation of the response variable. In the first place, this rules out issues associated with sample zeroes. Moreover, because a non-linear transformation has been avoided, when conditioning on $\hat{\mathbf{Y}}^M$, the expected value of the error terms may still be assumed as zero.

The linear approximation for the logarithm can be considered good as long as Y_{aj} and $\hat{Y}_{aj}^M$ are strictly positive and reasonably close. As it is shown in Lemma 12 (page 69), if Y_{aj} and $\hat{Y}_{aj}^M$ are greater than zero and

$$\delta_{aj} := \left| \frac{Y_{aj} - \hat{Y}_{aj}^M}{Y_{aj}} \right| < k,$$

then the absolute value of the remainder of the linear approximation, $|R|$, is smaller than $\frac{1}{2}\left(\frac{k}{1-k}\right)^2$. For instance, if $\delta_{aj} < 0.1$, then $|R| < 0.006$.

As previously described in equations (1.4) and (1.5) in section 1.2, for known $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$, the BLUP of $\mathbf{u}$ under model (3.11) is given by

$$\tilde{\mathbf{u}} = \boldsymbol{\Sigma}_u \mathbf{V}^{-1} \left(\boldsymbol{\eta} - \mathbf{Z}\tilde{\boldsymbol{\Psi}}\right), \tag{3.12}$$

with

$$\tilde{\boldsymbol{\Psi}} = \left(\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}\right)^{-1} \mathbf{Z}^T\mathbf{V}^{-1}\boldsymbol{\eta} \tag{3.13}$$

for $\mathbf{V} = (\boldsymbol{\Sigma}_e + \boldsymbol{\Sigma}_u)$. Notice that $\hat{\boldsymbol{\alpha}}^Y$ and hence $\hat{\mathbf{Y}}^M$ have been assumed fixed. In our set-up, it has been assumed that $\boldsymbol{\Sigma}_e$ is known but the vector of variance components $\boldsymbol{\sigma}$ that governs $\boldsymbol{\Sigma}_u$, and hence $\boldsymbol{\Sigma}_u$ are unknown. Given an estimate $\hat{\boldsymbol{\sigma}}^2 = [\hat{\sigma}_1^2, \ldots, \hat{\sigma}_J^2]$ of $\boldsymbol{\sigma}^2$, the estimate

$$\hat{\boldsymbol{\Sigma}}_u = (\mathbf{C}_{(J)} \mathrm{diag}(\hat{\boldsymbol{\sigma}}^2) \mathbf{C}_{(J)}) \otimes \mathbf{C}_{(A)}$$

implies the estimate $\hat{\mathbf{V}} = \boldsymbol{\Sigma}_e + \boldsymbol{\Sigma}_u$. With $\hat{\boldsymbol{\Psi}} = \left(\mathbf{Z}^T \hat{\mathbf{V}}^{-1} \mathbf{Z}\right)^{-1} \mathbf{Z}^T \hat{\mathbf{V}}^{-1} \boldsymbol{\eta}$, the EBLUP of $\mathbf{u}$ is given by

$$\hat{\mathbf{u}} = \hat{\boldsymbol{\Sigma}}_u \hat{\mathbf{V}}^{-1} \left(\boldsymbol{\eta} - \mathbf{Z}\hat{\boldsymbol{\Psi}}\right). \tag{3.14}$$

Because the matrix $\hat{\boldsymbol{\Sigma}}_u$ is not full rank, the predicted set of random effects using the equation above satisfies $\hat{\mathbf{u}}_{a+} = \hat{\mathbf{u}}_{+j} = 0$ automatically.

Assuming normality of both $\mathbf{u}$ and $\boldsymbol{e}$, a Fisher-scoring algorithm can be used to obtain ML or REML estimates $\hat{\boldsymbol{\Psi}}$ and $\hat{\mathbf{u}}$, as described in section 1.2. It is straightforward to show that, given the structure of random effects assumed, the matrix $\mathbf{V}_{(j)}$ containing the first derivatives of $\mathbf{V}$ with respect to σ_j^2 is given by

$$(\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)})(\mathbb{1}_{(J)}^{jj} \otimes \mathbf{I}_{(A)})(\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)}),$$

where $\mathbb{1}_{(J)}^{jj}$ denotes a square matrix of dimension J with zeroes in all components except for the entry (j, j).

Without the normality assumption, the proposal is to carry out the estimation of the variance components using a moment-type estimator that we have developed for model (3.11), in the spirit of Henderson (1953), and which will be introduced next. Denote by $\tilde{e}_{aj}$, the residual corresponding to unit (a, j) from the Ordinary Least Squares (OLS) fit of model (3.24) and let $SSR_j = \sum_{a=1}^{A} \tilde{e}_{aj}^2$ be the sum of squares of the residuals for column j. The proposed estimator is given by the expression

$$\hat{\sigma}_j^2 = \frac{1}{(A-1)(J-1)(J-2)} \left(J(J-1)(SSR_j - \xi_j) - \sum_k (SSR_k - \xi_k) \right), \tag{3.15}$$

for $j = 1, \ldots, J$, where $\xi_j = \mathrm{Tr}\left[\mathbf{D}_j \boldsymbol{\Sigma}_e\right]$ with $\mathbf{D}_j = \left(\mathbf{C}_{(J)} \mathbb{1}_{(J)}^{jj} \mathbf{C}_{(J)}\right) \otimes \mathbf{C}_{(A)}$. $\mathbf{C}$ is the matrix defined in Lemma 5 (page 48) and $\mathbb{1}_{(J)}^{jj}$ was defined above. Theorem 16 (page 76), proves that $\hat{\sigma}_j^2$ is unbiased under model (3.11). Because an OLS rather than Weighted Least Squares (WLS) fitting has been used to develop the estimator, no iteration is required.

Unfortunately, ML, REML and the method proposed above, can produce negative solutions. In our experience, this is prone to happen when the variance components are very small, particularly if the survey sizes are small as well. A first possible solution is use truncated versions of the estimators, where negative estimates are replaced by zero but not recalculation of the remaining estimates is performed. Clearly, for an unbiased estimator such as the one proposed in equation (3.15), this would create a positive bias. Further study will be required to determine the impact of such bias in the MSE of the MMSPREE.

A second alternative to this situation is to truncate all negative estimates, eliminate the observations corresponding to those categories from model (3.11) and repeat the estimation for the remaining categories. Furthermore, in cases with more than one variance component with a negative estimate, this procedure can be performed in a progressive way, starting with the most negative estimate. This alternative, that may seem appealing in principle, has performed poorly in initial simulation studies. The recalculation of variance estimates has often derived in new negative estimates, sometimes up to the point where all variance components receive a zero estimate. A possible explanation for this behaviour arises from the use of interaction terms to model the relationship between the proxy and target compositions in SPREE-type estimators. Because a given set of interactions depends on all categories simultaneously, considering only a subset of them somehow seems to distort our observation of the relationships between the two compositions.

A final alternative, which has been attempted successfully in initial simulation studies and has also been used in the application that will be discussed in Chapter 5, is to perform the estimation of the variance components using hard sources, e.g., another proxy composition, rather than survey data. Even though the MMSPREE can lose some efficiency due to the lack of unbiasedness in such situation, gains can still be obtained in comparison with the synthetic version of the estimator.

So far it has been assumed that $\mathbf{\Sigma}_e$ is known. Given an estimate of the variance-covariance matrix of the direct estimators, $\mathbf{\Sigma}_{\hat{Y}}$, it is possible to approximate the variance of the error terms on the log-scale using the Taylor approximation (3.10), to obtain

$$\mathbf{\Sigma}_e \approx \mathbf{G}\mathbf{\Sigma}_{\hat{Y}}\mathbf{G},$$

for $\mathbf{G} = [\text{diag}(1/\text{vec}(\hat{\mathbf{Y}}^{M}))]$.

3.3 MSE estimation

In section 2.3 we presented alternatives to estimate two different quantities related to the uncertainty of the MSPREE estimator. An unconditional MSE where the expectation is taken both over $\mathbf{Y}$ and over the sampling mechanism, and a Finite-Population MSE that only considers the uncertainty associated with the selection of the sample. In an analogous way, to study the uncertainty of the MMSPREE we will be interested in obtaining estimates for the quantities:

$$\text{MSE}(\hat{\mathbf{Y}}^{MM}) = E(\hat{\mathbf{Y}}^{MM} - \mathbf{Y})^2,$$

and

$$\text{FP-MSE}(\hat{\mathbf{Y}}^{MM}) = E(\hat{\mathbf{Y}}^{MM} - \mathbf{Y}|\mathbf{Y}).$$

For each one of those targets of estimation, a bootstrap procedure will be proposed. Finally, notice that these procedures can also be used to obtain estimates of MSE and FP-MSE for the MSPREE that take into account possible misspecification of the MSPREE in the sense of (3.1).

3.3.1 Estimation of MSE($\hat{\mathbf{Y}}^{MM}$)

Estimation of MSE($\hat{\mathbf{Y}}^{MM}$) can be performed using the parametric bootstrap under assumption (3.2), given the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}_u$ obtained from the original sample, and assuming normality for the random effects u_{aj}. The proposed procedure follows:

For $b = 1, \dots, B$,

1. Generate independent random variables $\vartheta^b_{aj} \sim N(0, \hat{\sigma}^2_j)$ for $a = 1, \dots, A$; $j = 1, \dots, J$. Arrange them in the $(A \times J)$ matrix $\boldsymbol{\vartheta}^b$.
2. Calculate $\mathbf{u}^b = \text{vec}(\mathbf{C}_{(A)} \boldsymbol{\vartheta}^b \mathbf{C}_{(J)})$.
3. Make $\alpha^{Y^b}_{aj} = \sum_l \hat{\beta}_{jl} \alpha^X_{aj} + u^b_{aj}$, for $a = 1, \dots, A$; $j = 1 \dots, J$.
4. Obtain the population composition $\mathbf{Y}^b = \text{IPF}\left(\exp \alpha^{Y^b}_{aj}; \mathbf{Y}_{a+}, \mathbf{Y}_{+j}\right)$.
5. Select a sample from the population defined by the $\mathbf{Y}^b$.

6. Using the data in the sample, obtain the MMSPREE of $\mathbf{Y}^b$, $\hat{\mathbf{Y}}^{MM,b}$.

The quantity of interest can be estimated using the Monte Carlo approximation:

$$\widehat{\text{MSE}}(\hat{\mathbf{Y}}^{MM}) = \frac{1}{B}\sum_b \text{vec}\left(\hat{\mathbf{Y}}^{MM,b} - \mathbf{Y}^b\right)\text{vec}\left(\hat{\mathbf{Y}}^{MM,b} - \mathbf{Y}^b\right)^T.$$

Furthermore, having obtained estimates of the variance components σ_j^2 for $j = 1, \ldots, J$, it is possible to substitute the MMSPREE by the MSPREE in step 6 above, to obtain a bootstrap estimate of $\text{MSE}(\hat{\mathbf{Y}}^M)$ under a more general structural assumption than the estimator proposed in section 2.3.2. This seems practically appealing because it would provide protection against possible misspecification of the MSPREE.

3.3.2 Estimation of FP-MSE$(\hat{\mathbf{Y}}^{MM})$

The interest in this section is to obtain an estimate of the MSE where the population of interest has been set as fixed and only the variability associated with the sampling mechanism is taken into account. Under assumption (3.2), the MMSPREE estimate based on the original sample, $\hat{\mathbf{Y}}^{MM}$ is a suitable composition to play the role of fixed population for evaluation purposes. The procedure then is to repeatedly select samples from the population defined by the composition $\hat{\mathbf{Y}}^{MM}$, and calculate in each of them the MMSPREE, $\hat{\mathbf{Y}}^{MM}$. Then, the Monte Carlo approximation

$$\widehat{\text{FP-MSE}}(\hat{\mathbf{Y}}^{MM}) := \frac{1}{B}\sum_b \text{vec}\left(\hat{\mathbf{Y}}^{MM,b} - \hat{\mathbf{Y}}^{MM}\right)\text{vec}\left(\hat{\mathbf{Y}}^{MM,b} - \hat{\mathbf{Y}}^{MM}\right)^T,$$

can be used to estimate the FP-MSE$(\hat{\mathbf{Y}}^{MM}) = E(\hat{\mathbf{Y}}^{MM} - \mathbf{Y}|\mathbf{Y})$. As with the method proposed in section 3.3.1, this procedure can be used to obtain an estimate of the finite population MSE of the MSPREE that takes into account possible misspecification of the structural assumption of the MSPREE.

3.4 Proofs

Definition $\mathbf{u}$ and induced variance

This lemma shows how to obtain the random effects $\mathbf{u}$ of the MMSPREE, starting with the matrix of random variables ϑ. Its corollary specifies the variance of $\mathbf{u}$ as induced from ϑ. The lemma and its corollary are used in the proof of Theorem 15, and referenced in the definition of the MMSPREE, page 60, as well as in page 62.

Lemma 11. *Let* $\{\vartheta_{11}, \dots, \vartheta_{AJ}\}$ *be a set of independent random variables with* $\mathrm{E}(\vartheta_{aj}) = 0$ *and* $\mathrm{V}(\vartheta_{aj}) = \sigma_j^2$ *for* $a = 1, \dots, A$; $j = 1 \dots, J$, *arranged in a matrix of dimension* $(A \times J)$ *denoted by* $\boldsymbol{\vartheta}$. *Thus:*

1. *The product* $\mathbf{C}_{(A)}\boldsymbol{\vartheta}\mathbf{C}_{(J)}$, *with* $\mathbf{C}$ *defined as in Lemma 5, sums to zero by row and column.*

2. *Let* $\mathbf{u}$ *be the column vector of dimension* (AJ) *defined as* $\mathbf{u} = \text{vec}\left(\mathbf{C}_{(A)}\boldsymbol{\vartheta}\mathbf{C}_{(J)}\right)$, *where* vec *represents the vector operator. Then*

$$\boldsymbol{\Sigma}_u := \mathrm{V}(\mathbf{u}) = (\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)})\boldsymbol{\Sigma}_\vartheta(\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)})$$

for $\boldsymbol{\Sigma}_\vartheta := \mathrm{V}(\text{vec}(\boldsymbol{\vartheta})) = \text{diag}(\boldsymbol{\sigma}^2) \otimes \mathbf{I}_A$ *and* $\boldsymbol{\sigma}^2 = [\sigma_1^2, \dots, \sigma_J^2]^T$

Proof. The proof of the first item is a direct application of Lemma 5. Notice that this item implies that the row and column margins of $\mathbf{C}_{(A)}\boldsymbol{\vartheta}\mathbf{C}_{(J)}$ are non random and equal to zero for any realisation of $\boldsymbol{\vartheta}$.

To prove the second item we will use a relationship between the vector operator and matrix multiplication. According to Theorem 16.2.1 in Harville (1997), for any matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ such that the product $\mathbf{ABC}$ is defined,

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\text{vec}(\mathbf{B}).$$

An application of this result in the case of $\mathbf{u}$, leads to

$$\mathbf{u} = (\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)})\text{vec}(\boldsymbol{\vartheta}),$$

given that the matrix $\mathbf{C}$ is symmetric, as proven in Lemma 5. Moreover, since the Kronecker product of two symmetric matrices is symmetric (see for instance Harville, 1997, equation 1.15, page 336) it is possible to conclude that $(\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)})$ is a symmetric matrix of constants. The desired expression is then obtained using the bilinearity property of the variance. □

Corollary 11.1. *The covariance assumed between components* u_{aj} *and* u_{al} *defined as* *in Lemma 11 is:*

$$\mathrm{Cov}(u_{aj}, u_{rl}) = \begin{cases} \frac{(A-1)}{AJ}\left[(J-2)\sigma_j^2 + \bar{\sigma}^2\right] & \textit{for } a = r, j = l \\ \frac{-1}{AJ}\left[(J-2)\sigma_j^2 + \bar{\sigma}^2\right] & \textit{for } a \neq r, j = l \\ \frac{-(A-1)}{AJ}\left[\sigma_j^2 + \sigma_l^2 - \bar{\sigma}^2\right] & \textit{for } a = r,\ j \neq l \\ \frac{1}{AJ}\left[\sigma_j^2 + \sigma_l^2 - \bar{\sigma}^2\right] & \textit{for } a \neq r,\ j \neq l \end{cases}$$

with $\bar{\sigma}^2 = J^{-1}(\sigma_+^2)$.

Remainder of the linear approximation to $\log Y_{aj}$

This lemma presents a bound for the remainder of the linear approximation applied to $\log Y_{aj}$ as part of the process of estimation of the MMSPREE. It is referenced in page 63.

Lemma 12. *For* Y_{aj} *and* $\hat{Y}_{aj}^M$ *strictly positive such that*

$$\left|\frac{Y_{aj} - \hat{Y}_{aj}^M}{Y_{aj}}\right| \leqslant k,$$

a bound for the remainder of the Taylor approximation of order 1 to $\log Y_{aj}$ *around* $\hat{Y}_{aj}^M$ *is given by*

$$\left(-\frac{1}{2}\left(\frac{k}{1-k}\right)^2, 0\right).$$

Proof. The first part of this proof uses the so-called *Lagrange's form of the remainder*, which can be found in many Calculus textbooks (see for instance Stewart, 2008, page 738). For $f(x)$ a continuous function, with derivatives of all orders, if the $n+1$ derivative of f is continuous on an open interval I that contains a, and x is in I, then the remainder of the Taylor approximation of order n of $f(x)$ around a, $R(x)$, satisfies:

$$R(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}, \tag{3.16}$$

for some number c between a and x.

The linear approximation in consideration satisfies the requirements of the formula above as long as Y_{aj} and $\hat{Y}_{aj}^M$ are strictly positive. Hence, under those conditions, there exists a number c between Y_{aj} and $\hat{Y}^M_{aj}$ such that

$$R(Y_{aj}) = -\frac{1}{2c^2}(Y_{aj} - \hat{Y}^M_{aj})^2. \tag{3.17}$$

Notice that equation (3.17) corresponds to an inverted parabola with vertex in $\left(\hat{Y}^M_{aj}, 0\right)$. Moreover, as $R(Y_{aj}) \leqslant 0$, the linear approximation overestimates the value of the function for all values of Y_{aj}.

For each Y_{aj}, the actual value of c for which equation (3.17) holds is unknown. However, it is possible to build a lower bound for $R(Y_{aj})$ considering the values of c that would lead to a minimum value. For fixed Y_{aj} and $\hat{Y}^M_{aj}$, the minimum of $R(Y_{aj})$ is obtained when $c \to 0$. However, as c is between Y_{aj} and $\hat{Y}^M_{aj}$, the minimum is obtained when $c = \min(Y_{aj}, \hat{Y}^M_{aj})$, i.e., for $c = Y_{aj}$ when $Y_{aj} < \hat{Y}^M_{aj}$ and for $c = \hat{Y}^M_{aj}$ when $\hat{Y}^M_{aj} < Y_{aj}$.

Now we will connect the lower bound of the reminder with the relative error δ_{aj} given in the enunciate of the lemma. For the case where $Y_{aj} < \hat{Y}^M_{aj}$, we have

$$0 \geqslant R(Y_{aj}) \geqslant -\frac{1}{2}\left(\frac{Y_{aj} - \hat{Y}^M_{aj}}{Y_{aj}}\right)^2 \geqslant -\frac{1}{2}k^2. \tag{3.18}$$

On the other hand, from the enunciate of the lemma,

$$\left|\frac{Y_{aj} - \hat{Y}^M_{aj}}{Y_{aj}}\right| \leqslant k \quad \Leftrightarrow \quad \left(\frac{Y_{aj} - \hat{Y}^M_{aj}}{\hat{Y}^M_{aj}}\right)^2 \leqslant \left(\frac{kY_{aj}}{\hat{Y}^M_{aj}}\right)^2, \tag{3.19}$$

for $\hat{Y}^M_{aj} > 0$. Moreover,

$$-k \leqslant \frac{Y_{aj} - \hat{Y}^M_{aj}}{Y_{aj}} \leqslant k \quad \Leftrightarrow \quad -k \leqslant 1 - \frac{\hat{Y}^M_{aj}}{Y_{aj}} \leqslant k.$$

Subtracting 1 in all sides of the inequality, inverting and multiplying by (-1) it is possible to obtain

$$\frac{k}{k+1} \leqslant \frac{kY_{aj}}{\hat{Y}^M_{aj}} \leqslant \frac{k}{1-k},$$

for $k > 0$; $k \neq 1$. Applying the square and taking the right hand side of the inequality, we arrive to the inequality

$$\left(\frac{kY_{aj}}{\hat{Y}^M_{aj}}\right)^2 < \left(\frac{k}{1-k}\right)^2. \tag{3.20}$$

Finally, substituting equation (3.20) in the right hand side of (3.19) leads to

$$\left(\frac{Y_{aj}-\hat{Y}^M_{aj}}{\hat{Y}^M_{aj}}\right)^2 < \left(\frac{k}{1-k}\right)^2.$$

Going back to the case where $\hat{Y}^M_{aj} < Y_{aj}$, the previous equation together with the bound for $R(Y_{aj})$ obtained when $c = \hat{Y}^M_{aj}$ allows us to conclude that

$$0 \geqslant R(Y_{aj}) \geqslant -\frac{1}{2}\left(\frac{Y_{aj}-\hat{Y}^M_{aj}}{\hat{Y}^M_{aj}}\right)^2 > -\frac{1}{2}\left(\frac{k}{1-k}\right)^2. \tag{3.21}$$

Between the two bounds proposed for $R(Y_{aj})$ in equations (3.18) and (3.21), the wider bound is then chosen. □

Unbiasedness of $\hat{\sigma}^2_j$

The remaining lemmas and theorems of this chapter intend to show the unbiasedness of the estimator for the variance components proposed in equation (3.15) (page 64). Lemma 13 shows the specific form of the projection matrix $\mathbf{P} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ under model (3.11). Lemma 14 shows the specific form of the matrix $\mathbf{D}_j$ induced by $\mathbf{P}$. Theorem 15 uses the previous lemmas to calculate the expected value of the sum of squares of the residuals of model (3.11), for a given column. Finally, Theorem 16 shows the unbiasedness of the proposed estimator. Only the latter theorem is referenced in the document, in page 64.

Lemma 13. *Let* $\mathbf{Z}$ *be a design matrix of dimension* $(AJ \times (A + J - 1))$ *given by*

$$\mathbf{Z} = \left[\begin{array}{c;{2pt/2pt}c} \mathbf{1}_{(J\times 1)} \otimes \mathbf{I}_{(A)} & \mathbf{T} \otimes \mathbf{1}_{(A\times 1)} \end{array}\right]$$

with $\mathbf{T}_{(J\times(J-1))} = \left[\begin{array}{c;{2pt/2pt}c} \mathbf{I} & -\mathbf{1} \end{array}\right]^T$. *The projection matrix* $\mathbf{P} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ *is given by the Kronecker product of matrices*

$$\mathbf{P} = \mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)},$$

with $\mathbf{C}$ *the matrix defined in Lemma 5 (page 48), i.e.,* $\mathbf{C}_{(K)} = \mathbf{I}_{(K)} - \frac{1}{K}\mathbf{1}_{(K)}$.

Proof. It is straightforward to see that for Z defined as above,

$$\mathsf{Z}^{\mathsf{T}}\mathsf{Z} = \left[\begin{array}{c:c} J\mathbf{I}_{(A)} & \mathbf{0}_{(A\times(J-1))} \\ \hdashline \mathbf{0}_{((J-1)\times A)} & A(\mathbf{1}+\mathbf{I})_{(J-1)} \end{array}\right].$$

Moreover, for a matrix $\mathbf{M} = (\mathbf{1}+\mathbf{I})$ of dimension $(J-1)$, $\mathbf{M}^{-1} = \mathbf{I} - \frac{1}{J}\mathbf{1}$. Using that argument and a lemma for the inverse of block diagonal matrices, (see for instance Harville, 1997, sec. 8.5),

$$\left(\mathsf{Z}^{\mathsf{T}}\mathsf{Z}\right)^{-1} = \frac{1}{J}\left[\begin{array}{c:c} \mathbf{I}_{(A)} & \mathbf{0}_{(A\times(J-1))} \\ \hdashline \mathbf{0}_{((J-1)\times A)} & \frac{1}{A}\left(J\mathbf{I}-\mathbf{1}\right)_{(J-1)} \end{array}\right].$$

Furthermore,

$$\left(\mathsf{Z}^{\mathsf{T}}\mathsf{Z}\right)^{-1}\mathsf{Z}^{\mathsf{T}} = \left[\begin{array}{c} \frac{1}{J}\left(\mathbf{1}_{(1\times J)} \otimes \mathbf{I}_{(A)}\right) \\ \hdashline \frac{1}{A}\left[\left(\mathbf{I}_{(J-1)} \otimes \mathbf{1}_{(1\times A)}\right) - \frac{1}{J}\mathbf{1}_{(J-1)\times(A(J-1))}\right] \;\Big|\; \frac{-1}{AJ}\mathbf{1}_{(J-1)\times A} \end{array}\right].$$

The *hat* matrix $\mathsf{H} = \mathsf{Z}(\mathsf{Z}^{\mathsf{T}}\mathsf{Z})^{-1}\mathsf{Z}^{\mathsf{T}}$ is in this case given by

$$\mathsf{H} = \frac{-1}{AJ}\mathbf{1}_{(AJ)} + \frac{1}{J}\left(\mathbf{1}_{(J)} \otimes \mathbf{I}_{A}\right) + \frac{1}{A}\left(\mathbf{I}_{(J)} \otimes \mathbf{1}_{(A)}\right). \tag{3.22}$$

writing $\mathbf{1}_{(AJ)}$ as $\mathbf{1}_{(A)} \otimes \mathbf{1}_{(J)}$, and $\mathbf{I}_{(AJ)}$ as $\mathbf{I}_{(A)} \otimes \mathbf{I}_{(J)}$ and using the property of the mixed product of the kronecker product (see for instance Harville, 1997, section 16.1) we obtain,

$$\mathsf{P} = \left(\mathbf{I}_{(J)} - \frac{1}{J}\mathbf{1}_{J}\right) \otimes \left(\mathbf{I}_{(A)} - \frac{1}{A}\mathbf{1}_{A}\right),$$

which corresponds to the desired expression according to the definition of $\mathbf{C}$. Notice that P is symmetric. □

Lemma 14. *Consider* P *defined as in Lemma 13 and denote* $\boldsymbol{\Delta}_j = \left(\mathbb{1}^{j}_{(1\times J)} \otimes \mathbf{I}_{A}\right)$, *where* $\mathbb{1}^{j}_{(1\times J)}$ *denotes a row vector of dimension* J *with value of 1 on the* j*-th component and zero everywhere else. Then:*

$$\mathbf{D}_j := \mathsf{P}\,\boldsymbol{\Delta}_j^{\mathsf{T}}\boldsymbol{\Delta}_j\mathsf{P} = \left(\mathbf{C}_{(J)}\mathbb{1}^{jj}_{(J)}\mathbf{C}_{(J)}\right) \otimes \mathbf{C}_{(A)},$$

where $\mathbb{1}^{jj}_{(J)}$ *and* $\mathbb{1}^{\circ j}_{(J)}$ *represent square matrices of dimension* J*, with zeroes in all components except for the one indicated by the superscript, that receives the value 1. The* *symbol* $\circ$ *indicates all elements on the index which substitutes, e.g,* $\circ j$ *indicates column j.*

Proof. According to the definition,

$$\begin{aligned}\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j &= \left(\mathbb{1}^j_{(1\times J)}\otimes \mathbf{I}_A\right)^T\left(\mathbb{1}^j_{(1\times J)}\otimes \mathbf{I}_A\right)\\ &= \mathbb{1}^{jj}_{(J)}\otimes \mathbf{I}_A,\end{aligned} \tag{3.23}$$

given the mixed product property of the Kronecker product and the identity $\mathbb{1}^{jj}_{(J)} = \mathbb{1}^j_{(J\times 1)}\mathbb{1}^j_{(1\times J)}$. Notice that $\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j = \boldsymbol{\Delta}_j\boldsymbol{\Delta}_j^T$. Using again the mixed product property, the definition of $\mathbf{P}$ and equation (3.23) we have

$$\mathbf{P}\,\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j = \left(\mathbf{C}_{(J)}\mathbb{1}^{jj}_{(J)}\right)\otimes \mathbf{C}_{(A)},$$

The desired result is obtained multiplying the equation above on the right by $\mathbf{P}$, using the mixed product property of the Kronecker product and the idempotence of $\mathbf{C}$ shown in Lemma 5. Notice that as $\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j$ and $\mathbf{P}$ are both symmetric, $\mathbf{D}_j$ is also symmetric. □

Theorem 15. *Consider the model*

$$\eta_{aj} = \gamma_a + \lambda_j + u_{aj} + e_{aj} \tag{3.24}$$

for $a = 1,\ldots,A$; $j = 1,\ldots,J$, *with* γ_a *and* λ_j *sets of unknown parameters satisfying* $\lambda_+ = 0$, *and where* u_{aj} *are random effects and* e_{aj} *random errors. It will be assumed that the terms* u_{aj} *are the components of the vector* $\mathbf{u}$ *defined in Lemma 11, hence* $\mathrm{E}(\mathbf{u}) = \mathbf{0}$ *and* $\mathrm{V}(\mathbf{u}) = \boldsymbol{\Sigma}_u$ *unknown, determined by the vector of variance components* $\boldsymbol{\sigma}^2 = (\sigma_1^2,\ldots,\sigma_J^2)$. *The error terms* e_{aj} *are arranged as* $\mathbf{e} = [e_{11}, e_{21},\ldots,e_{1J},\ldots,e_{AJ}]^T$, *with* $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ *and* $\mathrm{V}(\mathbf{e}) = \boldsymbol{\Sigma}_e$ *known. Furthermore, it will be assumed that* $\mathrm{Cov}(\mathbf{u},\mathbf{e}) = \mathbf{0}$. *Denote by* $\tilde{e}_{aj}$, *the residual corresponding to unit* (a,j) *on the Ordinary Least Squares (OLS) fit of model* (3.24) *and let* $\mathrm{SSR}_j = \sum_{a=1}^{A}\tilde{e}^2_{aj}$ *be the sum of squares of the residuals for column j. Then,*

$$\mathrm{E}\left[\mathrm{SSR}_j\right] = \frac{A-1}{J^2}\left[J(J-2)\sigma_j^2 + \sigma_+^2\right] + \xi_j$$

where $\sigma_+^2 = \sum_j \sigma_j^2$ *and* $\xi_j = \mathrm{Tr}\left[\mathbf{D}_j\boldsymbol{\Sigma}_e\right]$ *with* $\mathbf{D}_j = \left(\mathbf{C}_{(J)}\mathbb{1}^{jj}_{(J)}\mathbf{C}_{(J)}\right)\otimes \mathbf{C}_{(A)}$ *and* $\mathbf{C}$ *is the matrix defined in Lemma 5.*

Proof. Equation (3.24) can be written in matrix notation as:

$$\eta = \mathbf{Z}\boldsymbol{\Psi} + \mathbf{u} + \mathbf{e} \tag{3.25}$$

where $\boldsymbol{\eta}$, $\mathbf{u}$ and $\boldsymbol{e}$ are column vectors of dimension AJ, $\mathbf{Z}$ is the matrix defined in Lemma 13, and $\boldsymbol{\Psi}$ is a column vector of dimension $A + J - 1$ given by

$$\boldsymbol{\Psi} = \begin{bmatrix} \gamma_1 & \dots & \gamma_A & \lambda_1 & \dots & \lambda_{J-1} \end{bmatrix}^T.$$

In order to derive an unbiased estimator for the variance components, we will study the expectation of the sum by column of the squares of the residuals (SSR) of an Ordinary Least Squares (OLS) fit of (3.25). We will start by remembering that the OLS estimator of $\boldsymbol{\Psi}$ under this model is $\tilde{\boldsymbol{\Psi}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\eta}$. The expected value of the SSR corresponding to column j is given by

$$\mathrm{E}\left[\mathrm{SSR}_j\right] = \mathrm{E}\left[\left(\boldsymbol{\eta}_j - \mathbf{Z}_j\tilde{\boldsymbol{\Psi}}\right)^T\left(\boldsymbol{\eta}_j - \mathbf{Z}_j\tilde{\boldsymbol{\Psi}}\right)\right], \tag{3.26}$$

where $\boldsymbol{\eta}_j = \boldsymbol{\Delta}_j\boldsymbol{\eta}$ and $\mathbf{Z}_j = \boldsymbol{\Delta}_j\mathbf{Z}$ are the corresponding submatrices when only the rows corresponding to the j-th column are selected, and $\boldsymbol{\Delta}_j$ is defined as in Lemma 14. Using matrix $\boldsymbol{\Delta}_j$ and the property of cyclic permutations of the trace (see for instance Harville, 1997, section 5.2), equation (3.26) can be written as

$$\mathrm{E}\left[\mathrm{SSR}_j\right] = \mathrm{Tr}\left[\mathrm{E}\left[\left(\boldsymbol{\eta} - \mathbf{Z}\tilde{\boldsymbol{\Psi}}\right)\left(\boldsymbol{\eta} - \mathbf{Z}\tilde{\boldsymbol{\Psi}}\right)^T\right]\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j\right],$$

or, by using the identity $\boldsymbol{\eta} - \mathbf{Z}\tilde{\boldsymbol{\Psi}} = \mathbf{P}\boldsymbol{\eta}$, with $\mathbf{P}$ the projection matrix defined by $\mathbf{P} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$, as

$$\mathrm{E}\left[\mathrm{SSR}_j\right] = \mathrm{Tr}\left[\mathbf{P}\,\mathrm{E}[\boldsymbol{\eta}\boldsymbol{\eta}^T]\mathbf{P}\,\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j\right]. \tag{3.27}$$

Regarding $\mathrm{E}[\boldsymbol{\eta}\boldsymbol{\eta}^T]$, we have

$$\begin{aligned}\mathrm{E}[\boldsymbol{\eta}\boldsymbol{\eta}^T] &= \mathrm{E}\left[(\mathbf{Z}\boldsymbol{\Psi} + \mathbf{u} + \boldsymbol{e})(\mathbf{Z}\boldsymbol{\Psi} + \mathbf{u} + \boldsymbol{e})^T\right] \\ &= \mathbf{Z}\boldsymbol{\Psi}\boldsymbol{\Psi}^T\mathbf{Z}^T + \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e, \end{aligned} \tag{3.28}$$

due to the assumptions of zero mean for both $\mathbf{u}$ and $\boldsymbol{e}$ and independence between them. Substituting (3.28) in (3.27), we obtain

$$\begin{aligned}\mathrm{E}\left[\mathrm{SSR}_j\right] &= \mathrm{Tr}\left[\mathbf{P}\mathbf{Z}\boldsymbol{\Psi}\boldsymbol{\Psi}^T\mathbf{Z}^T\mathbf{P}\,\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j\right] + \mathrm{Tr}\left[\mathbf{P}\,(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e)\mathbf{P}\,\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j\right] \\ &= \mathrm{Tr}\left[(\boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e)\mathbf{P}\,\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j\mathbf{P}\right] \\ &= \mathrm{Tr}\left[\mathbf{D}_j\boldsymbol{\Sigma}_u\right] + \mathrm{Tr}\left[\mathbf{D}_j\boldsymbol{\Sigma}_e\right], \end{aligned} \tag{3.29}$$

with $\mathbf{D}_j = \mathbf{P}\,\boldsymbol{\Delta}_j^T\boldsymbol{\Delta}_j\mathbf{P}$. The second line is obtained using the property of cyclic permutation of the trace and the orthogonality between the projection matrix

$\mathbf{P}$ and $\mathbf{Z}$.

The particular form of $\mathbf{D}_j$ under model (3.24) is given by Lemma 14,

$$\mathbf{D}_j = \left(\mathbf{C}_{(J)}\mathbb{1}^{jj}_{(J)}\mathbf{C}_{(J)}\right) \otimes \mathbf{C}_{(A)}, \tag{3.30}$$

and the form of $\mathbf{\Sigma}_u$ is given by Lemma 11,

$$\mathbf{\Sigma}_u = (\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)})\mathbf{\Sigma}_\vartheta(\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)}) \tag{3.31}$$

with $\mathbf{\Sigma}_\vartheta = \text{diag}(\boldsymbol{\sigma}^2) \otimes \mathbf{I}_A$ and $\boldsymbol{\sigma}^2 = [\sigma^2_1, \dots, \sigma^2_J]^T$. Substituting (3.31) into $\mathbf{D}_j\mathbf{\Sigma}_u$ and using the property of cyclic permutations of the trace,

$$\text{Tr}\left[\mathbf{D}_j\mathbf{\Sigma}_u\right] = \text{Tr}\left[(\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)})\mathbf{D}_j(\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)})\mathbf{\Sigma}_\vartheta\right].$$

Using the definition of $\mathbf{D}_j$ given by equation (3.30), the mixed product property of the Kronecker product and the idempotence of $\mathbf{C}$, it is possible to write

$$(\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)})\mathbf{D}_j(\mathbf{C}_{(J)} \otimes \mathbf{C}_{(A)}) = \mathbf{D}_j,$$

hence,

$$\text{Tr}\left[\mathbf{D}_j\mathbf{\Sigma}_u\right] = \text{Tr}\left[\mathbf{D}_j\mathbf{\Sigma}_\vartheta\right]. \tag{3.32}$$

Moreover, according to the definition of $\mathbf{D}_j$ and $\mathbf{\Sigma}_\vartheta$,

$$\begin{aligned}
\text{Tr}\left[\mathbf{D}_j\mathbf{\Sigma}_\vartheta\right] &= \text{Tr}\left[\left((\mathbf{C}_{(J)}\mathbb{1}^{jj}_{(J)}\mathbf{C}_{(J)}) \otimes \mathbf{C}_{(A)}\right)\left(\text{diag}(\boldsymbol{\sigma}^2) \otimes \mathbf{I}_A\right)\right] \\
&= \text{Tr}\left[\left(\mathbf{C}_{(J)}\mathbb{1}^{jj}_{(J)}\mathbf{C}_{(J)}\text{diag}(\boldsymbol{\sigma}^2)\right) \otimes \mathbf{C}_{(A)}\right] \\
&= \text{Tr}\left[\left(\mathbf{C}_{(J)}\mathbb{1}^{jj}_{(J)}\mathbf{C}_{(J)}\text{diag}(\boldsymbol{\sigma}^2)\right)\right]\text{Tr}\left[\mathbf{C}_{(A)}\right].
\end{aligned} \tag{3.33}$$

The mixed product property is used to go from line 1 to 2. Line 3 is obtained using the property: $\text{Tr}(A \otimes B) = \text{Tr}(A)\text{Tr}(B)$ (Harville, 1997, chapter 16, equation 1.26).

Now we will focus on the first term on the right hand side. Using the property of cyclic permutations of the trace we can write

$$\text{Tr}\left[\left(\mathbf{C}_{(J)}\mathbb{1}^{jj}_{(J)}\mathbf{C}_{(J)}\text{diag}(\boldsymbol{\sigma}^2)\right)\right] = \text{Tr}\left[\left(\mathbb{1}^{jj}_{(J)}\mathbf{C}_{(J)}\text{diag}(\boldsymbol{\sigma}^2)\mathbf{C}_{(J)}\right)\right].$$

It is straightforward to see that if $\tilde{\mathbf{M}}$ is a square matrix of dimension (J), the product $\mathbb{1}^{jj}_{(J)}\tilde{\mathbf{M}}$ has as result a matrix with all components zero except for the j-th row, that remains as in $\tilde{\mathbf{M}}$. Hence, $\mathrm{Tr}(\mathbb{1}^{jj}_{(J)}\tilde{\mathbf{M}}) = \tilde{M}_{jj}$. Similarly, it is straightforward to see that the product $\tilde{\mathbf{M}} = \mathbf{CMC}$ has components

$$\tilde{M}_{jl} = M_{jl} - \bar{M}_{j+} - \bar{M}_{+l} - \bar{M}_{++}.$$

Taking $\mathbf{M} = \mathrm{diag}(\boldsymbol{\sigma})$ we obtain

$$\mathrm{Tr}\left[\left(\mathbb{1}^{jj}_{(J)}\mathbf{C}_{(J)}\mathrm{diag}(\boldsymbol{\sigma}^2)\mathbf{C}_{(J)}\right)\right] = \frac{1}{J^2}\left(J(J-2)\sigma_j^2 + \sigma_+^2\right), \tag{3.34}$$

which, substituted in (3.33) and then in (3.32) leads to

$$\mathrm{Tr}\left[\mathbf{D}_j\boldsymbol{\Sigma}_u\right] = \frac{(A-1)}{J^2}\left(J(J-2)\sigma_j^2 + \sigma_+^2\right), \tag{3.35}$$

given that $\mathrm{Tr}\left[\mathbf{C}_{(A)}\right] = (A-1)$. The proof is completed by substituting (3.35) in (3.29). □

Theorem 16. *In the same conditions of Theorem 15, an unbiased estimator of the variance component* σ_j^2 *is given by*

$$\hat{\sigma}_j^2 = \frac{1}{(A-1)(J-1)(J-2)}\left(J(J-1)(SSR_j - \xi_j) - \sum_k (SSR_k - \xi_k)\right), \tag{3.36}$$

for $j = 1, \dots, J$.

Proof. The proof is immediate taking the expectation on both sides of (3.36) and using Theorem 15. □

Chapter 4

Simulation Exercise

This chapter presents a simulation exercise that was carried out with the purpose of illustrating the behaviour of the MSPREE and MMSPREE, as well as to test the proposed methodologies for the estimation of their MSEs. In an attempt to reflect a realistic situation, a specific estimation problem was identified, and real data was used to set up the characteristics of the simulation.

The target of estimation for this exercise is a population composition conformed by individuals, ages 16 and over, disaggregated according to the highest academic qualification they have obtained and the Local Authority (LA) they reside in. Real versions of this composition, can be obtained from the 2001 and 2011 population censuses in England and Wales, through the website NOMIS provided by the ONS, https://www.nomisweb.co.uk. Those two census compositions were used to generate the target and proxy compositions for the simulation exercise as it will be explained next.

The row and column margins of the 2011 census composition, denoted $\mathbf{Y}_{a+}$ and $\mathbf{Y}_{+j}$ for $a = 1, \ldots, A$ and $j = 1, \ldots, J$, constitute the column and row margins of the target composition and are kept fixed through all the simulation exercise. The inner cells of the target composition, on the other hand, are generated using one out of three possible scenarios of association structure: 1) the MSPREE structural assumption, 2) the MMSPREE structural assumption, or 3) the association structure observed in the census 2011 composition. In order to provide some degree of comparability across scenarios, the observed composition of the 2001 census, denoted $\mathbf{X}$, was used to generate the target association structures for scenarios 1 and 2, and also used as proxy for the calculation of the SPREE-type estimators for all three scenarios. Moreover, the two census compositions were used to set the matrix of coefficients $\boldsymbol{\beta}$ of the

MSPREE and MMSPREE structural assumptions, i.e., for scenarios 1 and 2. The variance components required for scenario 2, were also motivated by the real data.

In order to evaluate the characteristics of the point estimators as well as the proposed methodologies for the estimation of their MSE's, a double bootstrap approach was implemented. In each iteration, a population composition $\mathbf{Y}$ was generated, with association structure given by one of the scenarios and with margins $\mathbf{Y}_{a+}$ and $\mathbf{Y}_{+j}$. Then, a sample composition $\mathbf{y}$ was selected from $\mathbf{Y}$ and the SPREE-type estimators were calculated. Furthermore, bootstrap samples were selected from each $\mathbf{y}$, using the procedures described in sections 2.3 and 3.3 of this document, in order to obtain estimates of the MSEs of the MSPREE and MMSPREE for each $\mathbf{Y}$.

The simulation exercise provides point estimates of the different SPREE-type estimators, as well as estimates of their MSEs. Point estimates are analysed to study the existence of bias and also used to calculate Monte Carlo estimates of the true MSEs, which are then used to compare the performance of different SPREE-type estimators, and to study the bias of the methodologies proposed previously in sections 2.3 and 3.3 for MSE estimation.

This chapter is divided into five sections as follows. Section 4.1 presents a brief description of the estimation problem according to the real data provided by the 2001 and 2011 census compositions, as well as some preliminary analysis using SPREE-type estimators, with the aim of providing background for the analysis of results. Section 4.2 presents in more detail the three scenarios used to generate the association structure of the target compositions. Section 4.3 defines the criteria used to compare the different estimators and their MSE's. Section 4.4 describes and discusses the main results of the simulation. Finally, section 4.5 presents complementary tables and figures that may be helpful to further understand the results of this exercise.

4.1 The estimation problem in the 2001 and 2011 census compositions

It was previously mentioned that, in this simulation exercise, we are interested in the estimation of a composition that disaggregates the population according to the highest academic qualification they have obtained and the LA they reside in. Data regarding all academic qualifications possessed by a person is collected for all persons living in England or Wales and aged 16-74, in questions 16 and 17 on the 2001 population census, and for all persons aged 16 or over, in question 25 on the 2011 population census. ONS makes available data regarding the Highest Qualification attained by each individual. Despite some differences between the questions used in the two censuses, the data are considered broadly comparable (Office for National Statistics, 2012). The items considered in the census questions are hereby grouped in six categories: *No qualifications (NQ)*, *Level 1 qualifications (L1)*, *Level 2 qualifications (L2)*, *Level 3 qualifications (L3)*, *Level 4 qualifications or above (L4+)* and *Other qualifications (OQ)*. The questions used in each census to collect these data, as well as the detail of the categorisation above mentioned, are presented at the end of this chapter, in Section 4.5, tables 4.12 and 4.13.

Preliminary analysis

England and Wales are geographically disaggregated into 348 LAs. Two of them, *Isles of Scilly* and *City of London*, have less than 7000 persons aged 16 and over according to the 2011 Census, and are discarded from the simulation exercise. Table 4.1 contains descriptive statistics of the population size and the distribution of the variable of interest for the remaining 346 LA in consideration. The most noticeable differences between the two censuses lay in the two extremes of the classification: between 2001 and 2011, there is a reduction in the proportion of people with non or at most level 1 qualifications that is compensated by an increase in the proportion of people with L3 and L4+ qualifications.

Because SPREE-type estimators are built on the assumption of a linear relationship between the association structures of the target and proxy compositions, we will study the relationship between the two sets of interaction terms, α^Y_{aj} and α^X_{al}, defined as in equation (1.16). Figure 4.1 shows the matrix scatter plot between pairs of variables α^Y_{aj} and α^X_{al}, for $j, l = 1, \dots, 6$, as well as the lines that represent the SPREE and GSPREE structural assumptions for these

Table 4.1: Descriptive statistics for the composition of Higher Qualifications by LA. Population censuses 2001 and 2011 in England and Wales.

Statistic	Census	LA Size	Proportion of individuals					
			NQ	L1	L2	L3	L4+	OQ
Min.	2001	21 759	0.2609	0.1891	0.2496	0.0758	0.1479	0.0767
	2011	26 893	0.2028	0.1387	0.2299	0.1186	0.2697	0.0403
Q1	2001	62 145	0.2779	0.1802	0.2143	0.0734	0.1767	0.0775
	2011	75 544	0.2198	0.142	0.2074	0.1231	0.261	0.0468
Median	2001	83 776	0.2826	0.1731	0.2066	0.0774	0.188	0.0724
	2011	101 634	0.2246	0.1357	0.1979	0.1239	0.2689	0.049
Mean	2001	108 669	0.2909	0.1657	0.1938	0.0827	0.1976	0.0694
	2011	131 469	0.2266	0.1329	0.1884	0.1235	0.2721	0.0565
Q3	2001	119 067	0.317	0.1816	0.2141	0.0101	0.2034	0.0736
	2011	155 873	0.2299	0.1354	0.186	0.1257	0.2677	0.0552
Max.	2001	620 059	0.4068	0.164	0.1795	0.0005	0.1823	0.067
	2011	728 363	0.321	0.1512	0.1807	0.0109	0.2613	0.0748

data. In all plots, the Y-axis corresponds to α_{aj}^{Y} and the X-axis to α_{al}^{X}. The different categories of the variable of interest are indicated in the left and bottom corners of the plot. The red line corresponds to an OLS fit between the two variables in each cell of the matrix-plot. The black continuous line and the dashed black line in the plots on the diagonal of the matrix, represent the structural assumptions of the SPREE and GSPREE, respectively. The red line represents the corresponding OLS fit.

According to the plots in the diagonal of Figure 4.1, there is a fairly strong linear relationship between α_{aj}^{Y} and α_{aj}^{X}, for all categories except OQ. This is the kind of relationship that is assumed by estimators SPREE and GSPREE. Moreover, notice that the SPREE assumption, i.e., the black line, may hold reasonably for categories NQ, L1 and L4+, whereas it does not seem to hold for categories L2, L3 and OQ. On the other hand, the common slope assumed by the GSPREE for all categories, which for this dataset turns out to be $\beta = 0.74$ and is represented by the dashed line, does not improve considerably the fit for those categories.

Furthermore, notice that any relationship between pairs of interactions corresponding to different categories, i.e., the off-diagonal plots in Figure 4.1, is ignored by the SPREE and GSPREE estimators but could be taken into account by the new proposed estimators, MSPREE and MMSPREE, due to their multivariate nature.

Figure 4.1: Interaction terms in the compositions of LA by Higher Qualifications. Population censuses 2001 and 2011 in England and Wales.
Y-axis: α^{Y}_{aj}. X-axis: α^{X}_{al}. Black continuous line: SPREE structural assumption. Black dashed line: GSPREE structural assumption. Red line: OLS fit.

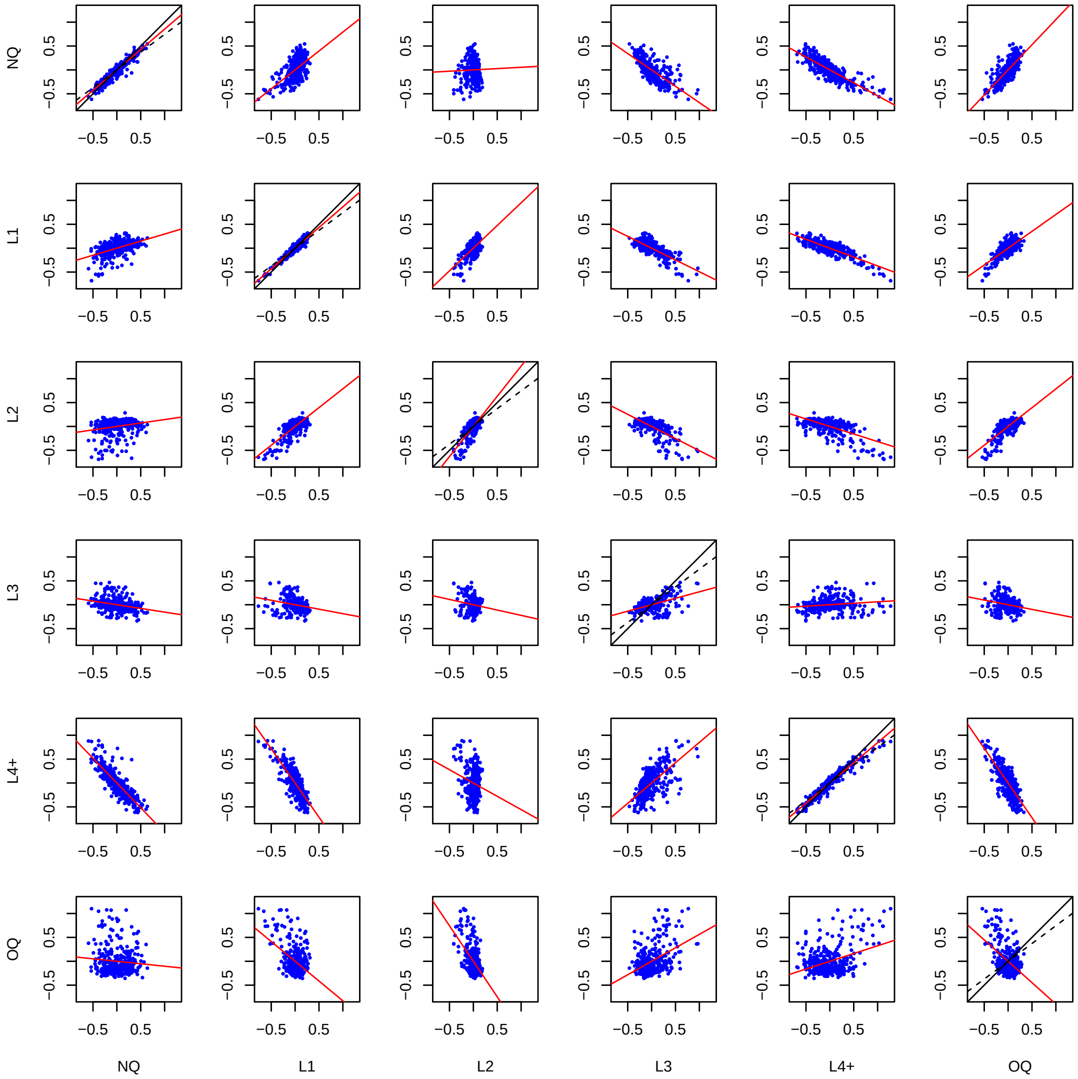


Table 4.2 presents the rescaled coefficients of the MSPREE for this dataset (matrix $\mathbf{B}^{p}$ according to section 2.4.1). Notice that, except for category OQ, the diagonal elements of $\mathbf{B}^{p}$ do not differ substantially from the GSPREE estimated parameter. However, the coefficient estimated using the MSPREE for category OQ seems more sensible considering the shape of the cloud in the bottom-right corner of Figure 4.1, than the one estimated using the GSPREE.

As it will be explained in the next section, the matrix $\mathbf{B}^p$ presented in Table 4.2 will also be used to generate the association structures of the target composition under scenarios 1 and 2.

Table 4.2: Matrix $\mathbf{B}^p$ of estimated coefficients for the MSPREE. Composition of LA by Higher Qualifications. Population censuses 2001 and 2011 in England and Wales.
Rows: Census 2011. Columns: Census 2001.

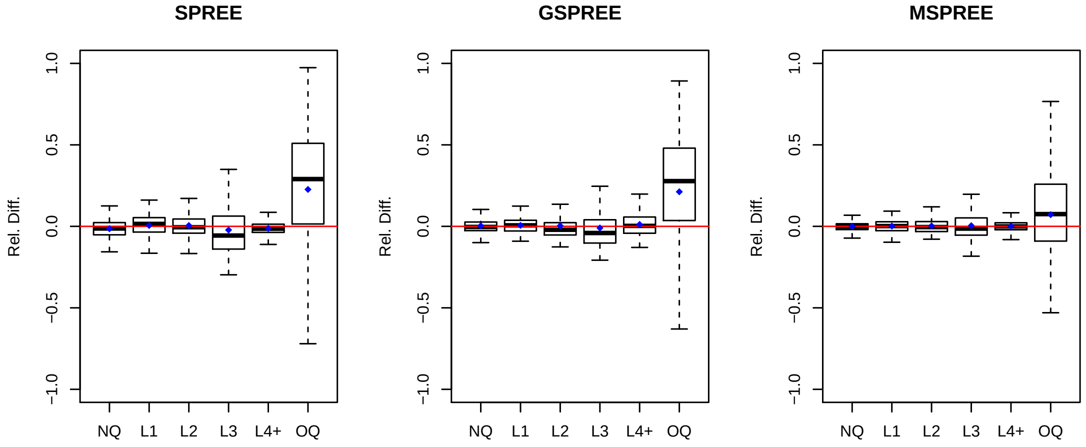
Category	NQ	L1	L2	L3	L4+	OQ
NQ	0.7306	0.0740	-0.1978	-0.0870	-0.1407	0.3516
L1	-0.2314	0.6743	0.0180	-0.0310	0.0221	0.2223
L2	0.2284	0.3067	0.6305	0.0621	0.1774	-0.7746
L3	0.0825	0.0567	0.0632	0.7832	0.0608	-0.2633
L4+	-0.0998	-0.0517	-0.1480	-0.1644	0.8524	0.4639
OQ	0.0203	-0.3857	0.2646	0.2203	-0.1196	-1.3538

The relative differences between the observed value and each of the SPREE-type estimates were calculated for each cell in the 2011 composition, with the purpose of studying the degree of model misspecification of each one of the SPREE-type estimators for the 2001 and 2011 compositions. For instance, for the SPREE we defined,

$$\text{Rel. Diff.} = \frac{(\hat{Y}^S_{aj} - Y_{aj})}{Y_{aj}}.$$

These relative differences are summarized by category, for the SPREE, GSPREE and MSPREE, in Figure 4.2. Notice that the GSPREE does not improve substantially over the SPREE regarding the average difference by column in this case. The MSPREE on the other hand, reduces considerably the size of the differences in general and the average differences in absolute value, particularly for the category OQ. Results for the MMSPREE are not presented here because, due to the huge sample size (the 2011 population composition is being used as sample), the MMSPREE is virtually equal to the sample estimate, i.e., fits almost perfectly all cells.

Figure 4.2: Relative differences between the observed composition and SPREE-type estimates. Composition of LA by Higher Qualifications. Census 2001 and 2011 in England and Wales.
Red line: Rel. Diff. = 0. Blue diamond: Mean



4.2 Simulation Scenarios

As mentioned at the beginning of this chapter, three different scenarios are used in this simulation exercise to generate the association structure of the target composition:

- **Scenario 1. MSPREE structural assumption.**
As in equation (2.2), the association structure of the target composition is given by $\alpha^{Y}_{aj} = \sum_{l} \beta_{jl}\alpha^{X}_{al}$, for $a = 1, \ldots, A$ and $j, l = 1, \ldots, J$. The coefficients β_{jl} correspond to the fitting of the real data and were presented in table 4.2. The terms α^{X}_{al} correspond to the association structure of the 2001 composition.

- **Scenario 2. MMSPREE structural assumption.**
As in equation (3.2), the association structure of the target composition is given by $\alpha^{Y}_{aj} = \sum_{l} \beta_{jl}\alpha^{X}_{al} + u_{aj}$, for $a = 1, \ldots, A$ and $j, l = 1, \ldots, J$, with β_{jl} and α^{X}_{al} as in scenario 1 and u_{aj} denoting a set of $A \times J$ random variables with expectation zero, defined as in equation (3.3):

$$u_{aj} = \vartheta_{aj} - \frac{1}{A}\vartheta_{+j} - \frac{1}{J}\vartheta_{a+} + \frac{1}{AJ}\vartheta_{++},$$

for ϑ_{aj} a set of $A \times J$ independent random variables normally distributed with expectation zero and variance σ^2_j fixed, for $j = 1, \ldots, J$.

- **Scenario 3. 2011 association structure.**
  The association structure of the target composition is that of the observed census 2011 composition.

To motivate a vector of variance components $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_J^2)$, for scenario 2, we considered the variance of the misspecification terms in the interaction scale, for each one of the SPREE-type estimators, when using the 2001 composition as proxy and the 2011 composition as sample estimate. For instance, $m_{aj}^G = (\hat{\alpha}_{aj}^{G,Y} - \alpha_{aj}^Y)$, with $\hat{\alpha}_{aj}^{G,Y}$ the estimated interaction term for cell (a, j) when using the GSPREE, denotes the misspecification term for that estimator and cell.

Table 4.3 presents the observed variances of the interaction terms α_{aj}^X, as well as the variances of the misspecification terms for different SPREE-type estimators. For the simulations, we decided to set the variance components σ_j^2 to the values in Table 4.4, which results in the variances of the random effects u_{aj} that are presented in the last row of Table 4.3.

Table 4.3: Variance calculations. Composition of LA by Higher Qualifications. Population censuses 2001 and 2011 in England and Wales.

	NQ	L1	L2	L3	L4+	OQ
$V(\alpha_{aj}^X)$	0.0569	0.0264	0.0104	0.0522	0.1122	0.0252
$V(m_{aj}^S)$	0.0053	0.0024	0.0123	0.04	0.0077	0.1543
$V(m_{aj}^G)$	0.0048	0.0023	0.0143	0.024	0.0061	0.1315
$V(m_{aj}^M)$	0.002	0.0016	0.0051	0.0106	0.003	0.0573
$V(u_{aj})$	0.0152	0.0152	0.0152	0.0412	0.0152	0.1555

Table 4.4: Variance components for the simulation exercise. Scenario 2.

	NQ	L1	L2	L3	L4+	OQ
σ^2	0.0100	0.0100	0.0100	0.0491	0.0100	0.2210

The target association structure under scenario 1 does not involve random components, so it was calculated only once. Afterwards, IPF was used to obtain a composition of expected values with margins $\mathbf{Y}_{a+}$ and $\mathbf{Y}_{+j}$ from which fixed populations were repeatedly generated using independent multinomial distributions in each area. Analogously, under scenario 3, fixed populations were generated using independent multinomial distributions in each area, with the 2011 composition as expected value. Finally, under scenario 2, random variables ϑ_{aj} and hence new target association structures, were generated in each iteration. The simulation procedure is presented in detail at the end of this chapter, in section 4.5.2.

Under scenarios 1 and 3, samples were selected using three possible sampling fractions: $f = n_a/Y_{a+} \in \{0.1, 0.05, 0.01\}$. In each scenario, 1000 replications were obtained, i.e., equal number of populations and samples were selected. Further 200 bootstrap subsamples were generated from each sample, in order to obtain estimates for the MSE of MSPREE and MMSPREE. Under scenario 2, a sample fraction $f = 0.05$ was assumed and 800 replications with 100 bootstrap subsamples each, were selected.

4.3 Evaluation criteria

In the first simulation scenario, we generate populations that satisfy the structural assumption of the MSPREE. Hence, we are interested in verifying if there is evidence of bias in that estimator; in comparing the relative efficiency of the different SPREE-type estimators and in studying the potential bias of: i) $\widehat{AV}(\hat{\mathbf{Y}}^M)$, the analytical approximation to the variance of the MSPREE; and ii) the bootstrap methodology proposed to estimate $\text{MSE}(\hat{\mathbf{Y}}^M)$. Furthermore, point estimates of the MMSPREE are also included in the evaluation, in order to illustrate its behaviour in a situation where the population does not exhibit unexplained heterogeneity.

In the second simulation scenario, the populations satisfy the MMSPREE structural assumption. Hence, were are interested in verifying if there is evidence of bias for that estimator and in illustrating the behaviour of the MSPREE in the presence of unexplained heterogeneity. Moreover, potential bias in the bootstrap methodologies proposed for the estimation of the unconditional MSE of the MSPREE and MMSPREE are also studied.

Finally, populations generated under the third simulation scenario do not satisfy any of the previously mentioned structural assumptions. Therefore, we are interested in studying the bias shown by the different SPREE-type estimators and in comparing their relative efficiency.

Next, we will describe the indicators that are used to summarize the results of the simulation. For convenience, in the rest of this chapter we will denote by $\hat{Y}^{K}_{aj}$ an estimator of Y_{aj}, with $K \in \{D, S, G, M, MM\}$ indicating the Direct estimator, SPREE, GSPREE, MSPREE and MMSPREE respectively.

4.3.1 Bias. Point estimators.

The existence of bias, $B(\hat{Y}^{K}_{aj}) = E\left[\hat{Y}^{K}_{aj} - Y_{aj}\right]$, is analysed using the Monte Carlo estimate

$$\tilde{B}(\hat{Y}^{K}_{aj}) = \frac{1}{S}\sum_{s=1}^{S}(\hat{Y}^{K,s}_{aj} - Y^{s}_{aj}).$$

where $\hat{Y}^{K,s}_{aj}$ and Y^{s}_{aj} denote the population value and its estimate for the s-th replication, and S is the total number of Monte Carlo replications. In order to account for the bootstrap variation, $\tilde{B}(\hat{Y}^{K}_{aj})$, its standard error and a normal approximation are used to build a 95% prediction interval for $B(\hat{Y}^{K}_{aj})$. A number of prediction intervals which do not contain zero considerably higher than 5% can be interpreted as evidence against the unbiasedness of a given SPREE-type estimator.

The size and direction of the bias are studied using the Relative Bias:

$$\text{Rel. Bias}(\hat{Y}^{K}_{aj}) = \frac{\tilde{B}(\hat{Y}^{K}_{aj})}{\frac{1}{S}\sum_{s=1}^{S} Y^{s}_{aj}}.$$

Finally, we study the absolute size of the bias is analysed considering averages of the absolute value of Rel. Bias$(\hat{Y}^{K}_{aj})$, aggregating by category, and over all cells.

4.3.2 MSE. Point estimators.

The relative performance of different SPREE-type estimators is studied using the Relative Square Root MSE (RSRMSE):

$$\text{RSRMSE}(\hat{Y}^{K}_{aj}) = \frac{\sqrt{\widetilde{\text{MSE}}(\hat{Y}^{K}_{aj})}}{\frac{1}{S}\sum_{s=1}^{S} Y^{s}_{aj}},$$

where $\widetilde{\text{MSE}}(\hat{Y}^K_{aj})$ is a Monte Carlo estimator of the unconditional MSE of $\hat{Y}^K_{aj}$ given by

$$\widetilde{\text{MSE}}(\hat{Y}^K_{aj}) = \frac{1}{S}\sum_{s=1}^{S}(\hat{Y}^{K,s}_{aj} - Y^s_{aj})^2.$$

Moreover, the Monte Carlo variance of $\hat{Y}^M_{aj}$, denoted by $\tilde{V}(\hat{Y}^M_{aj})$, is also calculated.

4.3.3 Variance estimators.

Several variance estimators are considered in this simulation study: i) the bootstrap estimators $\widehat{\text{MSE}}(\hat{Y}^M_{aj})$ and $\widehat{\text{MSE}}(\hat{Y}^{MM}_{aj})$; ii) the estimator for the analytic approximation to the variance of the MSPREE, $\widehat{\text{AV}}(\hat{\mathbf{Y}}^M)$ and iii) the estimators for the variance components $\hat{\sigma}^2_j$ for $j = 1, \dots, J$.

For estimators in i) and ii), the bias is estimated using the deviations respect to the corresponding Monte Carlo estimates (defined in section 4.3.2), i.e, respect to $\widetilde{\text{MSE}}(\hat{Y}^K_{aj})$ for the estimators in i) and to the Monte Carlo variance $\tilde{V}(\hat{Y}^M_{aj})$ for the estimator in ii). For the variance components, results regarding bias (evaluated respect to the fixed values in table 4.4) as well as Monte Carlo estimates of their MSEs, are presented.

4.4 Results

4.4.1 Scenario 1

Bias. Point estimators.
Table 4.5 presents the proportion of areas for which the 95% prediction interval for the bias of the point estimator, does not include the zero, aggregated by column, for the MSPREE and MMSPREE, in all three scenarios of sample size. The proportions are close to the coverage level, more for the MSPREE than for the MMSPREE, suggesting that there is no strong evidence of bias for those estimators under this scenario. Moreover, when considering the average size of the absolute value of the relative bias, it is possible to say that any potential bias in those two estimators is negligible.

On the other hand, there is evidence of bias of around 10% for the SPREE and 9% for the GSPREE in all three sub-scenarios of sampling fraction (see tables

4.14 and 4.15 in section 4.5.3). Because the SPREE does not use sample information, its behaviour is the same for all three sampling fraction sub-scenarios under study.

Table 4.5: Summary of bias results, MSPREE and MMSPREE, by category and sampling fraction. Scenario 1.

Estimator	Category	Proportion of areas with 0 not in the 95% PI			|Relative bias| Mean		
		f = 0.1	f = 0.05	f = 0.01	f = 0.1	f = 0.05	f = 0.01
MSPREE	NQ	0.0318	0.0202	0.0145	0.0001	0.0001	0.0002
	L1	0.0983	0.0838	0.0434	0.0002	0.0002	0.0003
	L2	0.0491	0.0434	0.0145	0.0002	0.0002	0.0002
	L3	0.0462	0.0289	0.0058	0.0002	0.0002	0.0002
	L4+	0.0751	0.0549	0.0723	0.0001	0.0001	0.0002
	OQ	0.0405	0.0231	0.0780	0.0003	0.0003	0.0005
	Average	0.0568	0.0424	0.0381	0.0002	0.0002	0.0003
MMSPREE	NQ	0.0578	0.0318	0.0173	0.0002	0.0002	0.0002
	L1	0.1012	0.0925	0.0665	0.0003	0.0003	0.0003
	L2	0.0665	0.0723	0.0318	0.0002	0.0002	0.0003
	L3	0.0434	0.0231	0.0376	0.0002	0.0002	0.0003
	L4+	0.0780	0.0549	0.0896	0.0002	0.0001	0.0002
	OQ	0.0809	0.0607	0.2890	0.0004	0.0004	0.0010
	Average	0.0713	0.0559	0.0886	0.0002	0.0002	0.0004

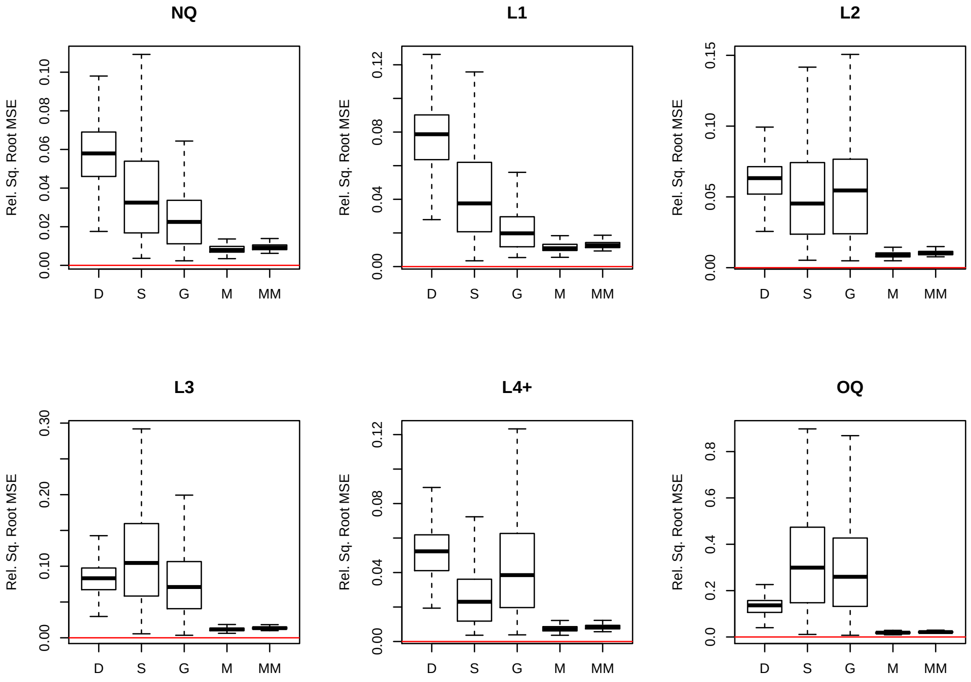
MSE. Point estimators.

The proposed estimators improve considerably over the SPREE and GSPREE in terms of MSE, for all categories. The RSRMSE of the MSPREE is approximately 11% the RSRMSE of the SPREE, regardless of the sampling fraction. Moreover, possibly because there is not unexplained heterogeneity under this scenario, the RSRMSE of the MMSPREE is slightly bigger (around 15%) than the one of the MSPREE, for all categories and sub-scenarios of sampling fraction. For illustration, results for $f = 0.01$ are presented in Table 4.6 and Figure 4.3. Additional results and those corresponding to other sampling fractions can be found in Figure 4.7 and Table 4.17 in section 4.5.3.

Table 4.6: Mean RSRMSE of estimators of **Y**, by category. f=0.01. Scenario 1.

Category	Direct	SPREE	GSPREE	MSPREE	MMSPREE
NQ	0.0583	0.0396	0.0270	0.0084	0.0096
L1	0.0782	0.0469	0.0226	0.0115	0.0132
L2	0.0627	0.0509	0.0620	0.0093	0.0107
L3	0.0835	0.1160	0.0821	0.0120	0.0139
L4+	0.0524	0.0275	0.0450	0.0075	0.0085
OQ	0.1334	0.3415	0.3014	0.0187	0.0214
Average	0.0781	0.1037	0.0900	0.0112	0.0129

Figure 4.3: RSRMSE of estimators of **Y**, by category. f=0.01. Scenario 1.
Estimators: D=Direct, S=SPREE, G=GSPREE, M=MSPREE, M=MMSPREE.
Red line: RSRMSE = 0.



Variance estimators.
The main results for the variance estimators are presented in Table 4.7. Starting with $\widehat{\mathrm{AV}}(\hat{\mathbf{Y}}^{\mathrm{M}})$, there is evidence of bias. Considering the process of building this estimator, possible causes of such bias are: i) a poor approximation from the first iteration of the IPF to the estimates after convergence; ii) a bad performance of the Taylor linearisation used to derive the estimator or iii) a bad performance of the estimator of $\mathbf{V}_{\hat{\alpha}}$. Further research would be necessary to clarify which of these reasons could be responsible for the bias in this simulation exercise.

When considering the average of the absolute values of the relative bias for this estimator, it is observed an over/under estimation of approximately 4% on the square root of the variance approximation for the three sub-scenarios of sampling fraction. However, it is important to notice that the estimator is not always conservative. Table 4.18 and Figure 4.8, in the complementary material at the end of this chapter, suggest a more conservative behaviour when the sample size decreases, but this evidence is not conclusive.

There is also evidence of bias for $\widehat{\mathrm{MSE}}(\hat{\mathbf{Y}}^{\mathrm{M}})$. Without taking into account the direction of the bias, the estimator exhibits an average relative bias of around 2% on the square root of the MSE, irrespective of category or sampling fraction sub-scenario, which cannot be attributed to bootstrap variation. As with the analytical approximation above, the MSE estimator is not always conservative. Furthermore, its average for a given category is almost zero (see Table 4.19 and Figure 4.9 in section 4.5.3).

4.4.2 Scenario 2

Bias. Point estimators.
Considering the proportion of cells for which zero is not contained in a 95% prediction interval for the estimator of the bias, there is no strong evidence of bias for the MSPREE, but there is weak evidence of bias for the MMSPREE and there is strong evidence of bias for both SPREE and GSPREE estimators under this scenario. However, notice that the relative size of the bias of the MMSPREE in absolute value is, on average, less than 0.3% and smaller than the bias of any other of the SPREE-type estimators, especially for the category OQ, as can be seen in Table 4.8 and Figure 4.4.

Table 4.7: Summary of bias results, $\sqrt{\widehat{\mathrm{AV}}(\hat{\mathbf{Y}}^{\mathrm{M}})}$ and $\sqrt{\widehat{\mathrm{MSE}}(\hat{\mathbf{Y}}^{\mathrm{M}})}$, by category and sampling fraction. Scenario 1.

Estimator	Category	Proportion of areas with 0 not in the 95% PI for the bias			|Relative bias| Mean		
		$f = 0.1$	$f = 0.05$	$f = 0.01$	$f = 0.1$	$f = 0.05$	$f = 0.01$
$\sqrt{\widehat{\mathrm{AV}}(\hat{\mathbf{Y}}^{\mathrm{M}})}$	NQ	1.0000	1.0000	0.9971	0.0295	0.0379	0.0412
	L1	1.0000	0.9971	0.9971	0.0412	0.0390	0.0455
	L2	1.0000	0.9942	0.9971	0.0536	0.0316	0.0315
	L3	1.0000	0.9942	0.9942	0.0368	0.0304	0.0394
	L4+	1.0000	1.0000	0.9971	0.0494	0.0366	0.0432
	OQ	0.9971	0.9942	0.9942	0.0297	0.0332	0.0506
	Average	0.9995	0.9966	0.9961	0.0400	0.0348	0.0419
$\sqrt{\widehat{\mathrm{MSE}}(\hat{\mathbf{Y}}^{\mathrm{M}})}$	NQ	0.9104	0.9075	0.8988	0.0201	0.0208	0.0196
	L1	0.9075	0.9133	0.8873	0.0193	0.0194	0.0207
	L2	0.9017	0.8931	0.8873	0.0189	0.0198	0.0195
	L3	0.9104	0.8815	0.8815	0.0163	0.0162	0.0174
	L4+	0.9133	0.9075	0.9075	0.0214	0.0215	0.0200
	OQ	0.9191	0.9075	0.8757	0.0202	0.0202	0.0182
	Average	0.9104	0.9017	0.8897	0.0194	0.0196	0.0192

It was argued in section 3.2 that the introduction of random effects in the structural equation of the MSPREE, under the assumption of normality, has a column-specific multiplicative effect on the expected value of $\mathbf{Y}$. This could affect the size of the bias of the SPREE-type estimators. However, perhaps because the estimators are benchmarked to the known column and row margins, the bias does not seem to increase. In fact, comparing with the results obtained under scenario 1 we can notice that the biases of the SPREE and GSPREE estimators under both scenarios have approximately the same relative size (see Table 4.15 for $f = 0.05$ in section 4.5.3, and Table 4.8 below).

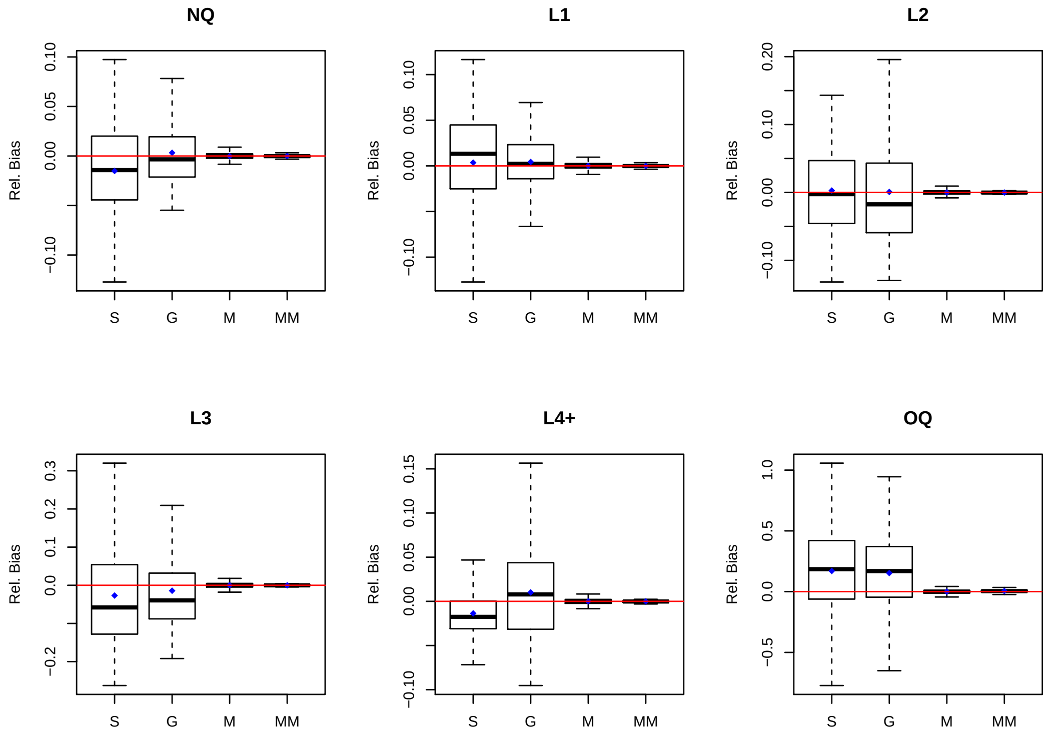
MSE. Point estimators.

Table 4.9 presents the Relative Square Root MSE of the SPREE-type estimators of $\mathbf{Y}$ under this scenario. As expected, given that the target composition has been generated satisfying the MMSPREE structural assumption, such estimator over-performs all fixed-effects SPREE-type estimators in consideration. More interesting is to notice that, even though the random extra heterogeneity added to the target association structure in this case does not seem to affect the size of the bias of the fixed effects estimators, as it was mentioned above, it does increase their RSRMSE substantially (see Tables 4.17 for $f = 0.05$ in section 4.5.3 and Table 4.9 below).

Table 4.8: Summary of bias results for estimators of **Y**, by category. Scenario 2.

Category	Proportion of areas with 0 not in the 95% PI for B($\hat{\mathbf{Y}}$)				|Relative bias| Mean			
	SPREE	GSPREE	MSPREE	MMSPREE	SPREE	GSPREE	MSPREE	MMSPREE
NQ	0.9306	0.8179	0.0578	0.1590	0.0393	0.0253	0.0027	0.0011
L1	0.9017	0.7919	0.0578	0.1098	0.0459	0.0206	0.0029	0.0012
L2	0.9364	0.9277	0.0434	0.1214	0.0501	0.0609	0.0028	0.0010
L3	0.9595	0.9017	0.0723	0.0838	0.1154	0.0808	0.0060	0.0015
L4+	0.8671	0.9277	0.0607	0.0723	0.0269	0.0441	0.0027	0.0009
OQ	0.9422	0.9451	0.0751	0.1445	0.3367	0.2958	0.0140	0.0103
Average	0.9229	0.8854	0.0612	0.1151	0.1024	0.0879	0.0052	0.0027

Figure 4.4: Relative Bias of estimators of **Y**, by category. Scenario 2.
Estimators: D=Direct, S=SPREE, G=GSPREE, M=MSPREE, M=MMSPREE. Red line: Relative bias = 0.



Variance estimators.

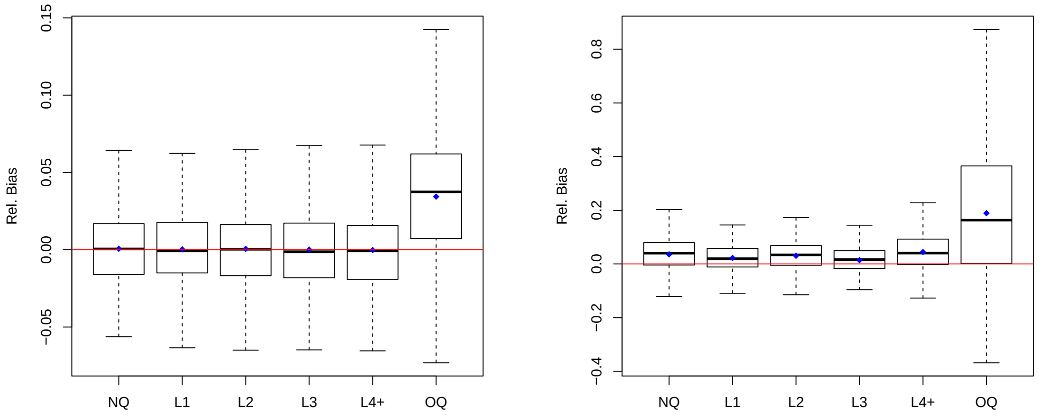
Considering the absolute value of the relative biases, there is evidence of an average bias of around 2% in them square root of the bootstrap estimator of

Table 4.9: Mean RSRMSE of estimators of $\mathbf{Y}$, by category. Scenario 2.

Category	SPREE	GSPREE	MSPREE	MMSPREE
NQ	0.1045	0.0989	0.0932	0.0290
L1	0.1172	0.1063	0.1029	0.0361
L2	0.1125	0.1206	0.0965	0.0307
L3	0.2408	0.2237	0.2005	0.0464
L4+	0.0947	0.1027	0.0885	0.0271
OQ	0.5999	0.5730	0.4585	0.2552
Average	0.2116	0.2042	0.1734	0.0708

MSE($\hat{\mathbf{Y}}^{M}$), and around 8% in the bootstrap estimator of MSE($\hat{\mathbf{Y}}^{MM}$), as can be seen in Table 4.21 in section 4.5.4. However, when including the direction of the bias it can be seen that if the areas are averaged for a given column, the bias is close to zero for all columns, except OQ for the MSPREE, see Figure 4.5 below and Table 4.22 in section 4.5.4.

Figure 4.5: Relative Bias of estimators of $\sqrt{\mathrm{MSE}(\hat{\mathbf{Y}})}$, by category. Scenario 2. Left: MSPREE. Right: MMSPREE. Red line: Relative bias = 0.



Finally, we will turn to the estimator of the variance components. The main results for this estimator are presented in Table 4.10. There is no evidence of bias in the estimator of the variance components, except for the category OQ which shows an overestimation of around 5%. Given that only 800 replications were used in this scenario, additional simulations might be required before making a conclusion in this matter. Notice that for all categories, the RSRMSE is quite high considering the sampling fraction of 0.05 used under this scenario. However, notice that despite the variability of the estimation of the variance components, the MMSPREE still over performs all other estimators in terms of their RSRMSE.

Table 4.10: Summary of results, estimation of the variance components. Scenario 2.

		Category					
		NQ	L1	L2	L3	L4+	OQ
σ_j^2		0.0100	0.0100	0.0100	0.0491	0.0100	0.2210
$E[\hat{\sigma}_j^2]$		0.0099	0.0100	0.0100	0.0490	0.0099	0.2326
Rel. Bias ($\hat{\sigma}_j^2$)		-0.0072	0.0034	-0.0026	-0.0009	-0.0065	0.0529
95% CI($\hat{\sigma}_j^2$) (Var)	Lim. Inf.	0.0098	0.0099	0.0099	0.0487	0.0098	0.2309
	Lim. Sup.	0.0100	0.0102	0.0101	0.0494	0.0101	0.2344
95% CI($\hat{\sigma}_j^2$) (MSE)	Lim. Inf.	0.0067	0.0067	0.0066	0.0395	0.0067	0.1780
	Lim. Sup.	0.0131	0.0134	0.0133	0.0586	0.0132	0.2873
$\sqrt{\text{MSE}(\hat{\sigma}_j^2)}$		0.0016	0.0017	0.0017	0.0049	0.0017	0.0279
RSRMSE($\hat{\sigma}_j^2$)		0.1634	0.1712	0.1700	0.0996	0.1674	0.1261

4.4.3 Scenario 3

Bias and MSE. Point estimators.
Remember that under this scenario, the target composition has the association structure observed in the census 2011 composition, i.e., does not satisfy the structural assumption of any of the SPREE-type estimates. To illustrate the results obtained under Scenario 3, the main findings for $f = 0.01$ are presented in Table 4.11. Results for other sampling fractions are very similar.

All SPREE-type estimators are biased under this scenario, and such bias cannot be attributed to Monte Carlo variation. Considering the average of the absolute value of the relative biases, thee GSPREE represents a modest improvement over the SPREE, whereas the proposed estimators, MSPREE and MMSPREE, have a substantially smaller bias. Moreover, comparing only the latter two, it can be observed that the bias of the MMSPREE is about the same size or smaller for all categories and the improvement is considerably for OQ. Regarding the RSRMSE, similar conclusions could be drawn for all fixed-effects SPREE-type estimators. On the other hand, the MMSPREE, despite having a smaller bias, presents a higher variance. Both things combined result in an RSRMSE that is almost comparable to the one of the MSPREE accross all categories, except for OQ, category for which the mixed effects estimator still performs substantially better.

Table 4.11: Summary of bias results, estimators of $\mathbf{Y}$, by category. $f = 0.01$. Scenario 3.

	Category	Direct	SPREE	GSPREE	MSPREE	MMSPREE
|Relative Bias|	NQ	0.0017	0.0467	0.0352	0.0242	0.0221
	L1	0.0024	0.0547	0.0369	0.0319	0.0325
	L2	0.0018	0.0554	0.0651	0.0367	0.0179
	L3	0.0021	0.1322	0.1024	0.0654	0.0241
	L4+	0.0012	0.0365	0.0561	0.0279	0.0219
	OQ	0.0033	0.3916	0.3620	0.2175	0.0383
	Average	0.0021	0.1195	0.1096	0.0673	0.0261
RSRMSE	NQ	0.0582	0.0475	0.0363	0.0267	0.0279
	L1	0.0784	0.0559	0.0384	0.0352	0.0379
	L2	0.0627	0.0564	0.0660	0.0389	0.0342
	L3	0.0835	0.1329	0.1034	0.0676	0.0561
	L4+	0.0525	0.0375	0.0568	0.0299	0.0285
	OQ	0.1368	0.3922	0.3627	0.2196	0.1175
	Average	0.0787	0.1204	0.1195	0.0697	0.0504

4.5 Complementary Material

4.5.1 Qualifications data on the 2001 and 2011 population censuses in England and Wales

Table 4.12: Qualifications on the 2001 and 2011 population censuses in England and Wales

Census 2001	Census 2011
16. Which of these qualifications do you have? *✓all the qualifications that apply or, if not specified, the nearest equivalent.* □ 1 + O levels/CSEs/GCSEs (any grades) □ 5 + O levels, 5 + CSEs (grade 1), 5+ GCSEs (grades A-C), School Certificate □ 1 + A levels/AS levels □ 2 + A levels, 4 + AS levels, Higher School Certificate □ First Degree (eg BA, BSc) □ Higher Degree (eg MA, PhD, PGCE, postgraduate certificates/diplomas) □ NVQ Level 1, Foundation GNVQ □ NVQ Level 2, Intermediate GNVQ □ NVQ Level 3, Advanced GNVQ □ NVQ Levels 4-5, HNC, HND □ Other Qualifications (eg City and Guilds, RSA/OCR, BTEC/Edexcel) □ No Qualifications 17. Do you have any of the following professional qualifications? □ No Professional Qualifications □ Qualified Teacher Status (for schools) □ Qualified Medical Doctor □ Qualified Dentist □ Qualified Nurse, Midwife, Health Visitor □ Other Professional Qualifications	25. Which of these qualifications do you have? ➪ *Tick **every** box that applies if you have **any** of the qualifications listed* ➪ *If your UK qualification is not listed, tick the box that contains its nearest equivalent* ➪ *If you have qualifications gained outside the UK, tick the 'Foreign qualifications' box and the nearest UK equivalents (if known)* □ 1-4 O levels/CSEs/GCSEs (any grades), Entry Level, Foundation Diploma □ NVQ Level 1, Foundation GNVQ, Basic Skills □ 5 + O levels (passes)/CSEs (grade 1)/GCSEs (grades A*-C), School Certificate, 1 A level/2-3 levels/VCEs, Higher Diploma, Welsh Baccalaureate Intermediate Diploma (Wales) □ NVQ Level 2, Intermediate GNVQ, City and Guilds Craft, BTEC First/General Diploma, RSA Diploma □ Apprenticeship □ 2 + A levels/VCEs, 4 + AS levels, Higher School Certificate, Progression/Advanced Diploma(England), Welsh Baccalaureate Advanced Diploma (Wales) □ NVQ Level 3, Advanced GNVQ, City and Guilds Advanced Craft, ONC, OND, BTEC National, RSA Advanced Diploma □ Degree (for example BA, BSc), Higher degree (for example MA, PhD, PGCE) □ NVQ Level 4-5, HNC, HND, RSA Higher Diploma, BTEC Higher Level □ Professional qualifications (for example teaching, nursing, accountancy) □ Other vocational/work-related qualifications □ Foreign qualifications □ No qualifications

Table 4.13: Categories of Highest Qualifications for the 2011 population census in England and Wales

Category	Items
No qualifications (NQ)	No qualifications.
Level 1 (L1)	1-4 O levels/CSEs/GCSEs (any grades), Entry Level, Foundation Diploma. NVQ Level 1, Foundation GNVQ, Basic Skills.
Level 2 (L2)	5 + O levels (passes)/CSEs (grade 1)/GCSEs (grades A*-C), School Certificate, 1 A level/2-3 levels/VCEs, Higher Diploma, Welsh Baccalaureate Intermediate Diploma (Wales). NVQ Level 2, Intermediate GNVQ, City and Guilds Craft, BTEC First/General Diploma, RSA Diploma. Apprenticeship.
Level 3 (L3)	2 + A levels/VCEs, 4 + AS levels, Higher School Certificate, Progression/Advanced Diploma(England), Welsh Baccalaureate Advanced Diploma (Wales). NVQ Level 3, Advanced GNVQ, City and Guilds Advanced Craft, ONC, OND, BTEC National, RSA Advanced Diploma.
Level 4 and above (L4+)	Degree (for example BA, BSc), Higher degree (for example MA, PhD, PGCE). NVQ Level 4-5, HNC, HND, RSA Higher Diploma, BTEC Higher Level. Professional qualifications (for example teaching, nursing, accountancy).
Other Qualifications (OQ)	Other vocational/work-related qualifications. Foreign qualifications (If the level is unknown).

4.5.2 Procedures for the double-bootstrap simulation

4.5.2.1 Scenario 1

The simulation under scenario 1 starts by generating the assumed association structure of the target composition

$$\alpha^{Y}_{aj} = \sum_{l} \beta_{jl} \alpha^{X}_{al},$$

a composition of expected values

$$\boldsymbol{\mu} = \text{IPF}\left(\exp\left\{\boldsymbol{\alpha}^{Y}\right\} ; \mathbf{Y}_{a+} ; \mathbf{Y}_{+j}\right),$$

and calculating the probabilities $\boldsymbol{\pi}_a = \boldsymbol{\mu}_a / Y_{a+}$.
For $s = 1, \ldots, S$:

1. Generate a population composition $\mathbf{Y}^s$, by sampling independently in each area a using a multinomial distribution with vector of probabilities $\boldsymbol{\pi}_a$ and sample size Y_{a+}. For each area a, calculate $\theta^s_a = \mathbf{Y}^s_a / Y_{a+}$.

2. Generate a sample composition $\mathbf{y}^s$, by sampling independently in each area a using a multinomial distribution with vector of probabilities θ^s_a

and sample size $y_{a+} = Y_{a+} \times f$; with $f \in \{0.1, 0.05, 0.01\}$.

3. Calculate $\hat{\mathbf{Y}}^{S,s}$, $\hat{\mathbf{Y}}^{G,s}$, $\hat{\mathbf{Y}}^{M,s}$ and $\hat{\mathbf{Y}}^{MM,s}$, the SPREE, GSPREE, MSPREE and MMSPREE of Y^s based on $\mathbf{y}^s$. For each area a, calculate $\hat{\boldsymbol{\theta}}_a^{M,s} = \hat{\mathbf{Y}}^{M,s}/Y_{a+}$.
4. Calculate the estimator of the variance of the MSPREE, $\widehat{AV}\left(\hat{\mathbf{Y}}^{M,s}\right)$ defined in section 2.3.1.
5. For $b = 1, \ldots, B$:
   - 5.1. Generate a population composition $\mathbf{Y}^{s,b}$, by sampling independently in each area a using a multinomial distribution with vector of probabilities $\hat{\boldsymbol{\theta}}_a^{M,s}$ and sample size Y_{a+}. For each area a, calculate $\boldsymbol{\theta}_a^{s,b} = \mathbf{Y}^{s,b}/Y_{a+}$.
   - 5.2. Generate a sample composition $\mathbf{y}^{s,b}$, by sampling independently in each area a using a multinomial distribution with vector of probabilities $\boldsymbol{\theta}_a^{s,b}$ and sample size y_{a+}.
   - 5.3. Calculate $\hat{\mathbf{Y}}^{M,s,b}$, the MSPREE of $\mathbf{Y}^{s,b}$ based on $\mathbf{y}^{s,b}$.
6. Use the bootstrap estimates $\hat{\mathbf{Y}}^{M,s,b}$ for $b = 1, \ldots, B$, to calculate the estimate of $\mathrm{MSE}(\hat{\mathbf{Y}}^{M,s})$ proposed in section 2.3.2.

4.5.2.2 Scenario 2

For $s = 1, \ldots, S$:

1. For $j = 1, \ldots, J$, generate A independent realisations of a normally distributed random variable with mean zero and variance σ_j^2. Denote them by ϑ_{aj}^s, for $a = 1, \ldots, A$. Calculate $u_{aj}^s = \vartheta_{aj}^s - \frac{1}{A}\vartheta_{+j}^s - \frac{1}{J}\vartheta_{a+}^s + \frac{1}{AJ}\vartheta_{++}^s$.
2. Set the association structure of the target composition

$$\alpha_{aj}^{Y,s} = \sum_l \beta_{jl}\alpha_{al}^X + u_{aj}^s;$$

   calculate the composition of expected values

$$\boldsymbol{\mu}^s = \mathrm{IPF}\left(\exp\left\{\boldsymbol{\alpha}^{Y,s}\right\}; \mathbf{Y}_{a+}; \mathbf{Y}_{+j}\right),$$

   and the probabilities $\boldsymbol{\pi}_a^s = \boldsymbol{\mu}_a^s/Y_{a+}$.

3. Generate a population composition $\mathbf{Y}^s$, by sampling independently in each area a using a multinomial distribution with vector of probabilities $\boldsymbol{\pi}_a^s$ and sample size Y_{a+}. For each area a, calculate $\boldsymbol{\theta}_a^s = \mathbf{Y}_a^s/Y_{a+}$.

4. Generate a sample composition $\mathbf{y}^s$, by sampling independently in each area a using a multinomial distribution with vector of probabilities $\boldsymbol{\theta}_a^s$ and sample size y_{a+}.

5. Calculate $\hat{\mathbf{Y}}^{S,s}$, $\hat{\mathbf{Y}}^{G,s}$, $\hat{\mathbf{Y}}^{M,s}$ and $\hat{\mathbf{Y}}^{MM,s}$, the SPREE, GSPREE, MSPREE and MMSPREE of Y^s. Denote by $\hat{\boldsymbol{\sigma}}^s$ the estimator of the vector of variance components based on $\mathbf{y}^s$. Denote by $\hat{\boldsymbol{\beta}}^s$ the estimated matrix of parameters of the MSPREE based on $\mathbf{y}^s$. For each area a, calculate $\hat{\boldsymbol{\theta}}_a^s = \hat{\mathbf{Y}}^{MM,s}/Y_{a+}$.

6. For $b = 1, \dots, B$:

   6.1. For $j = 1, \dots, J$, generate A independent realisations of a normally distributed random variable with mean zero and variance $\hat{\sigma}_j^{2,s}$. Denote them by $\vartheta_{aj}^{s,b}$, for $a = 1, \dots, A$. Calculate $u_{aj}^{s,b} = \vartheta_{aj}^{s,b} - \frac{1}{A}\vartheta_{+j}^{s,b} - \frac{1}{J}\vartheta_{a+}^{s,b} + \frac{1}{AJ}\vartheta_{++}^{s,b}$.

   6.2. Calculate
   $$\alpha_{aj}^{Y,s,b} = \sum_{l} \hat{\beta}_{jl}^{s} \alpha_{al}^{X} + u_{aj}^{s,b},$$
   calculate
   $$\mathbf{Y}^{s,b} = \mathrm{IPF}\left(\exp\left\{\boldsymbol{\alpha}^{Y,s,b}\right\} ; \mathbf{Y}_{a+} ; \mathbf{Y}_{+j}\right),$$
   and $\boldsymbol{\theta}_a^{s,b} = \mathbf{Y}_a^{s,b}/Y_{a+}$.

   6.3. Generate a sample composition $\mathbf{y}^{s,b}$, by sampling independently in each area a using a multinomial distribution with vector of probabilities $\boldsymbol{\theta}_a^{s,b}$ and sample size y_{a+}.

   6.4. Calculate $\hat{\mathbf{Y}}^{M,s,b}$ and $\hat{\mathbf{Y}}^{MM,s,b}$, the MSPREE and MMSPREE of $\mathbf{Y}^{s,b}$ based on $\mathbf{y}^{s,b}$.

7. Use the bootstrap estimates $\hat{\mathbf{Y}}^{M,s,b}$ and $\hat{\mathbf{Y}}^{MM,s,b}$ for $b = 1, \dots, B$, to calculate the estimates of MSE($\hat{\mathbf{Y}}^{M,s}$) and MSE($\hat{\mathbf{Y}}^{MM,s}$) proposed in section 3.3.1.

4.5.2.3 Scenario 3

The simulation under scenario 3 uses as target association structure the one observed in the Census 2011 composition. The procedure is analogous to the one followed for Scenario 1, replacing the composition of expected values $\boldsymbol{\mu}$ by the 2011 composition. Only point estimates are calculated under this scenario.

4.5.3 Additional results. Scenario 1.

Table 4.14: Proportion of cells for which zero is not included in a 95% prediction interval for $\mathrm{B}(\hat{\mathbf{Y}})$, by category. Scenario 1.

f	Estimator	Category						Average
		NQ	L1	L2	L3	L4+	OQ	
0.1	Direct	0.0694	0.1127	0.1098	0.0578	0.0665	0.0751	0.0819
	SPREE	0.9913	1.0000	0.9971	1.0000	0.9855	1.0000	0.9957
	GSPREE	0.9913	0.9971	0.9942	0.9942	0.9971	1.0000	0.9957
	MSPREE	0.0318	0.0983	0.0491	0.0462	0.0751	0.0405	0.0568
	MMSPREE	0.0578	0.1012	0.0665	0.0434	0.078	0.0809	0.0713
0.05	Direct	0.0665	0.1358	0.1156	0.0549	0.0751	0.0462	0.0824
	SPREE	0.9913	1.0000	0.9971	1.0000	0.9855	1.0000	0.9957
	GSPREE	0.9913	0.9971	0.9942	0.9942	0.9971	1.0000	0.9957
	MSPREE	0.0202	0.0838	0.0434	0.0289	0.0549	0.0231	0.0424
	MMSPREE	0.0318	0.0925	0.0723	0.0231	0.0549	0.0607	0.0559
0.01	Direct	0.0318	0.1098	0.0925	0.0751	0.0289	0.0318	0.0617
	SPREE	0.9913	1.0000	0.9971	1.0000	0.9855	1.0000	0.9957
	GSPREE	0.9913	0.9971	0.9942	0.9942	0.9971	1.0000	0.9957
	MSPREE	0.0145	0.0434	0.0145	0.0058	0.0723	0.0780	0.0381
	MMSPREE	0.0173	0.0665	0.0318	0.0376	0.0896	0.2890	0.0886

Table 4.15: Mean of the Relative Bias, in absolute value, of estimators of $\mathbf{Y}$, by category and sampling fraction. Scenario 1.

f	Estimator	Category						Average
		NQ	L1	L2	L3	L4+	OQ	
0.1	Direct	0.0006	0.0008	0.0006	0.0007	0.0005	0.0011	0.0007
	SPREE	0.0387	0.0455	0.0501	0.1154	0.0264	0.3407	0.1028
	GSPREE	0.0256	0.0202	0.0612	0.0811	0.0442	0.3008	0.0888
	MSPREE	0.0001	0.0002	0.0002	0.0002	0.0001	0.0003	0.0002
	MMSPREE	0.0002	0.0003	0.0002	0.0002	0.0002	0.0004	0.0002
0.05	Direct	0.0008	0.0013	0.0009	0.0010	0.0006	0.0014	0.0010
	SPREE	0.0387	0.0455	0.0501	0.1154	0.0264	0.3407	0.1028
	GSPREE	0.0256	0.0202	0.0612	0.0811	0.0442	0.3008	0.0888
	MSPREE	0.0001	0.0002	0.0002	0.0002	0.0001	0.0003	0.0002
	MMSPREE	0.0002	0.0003	0.0002	0.0002	0.0001	0.0004	0.0002
0.01	Direct	0.0017	0.0023	0.0018	0.0020	0.0013	0.0033	0.0021
	SPREE	0.0387	0.0455	0.0501	0.1154	0.0264	0.3407	0.1028
	GSPREE	0.0256	0.0203	0.0612	0.0811	0.0442	0.3007	0.0888
	MSPREE	0.0002	0.0003	0.0002	0.0002	0.0002	0.0005	0.0003
	MMSPREE	0.0002	0.0003	0.0003	0.0003	0.0002	0.0010	0.0004

Figure 4.6: Relative Bias of estimators of **Y**, by category and sampling fraction. Scenario 1.
Estimators: D = Direct, S = SPREE, G = GSPREE, M = MSPREE, MM = MMSPREE. Red line: Relative bias = 0. Blue diamond: Mean. First row: f=0.1. Second row: f=0.05. Third row: f=0.01.

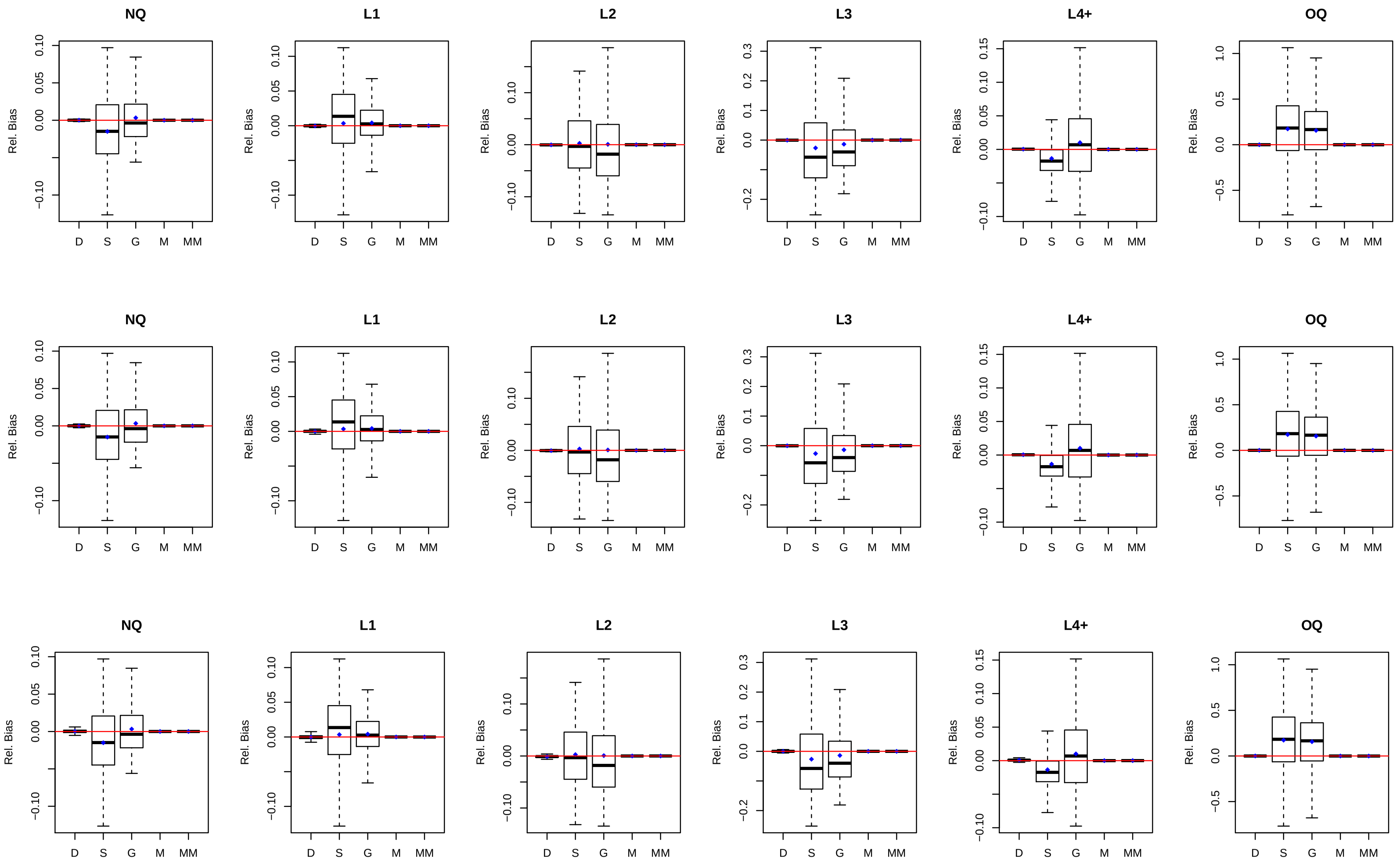


Table 4.16: Relative Bias of estimators of **Y**, by category and sampling fraction. Scenario 1.

Category	Statistic	f=0.1					f=0.05					f=0.1				
		Direct	SPREE	GSPREE	MSPREE	MMSPREE	Direct	SPREE	GSPREE	MSPREE	MMSPREE	Direct	SPREE	GSPREE	MSPREE	MMSPREE
NQ	Q1	-0.0004	-0.0447	-0.0217	-0.0001	-0.0001	-0.0007	-0.0447	-0.0217	-0.0001	-0.0001	-0.0013	-0.0447	-0.0217	-0.0001	-0.0001
	Median	0.0001	-0.0148	-0.0037	0.0000	0.0000	0.0001	-0.0148	-0.0037	0.0000	0.0000	0.0001	-0.0148	-0.0037	0.0000	0.0000
	Mean	0.0001	-0.0150	0.0033	0.0000	0.0000	0.0001	-0.0150	0.0033	0.0000	0.0000	0.0002	-0.0150	0.0033	0.0000	0.0000
	Q3	0.0007	0.0203	0.0213	0.0001	0.0002	0.0008	0.0203	0.0213	0.0001	0.0002	0.0018	0.0203	0.0214	0.0002	0.0002
L1	Q1	-0.0008	-0.0251	-0.0136	-0.0002	-0.0002	-0.0010	-0.0251	-0.0136	-0.0002	-0.0002	-0.0022	-0.0251	-0.0136	-0.0002	-0.0003
	Median	-0.0001	0.0136	0.0027	0.0000	0.0000	0.0000	0.0136	0.0027	0.0000	0.0000	-0.0002	0.0136	0.0027	0.0000	-0.0001
	Mean	-0.0002	0.0034	0.0042	0.0000	0.0000	-0.0001	0.0034	0.0042	0.0000	0.0000	-0.0002	0.0034	0.0042	0.0000	0.0000
	Q3	0.0005	0.0451	0.0223	0.0002	0.0002	0.0011	0.0451	0.0224	0.0002	0.0002	0.0018	0.0451	0.0224	0.0002	0.0002
L2	Q1	-0.0007	-0.0447	-0.0597	-0.0001	-0.0002	-0.0011	-0.0447	-0.0598	-0.0001	-0.0002	-0.0026	-0.0447	-0.0598	-0.0001	-0.0002
	Median	-0.0003	-0.0032	-0.0181	0.0000	0.0000	-0.0004	-0.0032	-0.0181	0.0000	0.0000	-0.0011	-0.0032	-0.0181	0.0000	0.0000
	Mean	-0.0002	0.0027	0.0007	0.0000	0.0000	-0.0005	0.0027	0.0007	0.0000	0.0000	-0.0012	0.0027	0.0007	0.0000	0.0000
	Q3	0.0001	0.0457	0.0381	0.0001	0.0001	0.0001	0.0457	0.0381	0.0001	0.0001	0.0001	0.0457	0.0381	0.0002	0.0002
L3	Q1	-0.0009	-0.1273	-0.0865	-0.0002	-0.0003	-0.0010	-0.1273	-0.0865	-0.0002	-0.0002	-0.0016	-0.1273	-0.0865	-0.0002	-0.0004
	Median	-0.0003	-0.0577	-0.0401	0.0000	0.0000	-0.0003	-0.0577	-0.0401	0.0000	0.0000	0.0000	-0.0577	-0.0401	0.0000	-0.0001
	Mean	-0.0003	-0.0266	-0.0139	0.0000	-0.0001	-0.0002	-0.0266	-0.0139	0.0000	0.0000	-0.0001	-0.0266	-0.0139	0.0000	-0.0001
	Q3	0.0003	0.0574	0.0340	0.0001	0.0001	0.0005	0.0574	0.0340	0.0002	0.0002	0.0015	0.0574	0.0340	0.0002	0.0002
L4+	Q1	0.0001	-0.0313	-0.0327	-0.0001	-0.0001	0.0000	-0.0313	-0.0327	-0.0001	-0.0001	-0.0001	-0.0313	-0.0327	-0.0001	-0.0001
	Median	0.0003	-0.0174	0.0069	0.0000	0.0000	0.0004	-0.0174	0.0069	0.0000	0.0000	0.0008	-0.0174	0.0069	0.0000	0.0001
	Mean	0.0004	-0.0135	0.0101	0.0000	0.0000	0.0004	-0.0135	0.0101	0.0000	0.0000	0.0008	-0.0135	0.0101	0.0000	0.0001
	Q3	0.0006	-0.0006	0.0456	0.0001	0.0002	0.0008	-0.0006	0.0456	0.0001	0.0001	0.0017	-0.0006	0.0457	0.0002	0.0003
OQ	Q1	-0.0008	-0.0629	-0.0517	-0.0003	-0.0006	-0.0009	-0.0629	-0.0517	-0.0002	-0.0005	-0.0025	-0.0629	-0.0517	-0.0007	-0.0013
	Median	-0.0001	0.1830	0.1661	-0.0001	-0.0003	0.0003	0.1830	0.1661	0.0000	-0.0002	-0.0004	0.1830	0.1661	-0.0003	-0.0008
	Mean	0.0001	0.1733	0.1563	-0.0001	-0.0002	0.0003	0.1733	0.1563	0.0000	-0.0002	0.0000	0.1733	0.1563	-0.0002	-0.0006
	Q3	0.0010	0.4264	0.3636	0.0002	0.0001	0.0015	0.4264	0.3636	0.0002	0.0001	0.0025	0.4264	0.3635	0.0002	-0.0001

Figure 4.7: RSRMSE of estimators of **Y**, by category and sampling fraction. Scenario 1.
Estimators: D = Direct, S = SPREE, G = GSPREE, M = MSPREE, MM = MMSPREE. Red line: RSRMSE = 0. First row: f=0.1. Second row: f=0.05. Third row: f=0.01.

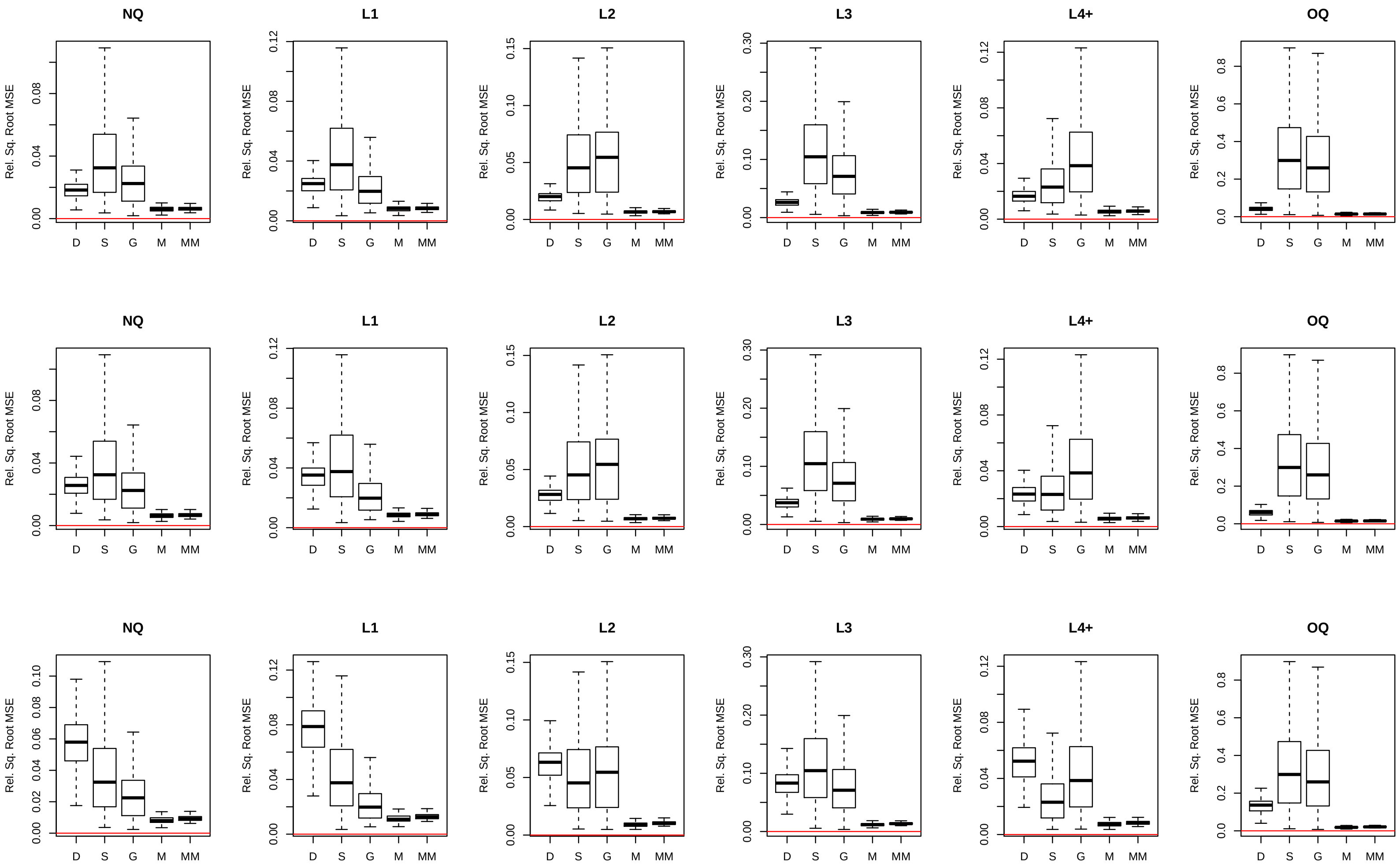


Table 4.17: RSRMSE of estimators of $\mathbf{Y}$, by category and sampling fraction. Scenario 1.

Category	Statistic	f=0.1					f=0.05					f=0.01				
		Direct	SPREE	GSPREE	MSPREE	MMSPREE	Direct	SPREE	GSPREE	MSPREE	MMSPREE	Direct	SPREE	GSPREE	MSPREE	MMSPREE
NQ	Q1	0.0146	0.0168	0.0112	0.0049	0.0055	0.0207	0.0168	0.0112	0.0052	0.0058	0.0460	0.0168	0.0112	0.0069	0.0082
	Median	0.0183	0.0325	0.0224	0.0060	0.0062	0.0257	0.0325	0.0224	0.0063	0.0066	0.0580	0.0325	0.0225	0.0080	0.0092
	Mean	0.0184	0.0396	0.0269	0.0061	0.0064	0.0259	0.0396	0.0270	0.0064	0.0068	0.0583	0.0396	0.0270	0.0084	0.0096
	Q3	0.0219	0.0538	0.0336	0.0072	0.0072	0.0308	0.0538	0.0336	0.0075	0.0076	0.0690	0.0538	0.0336	0.0098	0.0106
L1	Q1	0.0201	0.0208	0.0118	0.0067	0.0075	0.0284	0.0208	0.0118	0.0073	0.0080	0.0636	0.0208	0.0118	0.0096	0.0113
	Median	0.0249	0.0376	0.0198	0.0083	0.0085	0.0352	0.0376	0.0198	0.0086	0.0089	0.0787	0.0376	0.0198	0.0109	0.0127
	Mean	0.0246	0.0469	0.0225	0.0082	0.0086	0.0347	0.0469	0.0225	0.0087	0.0091	0.0782	0.0469	0.0226	0.0115	0.0132
	Q3	0.0283	0.0620	0.0296	0.0093	0.0093	0.0399	0.0620	0.0296	0.0097	0.0100	0.0901	0.0620	0.0296	0.0132	0.0143
L2	Q1	0.0164	0.0239	0.0240	0.0055	0.0061	0.0231	0.0239	0.0240	0.0059	0.0066	0.0520	0.0239	0.0240	0.0076	0.0092
	Median	0.0201	0.0453	0.0545	0.0066	0.0068	0.0282	0.0453	0.0545	0.0069	0.0073	0.0632	0.0453	0.0545	0.0088	0.0102
	Mean	0.0198	0.0509	0.0620	0.0066	0.0069	0.0281	0.0509	0.0620	0.0070	0.0075	0.0627	0.0509	0.0620	0.0093	0.0107
	Q3	0.0226	0.0742	0.0765	0.0075	0.0075	0.0318	0.0742	0.0765	0.0079	0.0081	0.0713	0.0742	0.0765	0.0104	0.0115
L3	Q1	0.0214	0.0584	0.0407	0.0073	0.0083	0.0301	0.0584	0.0407	0.0076	0.0088	0.0673	0.0584	0.0407	0.0100	0.0122
	Median	0.0262	0.1046	0.0709	0.0088	0.0092	0.0374	0.1046	0.0709	0.0092	0.0097	0.0832	0.1046	0.0709	0.0119	0.0135
	Mean	0.0264	0.1160	0.0821	0.0088	0.0093	0.0373	0.1160	0.0821	0.0092	0.0099	0.0835	0.1160	0.0821	0.0120	0.0139
	Q3	0.0309	0.1595	0.1065	0.0101	0.0102	0.0432	0.1595	0.1065	0.0105	0.0108	0.0975	0.1595	0.1065	0.0135	0.0150
L4+	Q1	0.0130	0.0119	0.0197	0.0044	0.0051	0.0183	0.0119	0.0197	0.0047	0.0054	0.0411	0.0119	0.0197	0.0061	0.0073
	Median	0.0165	0.0230	0.0384	0.0055	0.0058	0.0233	0.0230	0.0385	0.0057	0.0062	0.0523	0.0230	0.0385	0.0072	0.0081
	Mean	0.0167	0.0275	0.0449	0.0055	0.0059	0.0237	0.0275	0.0449	0.0058	0.0063	0.0524	0.0275	0.0450	0.0075	0.0085
	Q3	0.0199	0.0361	0.0626	0.0064	0.0066	0.0280	0.0361	0.0626	0.0067	0.0070	0.0618	0.0361	0.0626	0.0086	0.0094
OQ	Q1	0.0333	0.1478	0.1321	0.0112	0.0127	0.0468	0.1478	0.1321	0.0119	0.0135	0.1061	0.1478	0.1321	0.0159	0.0192
	Median	0.0429	0.2992	0.2597	0.0142	0.0143	0.0601	0.2992	0.2597	0.0148	0.0153	0.1365	0.2992	0.2597	0.0185	0.0210
	Mean	0.0421	0.3415	0.3015	0.0140	0.0147	0.0594	0.3415	0.3015	0.0147	0.0156	0.1334	0.3415	0.3014	0.0187	0.0214
	Q3	0.0500	0.4729	0.4263	0.0165	0.0161	0.0705	0.4729	0.4263	0.0170	0.0172	0.1573	0.4729	0.4263	0.0212	0.0233

Figure 4.8: Relative Bias of $\sqrt{\widehat{AV}(\hat{Y}^M)}$, by category and sampling fraction. Scenario 1.
Red line: Relative bias = 0. Blue diamond: Mean.

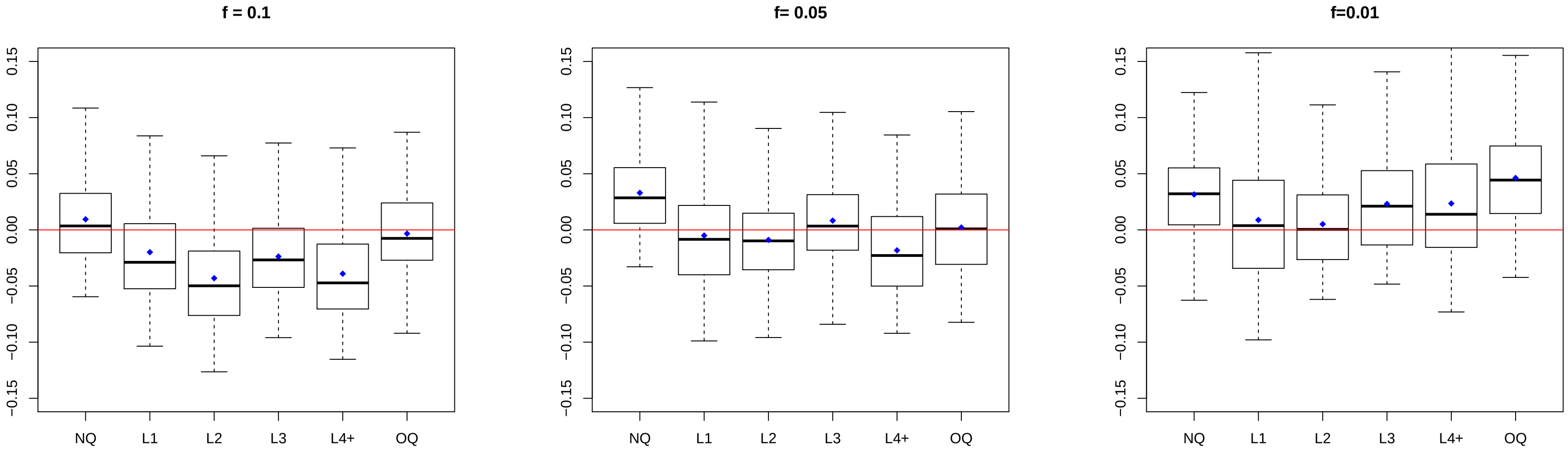


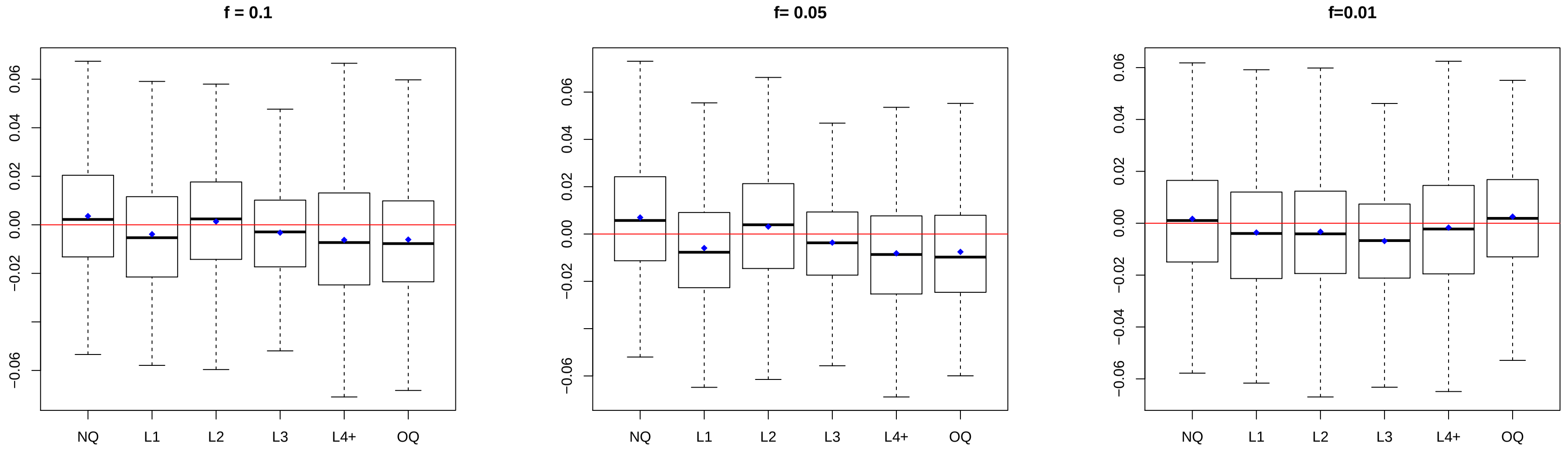
Table 4.18: Relative Bias of $\sqrt{\widehat{\mathrm{AV}}(\hat{\mathrm{Y}}^{\mathrm{M}})}$, by category and sampling fraction. Scenario 1.

Category	f=0.1				f=0.05				f=0.01			
	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
NQ	-0.0204	0.0035	0.0094	0.0324	0.0060	0.0285	0.0330	0.0554	0.0045	0.0321	0.0316	0.0552
L1	-0.0524	-0.0289	-0.0199	0.0052	-0.0399	-0.0084	-0.0050	0.0218	-0.0340	0.0038	0.0088	0.0440
L2	-0.0761	-0.0498	-0.0430	-0.0189	-0.0355	-0.0098	-0.0090	0.0146	-0.0264	0.0004	0.0051	0.0312
L3	-0.0512	-0.0267	-0.0238	0.0013	-0.0180	0.0034	0.0083	0.0310	-0.0133	0.0211	0.0231	0.0527
L4+	-0.0704	-0.0472	-0.0390	-0.0127	-0.0501	-0.0228	-0.0182	0.0118	-0.0155	0.0139	0.0235	0.0586
OQ	-0.0270	-0.0076	-0.0033	0.0235	-0.0307	0.0009	0.0021	0.0318	0.0153	0.0443	0.0462	0.0746

Table 4.19: Relative Bias of $\sqrt{\widehat{\mathrm{MSE}}(\hat{\mathbf{Y}}^{\mathrm{M}})}$, by category and sampling fraction. Scenario 1.

Category	f=0.1				f=0.05				f=0.01			
	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
NQ	-0.0132	0.0022	0.0036	0.0204	-0.0112	0.0057	0.0070	0.0242	-0.0149	0.0010	0.0017	0.0165
L1	-0.0214	-0.0053	-0.0039	0.0116	-0.0227	-0.0077	-0.0059	0.0090	-0.0213	-0.0039	-0.0036	0.0120
L2	-0.0142	0.0024	0.0014	0.0175	-0.0145	0.0039	0.0031	0.0212	-0.0194	-0.0041	-0.0033	0.0122
L3	-0.0172	-0.0029	-0.0032	0.0101	-0.0174	-0.0037	-0.0036	0.0093	-0.0211	-0.0067	-0.0069	0.0074
L4+	-0.0247	-0.0073	-0.0062	0.0131	-0.0253	-0.0087	-0.0082	0.0076	-0.0194	-0.0022	-0.0017	0.0146
OQ	-0.0235	-0.0077	-0.0061	0.0098	-0.0246	-0.0098	-0.0075	0.0079	-0.0130	0.0019	0.0025	0.0168

Figure 4.9: Relative Bias of $\sqrt{\widehat{\text{MSE}}(\hat{\mathbf{Y}}^{M})}$, by category and sampling fraction. Scenario 1.
Red line: Relative bias = 0. Blue diamond: Mean.



4.5.4 Additional results. Scenario 2.

Table 4.20: Relative Bias and RSRMSE of estimators of $\mathbf{Y}$, by category. Scenario 2.

Category	Statistic	Relative Bias				RSRMSE			
		SPREE	GSPREE	MSPREE	MMSPREE	SPREE	GSPREE	MSPREE	MMSPREE
NQ	Q1	-0.0443	-0.0212	-0.0023	-0.0010	0.0924	0.0913	0.0882	0.0251
	Median	-0.0143	-0.0034	-0.0001	-0.0002	0.0984	0.0948	0.0927	0.0289
	Mean	-0.0151	0.0032	-0.0001	0.0000	0.1045	0.0989	0.0932	0.0290
	Q3	0.0200	0.0193	0.0022	0.0007	0.1096	0.1004	0.0983	0.0329
L1	Q1	-0.0250	-0.0141	-0.0023	-0.0012	0.1047	0.1019	0.1000	0.0317
	Median	0.0133	0.0023	0.0001	-0.0002	0.1089	0.1048	0.1021	0.0365
	Mean	0.0036	0.0043	0.0001	-0.0001	0.1172	0.1063	0.1029	0.0361
	Q3	0.0448	0.0232	0.0025	0.0007	0.1203	0.1093	0.1049	0.0404
L2	Q1	-0.0456	-0.0592	-0.0024	-0.0009	0.0987	0.0994	0.0928	0.0271
	Median	-0.0025	-0.0175	0.0000	-0.0002	0.1064	0.1104	0.0953	0.0310
	Mean	0.0028	0.0008	0.0000	-0.0001	0.1125	0.1206	0.0965	0.0307
	Q3	0.0466	0.0431	0.0023	0.0005	0.1227	0.1234	0.0989	0.034
L3	Q1	-0.1280	-0.0879	-0.0047	-0.0013	0.2084	0.2048	0.1963	0.0420
	Median	-0.0581	-0.0396	0.0005	-0.0002	0.2265	0.2144	0.2013	0.0466
	Mean	-0.0269	-0.0143	0.0001	-0.0001	0.2408	0.2237	0.2005	0.0464
	Q3	0.0538	0.0311	0.0049	0.0010	0.2557	0.2279	0.2049	0.0507
L4+	Q1	-0.0309	-0.0315	-0.0022	-0.0008	0.0867	0.0906	0.0836	0.0232
	Median	-0.0176	0.0079	-0.0001	-0.0002	0.0933	0.0959	0.0894	0.0271
	Mean	-0.0136	0.0101	0.0000	-0.0001	0.0947	0.1027	0.0885	0.0271
	Q3	0.0001	0.0437	0.0022	0.0006	0.0993	0.1072	0.0937	0.0304
OQ	Q1	-0.0602	-0.0452	-0.0113	-0.0031	0.4838	0.4777	0.4483	0.2062
	Median	0.1848	0.1691	0.0002	0.0042	0.5478	0.5322	0.4583	0.2463
	Mean	0.1699	0.1530	0.0002	0.0045	0.5999	0.5730	0.4585	0.2552
	Q3	0.4202	0.3704	0.0108	0.0120	0.6667	0.6259	0.4716	0.2984

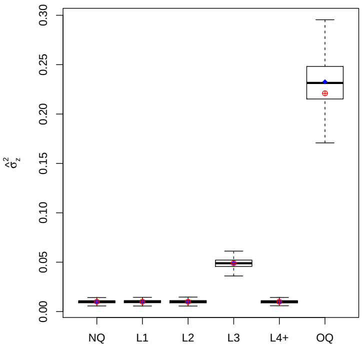
Table 4.21: Summary of Bias results for $\sqrt{\widehat{\text{MSE}}(\hat{\mathbf{Y}}))}$ by category. Scenario 2.

Category	Proportion of areas with 0 not in the 95% PI		|Relative bias| Mean	
	MSPREE	MMSPREE	MSPREE	MMSPREE
NQ	0.7746	0.8468	0.0192	0.0614
L1	0.7688	0.8468	0.0201	0.0438
L2	0.8121	0.8353	0.0200	0.0523
L3	0.7919	0.8671	0.0214	0.0427
L4+	0.8410	0.8642	0.0213	0.0704
OQ	0.8728	0.8613	0.0442	0.2450
Average	0.8319	0.8319	0.0244	0.0860

Table 4.22: Relative Bias of $\sqrt{\widehat{\text{MSE}}(\hat{\mathbf{Y}}))}$, by category. Scenario 2.

Category	MSPREE				MMSPREE			
	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
NQ	-0.0158	0.0005	0.0006	0.0168	-0.0038	0.0403	0.0357	0.0796
L1	-0.0150	-0.0008	0.0003	0.0177	-0.0113	0.0194	0.0223	0.0579
L2	-0.0167	0.0003	0.0006	0.0161	-0.0044	0.0335	0.0305	0.0687
L3	-0.0179	-0.0013	0.0001	0.0171	-0.0169	0.0159	0.0141	0.0493
L4+	-0.0191	-0.0007	-0.0002	0.0156	-0.0013	0.0405	0.0445	0.0924
OQ	0.0073	0.0374	0.0343	0.0618	0.0022	0.1635	0.1891	0.3653
Average	-0.0147	0.0028	0.0060	0.0233	-0.0087	0.0353	0.0560	0.0840

Figure 4.10: Estimates of the variance components $\hat{\sigma}_j^2$, by category. Scenario 2.
Blue diamond: Mean. Red dot: True σ_j^2



Chapter 5

Application. Estimation of the distribution of ethnic group by Local Authority in England

This chapter describes an application of the proposed methodology that is ongoing joint work with the Small Area Estimation team at the Office for National Statistics (ONS). Most of the material presented in sections 5.1 and 5.2 as well as some of the material at the beginning of section 5.3 has already been published as part of Luna et al. (2015). However, the scope of the analysis presented here exceeds what is covered in that paper, and results for the MSPREE and MMSPREE estimators are shown here for the first time. The main contributions of the co-authors in Luna et al. (2015) were the access to the datasets and excerpts in the description of the data sources (section 5.2). The data analysis is solely the responsibility of the author of this thesis.

5.1 Motivation

Estimates of demographic characteristics are among the main outputs of National Statistical Institutes. In addition to national and regional estimates, for topics such as Labour Force, Household composition or Ethnicity, periodic estimates at lower levels of geographic aggregation are in high demand both for public policy and research purposes. In census years, given the availability of data for almost all individuals in the population, reliable estimates for small geographic domains can be produced in a simply way. During the inter-censal period, updated socio-demographic data can only be obtained via sample surveys or administrative systems. It is generally difficult to obtain reliable direct estimates for small geographic domains from sample surveys due to the small sample sizes. Data from administrative systems do not have this problem but in contrast, may not cover the topics of interest. Moreover, definitions of the variables and domains in administrative sources respond to different purposes than those of the statistical interest. This can result in comparability issues with figures obtained from population censuses or household surveys.

Despite the suitability of SAE methods to address this type of problem, few official figures in the region are being produced using this approach. In the UK case, the ONS currently disseminates periodically small area estimates regarding three main topics: *population estimates*, combining data from the Patient Register and other sources; *households in poverty*, using the Family Resources Survey (FRS) and administrative data maintained by the Department for Work and Pensions; and *unemployment*, making use of the Annual Population Survey and the administrative register of jobseekers allowance.

As previously mentioned, the application hereby presented is the result of an ongoing joint work with the Small Area Estimation team at ONS. It addresses the problem of how to obtain estimates of the distribution of the population by ethnic group, in each LA of England, using proxy and survey data. Studying the feasibility of producing such estimates during the inter-censal period is a topic of interest for ONS. Moreover, because population censuses are experiencing transformations in many European countries, with increasing emphasis being given to alternative operations based on demographic systems that use information from administrative sources alone or combined with survey data, the potential impact of this type of methodology is expected to grow considerably in future.

5.2 Data sources

Data for this application was obtained from four different sources of the Office for National Statistics. Because some of these are subject to disclosure control, it was necessary to perform all the data analysis in a Safe Room of the Virtual Microdata Laboratory (VML) of ONS. All the calculations hereby presented are the sole responsibility of the author of this thesis.

5.2.1 Proxy Information

A proxy of the composition of interest was obtained from the 2011 Population Census in England, which counts the persons and households considered as usual residents of England and Wales on the 27th March. The 2011 census has an initial estimated coverage rate for persons of 93% and the observed counts are adjusted by over and undercount.

5.2.2 Survey estimates

For this application, we used direct estimates obtained from the Annual Population Survey (APS) for the period July 2012-June 2013, with reference point the 31st of December 2012, for all LAs in England, excluding *Isles of Scilly* and *City of London*. With quarterly periodicity and a total sample size of approximately 250,000 individuals per year, the APS has the biggest sample size among all periodic demographic surveys conducted by ONS.

For any given year, the APS consists of waves 1 and 5 of four successive quarters from the Labour Force Survey (LFS), plus the Annual Local Area LFS (ALA LFS) boost. Notice that because only households in the waves 1 and 5 of the LFS are included in the quarterly APS, each respondent appears only once in any given yearly dataset. Both LFS and the ALA LFS boost cover mainly private households, therefore communal establishments, armed forces accommodation etc., are not included in the APS. An implicit sampling fraction for the APS was calculated dividing the observed sample size of the period July 2012-June 2013 by the corresponding projected population total in each LA. Such fraction varies between 0.05% and 2.5% across LA, with an average of 0.8%.

5.2.3 Benchmark totals

Estimates of the LA population sizes can be obtained from the official mid-year population estimates. These estimates are produced by ONS using the cohort component method, which uses information on components of population change to update the most recent census population. The 2012 and 2013 mid-year population estimates at LA level were used to calculate the row marginal. As the reference date of such estimates is 30th of June of the corresponding year, an average of the mid-year population estimates for 2012 and

2013 would provide an estimate of the population close to the 31$^{\text{st}}$ of December 2012, consistent with the reference period chosen for the APS.

Direct estimates of the total population size by category of ethnicity, obtained from the APS at the national level, were rescaled to agree with the row marginal above described, and then used as benchmarks for the columns in this application.

5.2.4 Categories of the variable

The variable Ethnic group is collected in England in a very detailed way. The APS collects information regarding 15 subcategories of Ethnicity, grouped in 7 main categories: White, Mixed/multiple ethnic groups, Asian/Asian British, Black/African/Caribbean/Black British, Chinese, Arab and Other ethnic group. The Census 2011, on the other hand, uses 18 subcategories grouped in 5 main categories, with Chinese included within Asian and Arab within Other. To use a classification that is fully harmonisable with both sources, the following six categories were chosen for this application: *White*; *Mixed/multiple ethnic groups*; *Asian/Asian British*; *Black/African/Caribbean/Black British*; *Chinese* and *Other*.

5.3 Main results

Census estimates

According to Census 2011 data, the variable ethnicity presents a very unequal distribution in this population. Aggregating over the areas in consideration, the category White is dominant with 85.42% of individuals, followed by Asian (7.10%), Black (3.48%), Mixed (2.25%), Other (1.03%) and finally Chinese (0.72%). How different LA deviate from that global distribution can be observed in Figure 5.1. Notice that for categories Asian and Black it is possible to find some areas with proportions considerable higher than the global proportion. Moreover, notice that in such areas, non-white individuals are predominantly from one of the two above mentioned categories instead of evenly distributed. Meanwhile, for the categories Mixed, Chinese and Other, the proportions are uniformly low in all LA.

Figure 5.1: Distribution of Ethnicity by LA. Population Census 2011 in England.
Left: Boxplot proportions in each category by LA. Red diamond: mean.
Right: Detail of the more frequent categories. Lines: White: continuous grey. Asian: dotted black. Black: continuous black. After sorting the LAs according to the proportion of White, one of each three was included in the plot.

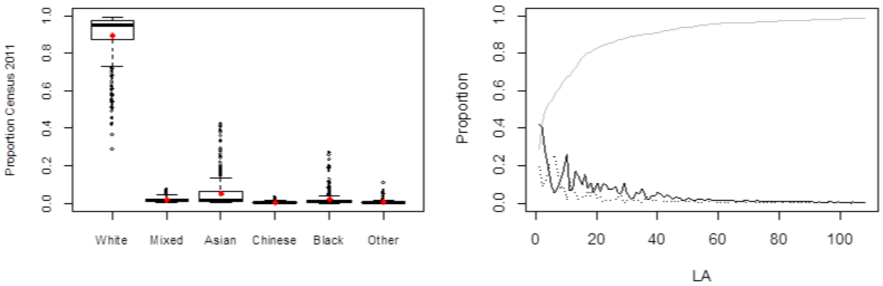
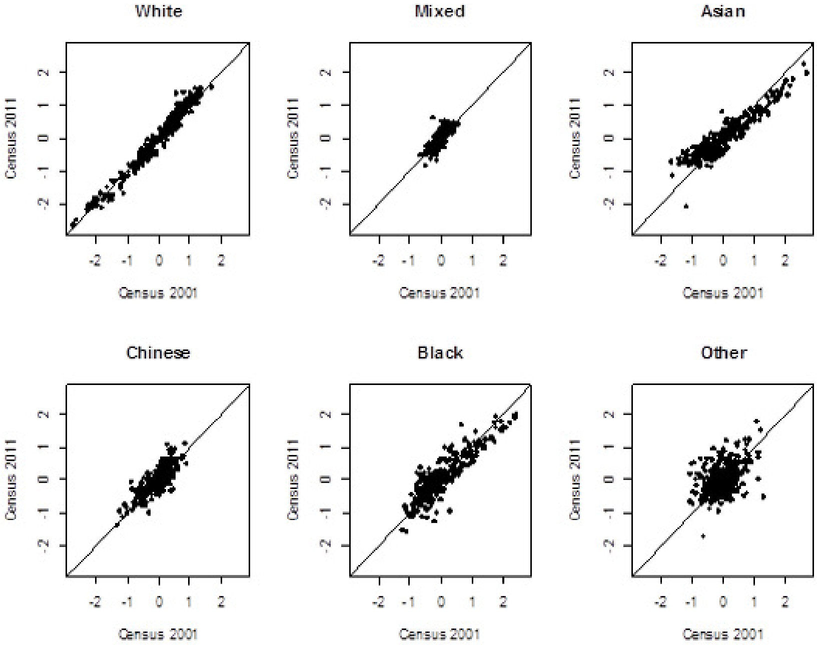


Figure 5.2: Interaction terms of the composition LA by Ethnicity. Population censuses 2001 and 2011 in England.
Line: Y=X.



To illustrate the linear relationship between pairs of interaction terms for the composition of ehtnicity by LA, Figure 5.2 shows the corresponding terms in the compositions of Census 2001 and Census 2011. Notice how, except for the category Other, interaction terms from the same composition 10 years before can still work fairly well as linear predictors. Given that the time lapse that is of our interest is considerably shorter, we expect then for interaction terms of the census 2011 composition to be good predictors in this case.

APS estimates

A characterisation of the distribution of ethnicity by LA using direct estimates is less straightforward due to the small sample sizes of the APS in some areas and the low frequency of some ethnic groups. Table 5.1 presents the distribution of LA according to the number of cells with zero estimates. Despite of the absence of structural zeroes in the proxy composition, 55.2% of the LA have at least one zero estimate according to the survey data. Moreover, for 4% of the LA only one ethnic group was observed in the APS.

Table 5.1: LA by number of cells with a direct estimate equal to zero. APS 2012-2013.

Number of cells with a zero estimate	Frequency	%
None	155	44.80
One	61	17.63
Two	55	15.90
Three	34	9.83
Four	27	7.80
Five	14	4.04
Total	346	100.00

Given the unequal distribution of the variable ethnicity that was previously mentioned, it is expected that some categories present a higher number of zero estimates than others. According to Table 5.2, which shows the number of LA with zero estimates by ethnic group, this problem is particularly accentuated for the category Chinese, with a zero estimate for more than 40% of the LA and, in a lesser degree, for Black and Other which are missing from around 26% of the LA.

Table 5.2: LA with a direct estimate equal to zero by Ethnic group. APS 2012-2013.

Category	Frequency	%
White	0	0.00
Mixed	58	15.93
Asian	43	11.81
Chinese	156	42.86
Black	96	26.37
Other	98	26.92

For the cells with a positive estimate, an Approximated Standard Error (ASE) was obtained as

$$\mathrm{ASE}(\hat{\theta}_{aj}) = \sqrt{\frac{\hat{\theta}_{aj}(1-\hat{\theta}_{aj})}{n_{a+}}} \times \mathrm{DEFT}_j$$

where $\hat{\theta}_{aj}$ is the estimate of the within LA proportion corresponding to category j in LA a; n_{a+} is the observed sample size on that LA; and DEFT_j is a category-specific design factor included to take into account, at least partially, the complexity of the sampling design. Because, up to our knowledge, design factors for the variable ethnicity in the APS are not available, we used a set of factors provided by ONS for the LFS (Office for National Statistics, 2011, page 146). Those design factors, which correspond to the LFS sample of the last quarter of 2010, for people aged 16 or more, are presented in Table 5.3. Particularly for the ethnic groups with lower frequencies, they can be considered conservative in the context of the LFS (see discussion in Office for National Statistics, 2011, Section 8.8).

Table 5.3: Design Factors for population aged 16 or more by ethnicity. LFS October-December 2010.

White	Mixed	Asian	Chinese	Black	Other
1.5441	1.2903	1.7351	1.6557	1.5809	1.6394

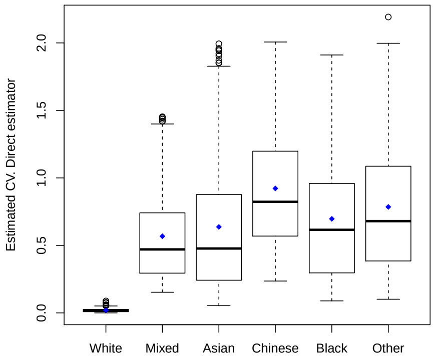
Using the $\mathrm{ASE}(\hat{\theta}_{aj})$ as numerator, an approximated Coefficient of Variation (CV) was calculated for each cell with positive estimate. Some descriptive statistics of these CVs are presented in Table 5.4 and a summary is provided in Figure 5.3. Considering an arbitrary threshold of CV of 0.2 or higher, only for category White the direct estimates could be considered accurate enough as to be useful. All other categories show a prohibitive high CV in most LA. However, in the evaluation of the accuracy of the direct estimates it is important to keep in mind that the information provided by the CV in this case can be irrelevant given the small sizes of the estimates of the proportions for most categories. Alternative measures of accuracy, such as the length of a confidence interval, might be required in this case in order to provide a more complete picture.

Before proceeding with the calculation of the SPREE-type estimators, it would be of interest to study the relationship between the association structures of the Census 2011 and APS compositions, in a similar way as it was done for the 2001 and 2011 censuses (see Figure 5.2). Unfortunately, this is infeasible due to the high number or cells with zero estimates obtained in the APS.

Table 5.4: Descriptive statistics. Approximated CVs for the estimators of the distribution of Ethnicity by LA. APS 2012-2013.

Statistic	Category					
	White	Mixed	Asian	Chinese	Black	Other
Q1	0.0087	0.2944	0.2418	0.5708	0.2965	0.3857
Median	0.0157	0.4702	0.4769	0.8228	0.6153	0.6797
Mean	0.0192	0.5675	0.6369	0.9219	0.6966	0.7850
Q3	0.0259	0.7399	0.8769	1.1947	0.9584	1.0828

Figure 5.3: Approximated CVs for the estimators of the distribution of Ethnicity by LA. APS 2012-2013.
Blue diamond: Mean.



SPREE, GSPREE and MSPREE

An estimate of the variance-covariance matrix of the APS estimates of the frequencies was obtained using the $\mathrm{ASE}(\hat{\theta}_{aj})$ previously defined, assuming independence between LAs and known total LA size, Y_{a+}. Such estimate and an IWLS algorithm were used to obtain estimates for the parameters of the GSPREE and MSPREE. The estimated coefficient of the GSPREE is 0.9976, i.e., basically coincides with the SPREE in this application. Therefore, only the GSPREE and MSPREE will considered in the following discussion.

Table 5.5 presents the estimated parameters of the MSPREE. Each row represents the coefficients of a predictor of an interaction, i.e., $\hat{\alpha}^{Y}_{aj} = \sum_{l} \hat{\beta}_{jl} \alpha^{X}_{al}$. The coefficients have been rescaled for an interpretation in terms of the proportional interactions, i.e., the diagonal elements are free and the sum of non-diagonal elements is zero by row and column (see Section 2.4.1). Notice that the free coefficients for categories White, Asian and Black are very close to 1 (SPREE, GSPREE), therefore, they may benefit less from the additional flexibility offered by the MSPREE than the remaining categories.

Table 5.5: Estimated coefficients of the MSPREE estimator. Matrix **B**.

	White	Mixed	Asian	Chinese	Black	Other
White	1.0338	0.0876	0.0804	-0.0743	0.1705	-0.2641
Mixed	0.0042	0.8121	-0.2256	-0.0353	-0.1989	0.4555
Asian	-0.0333	-0.0453	0.9946	0.1423	-0.0165	-0.0471
Chinese	-0.0015	-0.0320	0.0046	1.3014	-0.0062	0.0350
Black	0.0391	0.0454	0.1266	-0.0318	1.0383	-0.1793
Other	-0.0085	-0.0557	0.0140	-0.0009	0.0511	0.7272

Figure 5.4 compares Direct and MSPREE estimates. In general, there is no evidence of systematic departure of the MSPREE respect to the APS estimates, even though some over-shrinking is observed for categories Mixed and Chinese. As it can be seen in Figure 5.5, there are not big differences between GSPREE and MSPREE estimates for categories White, Asian and Black, even though some differences, principally for the larger proportions, are observed among the remaining categories. This behaviour was expected according to the estimated matrix of MSPREE discussed before.

MMSPREE

Initial estimates of the variance components required for the calculation of the MMSPREE were obtained using the estimator proposed in equation (3.15) in section 3.2, and the Census 2011 and APS compositions. Such estimates, denoted $\hat{\sigma}^2_{j,1}$, are presented in the first block of Table 5.6. Notice that negative estimates, afterwards truncated to zero, were obtained for categories White and Asian. Moreover, the estimated variance component for category Other is abnormally high and results in a set of predicted random effects which variance is, as large, as the variance of the corresponding synthetic estimates of the interaction terms (last column, rows 2 and 5 of Table 5.6). This is unexpected because, given that the synthetic estimates of the α^Y_{aj} and the predicted random effects are both centred around zero, variances of similar size for these two sets of terms would indicate that the relevance of the survey and synthetic estimates in the construction of the MMSPREE is also similar. This may be reasonable in a situation with big survey sample sizes, but certainly it does not seem to be the case of the APS.

Figure 5.4: Comparison Direct vs MSPREE estimators of the distribution of Ethnicity by LA.
Continuous line: $Y = X$

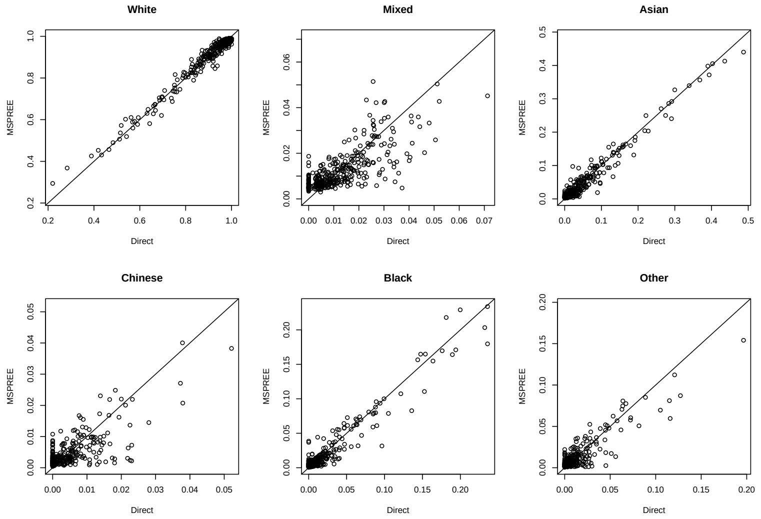
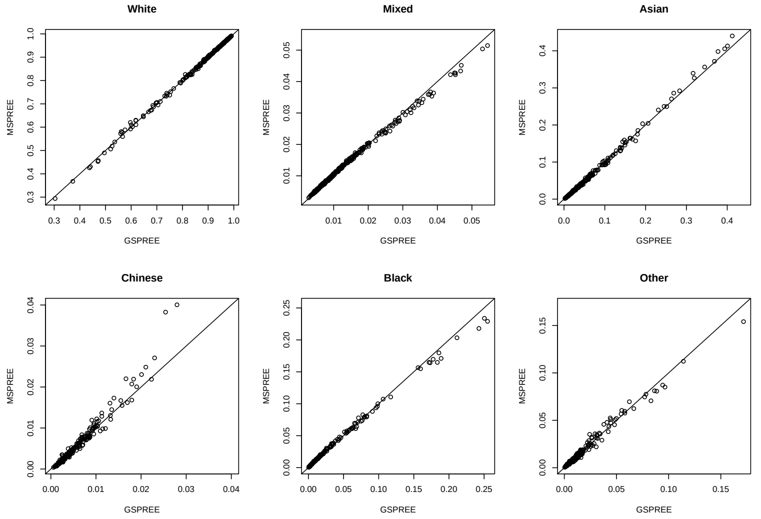


Figure 5.5: Comparison GSPREE vs MSPREE estimators of the distribution of Ethnicity by LA.
Continuous line: $Y = X$



Due to the instability exhibited by the estimates $\hat{\sigma}^2_{j,1}$, an alternative set of estimates was obtained using the two census compositions (2001 and 2011). These estimates are denoted by $\hat{\sigma}^2_{j,2}$ and presented in the second block of Table 5.6. Notice that such estimates do not show evidence of the issues discussed above, which may be caused by the small sample sizes of the APS. Unfortunately, they are not expected to be unbiased in this case.

Table 5.6: Results estimation of the variance components.

	Category					
	White	Mixed	Asian	Chinese	Black	Other
$\hat{\sigma}^2_{j,1}$	0.0000	0.0525	0.0000	0.0163	0.0335	0.9039
$\text{var}(\hat{u}_{aj,1})$	0.0126	0.0275	0.0126	0.0213	0.0278	0.1323
$\hat{\sigma}^2_{j,2}$	0.0243	0.0095	0.0454	0.0449	0.141	0.1357
$\text{var}(\hat{u}_{aj,2})$	0.0509	0.0051	0.0203	0.0111	0.0396	0.0325
$\text{var}(\hat{\alpha}_{aj})$	0.8403	0.0709	0.3624	0.1905	0.3885	0.1569

Results corresponding to the MMSPREE calculated using the variance components $\hat{\sigma}^2_{j,1}$ are presented in figures 5.6 and 5.7, and with the variance components $\hat{\sigma}^2_{j,2}$, in figures 5.8 and 5.9. As expected, there is less evidence of overshrinking with the MMSPREE than with the synthetic MSPREE, particularly with the second set of variance components, which also seems to improve over the MSPREE.

Figure 5.6: Comparison Direct vs MMSPREE($\hat{\sigma}^2_{j,1}$) estimators of the distribution of Ethnicity by LA.
Continuous line: $Y = X$

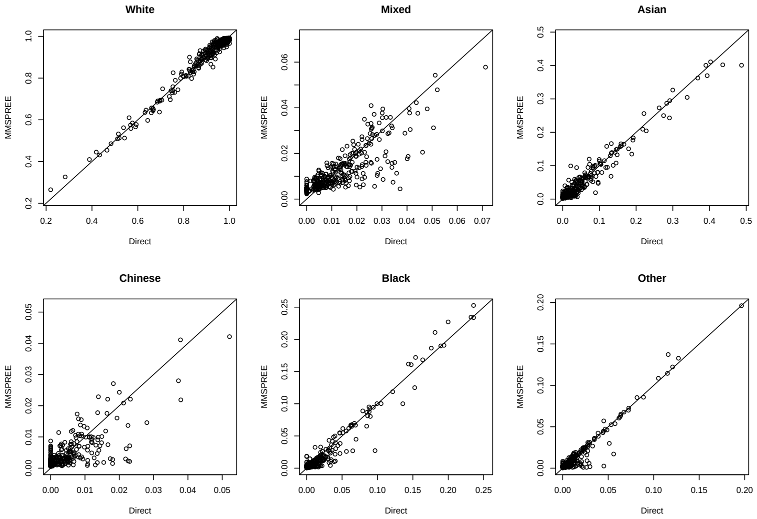


Figure 5.7: Comparison MSPREE vs MMSPREE($\hat{\sigma}^2_{j,1}$) estimators of the distribution of Ethnicity by LA.
Continuous line: $Y = X$

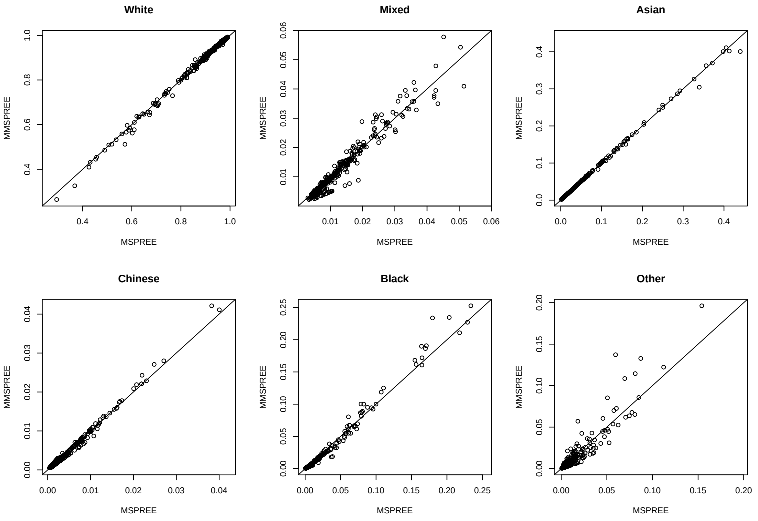


Figure 5.8: Comparison Direct vs MMSPREE($\hat{\sigma}^2_{j,2}$) estimators of the distribution of Ethnicity by LA.
Continuous line: $Y = X$

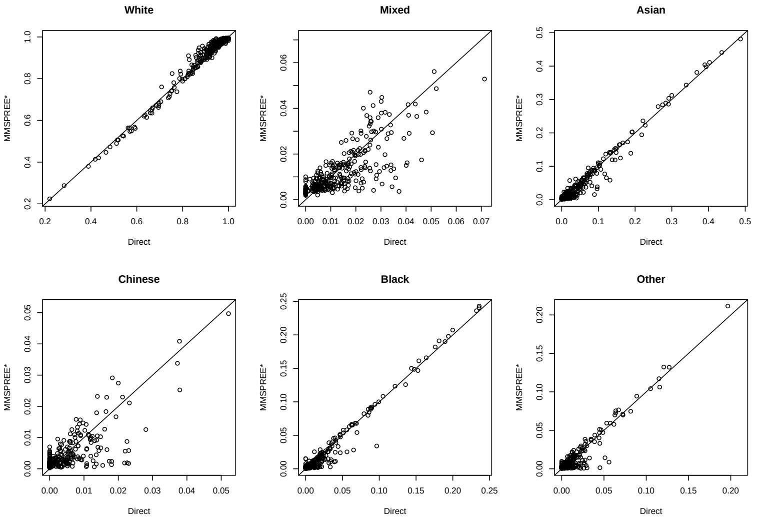
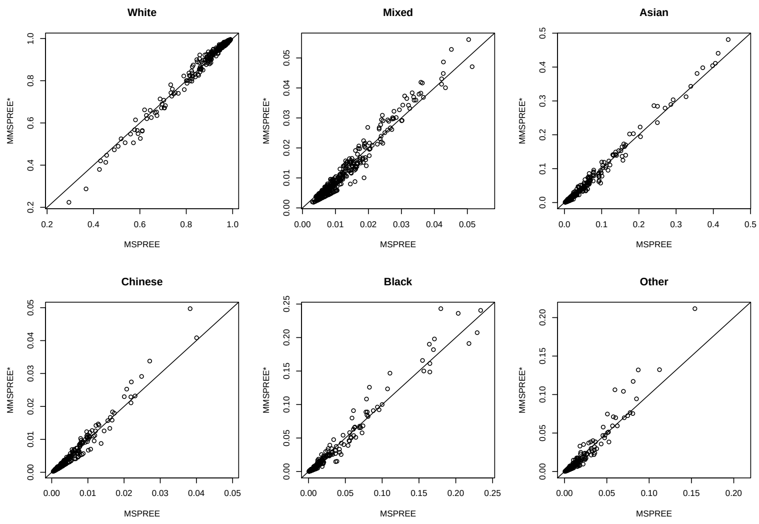


Figure 5.9: Comparison MSPREE vs MMSPREE($\hat{\sigma}^2_{j,2}$) estimators of the distribution of Ethnicity by LA.
Continuous line: $Y = X$



Finally, given the multivariate nature of the MMSPREE and the sum to zero constraints satisfied by the predicted random effects, it is not straightforward to provide results regarding the so-called *shrinking factors*, or in general, the relative sizes of the residual terms respect to the predicted random effects. Remember that according to equation (3.12),

$$\hat{\mathbf{u}} = \hat{\mathbf{\Sigma}}_u \hat{\mathbf{V}}^{-1} \left(\boldsymbol{\eta} - \mathbf{Z}\hat{\boldsymbol{\Psi}} \right)$$

i.e, the predicted random effects are linear combinations of the residual terms $\hat{\mathbf{e}} = \boldsymbol{\eta} - \mathbf{Z}\hat{\boldsymbol{\Psi}}$, with coefficients given by the product of matrices $\hat{\mathbf{\Sigma}}_u \hat{\mathbf{V}}^{-1}$. Even if the sampling design used to obtain the survey data induces independence across areas, i.e., if $\hat{\mathbf{V}}$ is sparse, the sum to zero constraints imposed on $\hat{\mathbf{u}}$ introduce non-zero entries in all the cells of $\hat{\mathbf{\Sigma}}_u$. In practice, this means that all residual terms may have a coefficient different from zero in the equation used to predict any random effect. Clearly, it seems natural that residual terms corresponding to the same area or category have a higher weight than other terms. Here, we propose to study the absolute value of the coefficient of $\hat{e}_{rs}$ when predicting $\hat{u}_{aj}$, denoted by $\hat{\gamma}_{rs,aj}$, for $a, r = 1, \ldots, A$; and $j, s = 1, \ldots, J$.

Tables 5.7 and 5.8 present the averages of $\hat{\gamma}_{rs,aj}$ obtained when using the estimates $\hat{\sigma}^2_{j,1}$ and $\hat{\sigma}^2_{j,2}$ respectively. In each table, the first block of rows correspond to residual terms in the same area and the last row averages all terms in different areas.

Table 5.7: Average of $\gamma_{rs,aj}$ when estimated from $\hat{\sigma}^2_{j,1}$. Rows: a, j. Columns: r, s.

		Category					
		White	Mixed	Asian	Chinese	Black	Other
$r = a$	White	0.8794	0.034	0.1043	0.0255	0.0366	0.0313
	Mixed	0.5429	0.318	0.0654	0.1347	0.118	0.2062
	Asian	0.8794	0.034	0.1043	0.0255	0.0366	0.0313
	Chinese	0.315	0.0433	0.0278	0.4594	0.0391	0.0684
	Black	0.3708	0.0614	0.0435	0.0678	0.3916	0.1176
	Other	1.9025	0.211	0.1745	0.2352	0.2358	0.4021
$r \neq a$		0.0005	0.0028	0.0005	0.003	0.0056	0.0047

Table 5.8: Average of $\gamma_{rs,aj}$ when estimated from $\hat{\sigma}^2_{j,2}$. Rows: a, j. Columns: r, s.

		Category					
		White	Mixed	Asian	Chinese	Black	Other
$r = a$	White	1.0014	0.0252	0.0353	0.017	0.0247	0.0226
	Mixed	0.1243	0.246	0.032	0.0384	0.0726	0.0692
	Asian	0.1701	0.0658	0.3395	0.1342	0.143	0.141
	Chinese	0.1242	0.0308	0.0775	0.4823	0.0965	0.0902
	Black	0.3858	0.0792	0.1138	0.1543	0.4749	0.1441
	Other	0.4196	0.0955	0.145	0.171	0.1767	0.4326
$r \neq a$		0.0006	0.0006	0.0029	0.0019	0.0024	0.002

The results for the variance components estimated from the two census compositions $\hat{\sigma}^2_{j,2}$, presented in Table 5.8 show the expected behaviour. For residual terms in the same area, the diagonal terms, i.e., those corresponding to the same category, are dominant. Nevertheless, White, the category with the biggest sample size, plays an important role in the prediction of all other random effects, even though having a smaller average coefficient than the diagonal term. Moreover, residual terms in other areas have little effect on the prediction of a given random effect.

On the other hand, the results for the variance components estimated from the APS, $\hat{\sigma}^2_{j,1}$, presented in Table 5.7 behave in an unexpected way: White has a big coefficient in the prediction of all random effects in the same area, even considerably higher than the corresponding diagonal term, as it is the case for categories Mixed, Asian and Other. This behaviour is a further indication of issues with that set of estimates of the variance components.

5.4 Discussion

Four different SPREE-type estimators were applied in this application: SPREE, GSPREE, MSPREE and MMSPREE. An evaluation of the performance of all them would necessary involve an study of the point estimates, as well as their estimated MSE. Unfortunately, due to time constraints it was not possible to obtain MSE estimates, hence we will focus on the behaviour of the point estimates.

As the estimated coefficient of the GSPREE was very close to 1, i.e., the GSPREE was practically equivalent to the SPREE. We decided to drop one of them from all analysis in order to avoid repetitions. The matrix of coefficients of the MSPREE, rescaled in order to be interpreted in terms of the proportional interactions assumption, shown coefficients far from 1 in absolute value for the categories Mixed, Chinese and Other, indicating that those columns are likely to benefit from the extra flexibility provided by the additional parameters of the MSPREE. Indeed, particularly for category Chinese, the MSPREE seems to present less over-shrinking even though any improvement is modest.

On the other hand, MMSPREE estimates were obtained using two different sets of variance component estimates: $\hat{\sigma}^{1}_{j,2}$, obtained from the APS and Census 2011 data and $\hat{\sigma}^{2}_{j,2}$, obtained using data from censuses 2001 and 2011. This approach was adopted because the estimates obtained using APS data exhibited instability. Such behaviour could be attributed to the small sample sizes of the APS, therefore, a second set of estimates was calculated using only census information. Unfortunately, because those estimates correspond to a different reference period, it is not possible to assume unbiasedness unless the population variance components are stable across time.

The main improvements observed for the MMSPREE respect to the MSPREE in this application were: i) an estimate that is closer to the unbiased direct estimator for the cells with biggest sample size, as category White or the highest proportions of Asian and Black; and ii) a reduction in the over-shrinking for the remaining categories.

Chapter 6

Summary and Outline for Future Research

This document proposes the MSPREE, a SPREE-type estimator for small area compositions that generalizes the SPREE of Purcell and Kish (1980) and the GSPREE of Zhang and Chambers (2004). The behaviour of the estimator is illustrated via simulation in chapter 4, where it exhibited a better performance than the other two estimators above mentioned in terms of Bias and MSE, in all scenarios in consideration. Proposing a more flexible fixed effects estimator is an important contribution because the usual approach of reducing the bias of a synthetic estimator via the inclusion of random effects is not always feasible, and can lead to estimates with high MSE if the sample sizes are very small. Considering the current context of transformation of the population censuses being undertaken by many NSIs, the proposed estimator emerges as a simple an flexible alternative with big potential of applicability.

An extension of the MSPREE including cell specific random effects is also proposed, in order to reduce the bias of the MSPREE if the sample size allows for the prediction of random effects. The proposed mixed effects estimator, MMSPREE, goes one step ahead of the GSPREE with mixed effects, because it imposes sum-to-zero row and column constraints in the set of predicted random efects, in order to ensure a well defined model on the interactions scale. Furthermore, such constraints are imposed without increasing considerably the computational requirements of the estimation process.

An unbiased estimator for the variance components of the random effects for the MMSPREE is also proposed as part of this thesis. Such an estimator makes no assumptions regarding the structure of the variance covariance matrix of the sampling errors, and can be used to derive predicted random effects that satisfy the required constraints.

Potential areas of future development for the proposed estimators have already been identified, particularly from the experience of fitting real data that was discussed in Chapter 5. They can be summarized under the name of *working with proxy or sample compositions that are somehow different from the target composition*. Examples of this situation are:

- **Compositions which refer to a slightly different set of areas.** The interaction terms α_{aj} which are the pivot for building SPREE-type estimators are defined relatively to the set of areas and categories that are included in the composition. In that sense, for a given area, a different set of interaction terms $\boldsymbol{\alpha}_a$, would be obtained depending on the subset of areas that is taken into consideration. It would be of interest to understand the impact of small changes in the geographical classification, or the case where not all areas are included in the sample.

- **Availability of more than one source of proxy data.** So far, the proposed estimators have assumed the existence of only one source of proxy data. In practice, several sources, perhaps covering different subgroups of the population, may be available. How to better combine different sources in order to produce an estimate for small area compositions, is of practical relevance.

- **Availability of more than one source of survey data.** In analogous way, it is possible in some cases to obtain more than one survey estimate of the composition of interest. More commonly, series of sample estimates for different periods of time might be obtained from periodic surveys. Borrowing strength from several sources or across time may have a big impact in the performance of the estimators. An interesting extension of the proposed estimators could use some time-dependent structure, either on the fixed or the random part of the model, in order to take advantage of the additional information.

- **Use of auxiliary information not in the form of a composition.** SPREE type estimators seem somewhat restrictive when compared with other estimators for small area compositions as those introduced in section 1.4 because auxiliary information not in the form of a proxy composition cannot, in principle, be included in the estimation process. Developing an extension of the MSPREE to handle other type of information would increase considerably the range of applicability of this estimator.

Bibliography

Agresti, A. (2013). *Categorical data analysis*. John Wiley & Sons, 3rd edition.

Aitchison, J. (2003). *The statistical analysis of compositional data*. The Blackburn Press.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83:28–36.

Berg, E. J. and Fuller, W. A. (2014). Small area prediction of proportions with applications to the Canadian labour force survey. *Journal of Survey Statistics and Methodology*, 2(3):227–256.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society. Series B*, 25(1):220–233.

Bousfield, M. V. (1977). Intercensal estimation using a current sample and census data. *Review of Public Data Use*, 5(6):6–15.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Chambers, R. and Clark, R. (2012). *An introduction to model-based survey sampling with applications*. Oxford University Press.

Chambers, R. and Tzavidis, N. (2006). M-Quantile models for small area estimation. *Biometrika*, 93(2):255–268.

Chambers, R. L. and Fenney, G. (1977). Log-linear models for small area estimation. In *Proceedings of the Joint Conference of the CSIRO Division of Mathematics and Statistics and the Australian Region of the Biometrics Society*.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons, 2nd edition.

Datta, G., Fay, R., and Ghosh, M. (1991). Hierarchical and empirical multivariate Bayes analysis in small area estimation. In *Proceedings of the Bureau of the Census Annual Research Conference*, pages 63–79.

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. *The Annals of Mathematical Statistics*, 11(4):427–444.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association*, 76:341–353.

Dostál, L., Münnich, R., Gabler, S., and Ganninger, M. (2016). Frame correction modelling with applications to the German register-assisted census 2011. *Scandinavian Journal of Statistics*, pages n/a–n/a.

ESSNet (2012). Report on workpackage 6 ESSNet on small area estimation. Unpublished document.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277.

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.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. Springer-Verlag.

Harville, D. A. and Jeske, D. R. (1992). Mean squared error of estimation or prediction under general linear model. *Journal of the American Statistical Association*, 87:724–731.

Henderson, C. R. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics*, 21:309–310.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9(2):226–252.

Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55(1):179–188.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Jiang, J., Luan, Y., and Wang, Y.-G. (2007). Iterative estimating equations: Linear convergence and asymptotic properties. *The Annals of Statistics*, 35(5):2233–2260.

Kackar, R. N. and Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics. Series A*, 10(13):1249–1261.

Kish, L. (1965). *Survey sampling*. John Wiley & Sons.

Lahiri, P. and Rao, J. N. K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90(430):758–766.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

López-Vizcaíno, E., Lombardía, M. J., and Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical modelling*, 13(2):153–178.

López-Vizcaíno, E., Lombardía, M. J., and Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Society: Series A*, 178(3):535–565.

Luna, A., Zhang, L.-C., Whitworth, A., and Piller, K. (2015). Small area estimates of the population distribution by ethnic group in England: A proposal using structure preserving estimators. *Statistics in Transition New Series*, 16(4):585–602.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, 11(1):59–67.

Molina, I., Saei, A., and Lombardía, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society: Series A*, 170(4):975–1000.

Noble, A., Haslett, S., and Arnold, G. (2002). Small area estimation via generalized linear models. *Journal of Official Statistics*, 18(1):45–60.

Office for National Statistics (2011). Labour force survey user guide - Volume 1: Background and methodology. Retrieved on 1st October 2015 from http://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/labour-market-statistics/volume-1—2011.pdf.

Office for National Statistics (2012). 2001-2011 Census in England and Wales. Questionnaire comparability. Retrieved on 1st October 2015 from http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/comparability-over-time/2011-2001-census-questionnaire-comparability.pdf.

Pawitan, Y. (2013). *In all likelihood*. Oxford University Press.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2):317–337.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1):40–68.

Prasad, N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85(409):163–171.

Purcell, N. J. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48(1):3–18.

Rao, J. N. K. (1986). Synthetic estimators, SPREE and the best model based predictors. In *Proceedings of the Conference on Survey Research Methods in Agriculture*, pages 1–16. U.S. Department of Agriculture.

Rao, J. N. K. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons, 2nd edition.

Rao, J. N. K. and Scott, A. J. (1981). The analysis od categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76:221–230.

Saei, A. and Taylor, A. (2012). Labour force status estimates under a bivariate random components model. *Journal of the Indian Society of Agricultural Statistics*, 66(1):187–201.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer.

Scealy, J. (2010). Small area estimation using a multinomial logit mixed model with category specific random effects. Research paper, Australian Bureau of Statistics. Retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/cat/1351.0.55.029.

Stewart, J. (2008). *Calculus. Early transcendentals*. Thomson Brooks, 6th edition.

Tzavidis, N., Zhang, L., Luna-Hernandez, A., Schmid, T., and Rojas-Perilla, N. (2016). From start to finish: A framework for the production of small area official statistics. *Working document*.

Valliant, R., Dorfman, A., and Royall, R. (2000). *Finite population sampling and inference: A prediction approach*. Wiley.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447.

Wolter, K. M. (2007). *Introduction to variance estimation*. Springer Science & Business Media, 2nd edition.

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.