*Journal of Survey Statistics and Methodology (2024) **12**, 1531–1566*

AREA-LEVEL MODEL-BASED SMALL AREA ESTIMATION OF DIVERGENCE INDEXES IN THE SPANISH LABOUR FORCE SURVEY

ESTEBAN CABELLO *
DOMINGO MORALES
AGUSTÍN PÉREZ

This article develops model-based predictors for area-level proportions of employed men and women by occupation sectors and for entropies and divergence indexes (DIs) within and between sex groups. Since the direct estimators of the proportions add up to one in the occupational sections, they are compositions that can be imprecise if the sample sizes are small. We fit a multivariate Fay–Herriot model to logratio transformations of the direct estimators of the proportions. Small area estimators of the proportions, entropies, and DIs are derived from the fitted model and the corresponding mean squared errors are estimated by parametric bootstrap. Several simulation experiments designed to analyze the behavior of the introduced model-based predictors are carried out. We give an application to Spanish Labour Force Survey data from 2022. The target is to investigate the state of sex occupational entropies and divergences in Spanish provinces.

KEY WORDS: Bootstrap; Compositional data; Divergence index; Labour Force Survey; Multivariate Fay–Herriot model; Occupation sectors; Small area estimation.

ESTEBAN CABELLO is a Research Fellow at the Center of Operations Research, Miguel Hernández University of Elche, Edificio Torretamarit—Avda. de la Universidad s/n, Elche (Alicante) 03202, Spain. DOMINGO MORALES is Professor at the Center of Operations Research, Miguel Hernández University of Elche, Edificio Torretamarit—Avda. de la Universidad s/n, Elche (Alicante) 03202, Spain. AGUSTÍN PÉREZ is Professor at the Departament of Economic and Financial Studies, Miguel Hernández University of Elche, Edificio La Galia—Avda. de la Universidad s/n, Elche (Alicante) 03202, Spain.

We are grateful to the referees, the Associate Editor, and the Editor-in-Chief for valuable comments and suggestions.This work was supported by the Ministry of Science and Innovation and the State Research Agency of the Spanish Government through the European Regional Development Fund (PID2022-136878NB-I00) and by the Conselleria d'Innovació, Universitats, Ciéncia i Societat Digital of the Generalitat Valenciana (Prometeo/2021/063).

*Address correspondence to Esteban Cabello, Center of Operations Research, Miguel Hernández University of Elche, Edificio Torretamarit—Avda. de la Universidad s/n, Elche (Alicante) 03202, Spain; E-mail: ecabello@umh.es.

https://doi.org/10.1093/jssam/smae023 Advance access publication 18 June 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Statement of Significance

We analyze data from the Spanish Labour Force Survey 2022 using a multivariate linear mixed model to investigate the state of sex occupational divergences in Spanish provinces. It is proposed a partitioned version of the multivariate Fay–Herriot model, with different parameterizations for men and women, and the ability to model logratio-transformed data while capturing sample correlations for each group separately. Under the model are derived the domain-level predictors of proportions, counts, entropies, and divergence indexes. This approach enables mapping entropies and divergences at different levels of aggregation, providing valuable insights for policymakers in deciding where to implement specific equality policies.

1. INTRODUCTION

The Kullback–Leibler divergence, or divergence index (DI), was introduced by Kullback and Leibler (1951) and can be used to measure how the distribution by occupation sectors of a first group of people is different from a second that serves as a reference. The DI is widely used in sociological studies (Belov and Armstrong 2010; Vilhena et al. 2014; Symeonaki 2022) because it is a measure of distance between two groups of individuals classified by sex, race, origin, religion, or culture, among others. The DI has been also applied to other fields such as bioinformatics (Volkau et al. 2006; Lin et al. 2007; Cabella et al. 2008), biology (Li and Wang 2008), and linguistics (Froud et al. 2010) among others. In addition to occupation sectors, the distributions in the two groups of interest might be defined by other categorical variables, like residential areas or education and income levels. We explore the use of the DI to measure occupational divergence by sex, where the group variable is sex and the distribution or classification variable is occupation sector.

In this context, the DI measures the distance between the distributions of men and women in different sectors of labor occupation. To do so, it compares the proportion of men and women employed in each occupation sector and provides a numerical value greater than or equal to zero. On the one side, the DI is zero if the two occupational distributions area equal. On the opposite side, the DI will take larger and larger values as the proportions of the categories in both groups differ from each other. Cover (1999) describes the basic properties of the DI. Pardo (2018) reviews the statistical inference methodologies based on divergence measures, including the DI.

The DI is also called relative (Shannon) entropy as it compares the uncertainties of the possible outcomes of the two compared distributions. Some authors have explored the problem of estimating divergences and entropies

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

and have studied the properties of the estimators in several contexts. Among some interesting contributions, Morales et al. (1994) studied the asymptotic behavior of a collection of divergence estimators under stratified random sampling. Menéndez et al. (1995a) gave an unified study for divergence statistics under mutinomial distributions, Menéndez et al. (1995b) extended their study to divergence-based estimation and testing of statistical models of classification. Pardo et al. (1997) considered the problem of estimating entropy measures. Morales et al. (2000) gave divergence test statistics in directed families of exponential experiments. These contributions assume that the available information is highly reliable. This can occur when the data come from administrative records or from surveys with large sample sizes. In such cases, the calculation of divergences does not present notable problems. However, the proportions of the categories of the classification variable are in practice unknown and must be estimated from smaller samples. It is at this point where the research problem that motivates the investigation of this work appears.

This article estimates DIs and entropies by areas using data from a labor force survey. For this purpose, we can calculate direct estimators of the proportions of employed men and women in each occupation sector which we directly substitute into the DI formula. If the sample sizes are small, the direct estimators, calculated using only the values of the target variable in the domain and the corresponding elevation factors, are not sufficiently precise and do not give good estimates of DIs or entropies. In such cases, it is preferable to compute model-based predictors to reduce variability by introducing auxiliary information into the inferential process. We can address the problem of domains with small sample sizes using small area estimation (SAE) techniques. That is, we can fit statistical models to data and obtain predictors based on them. This is a way of providing additional information from other domains, auxiliary variables, and dependency and correlation structures in hierarchical, spatial, or temporal data. See Molina and Rao (2010), Rao and Molina (2015), Pratesi (2016), and Morales et al. (2021) for introductory monographs.

Among the most used models in SAE for area-level data, the Fay–Herriot (FH) model introduced in Fay and Herriot (1979) stands out. This is a mixed linear model that incorporates both the sample weights through direct estimators and auxiliary information from explanatory variables. It is a model that is widely applied in practice due to its versatility and ease of use. The scientific literature presents generalizations of the FH model to data with temporal or hierarchical structure, highlighting the contributions of Pfeffermann and Burck (1990), Rao and Yu (1994), Ghosh et al. (1996), Datta et al. (1999, 2002), You and Rao (2000), Singh et al. (2005), Esteban et al. (2012), Marhuenda et al. (2014, 2016), Morales et al. (2015), and Burgard et al. (2020). The extension of the FH model to multivariate data goes one step further. Multivariate models allow addressing the estimation of multivariable indicators, considerably expanding the ability to model new correlation

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

structures. The multivariate Fay–Herriot (MFH) model and their variants have been discussed by Huang and Bell (2004), Porter et al. (2015), González-Manteiga et al. (2008), Arima et al. (2017), Benavent and Morales (2016, 2021), Ubaidillah et al. (2019), Esteban et al. (2020), Burgard et al. (2021, 2022), and Guha and Chandra (2024), among others.

In public statistics, many socioeconomic indicators are functions of the counts or proportions of units in the categories of a classification variable. That is, they are compositions that add up to one or to a known integer. The books by Aitchison (1982) and Pawlowsky-Glahn and Buccianti (2011), or the review paper by Egozcue and Pawlowsky-Glahn (2019), are basic references on compositional data analysis. The estimation of compositional parameters requires the use of multivariate models and procedures to estimate counts and proportions. For this purpose, there are contingency table methodologies that model cell counts using categorical auxiliary variables or procedures based on regression models that admit continuous auxiliary variables. Within the first approach, the contributions of Zhang and Chambers (2004) and Berg and Fuller (2012) stand out. The second approach mainly relies on Poisson, Binomial, or Multinomial regression mixed models. Based on this type of models, predictors of counts and proportions were introduced by Molina et al. (2007), Ferrante et al. (2010), López-Vizcaíno et al. (2013, 2015), Franco and Bell (2015), Chambers et al. (2016), Souza and Moura (2016), Fabrizi et al. (2016), Boubeta et al. (2016, 2017), Burgard et al. (2021, 2022), Erciulescu and Opsomer (2022), and Krause et al. (2022), among others. All of the above (nonexhaustive) collection of relevant papers utilize area-level SAE methodology for predicting domain proportions and totals. This is a partial step that this article also deals with.

The proportions of employed men, or women, in each occupation sector add up to one. The two vectors containing such proportions are compositions. The same is true for the two vectors containing the corresponding direct estimators of proportions. For this reason, this article proposes to fit an MFH model to logratio transformations of the compositions of direct estimators. In this way, more information is added to the construction of predictors of cell proportions, entropies, and divergence indices. The resulting model is called a compositional MFH model and jointly considers aggregated data across geographic areas, sex, and occupational categories. The new model generates model-based predictions of proportions that respect the constraints of being positive and adding one.

We give an application to data from the Spanish Labour Force Survey (SLFS) of 2022,[1] to study the entropy within and the divergence between

1. The data (at unit level) are publicly available on the website: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=resultados&idp=1254735976595#!tabs-1254736030639

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

the distributions of men and women in the different sectors of labor occupation in the Spanish provinces. This motivates research into statistical methodologies for mapping entropies and DIs at different levels of aggregation, which can be of great help to policymakers to decide where to implement specific equality policies. We present statistical methodology that is new in two main aspects: (i) the use of a partitioned version of the MFH models, with different parameterizations for men and women, and with the ability to model the logratio-transformed data and capture sample correlations for each group separately, and (ii) the derivation of domain-level predictors of proportions, counts, entropies, and DIs based on the partitioned MFH model fitted to the transformed data. In particular, the two last indicators are nonlinear and multivariate.

The paper is organized as follows. Section 2 introduces the probabilistic framework, the DI, the entropy, the data, and the SAE problem. Section 3 contains the new partitioned MFH model, the residual maximum likelihood (REML) estimators of the model parameters, the predictors of proportions of employed men and women by occupation sectors, the predictors of entropies and DIs, and the mean squared error (MSE) bootstrap estimators. Section 4 includes some simulation experiments to investigate the performance of the DI and entropy predictors and the effect of the number of bootstrap replicates in the accuracy of the MSE estimators. Section 5 deals with the application to real data. Section 6 provides some relevant conclusions, limitations, and future research. The supplementary material online contains five appendices. Appendix A in the supplementary data online gives the Fisher-scoring algorithm to calculate the REML estimators of the model parameters. Appendix B in the supplementary data online describes the additive, the centered, and the isometric logratio transformations of compositions. Appendix C in the supplementary data online gives additional results on the fit to the data and the diagnosis of the models resulting from applying the logratio transformations. Appendix D in the supplementary data online details the simulation steps of the two simulation experiments of section 4. Appendix E in the supplementary data online contains tables with additional information on the auxiliary variables and with estimates derived from applying different logratio transformations. Appendix F in the supplementary data online provides bar charts for direct estimates of the variances and covariances of the sampling error vector in the selected model.

2. THE PROBABILISTIC FRAMEWORK AND DATA

This article presents and applies the SAE methodology to estimate the divergence between the distributions of employed men and women in the

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

occupation sectors by Spanish provinces, and to estimate the entropies of both distributions. For this reason, this section establishes the probabilistic framework and describes the available data.

First, we provide the mathematical formulation needed to define the DIs between sex distributions and the entropies within sex distributions for every province. Consider the variables sex, province and occupation sector, with values $k = 1, 2$, $d = 1, \ldots, D$ and $r = 1, \ldots, R$, respectively, where D is the number of provinces and R is the number of occupation sectors. The finite population of interest (employed people) is divided by sex, $U = U_1 \cup U_2$, $U_1 \cap U_2 = \emptyset$, by province, $U_k = \cup_{d=1}^{D} U_{kd}$, $U_{kd_1} \cap U_{kd_2} = \emptyset, d_1 \neq d_2, k = 1, 2$, and by sector, $U_{kd} = \cup_{r=1}^{R} U_{kdr}, U_{kdr_1} \cap U_{kdr_2} = \emptyset, \quad r_1 \neq r_2, \quad k = 1, 2$, $d = 1, \ldots, D$. Let N, N_k, N_{kd} and N_{kdr} be the corresponding population sizes. The partition of U by occupation sectors is $U = \cup_{r=1}^{R} \Omega_r$, $\Omega_{r_1} \cap \Omega_{r_2} = \emptyset$, $r_1 \neq r_2, \quad r_1, r_2 = 1, \ldots, R$. For $k = 1, 2, \quad d = 1, \ldots, D, \quad r = 1, \ldots, R$, $j = 1, \ldots, N_{kd}$, the definition of the indicator variables, $z_{kdj,r}$, of occupation sector r in the sex-province U_{kd}, is $z_{kdj,r} = 1$ if $u_{kdj} \in \Omega_r$ and $z_{kdj,r} = 0$ otherwise, where u_{kdj} is the j^{th} individual of U_{kd}. Moreover, the definition of the indicator variables, $z_{kj,d}$, of province d in the sex U_k, is $z_{kj,d} = 1$ if $u_{kj} \in U_{kd}$ and $z_{kj,d} = 0$ otherwise, where u_{kj} is the j^{th} individual of U_k. Let $Z_{kdr} = \sum_{j \in U_{kd}} z_{kdj,r}$ and $\bar{Z}_{kdr} = Z_{kdr}/N_{kd}$ be the total and the proportion of employed people of sex k and province d that works in occupation sector r, respectively. It holds that $\sum_{r=1}^{R} Z_{kdr} = N_{kd}$ and $\sum_{r=1}^{R} \bar{Z}_{kdr} = 1$. We further define the vectors $Z_{kd} = (Z_{kd1}, \ldots, Z_{kdR-1})'$ and $\bar{Z}_{kd} = (\bar{Z}_{kd1} \ldots, \bar{Z}_{kdR-1})'$, $k = 1, 2$, $d = 1, \ldots, D$, which are compositions.

Second, we introduce the variable *occupation sector* (OC), which is categorical and allows introducing the probability vectors (or distributions) to be compared as arguments of the DIs and the entropies. The OC has $R = 7$ categories that cover a wide range of occupations across the national territory. They are

OC1 Directors and managers. Management of companies and public administrations,

OC2 Scientific and intellectual technicians and professionals,

OC3 Military occupations. Technicians and support professionals,

OC4 Accounting, administrative, and other office employees,

OC5 Workers in catering services, protection, and trade vendors,

OC6 Unskilled workers and skilled workers in the agricultural, livestock, forestry, and fishing sectors, and

OC7 Plant and machinery operators and assemblers. Artisans and skilled workers in the manufacturing, construction, and mining industries.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Third, we define the domain divergences and entropies. The DI of province d, between the OC distributions of employed men and women, is $K_d = K(\bar{Z}_{1d}, \bar{Z}_{2d}) = K(Z_{1d}, Z_{2d})$, where

$$K_d = \sum_{r=1}^{R} \bar{Z}_{1dr} \log \frac{\bar{Z}_{1dr}}{\bar{Z}_{2dr}} = \frac{1}{N_{1d}} \sum_{r=1}^{R} Z_{1dr} \log \frac{N_{2d} Z_{1dr}}{N_{1d} Z_{2dr}}, \quad d = 1, \ldots, D.$$

The Shannon entropy of the OC distribution for province d and sex k is

$$H_{kd} = -\sum_{r=1}^{R} \bar{Z}_{kdr} \log \bar{Z}_{kdr}, \quad k = 1, 2, d = 1, \ldots, D.$$

To estimate K_d and H_{kd}, $k = 1, 2$, $d = 1, \ldots, D$, we use data from the SLFS, which provides quarterly and annual data on the labour market in Spain. It is a two-stage survey with stratification in the first-stage units. The first-stage units are census sections, which are grouped by strata according to the size of the municipality to which they belong. The second stage units are inhabited family dwellings where all members aged 16 and over are interviewed, since this is the limit for compulsory schooling and the minimum legal age to work in Spain. The Spanish National Statistics Institute (INE) calculates the sampling design weights, makes corrections for nonresponses, and applies a calibration procedure to adjust the sample distribution to the population distribution of persons by provinces and by age groups and sex provided by the demographic projections unit of INE. For any sampled individual j of sex k and domain d, w_{kdj} denotes its calibrated sampling weight. INE calculates these weights (elevation factors) and includes them in the SLFS data file. We focus our attention on the groups of employed men ($k = 1$) and employed women ($k = 2$). For this analysis we have not included the autonomous cities of Ceuta and Melilla, so we have a total of $D = 50$ provinces. Finally, we use data from the fourth quarter of 2022 (SLFS2022.4), as they are the most recently published at the time of writing the first draft of this manuscript.

Let s_{kd}, $s_k = \cup_{d=1}^{D} s_{kd}$ and $s = \cup_{k=1}^{2} s_{kd}$ be subsets (samples) of U_{kd}, U_k and U, with sample sizes n_{kd}, n_k and n, respectively. The direct estimators of Z_{kdr}, N_{kd} and $\bar{Z}_{kdr}$ are

$$\hat{Z}_{kdr}^{dir} = \sum_{j \in s_{kd}} w_{kdj} z_{kdj,r}, \quad \hat{N}_{kd}^{dir} = \sum_{j \in s_{kd}} w_{kdj}, \quad \hat{\bar{Z}}_{kdr}^{dir} = \frac{\hat{Z}_{kdr}^{dir}}{\hat{N}_{kd}^{dir}},$$

where w_{kdj} is the sample weight of individual $u_{kdj} \in U_{kd}$. The direct estimator of K_d is

$$\hat{K}_d^{dir} = \sum_{r=1}^{R} \hat{\bar{Z}}_{1dr}^{dir} \log \frac{\hat{\bar{Z}}_{1dr}^{dir}}{\hat{\bar{Z}}_{2dr}^{dir}} = \frac{1}{\hat{N}_{1d}^{dir}} \sum_{r=1}^{R} \hat{Z}_{1dr}^{dir} \log \frac{\hat{N}_{2d}^{dir} \hat{Z}_{1dr}^{dir}}{\hat{N}_{1d}^{dir} \hat{Z}_{2dr}^{dir}}, \quad d = 1, \ldots, D.$$

The direct estimator of H_{kd} is

$$\hat{H}_{kd}^{dir} = -\sum_{r=1}^{R} \hat{\bar{Z}}_{kdr}^{dir} \log \hat{\bar{Z}}_{kdr}^{dir}, \quad k = 1, 2, d = 1, \ldots, D.$$

We make a first analysis of the distributions of the sample sizes for men and women. This gives us an idea of the precision that the direct estimators of the entropies and divergence indices can have. Let n_{kd}, n_{kdr}, and $\hat{N}_{kd}^{dir}$, $\hat{N}_{kdr}^{dir}$ be the sample sizes and the estimated population sizes of U_{kd} and U_{kdr}. Let $n_1 = \sum_{d=1}^{D} n_{1d} = 26922$, and $n_2 = \sum_{d=1}^{D} n_{2d} = 24428$ be the sample totals of employed men and women, respectively. Dividing by D, we have that $\frac{n_1}{D} \approx 538$ and $\frac{n_2}{D} \approx 488$ are the average employed men and women by province. Again dividing by R, we have that $\frac{n_1}{DR} \approx 77$ and $\frac{n_2}{DR} \approx 70$ are the average number of men and women employed by occupation sector and province. However, the sample sizes by province and by sector, n_{kd} and n_{kdr}, are not uniformly distributed, and neither are the corresponding sampling fractions in percent, $f_{kd} = 100\frac{n_{kd}}{\hat{N}_{kd}^{dir}}$, $f_{kdr} = 100\frac{n_{kdr}}{\hat{N}_{kdr}^{dir}}$, $k = 1, 2$, $d = 1, \ldots, 50$, $r = 1, \ldots, 7$. Table 1 presents the quantiles of the sets of sample sizes (to one decimal place) and sampling fractions, for men and women.

We observe that the averages of sample sizes, n_{kd} and n_{kdr}, for both sexes are between the $q_{0.6}$ and $q_{0.7}$ quantiles. This indicates that the distributions of these sample sizes are right-skewed. We also note that 50% of the provinces have $n_{1d} \leq 438.5$ for men and $n_{2d} \leq 379.5$ for women. At the province-occupation level, 50% of the sample sizes are such that $n_{1dr} \leq 54$ for men and $n_{2dr} \leq 51$ for women. As $f_{1d} \leq 1.22$ for men and $f_{2d} \leq 1.24$ for women at province level and $f_{1dr} \leq 1.38$ for men and $f_{2dr} \leq 1.66$ for women at province-occupation level, the representativeness of the sample sizes is small. Consequently, this is a SAE problem and direct estimators, such as the Hájek estimator, are not accurate enough. Table 2 shows the sample sizes $n_{k.r}$ and the direct estimates, $\hat{\bar{Z}}_{k.r}^{dir}$, of the proportions of employed men and women in each occupation sector at Spanish level, calculated from data of SLFS2022.4, that is,

$$n_{k.r} = \sum_{d=1}^{D} n_{kdr}, \quad \hat{Z}_{k.r}^{dir} = \sum_{d=1}^{D} \sum_{j \in s_{kd}} w_{kdj} z_{kdj,r}, \quad \hat{N}_{k.}^{dir} = \sum_{d=1}^{D} \sum_{j \in s_{kd}} w_{kdj}, \quad \hat{\bar{Z}}_{k.r}^{dir} = \frac{\hat{Z}_{k.r}^{dir}}{\hat{N}_{k.}^{dir}}.$$

Table 2 shows that the direct estimates of the proportions of men and women employed by sector of occupation are far from the uniform distribution and differ from each other.

Figure 1 (left) presents boxplots of direct estimates $\hat{\bar{Z}}_{kdr}^{dir}$ of the proportions of employed men and women in each occupation sector at the province level.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Table 1. Quantiles of Sample Sizes and Sampling Fractions in Percent for Employed Men and Women

q	q_0	$q_{0.1}$	$q_{0.2}$	$q_{0.3}$	$q_{0.4}$	$q_{0.5}$	$q_{0.6}$	$q_{0.7}$	$q_{0.8}$	$q_{0.9}$	q_1
n_{1d}	194.0	257.9	275.6	337.9	385.0	438.5	483.8	593.4	724.2	965.7	1,916.0
f_{1d}	0.09	0.17	0.20	0.23	0.28	0.36	0.40	0.46	0.54	0.62	1.22
n_{1dr}	3.0	17.9	25.0	35.0	47.6	54.0	64.8	90.3	118.2	153.0	424.0
f_{1dr}	0.09	0.17	0.20	0.23	0.29	0.36	0.41	0.47	0.55	0.66	1.38
n_{2d}	175.0	219.1	255.6	297.5	343.4	379.5	450.2	526.7	702.0	856.8	1,834.0
f_{2d}	0.09	0.18	0.22	0.25	0.31	0.37	0.44	0.49	0.57	0.67	1.24
n_{2dr}	1.0	10.0	17.8	27.7	38.0	51.0	66.0	80.0	105.2	151.3	522.0
f_{2dr}	0.09	0.19	0.22	0.25	0.31	0.37	0.44	0.50	0.58	0.72	1.66

Table 2. Sample Sizes and Direct Estimates of the Proportions of Employed Men and Women by Occupation Sector in SLFS2022.4

	OC	OC1	OC2	OC3	OC4	OC5	OC6	OC7
M	$n_{1.r}$	1,358	3,968	3,772	1,600	3,949	3,484	8,791
	$\hat{\bar{Z}}_{1.r}^{dir}$	0.051	0.148	0.143	0.062	0.154	0.128	0.311
W	$n_{2.r}$	703	5,887	2,327	3,728	6,833	3,780	1,170
	$\hat{\bar{Z}}_{2.r}^{dir}$	0.031	0.238	0.100	0.152	0.275	0.157	0.044

There are statistically significant differences between the second, fourth, fifth, and seventh occupation sectors. Figure 1 (center) plots the estimates of the coefficient of variation (CV) of the direct estimators of the proportions of employed men (dots) and women (asterisks) in each occupation sector at a province level, sorted by sample size. Figure 1 (right) plots the same set of dots but colored by occupation sector. We observe that as sample size increases, the CV decreases for both men and women and also by occupation sector. We underscore that the CVs that are greater than 50% are from Ávila ($d = 5$) and Huelva ($d = 21$) for employed men in sector 1, and from Badajoz ($d = 6$), Córdoba ($d = 14$), Huelva ($d = 21$), Lleida ($d = 25$), Palencia ($d = 34$), Segovia ($d = 40$), and Zamora ($d = 49$) for employed women in sector 1. As an extreme estimate of a CV, we point to the employed women of Palencia in sector 1, where the estimated proportion is 0.004 with an estimated standard deviation close to 0.0044, yielding a CV close to 1. That is reasonable if we observe that there is only one woman in that sector and province.

To overcome the lack of precision of the direct estimators $\hat{\bar{Z}}_{kdr}^{dir}$, we may fit a partitioned MFH model to the compositions $\hat{\bar{Z}}_{kd}^{dir} = (\hat{\bar{Z}}_{kd1}^{dir}, \ldots, \hat{\bar{Z}}_{kdR-1}^{dir})'$, $k = 1, 2$, $d = 1, \ldots, D$. However, such partitioned MFH model may not

Figure 1. Direct Estimates of Proportions of Employed Men and Women by Occupation Sector at Province Level (Left). Estimated CVs colored by sex (center) and OC (right).

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

produce predictions of $\bar{Z}_{kd}$ fulfilling the condition $\sum_{r=1}^{R} \bar{Z}_{kdr} = 1$. Therefore, we apply the centered log ratio (clr) transformation of R-compositions onto $\mathbb{R}^{R-1}$. We use the simpler notation $z_{kd} \triangleq \hat{\bar{Z}}_{kd}^{dir}$, $z_{kd} = (z_{kd1}, \ldots, z_{kdR-1})'$ for the direct estimators of compositions $\bar{Z}_{kd}$, and $\sigma_{z,k_1r_1,k_2r_2} = \mathrm{cov}_\pi(z_{k_1r_1}, z_{k_2r_2})$, $k_1, k_2 = 1, 2$, $r_1, r_2 = 1, \ldots, R-1$, for the design-based variances and covariances. In matrix form, we have

$$\mathrm{var}_\pi(z_{kd}) = (\sigma_{z,k_1r_1,k_2r_2})_{k_1,k_2=1,2;r_1,r_2=1,\ldots,R-1}.$$

From Särndal et al. (2003, p. 43, 185, and 391) as well as remark 2.3 in Morales et al. (2021), a direct estimator of the covariance of z_{kdr_1} and z_{kdr_2} is

$$\hat{\mathrm{cov}}_\pi(z_{kdr_1}, z_{kdr_2}) = \frac{1}{\hat{N}_{kd}^{dir}} \sum_{j \in s_{kd}} w_{kdj}(w_{kdj} - 1)(z_{kdj,r_1} - z_{kdr_1})(z_{kdj,r_2} - z_{kdr_2}). \quad (1)$$

This article uses direct estimates of the variance matrices $(R-1) \times (R-1)$, with elements defined in (1), without any smoothing. Appendix F in the supplementary data online provides bar charts of the relative frequencies for direct estimates of the variances and covariances, which show the absence of outliers and denote acceptably smooth behavior. There are no zeros in the variance estimates nor extremely high values in the variance or covariance estimates. A more sophisticated alternative approach is to adapt to logratio transformations and apply the projection-based smoothing procedure proposed by Erciulescu and Opsomer (2022). Depending on the results of the residual analysis and the diagnosis of the model, there will be cases in which smoothing will be appropriate.

We assume that $z_{kdr} > 0$, $k = 1, 2$, $d = 1, \ldots, D$, $r = 1, \ldots, R-1$, and we note that $z_{kd1} + \cdots + z_{kdR} = 1$. For $d = 1, \ldots, D$, let $y_{kd} = (y_{kd1}, \ldots, y_{kdR-1})' \in \mathbb{R}^{R-1}$ be the clr transformation of z_{kd}, that is, $y_{kd} = h(z_{kd}) = (h_1(z_{kd}), \ldots, h_{R-1}(z_{kd}))'$ with

$$y_{kdr} = h_r(z_{kd}) = \log\ z_{kdr} - \frac{1}{R}\sum_{r=1}^{R-1} \log\ z_{kdr} - \frac{1}{R}\log(1 - z_{kd1} - \ldots - z_{kdR-1}),$$
$$r = 1, \ldots, R-1.$$

The first partial derivatives of the clr transformation are

$$H_{rr}(z_{kd}) = \frac{\partial h_r(z_{kd})}{\partial z_{kdr}} = \frac{1}{z_{kdr}} - \frac{1}{q}\left(\frac{1}{z_{kdr}} - \frac{1}{z_{kdR}}\right),$$
$$H_{rt}(z_{kd}) = \frac{\partial h_r(z_{kd})}{\partial z_{kdt}} = -\frac{1}{q}\left(\frac{1}{z_{kdt}} - \frac{1}{z_{kdR}}\right), r \neq t, r, t = 1, \ldots, R-1.$$

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Let 1_a and 0_a be the $a \times 1$ vectors with all the components equal to one and to zero, respectively. In matrix form, we have

$$H(z_{kd}) = (H_{ij}(z_{kd}))_{i,j=1,\ldots,R-1} = \underset{1\le r\le R-1}{\text{diag}}\,(z_{kd}^{-1}) - \frac{1}{R}1_{R-1}(z_{kd1}^{-1} - z_{kdR}^{-1}, \ldots, z_{R-1}^{-1} - z_R^{-1}).$$

A Taylor series expansion of $h(z_{kd})$ around a given z_0 yields to

$$y_{kd} = h(z_{kd}) \approx h(z_0) + H(z_0)(z_{kd} - z_0). \tag{2}$$

If we take $z_0 = R^{-1}1_{R-1}$, then $h(z_0) = 0_{R-1}$, $H_0 = H(z_0) = RI_{R-1}$, where I_a is the $a \times a$ identity matrix. Alternatively, we can take $z_0 = \frac{1}{2D}\sum_{d=1}^{D}(z_{1d} + z_{2d})$ or we can select z_0 depending on d and close to z_{kd} for a better Taylor expansion approximation. From (2), we get the approximated covariance matrix

$$\text{var}_\pi(y_{kd}) \approx H_0 \text{var}_\pi(z_{kd}) H_0'. \tag{3}$$

For estimating $\bar{Z}_{kdr}$, $k = 1, 2$, $d = 1, \ldots, D$, $r = 1, \ldots, R-1$, this article proposes to fit a partitioned MFH model to $y_{kd} \overset{ind}{\sim} N_{R-1}(\mu_{kd}, V_{kd})$, where μ_{kd} is a mean vector depending on unknown regression parameters and auxiliary variables and V_{kd} is a covariance depending of some unknown parameters. Section 3 gives a flexible MFH model for y_{kd}, $k = 1, 2$, $d = 1, \ldots, D$, which allows positive and negative covariances.

Appendix B in the supplementary data online provides an expanded description of the clr transformation, as well as additive and isometric logratio transformations

3. MODEL-BASED INFERENCE

3.1 The PMFH Model

Let $y_{kd} = (y_{kd1}, \ldots, y_{kdR-1})'$ be the clr transformation of $z_{kd} = (z_{kd1}, \ldots, z_{kdR-1})'$. Let $\mu_{kd} = E_\pi(y_{kd}) = (\mu_{kd1}, \ldots, \mu_{kdR-1})'$ be the vector of design-based expectations of y_{kd}. The partitioned multivariate Fay–Herriot (PMFH) model is defined in two stages. The sampling model is

$$y_{kd} = \mu_{kd} + e_{kd}, \quad k = 1, 2, d = 1, \ldots, D, \tag{4}$$

where the vectors of random errors $e_{kd} \sim N(0, V_{ekd})$ are independent and the $(R-1) \times (R-1)$ covariance matrices V_{ekd} are known. Following (3), we take $V_{ekd} = H_0 \hat{\text{var}}_\pi(z_{kd}) H_0'$, where $\hat{\text{var}}_\pi(z_{kd})$ is a direct estimator of the design-based variance $\text{var}_\pi(z_{kd})$. We assume that μ_{kdr} is linearly related to the row vector of explanatory variables $x_{kdr} = (x_{kdr1}, \ldots, x_{kdrm_{kr}})$, and we define

$X_{kd} = \text{diag}(x_{kd1}, \ldots, x_{kdR-1})_{(R-1)\times m_k}$, where $m_k = \sum_{r=1}^{R-1} m_{kr}$. Let β_{kr} be a column vector of size m_{kr} containing the regression parameters for μ_{kdr} and let $\beta_k = (\beta_{k1}^{'}, \ldots, \beta_{kr}^{'})_{m_k\times 1}^{'}$. The linking model is

$$\mu_{kd} = X_{kd}\beta_k + u_{kd}, \quad k = 1, 2, d = 1, \ldots, D, \tag{5}$$

where the vectors of random effects u_{kd}s are independent of the vectors e_{kd}s, and

$$u_{kd} \sim N(0, V_{ukd}), V_{ukd} = \underset{1\le r\le R-1}{\text{diag}}(\sigma_{kr}^2), k = 1, 2, d = 1, \ldots, D.$$

Let $\theta = \text{col}_{1\le k\le 2}(\text{col}_{1\le r\le R-1}(\sigma_{kr}^2))$ be the column vector of variance parameters. Let I_n be the $n \times n$ identity matrix. At the level of U_k, we define the following vectors and matrices:

$$y_k = \underset{1\le d\le D}{\text{col}}(y_{kd}), u_k = \underset{1\le d\le D}{\text{col}}(u_{kd}), e_k = \underset{1\le d\le D}{\text{col}}(e_{kd}),$$
$$X_k = \underset{1\le d\le D}{\text{col}}(X_{kd}), Z_k = I_{D(R-1)}, V_{uk} = \underset{1\le d\le D}{\text{diag}}(V_{ukd}), V_{ek} = \underset{1\le d\le D}{\text{diag}}(V_{ekd}),$$

where col is the matrix operators stacking by columns. At the level of U, we define $m = m_1 + m_2$ and the following vectors and matrices

$$y = \underset{1\le k\le 2}{\text{col}}(y_k), u = \underset{1\le k\le 2}{\text{col}}(u_k), e = \underset{1\le k\le 2}{\text{col}}(e_k), \beta = \underset{1\le k\le 2}{\text{col}}(\beta_k),$$
$$X = \underset{1\le k\le 2}{\text{diag}}(X_k), Z = I_{2D(R-1)}, V_u = \underset{1\le k\le 2}{\text{diag}}(V_{uk}), V_e = \underset{1\le k\le 2}{\text{diag}}(V_{ek}).$$

In matrix form, the PMFH model [(4) and (5)] is

$$y = X\beta + Zu + e, \tag{6}$$

where e and u are independent with distributions $e \sim N(0, V_e)$ and $u \sim N(0, V_u)$.

The components β_{kr} of the vector β in (6) depend on k and r, but not on d. By changing diag to col in the definition of X_{kd}, or in the definition of X, we can obtain regression coefficients depending only on k, only on r, or not depending on k and r. That is, vector β may or may not be partitioned by groups $k = 1, 2$. This modeling flexibility is a difference from the standard MFH model. Additionally, note that V_u captures the variability of the distinct K-groups. Since these variants can be written in the form of (6), we have developed R software for them, but we do not treat them as different models. Appendix A in the supplementary data online develops the Fisher-scoring algorithm to calculate the REML estimators of the model parameters.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Under model (6), it holds that

$$E(y) = X\beta \quad \text{and}$$
$$V_y = \text{var}(y) = Z' V_u Z + V_e = V_u + V_e \triangleq \underset{1\le k\le 2}{\text{diag}}\,(\underset{1\le d\le D}{\text{diag}}\,(V_{ykd})),$$

where $V_{ykd} = V_{ukd} + V_{ekd}$, $k = 1, 2$, $d = 1, \ldots, D$.

Further, the best linear unbiased estimator (BLUE) of β and the best linear unbiased predictors (BLUP) of u and μ are

$$\begin{aligned}\hat{\beta}^{blue} &= (X' V_y^{-1} X)^{-1} X' V_y^{-1} y, \hat{u}^{blup} = V_u Z' V_y^{-1}(y - X\hat{\beta}^{blue}), \hat{\mu}^{blup} \\ &= X\hat{\beta}^{blue} + Z\hat{u}^{blup}. \end{aligned} \tag{7}$$

The output of REML Fisher-scoring algorithm, $\hat{\theta}$, is the REML estimator of θ. By plugging $\hat{\theta}$ in V_u, we get $\hat{V}_u = V_u(\hat{\theta})$ and $\hat{V}_y = \hat{V}_u + V_e$. By substituting $\hat{V}_u$ in (7), we obtain the empirical best linear unbiased predictor (EBLUP) of $\mu = X\beta + Zu$, that is,

$$\begin{aligned}\hat{\mu}^{eblup} &= X\hat{\beta}^{eblue} + Z\hat{u}^{eblup}, \hat{\beta}^{eblue} = (X' \hat{V}_y^{-1} X)^{-1} X' \hat{V}_y^{-1} y, \hat{u}^{eblup} \\ &= \hat{V}_u Z' \hat{V}_y^{-1}(y - X\hat{\beta}^{eblue}). \end{aligned}$$

We denote the predictors of μ and u as $\hat{\mu} = \hat{\mu}^{eblup}$ and $\hat{u} = \hat{u}^{eblup}$, respectively. The REML estimator of β is $\hat{\beta} = \hat{\beta}^{eblue}$. The asymptotic distributions of the REML estimators $\hat{\theta}$ and $\hat{\beta}$,

$$\hat{\theta} \sim N_{2(R-1)}(\theta, F^{-1}(\theta)), \quad \hat{\beta} \sim N_m(\beta, (X' V_y^{-1} X)^{-1}),$$

can be used to construct $(1-\alpha)$-level confidence intervals (CIs) for σ^2_{kr} and β_j, where $F(\theta)$ is the REML Fisher information matrix defined in appendix A in the supplementary data online.

3.2 The Predictors

We predict the inverse clr transformation of $\mu_{kdr} = E[y_{kdr}|u_{kdr}]$ under the PMFH model, that is,

$$p_{kdr} = \frac{\exp\{\mu_{kdr}\}}{\exp\{\mu_{kd1}\} + \cdots + \exp\{\mu_{kdR}\}}, p_{kdR} = -p_{kd1} - \ldots - p_{kdR-1},$$
$$r = 1, \ldots, R-1.$$

The plug-in predictors, $\hat{p}^{in}_{kd} = (\hat{p}^{in}_{kd1}, \ldots, \hat{p}^{in}_{kdR-1})'$ and $\hat{p}^{in}_{kdR}$, of the proportions $p_{kd} = (p_{kd1}, \ldots, p_{kdR-1})'$ and p_{kdR} are jointly obtained by applying the inverse

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

of clr transformation to the EBLUPs $\hat{\mu}_{kd}^{eblup} = (\hat{\mu}_{kd1}^{eblup}, \ldots, \hat{\mu}_{kdR-1}^{eblup})'$, that is, $\hat{p}_{kd}^{in} = \text{clr}^{-1}(\hat{\mu}_{kd}^{eblup})$ or equivalently

$$\hat{p}_{kdr}^{in} = \frac{\exp\{\hat{\mu}_{kdr}^{eblup}\}}{\exp\{\hat{\mu}_{kd1}^{eblup}\} + \cdots + \exp\{\hat{\mu}_{kdR}^{eblup}\}}, \hat{p}_{kdR}^{in} = -\hat{p}_{kd1}^{in} - \cdots - \hat{p}_{kdR-1}^{in},$$
$$r = 1, \ldots, R-1.$$

The plug-in predictors of the domain proportions $\bar{Z}_{kdr}$ and counts $Z_{kdr} = N_{kd}\bar{Z}_{kdr}$ are $\hat{\bar{Z}}_{kdr}^{in} = \hat{p}_{kdr}^{in}$ and $\hat{Z}_{kdr}^{in} = N_{kd}\hat{\bar{Z}}_{kdr}^{in}$, $r = 1, \ldots, R$, respectively. For every domain d, the predictors based on models for compositional data fulfill the two conditions: (i) $0 \leq p_{kdr} \leq 1$, $r = 1, \ldots, R$ and (ii) $\sum_{r=1}^{R} p_{kdr} = 1$. Estimating p_{kdr} with predictors based on univariate area-level mixed models, like the FH model, is not a good option because conditions (1) and (2) might not be fulfilled.

The best predictors of the proportions p_{kdr} are $\hat{p}_{kdr}^{bp} = \hat{p}_{kdr}^{bp}(\beta, \theta) = E_{\beta,\theta}[p_{kdr}|y_{kd}]$, $r = 1, \ldots, R-1$, and $\hat{p}_{kdR}^{bp} = 1 - \hat{p}_{kd1}^{bp} - \ldots - \hat{p}_{kdR-1}^{bp}$. It holds that

$$\hat{p}_{kd}^{bp} = \frac{\int_{\mathbb{R}^{R-1}} \frac{\exp\{\mu_{kdr}\}}{\exp\{\mu_{kd1}\}+\cdots+\exp\{\mu_{kdR}\}} f(y_{kd}|u_{kd}) f(u_{kd}) du_{kd}}{\int_{\mathbb{R}^{R-1}} f(y_{kd}|u_{kd}) f(u_{kd}) du_{kd}} \triangleq \frac{A_{kdr}}{B_{kd}},$$

where $\mu_{kdr} = x_{kdr}\beta_k + u_{kdr}$, $y_{kd}|u_{kd} \sim N_{R-1}(X_{kd}\beta_k + u_{kd}, V_{ekd})$, $u_{kd} \sim N_{R-1}(0, V_{ukd}(\theta))$, and

$$A_{kdr} = \int_{\mathbb{R}^{R-1}} \frac{\exp\{\mu_{kdr}\} \exp\{-\frac{1}{2}(y_{kd} - X_{kd}\beta - u_{kd})' V_{ekd}^{-1}(y_{kd} - X_{kd}\beta - u_{kd})\}}{\exp\{\mu_{kd1}\} + \cdots + \exp\{\mu_{kdR}\}} f_\theta(u_{kd}) du_{kd},$$
$$B_{kd} = \int_{\mathbb{R}^{R-1}} \exp\{-\frac{1}{2}(y_{kd} - X_{kd}\beta - u_{kd})' V_{ekd}^{-1}(y_{kd} - X_{kd}\beta - u_{kd})\} f_\theta(u_{kd}) du_{kd}.$$

The EBP of p_{kdr} is $\hat{p}_{kdr}^{ebp} = \hat{p}_{kdr}^{bp}(\hat{\beta}, \hat{\theta})$, $r = 1, \ldots, R-1$, and can be approximated by the following Monte Carlo algorithm:

(1) Calculate the PMFH model parameters $\hat{\beta}$ and $\hat{\theta}$.
(2) For $\ell = 1, \ldots, L$, $d = 1, \ldots, D$, $k = 1, 2$, generate $u_{kd}^{(\ell)}$ i.i.d. $N_{R-1}(0, V_{ukd}(\hat{\theta}))$ and do $u_{kd}^{(L+\ell)} = -u_{kd}^{(\ell)}$.
(3) Calculate $\hat{p}_{kdr}^{ebp} = \hat{A}_{kdr}/\hat{B}_{kd}$, where

$$\hat{A}_{kdr} = \frac{1}{2L}\sum_{\ell=1}^{2L} \frac{\exp\{\hat{\mu}_{kdr}^{(\ell)}\} \exp\{-\frac{1}{2}(y_{kd} - X_{kd}\hat{\beta} - u_{kd}^{(\ell)})' V_{ekd}^{-1}(y_{kd} - X_{kd}\hat{\beta} - u_{kd}^{(\ell)})\}}{\exp\{\hat{\mu}_{kd1}^{(\ell)}\} + \cdots + \exp\{\hat{\mu}_{kdR}^{(\ell)}\}},$$
$$\hat{B}_{kd} = \frac{1}{2L}\sum_{\ell=1}^{2L} \exp\{-\frac{1}{2}(y_{kd} - X_{kd}\hat{\beta} - u_{kd}^{(\ell)})' V_{ekd}^{-1}(y_{kd} - X_{kd}\hat{\beta} - u_{kd}^{(\ell)})\},$$

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

and $\hat{\mu}_{kdr}^{(\ell)} = x_{kdr}\hat{\beta}_{kd} + u_{kd}^{(\ell)}$, $d = 1, \ldots, D$, $k = 1, 2$, $r = 1, \ldots, R-1$, $\ell = 1, \ldots, 2L$.

(4) Calculate $\hat{p}_{kdR}^{ebp} = 1 - \hat{p}_{kd1}^{ebp} - \ldots - \hat{p}_{kdR-1}^{ebp}$, $d = 1, \ldots, D$, $k = 1, 2$.

The EBPs of the domain proportions $\bar{Z}_{kdr}$ are $\hat{\bar{Z}}_{kdr}^{ebp} = \hat{p}_{kdr}^{ebp}$, $r = 1, \ldots, R$. The EBPs of the counts $Z_{kdr} = N_{kdr}\bar{Z}_{kdr}$ are $\hat{Z}_{kdr}^{ebp} = N_{kdr}\hat{\bar{Z}}_{kdr}^{ebp}$, $r = 1, \ldots, R$.

For the alr and ilr transformations, the EBP of p_{kdr} is obtained by substituting the r^{th} component of the inverse clr transformation,

$$\mathrm{clr}_r^{-1}(\mu_{kd}) = \frac{\exp\{\mu_{kdr}\}}{\exp\{\mu_{kd1}\} + \cdots + \exp\{\mu_{kdR}\}},$$

by the corresponding r^{th} component of alr^{-1} and ilr^{-1} transformations, respectively.

The plug-in predictors of DIs and entropies are

$$\hat{K}_d^{in} = \sum_{r=1}^{R} \hat{p}_{1dr}^{in} \log \frac{\hat{p}_{1dr}^{in}}{\hat{p}_{2dr}^{in}}, \quad \hat{H}_{kd}^{in} = -\sum_{r=1}^{R} \hat{p}_{kdr}^{in} \log \hat{p}_{kdr}^{in}, \quad k = 1, 2,\ d = 1, \ldots, D.$$

The EBP-in predictors of DIs and entropies are

$$\hat{K}_d^{ebp.in} = \sum_{r=1}^{R} \hat{p}_{1dr}^{ebp} \log \frac{\hat{p}_{1dr}^{ebp}}{\hat{p}_{2dr}^{ebp}}, \quad \hat{H}_{kd}^{ebp.in} = -\sum_{r=1}^{R} \hat{p}_{kdr}^{ebp} \log \hat{p}_{kdr}^{ebp}, \quad k = 1, 2, d = 1, \ldots, D.$$

The best predictors of entropies

$$H_{kd}(\mu_{kd}) = -\sum_{r=1}^{R} p_{kdr} \log p_{kdr}$$

are $\hat{H}_{kd}^{bp} = \hat{H}_{kd}^{bp}(\beta, \theta) = E_{\beta,\theta}[H_{kd}|y_{kd}]$, $k = 1, 2$, $d = 1, \ldots, D$. It holds that

$$\hat{H}_{kd}^{bp} = \frac{\int_{\mathbb{R}^{R-1}} H_{kd}(\mu_{kd}) f(y_{kd}|u_{kd}) f(u_{kd}) du_{kd}}{\int_{\mathbb{R}^{R-1}} f(y_{kd}|u_{kd}) f(u_{kd}) du_{kd}} \triangleq \frac{A_{Hkd}}{B_{kd}},$$

where $\mu_{kdr} = x_{kdr}\beta_k + u_{kdr}$, $y_{kd}|u_{kd} \sim N_{R-1}(X_{kd}\beta_k + u_{kd}, V_{ekd})$, $u_{kd} \sim N_{R-1}(0, V_{ukd}(\theta))$, and

$$A_{Hkd} = \int_{\mathbb{R}^{R-1}} H_{kd}(\mu_{kd}) \exp\{-\frac{1}{2}(y_{kd} - X_{kd}\beta - u_{kd})' V_{ekd}^{-1}(y_{kd} - X_{kd}\beta - u_{kd})\} f_\theta(u_{kd}) du_{kd},$$

$$B_{kd} = \int_{\mathbb{R}^{R-1}} \exp\{-\frac{1}{2}(y_{kd} - X_{kd}\beta - u_{kd})' V_{ekd}^{-1}(y_{kd} - X_{kd}\beta - u_{kd})\} f_\theta(u_{kd}) du_{kd}.$$

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

The EBP of H_{kd} is $\hat{H}_{kd}^{ebp} = H_{kd}^{bp}(\hat{\beta}, \hat{\theta})$ and can be approximated by the Monte Carlo algorithm:

(1) Calculate the PMFH model parameters $\hat{\beta}$ and $\hat{\theta}$.
(2) For $\ell = 1, \ldots, L,\ d = 1, \ldots, D,\ k = 1, 2,$ generate $u_{kd}^{(\ell)}$ i.i.d. $N_{R-1}(0, V_{ukd}(\hat{\theta}))$ and let $u_{kd}^{(L+\ell)} = -u_{kd}^{(\ell)}$.
(3) Calculate $\hat{H}_{kd}^{ebp} = \hat{A}_{Hkd}/\hat{B}_{kd}$, where

$$\hat{A}_{Hkd} = \frac{1}{2L}\sum_{\ell=1}^{2L} H_{kd}(\hat{\mu}_{kdr}^{(\ell)}) \exp\{-\frac{1}{2}(y_{kd} - X_{kd}\hat{\beta} - u_{kd}^{(\ell)})' V_{ekd}^{-1}(y_{kd} - X_{kd}\hat{\beta} - u_{kd}^{(\ell)})\},$$

$$\hat{B}_{kd} = \frac{1}{2L}\sum_{\ell=1}^{2L} \exp\{-\frac{1}{2}(y_{kd} - X_{kd}\hat{\beta} - u_{kd}^{(\ell)})' V_{ekd}^{-1}(y_{kd} - X_{kd}\hat{\beta} - u_{kd}^{(\ell)})\},$$

and $\hat{\mu}_{kdr}^{(\ell)} = x_{kdr}\hat{\beta}_{kd} + u_{kd}^{(\ell)}, \quad d = 1, \ldots, D, \quad k = 1, 2, \quad r = 1, \ldots, R-1, \ \ell = 1, \ldots, 2L.$

The best predictors of

$$K_d(\mu_{1d}, \mu_{2d}) = \sum_{r=1}^{R} p_{1dr} \log \frac{p_{1dr}}{p_{2dr}}$$

are $\hat{K}_d^{bp} = \hat{A}_{Kd}^{bp}(\beta, \theta) = E_{\beta,\theta}[K_d | y_{1d}, y_{2d}],\ d = 1, \ldots, D$. It holds that

$$\hat{A}_{Kd}^{bp} = \frac{\int_{\mathbb{R}^{2(R-1)}} K_d(\mu_{1d}, \mu_{1d}) f(y_{1d}|u_{1d}) f(y_{2d}|u_{2d}) f(u_{1d}) f(u_{2d}) du_{1d} du_{2d}}{\int_{\mathbb{R}^{2(R-1)}} f(y_{1d}|u_{1d}) f(y_{2d}|u_{2d}) f(u_{1d}) f(u_{2d}) du_{1d} du_{2d}} \triangleq \frac{A_{Kd}}{B_{Kd}},$$

where $\mu_{kdr} = x_{kdr}\beta_k + u_{kdr},\ y_{kd}|u_{kd} \sim N_{R-1}(X_{kd}\beta_k + u_{kd}, V_{ekd}),\ u_{kd} \sim N_{R-1}(0, V_{ukd}(\theta))$ and

$$A_{Kd} = \int_{\mathbb{R}^{2(R-1)}} K_d(\mu_{1d}, \mu_{2d}) \exp\{-\frac{1}{2}(y_{1d} - X_{1d}\beta - u_{1d})' V_{e1d}^{-1}(y_{1d} - X_{1d}\beta - u_{1d})\} f_\theta(u_{1d})$$
$$\cdot \exp\{-\frac{1}{2}(y_{2d} - X_{2d}\beta - u_{2d})' V_{e2d}^{-1}(y_{2d} - X_{2d}\beta - u_{2d})\} f_\theta(u_{2d}) du_{1d} du_{2d},$$

$$B_{Kd} = \int_{\mathbb{R}^{2(R-1)}} \exp\{-\frac{1}{2}(y_{1d} - X_{1d}\beta - u_{1d})' V_{e1d}^{-1}(y_{1d} - X_{1d}\beta - u_{1d})\} f_\theta(u_{1d})$$
$$\cdot \exp\{-\frac{1}{2}(y_{2d} - X_{2d}\beta - u_{2d})' V_{e2d}^{-1}(y_{2d} - X_{2d}\beta - u_{2d})\} f_\theta(u_{2d}) du_{1d} du_{2d}.$$

The EBP of K_d is $\hat{K}_d^{ebp} = K_d^{bp}(\hat{\beta}, \hat{\theta})$, and can be calculated by the following Monte Carlo algorithm:

(1) Calculate the PMFH model parameters $\hat{\beta}$ and $\hat{\theta}$.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

(2) For $\ell = 1, \ldots, L$, $d = 1, \ldots, D$, $k = 1, 2$ generate $u_{kd}^{(\ell)}$ i.i.d. $N_{R-1}(0, V_{ukd}(\hat{\theta}))$ and let $u_{kd}^{(L+\ell)} = -u_{kd}^{(\ell)}$.

(3) Calculate $\hat{K}_d^{ebp} = \hat{A}_{Kd}/\hat{B}_{Kd}$, where

$$\hat{A}_{Kd} = \frac{1}{2L}\sum_{\ell=1}^{2L} K_d(\hat{\mu}_{1dr}^{(\ell)}, \hat{\mu}_{2dr}^{(\ell)}) \exp\{-\frac{1}{2}(y_{1d} - X_{1d}\hat{\beta} - u_{1d}^{(\ell)})' V_{e1d}^{-1}(y_{1d} - X_{1d}\hat{\beta} - u_{1d}^{(\ell)})\}$$

$$\cdot \exp\{-\frac{1}{2}(y_{2d} - X_{2d}\hat{\beta} - u_{2d}^{(\ell)})' V_{e2d}^{-1}(y_{2d} - X_{2d}\hat{\beta} - u_{2d}^{(\ell)})\},$$

$$\hat{B}_{Kd} = \frac{1}{2L}\sum_{\ell=1}^{2L} \exp\{-\frac{1}{2}(y_{1d} - X_{1d}\hat{\beta} - u_{1d}^{(\ell)})' V_{e1d}^{-1}(y_{1d} - X_{1d}\hat{\beta} - u_{1d}^{(\ell)})\}$$

$$\cdot \exp\{-\frac{1}{2}(y_{2d} - X_{2d}\hat{\beta} - u_{2d}^{(\ell)})' V_{e2d}^{-1}(y_{2d} - X_{2d}\hat{\beta} - u_{2d}^{(\ell)})\},$$

and $\hat{\mu}_{kdr}^{(\ell)} = x_{kdr}\hat{\beta}_{kd} + u_{kd}^{(\ell)}$, $d = 1, \ldots, D$, $k = 1, 2$, $r = 1, \ldots, R-1$, $\ell = 1, \ldots, 2L$.

3.3 Mean Squared Errors

The approximation of the MSEs of the plug-in predictors of K_d and H_{kd} is an important mathematical challenge. The literature mainly contains theoretical derivations for MSEs of EBLUPs of univariable linear indicators under linear mixed models. It is possible to find approximations of MSEs of EBPs of univariable linear indicators under generalized linear mixed models. However, we have not found any paper dealing with the approximation of the MSEs of EBPs of univariable nonlinear indicators under a linear or nonlinear mixed model. The DI and the entropy are multivariable nonlinear and our predictors are based on multivariate linear mixed models. The challenge of approximating the MSE of an EBP of a nonlinear indicator is consistent enough to merit own investigation. Such a challenge should be addressed first for nonlinear univariate indicators under univariate linear mixed models. For these reasons, this section proposes estimating the MSEs by parametric bootstrap.

There are several variants of parametric bootstrap for estimating MSEs under area-level mixed models. A first approach is the basic parametric bootstrap method for estimating MSEs, like the procedures given by González-Manteiga et al. (2008, 2010) or Erciulescu and Fuller (2019). A second approach deals with double bootstrap methods, highlighting the contributions of Hall and Maiti (2006) and Erciulescu and Fuller (2013). Due to the simplicity and lower computational cost, in the application to real data, we have chosen to implement a basic parametric bootstrap to estimate the MSE of K_d and H_{kd}.

The steps of the bootstrap resampling algorithm are as follows:

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

(1) Fit the PMFH model (6) to the data (y_{kd}, X_{kd}), $k = 1, 2$, $d = 1, \ldots, D$ and calculate $\hat{\beta}$ and $\hat{\theta}$.
(2) Repeat B times, $b = 1, \ldots, B$,

2.1. Generate $u_{kd}^{*(b)} \sim N_{R-1}(0, V_{ukd}(\hat{\theta}))$, $e_{kd}^{*(b)} \sim N_{R-1}(0, V_{ekd})$, $\mu_{kd}^{*(b)} = X_{kd}\hat{\beta} + u_{kd}^{*(b)}$, $y_{kd}^{*(b)} = \mu_{kd}^{*(b)} + e_{kd}^{*(b)}$, $p_{kd}^{*(b)} = \text{clr}^{-1}(\mu_{kd}^{*(b)})$, $k = 1, 2$, $d = 1, \ldots, D$.
2.2. For $k = 1, 2$, $d = 1, \ldots, D$, calculate

$$K_d^{*(b)} = \sum_{r=1}^{R} p_{1dr}^{*(b)} \log \frac{p_{1dr}^{*(b)}}{p_{2dr}^{*(b)}}, \quad H_{kd}^{*(b)} = -\sum_{r=1}^{R} p_{kdr}^{*(b)} \log p_{kdr}^{*(b)}.$$

2.3. Calculate $\hat{p}_{kd}^{*in(b)}$ based on the data $(y_{kd}^{*(b)}, X_{kd})$, $k = 1, 2$, $d = 1, \ldots, D$.
2.4. For $k = 1, 2$, $d = 1, \ldots, D$, calculate

$$\hat{K}_d^{*in(b)} = \sum_{r=1}^{R} \hat{p}_{1dr}^{*in(b)} \log \frac{\hat{p}_{1dr}^{*in(b)}}{\hat{p}_{2dr}^{*in(b)}}, \quad \hat{H}_{kd}^{*in(b)} = -\sum_{r=1}^{R} \hat{p}_{kdr}^{*in(b)} \log \hat{p}_{kdr}^{*in(b)}.$$

(3) For $k = 1, 2$, $d = 1, \ldots, D$, $r = 1, \ldots, R-1$, calculate the bootstrap MSE estimators

$$mse_{kdr}^{*} = \frac{1}{B}\sum_{b=1}^{B}(\hat{p}_{kdr}^{*in(b)} - p_{kdr}^{*(b)})^2, mse_{kd}^{*} = \frac{1}{B}\sum_{b=1}^{B}(\hat{H}_{kd}^{*in(b)} - H_{kd}^{*(b)})^2,$$

$$mse_{d}^{*} = \frac{1}{B}\sum_{b=1}^{B}(\hat{K}_{d}^{*in(b)} - K_{d}^{*(b)})^2.$$

Similar algorithms calculate the MSEs of the EBP and EBP-in predictors of K_d and H_{kd}.

4. SIMULATION EXPERIMENTS

This section presents two simulation experiments that mimic the application to SLFS2022.4 data. We use the same auxiliary data and generate the target vectors from the PMFH model selected in section 5. That is, we assume that the estimates of the model parameters in the application to SLFS2022.4 data are the true parameters. Simulation 1 examines the performance of the fitting algorithm that computes the REML estimators of the model parameters and examines the behavior of the DI and entropy predictors. See appendix A in the supplementary data online and section 3.2. Simulation 2 deals with the MSE bootstrap estimation and provides a recommendation on the number of replicates to be used. See section 3.3.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

4.1 Simulation 1

The goal of simulation 1 is to investigate the behavior of the fitting algorithm and the performance of the predictors of K_d and H_{kd}, $k = 1, 2$, $d = 1, \ldots, D$. We remind that there are $R = 7$ occupation sectors and $D = 50$ provinces. We run simulation 1 with $I = 100$ iterations. Appendix B.1 in the supplementary data online gives the steps of simulation 1 and the definitions of the absolute and relative performance measures.

Table 3 presents the biases (BIAS) and the root MSEs (RMSEs) of the REML estimators $\hat{\beta}$ and $\hat{\theta}$ of the regression and variance parameters of the PMFH model that generates the data. For $k = 1, 2$, $r = 1, \ldots, 6$, the regression parameters are denoted as follows. The intercept of sex k and occupation sector r is β^0_{kr}. The slope of an auxiliary variable x for sex k and occupation sector r is β^x_{kr}. In the case that x has the same slope for both sexes in occupation sector r, we write $\beta^x_{.r}$. For $k = 1, 2$, $r = 1, \ldots, 6$, the variance parameters are $\theta_{kr} = \sigma^2_{kr}$. Table 3 shows that the biases and RMSEs of the REML estimators of β and θ are very close to zero.

Table 3. Biases and Root MSEs of Estimators of Model Parameters

r	β	True	BIAS	RMSE	θ	True	BIAS	RMSE
1	β^0_{11}	−3.1525	0.0034	0.4237	θ_{11}	0.0719	−0.0046	0.0157
	β^0_{21}	−3.5876	0.0098	0.4059	θ_{21}	0.1533	0.0010	0.0342
	$\beta^{situ4}_{.1}$	3.2708	−0.0081	0.6516				
2	β^0_{12}	−0.5823	0.0073	0.1518	θ_{12}	0.0149	0.0002	0.0057
	β^0_{22}	−0.0869	0.0082	0.1775	θ_{22}	0.0179	0.0011	0.0083
	$\beta^{edu3}_{.2}$	1.1838	−0.0126	0.2438				
3	β^0_{13}	−1.2882	−0.0520	0.4027	θ_{13}	0.0215	0.0004	0.0067
	β^0_{23}	−1.6014	−0.0570	0.4107	θ_{23}	0.0452	0.0008	0.0117
	$\beta^{age2}_{.3}$	2.2689	0.0927	0.6639				
4	β^0_{14}	−1.3147	0.0379	0.2116	θ_{14}	0.0676	0.0013	0.0163
	β^0_{24}	−0.3945	0.0444	0.2563	θ_{24}	0.0326	−0.0005	0.0106
	$\beta^{edu3}_{.4}$	0.9196	−0.0629	0.3475				
5	β^0_{15}	1.2061	−0.0204	0.1981	θ_{15}	0.0407	0.0023	0.0120
	β^0_{25}	2.2151	−0.0219	0.2347	θ_{25}	0.0299	−0.0003	0.0100
	$\beta^{edu3}_{.5}$	−1.5961	0.0350	0.3157				
6	β^0_{16}	−0.1852	−0.0020	0.0773	θ_{16}	0.0880	0.0023	0.0195
	β^0_{26}	0.1963	0.0053	0.0681	θ_{26}	0.0544	0.0010	0.0168
	$\beta^{edu1}_{.6}$	7.9098	−0.0922	1.4376				

Table 4 shows the average absolute biases (ABIAS), the average absolute relative biases in percent (ARBIAS), the mean RMSEs, and the mean relative root MSEs in percent (RRMSEs) of the predictors of DIs and entropies. The relative absolute biases are quite small in all cases. However, the average RRMSEs are smaller for the predictors of entropies than for the predictors of DIs. In any case, the RRMSEs are less than 25 percent in all cases.

Figure 2 shows the boxplots of the biases and RMSEs of the plug-in, EBP, and EBP-in predictors and the direct estimators of the DIs. This figure shows that the EBP is the only one that is close to being unbiased. In terms of RMSE, the plug-in and the EBP-in predictors give almost identical results. However,

Table 4. Biases and Root MSEs of Predictors

	ABIAS	RMSE	ARBIAS	RRMSE
$\hat{K}^{dir}$	0.01994	0.1289	2.8796	19.3408
$\hat{K}^{in}$	0.01307	0.1144	− 1.7223	18.0576
$\hat{K}^{ebp}$	0.01167	0.1310	− 0.5748	20.4289
$\hat{K}^{ebp.in}$	0.01351	0.1146	− 1.8256	18.1080
$\hat{H}_1^{dir}$	0.00827	0.0598	− 0.4710	3.6787
$\hat{H}_1^{in}$	0.00712	0.0547	0.3585	3.3291
$\hat{H}_1^{ebp}$	0.00457	0.0543	− 0.0027	3.3226
$\hat{H}_1^{ebp.in}$	0.00737	0.0547	0.3963	3.3324
$\hat{H}_2^{dir}$	0.01304	0.0398	− 0.7733	2.3595
$\hat{H}_2^{in}$	0.00422	0.0252	0.2420	1.4774
$\hat{H}_2^{ebp}$	0.00200	0.0250	− 0.0516	1.4707
$\hat{H}_2^{ebp.in}$	0.00431	0.0253	0.2478	1.4842

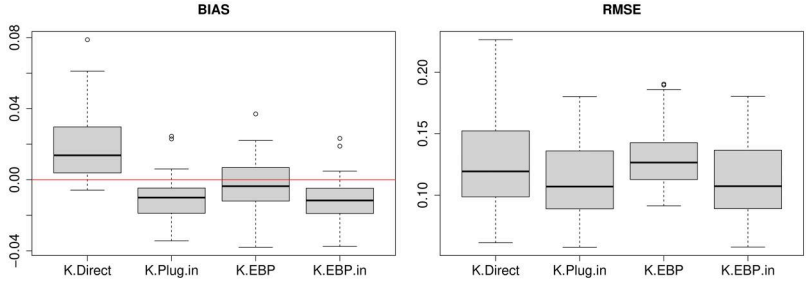

Figure 2. BIAS (Left) and RMSEs (Right) of Direct Estimators, Plug-In, EBP and EBP-In Predictors of DIs.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

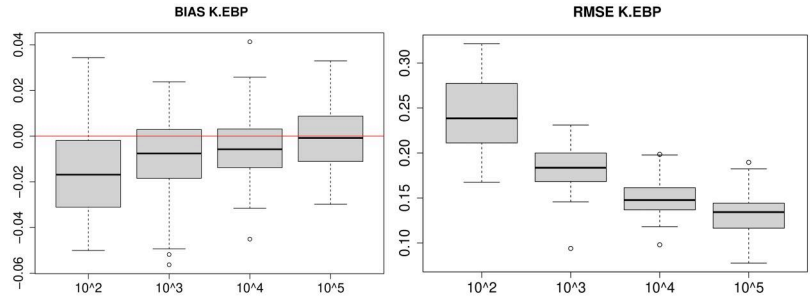

Figure 3. BIAS (Left) and RMSEs (Right) of EBPs with $L=10^2, 10^3, 10^4, 10^5$.

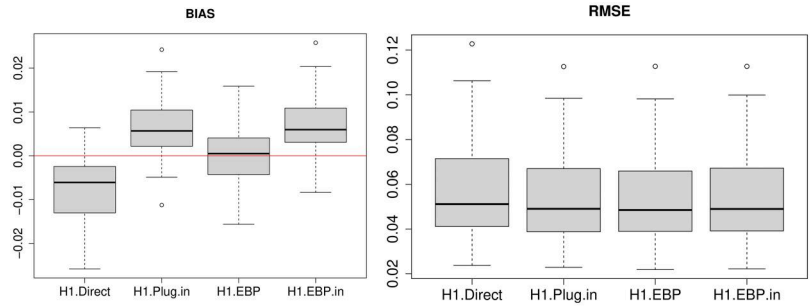

Figure 4. BIAS (Left) and RMSEs (Right) of Plug-In and Direct Estimators of Men Entropies.

the EBP is not able to beat the direct estimator in terms of RMSE. This is explained by the calculation of the EBP, whose algorithm introduces additional variance and can increase the MSE of the predictors. Although the EBP has some optimality properties, the computation times become very high compared to the plug-in and EBP-in predictors. In practice, we do not recommend computing EBPs of DIs.

Figure 3 shows the boxplots of the empirical bias and RMSE when the EBP is approximated with $L = 10^2, 10^3, 10^4$, and 10^5 Monte Carlo iterations. We see significant reductions in bias and RMSE as L increases. Simulation 1 was run on an Intel Xeon Platinum 8160 server running a Linux Debian Squeeze 64-bit operating system, 8 CPUs at 2.1 GHz, and 8 GB Ddr3 RAM, and implemented in R software. The obtained average running times (over the $I = 100$ iterations) are 0.923, 8.733, 85.292, 853.741, and $<10^{-3}$ seconds for $L = 10^2, 10^3, 10^4, 10^5$, and the plug-in predictor, respectively.

Figures 4 and 5 show the boxplots of the biases and RMSE of the direct estimator, plug-in, EBP, and EBP-in predictors of the entropies. These figures

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

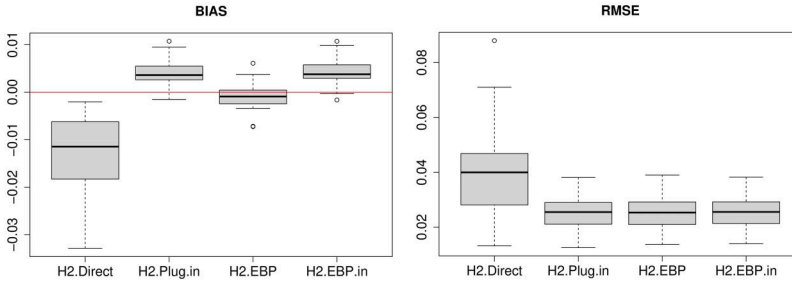

Figure 5. BIAS (Left) and RMSEs (Right) of Plug-In and Direct Estimators of Women Entropies.

show that the plug-in, EBP, and EBP-in predictors perform better than the direct estimator. In this case, we do not see much difference between these three predictors, so we recommend using the plug-in predictor.

We note that model-based predictions of proportions smooth the corresponding direct estimates. Consequently, model-based predictions of probability vectors will tend to be closer to the uniform distribution where entropy is at its maximum. For this reason, the plug-in and EBP-in predictors slightly overestimate the entropies, unlike the direct estimators. This also affects the DI, as the plug-in and EBP-in predictions of male and female proportions tend to be closer than the corresponding direct estimates. For this reason, the plug-in and EBP-in predictor of the DI has a slight negative bias, whereas the direct estimator had a positive bias.

4.2 Simulation 2

Simulation 2 studies the behavior of the estimator of the MSE of the predictors $mse^*(\hat{K}_d^{dir})$ and $mse^*(\hat{K}_d^{in})$. To do this, such estimators are compared with the empirical MSE of $\hat{K}_d$ obtained from the output of simulation 1. The procedure is described in appendix B.2 in the supplementary data online.

Table 5 presents the average absolute biases AB and the average RMSEs RE of the bootstrap estimators mse^{*in}. The REs decrease as B increases, but the ABs do not. Figure 6 presents the boxplots of the relative biases B_d (left) and the RRMSEs RE_d (right) of the bootstrap estimators mse_d^{*in} of the MSEs of the plug-in predictor $\hat{K}_d^{in}$ of the DIs. The performance measures of the bootstrap estimators have a marked decrease up to $B = 400$. For numbers of bootstrap iterations greater than $B = 200$, the decay of the measures is mitigated or stabilized.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Table 5. AB and RE of mse^{*in} and mse^{*dir} for $B=50, 100, 200, 400$

	$B = 50$	$B = 100$	$B = 200$	$B = 400$
AB^{dir}	0.0036	0.0037	0.0036	0.0037
AB^{in}	0.0024	0.0023	0.0023	0.0023
RE^{dir}	0.0067	0.0057	0.0051	0.0048
RE^{in}	0.0052	0.0043	0.0038	0.0035

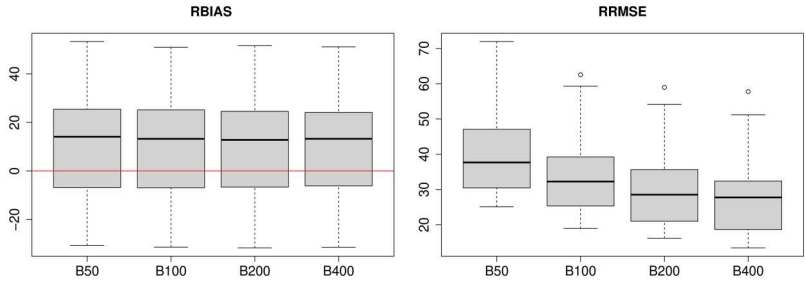

Figure 6. B_d (Left) and RE_d (Right) of mse^{*in} for $B=50, 100, 200, 400$.

5. APPLICATION TO SPANISH LABOUR FORCE SURVEY

This section applies the statistical methodology introduced to the SLFS2022.4 data described in section 2. To overcome the lack of precision of the direct estimators, we construct a data file with auxiliary aggregated variables by province, sex, and occupation sector. Unfortunately, the authors do not have access to demographic administrative records with such a level of disaggregation. This is why explanatory variables are constructed from surveys and therefore have measurement errors different from sampling errors. Extending our research to include the measurement error methods proposed by Ybarra and Lohr (2008) and Burgard et al. (2020, 2022) is beyond the scope of our current research objectives. To significantly reduce the variance of the aggregate variables, they are taken from 20 SLFSs, starting in the last quarter of 2017 (2017.4) and ending in the last quarter of 2022 (2022.4). Consequently, the direct estimators of the variances, by domain and sector, of the aggregated auxiliary variables are much smaller than the direct estimators of the variances of the corresponding target variables. For men, women, and both, appendix D in the supplementary data online calculates the averages of quotients between the direct estimators of the variances of the response variables and the corresponding aggregated auxiliary variables. In summary, the variances of the direct estimator of the target variables are between three and 120 times greater

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

than the variances of the auxiliary variables. These results allow us to analyze the data without having to resort to more complex multivariate models that include measurement errors in the auxiliary variables.

The selected auxiliary variables are direct estimates of the means (proportions) of unit-level binary variables measured on individuals of U_{kdr}. They are

Age group, with three categories: between 16 and 30 years (age1), between 30 and 50 years (age2) and over 50 years (age3).

Citizenship, with two categories: Spanish (nac1) and not Spanish (nac2).

Education, with three categories: primary or less (edu1), basic secondary education (edu2), bachelor or higher education, such as university (edu3).

Working hours, with two categories: full-time (work1) and part-time work (work2).

Professional status, with five categories: employer (with or without employees) or independent worker (st1), cooperative or family business (st2), public sector salaried employee (st3), private sector salaried employee (st4), and others (st5).

Appendix B in the supplementary data online defines the logratio transformations and provides the model diagnostics for each transformation. Based on the model diagnostics, we find that the model with the clr transformation performs best in terms of p-values for the model parameters, estimates of the variances of the random effects, and standardized residuals. Therefore, we apply the clr transformation to the target data, that is, $y_{kd} = \text{clr}(z_{kd})$, $k = 1, 2$, $d = 1, \ldots, D$. The PMFH model is fitted to the transformed data and to a subset of significant auxiliary aggregated variables. For each province and sex, the provisional population indicators of interest are the proportions of employed persons in the seven occupational sectors. We exclude the autonomous cities of Ceuta and Melilla from the statistical analysis. We are first interested in estimating the composition of the domains with $R = 7$ categories. As explanatory variables, we select those with a consistent actual interpretation and better in terms of p-value when fitting the PMFH model to the data.

Table 6 shows the selected auxiliary variables, the estimates of the regression parameters, the standard errors (SE), p-values, and asymptotic CIs.

Within each occupation sector, men and women have the same explanatory variables with regression parameters $\beta_{.1}^{situ4}$, $\beta_{.2}^{edu3}$, $\beta_{.3}^{age2}$, $\beta_{.4}^{edu3}$, $\beta_{.5}^{edu3}$, and $\beta_{.6}^{edu1}$. However, they have different intercepts, that is, $\beta_{1r}^{(0)} \neq \beta_{2r}^{(0)}$, $r = 1, \ldots, 6$. Therefore, the dimensions of y_{kd} and X_{kd} are 6×1 and 6×12, respectively, $k = 1, 2$, $d = 1, \ldots, D$.

As all the auxiliary variables are proportions, the sign and magnitude of the regression parameters provide interesting interpretations, assuming that the rest of the auxiliary variables remain constant. The clr transformation of the

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Table 6. PMFH Model Parameter Estimates

r	β	Estimate	SE	p-value	Asymp. 95% CI
1	β^0_{11}	−3.1525	0.3833	1.9601e–16	(−3.9038, −2.4012)
	β^0_{21}	−3.5876	0.3781	2.3714e–21	(−4.3288, −2.8465)
	$\beta^{situ4}_{.1}$	3.2708	0.5924	3.3754e–08	(2.1096, 4.4319)
2	β^0_{12}	−0.5823	0.1496	9.9971e–05	(−0.8757, −0.2890)
	β^0_{22}	−0.0869	0.1810	6.3093e–01	(−0.4417, 0.2678)
	$\beta^{edu3}_{.2}$	1.1838	0.2446	1.3016e–06	(0.7044, 1.6633)
3	β^0_{13}	−1.2882	0.4521	4.3796e–03	(−2.1743, −0.4021)
	β^0_{23}	−1.6014	0.4649	5.7307e–04	(−2.5127, −0.6900)
	$\beta^{age2}_{.3}$	2.2689	0.7449	2.3223e–03	(0.8087, 3.7290)
4	β^0_{14}	−1.3147	0.2116	5.2155e–10	(−1.7295, −0.8999)
	β^0_{24}	−0.3945	0.2539	1.2017e–01	(−0.8922, 0.1030)
	$\beta^{edu3}_{.4}$	0.9196	0.3456	7.7989e–03	(0.2421, 1.5971)
5	β^0_{15}	1.2061	0.2022	2.4815e–09	(0.8096, 1.6025)
	β^0_{25}	2.2151	0.2429	7.6813e–20	(1.7389, 2.6913)
	$\beta^{edu3}_{.5}$	−1.5961	0.3281	1.1494e–06	(−2.2393, −0.9530)
6	β^0_{16}	−0.1852	0.0914	4.2781e–02	(−0.3644, −0.0060)
	β^0_{26}	0.1963	0.0642	2.2480e–03	(0.0703, 0.3222)
	$\beta^{edu1}_{.6}$	7.9098	1.5636	4.2241e–07	(4.8451, 10.9745)

Table 7. Asymptotic 95% CIs for θ_{kr}, $k=1,2$, $r=1,\ldots,6$

	k	θ_{k1}	θ_{k2}	θ_{k3}	θ_{k4}	θ_{k5}	θ_{k6}
$\hat{\theta}$	1	0.0719	0.0149	0.0215	0.0676	0.0407	0.0880
CI inf		0.0413	0.0040	0.0080	0.0380	0.0182	0.0459
CI sup		0.1025	0.0257	0.0350	0.0972	0.0633	0.1301
$\hat{\theta}$	2	0.1533	0.0179	0.0452	0.0326	0.0299	0.0544
CI inf		0.0909	0.0031	0.0230	0.0139	0.0084	0.0255
CI sup		0.2157	0.0328	0.0674	0.0512	0.0513	0.0833

direct estimator of the employed people (men and women) in sector 1 is explained with a positive sign by *situ*4 (private sector employees). The proportion of people employed in sector 2 tends to increase with higher education (*edu*3). For men and women in sector 3, the proportions tend to increase with middle age (*age*2). Sector 4 shows a similar behavior to sector 2. Sectors 5

and 6 are similar in that higher levels of education (*edu*3) are associated with lower proportions of men and women in work, while lower levels of education (*edu*1) lead to higher proportions.

Table 7 gives the 95% CIs of the variance parameters $\theta_{kr} = \sigma^2_{ukr}$, $k = 1, 2$, $r = 1, \ldots, 6$. We observe that all variances are significantly greater than zero.

Following (4), the raw residuals and the unit-level standardized residuals (u.sresiduals) of the fitted PMFH model are

$$\varepsilon_{kdr} = y_{kdr} - \hat{\mu}_{kdr}, \quad \varepsilon^{st}_{kdr} = \varepsilon_{kdr}/V^{1/2}_{ekdr}, \quad k = 1, 2, d = 1, \ldots, D, r = 1, \ldots, 6, \quad (8)$$

where V_{ekdr} is the r^{th} diagonal element of V_{ekd}. Alternatively, appendix C in the supplementary data online presents a definition of the raw residuals and the area-level standardized residuals (a.sresiduals) for the fitted PMFH model, based on the inverse clr transformation of the direct estimators and the EBLUPs. In addition, boxplots of the a.sresiduals for men and women, jointly and separately, are provided. However, (8) is a more natural residual definition under the PMFH model, according to the normality assumption.

Figure 7 plots the u.sresiduals for men and women (left), men (centre), and women (right). The u.sresiduals are fairly symmetrical around zero and show no relevant pattern except in sector 7, where they have a left-skewed distribution. The model-based predictions of the proportions of men employed in sector 7 tend to be higher than the corresponding direct estimates. There are no outliers u.sresiduals.

Figure 8 shows a histogram of the u.sresiduals and the corresponding kernel density estimate to check the normality hypothesis. It appears that the distribution of the u.sresiduals is slightly skewed to the left.

Figure 9 plots the direct estimates and plug-in predictions of the proportions of employed men and women for provinces sorted by sample size. For the first six categories, the plug-in predictors present smoother behavior across domains, especially for men.

Figure 10 plots the bootstrap RMSE estimates of $\hat{K}^{in}_d$ and $\hat{K}^{dir}_d$, sorted by province sample size $n_{.d} = n_{1d} + n_{2d}$, $d = 1, \ldots, D$. This figure shows that the estimated RMSEs of the plug-in predictor are lower than the corresponding ones of the direct estimator.

Figure 11 (left) presents a map of the Spanish provinces colored according to the DI predictions for the fourth quarter of 2022. Figure 11 (right) gives the maps of their RRMSEs. These maps allow us to analyze how sex divergence differs across Spanish provinces. First, we observe that the largest discrepancies in the distribution of men and women by main occupations are found in Ávila, Huelva, Teruel, Lérida, Baleares, Castellón, Asturias, Guipúzcoa, León, Huesca, Burgos, Badajoz, and Valladolid (in descending order). Second, there seems to be a pattern of gender divergence in the western region, although it is not very clear. Third, the north-eastern zone of Spain seems to have a lower divergence than the rest of the zones of the country in general.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

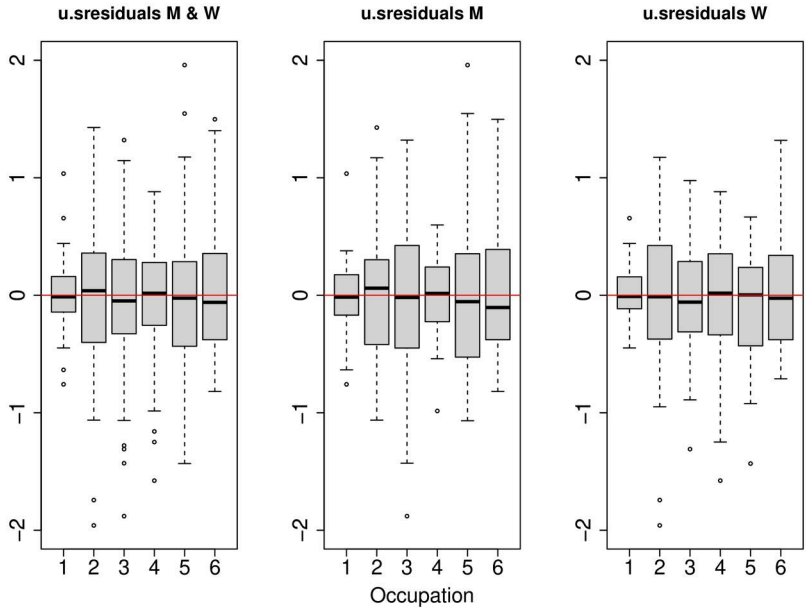

Figure 7. Boxplots of u.sresiduals.

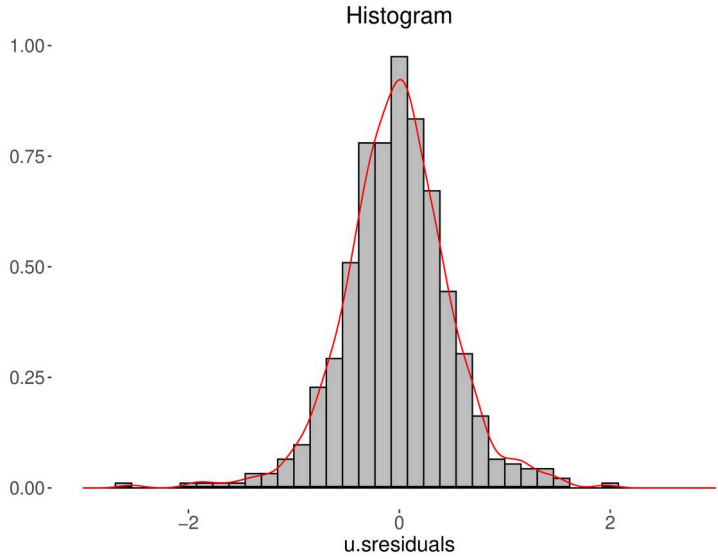

Figure 8. Histogram of u.sresiduals.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

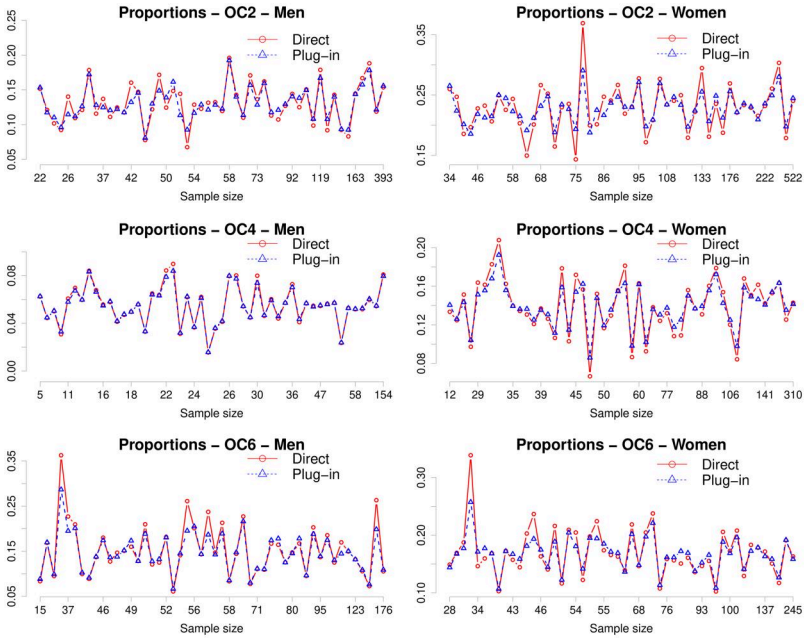

Figure 9. Direct and Plug-In Estimated Proportions of Employed Men and Women by Sectors for Provinces Sorted by Sample Size.

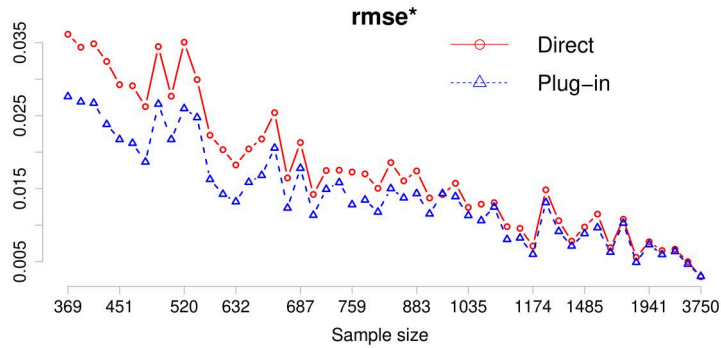

Figure 10. Root MSE Estimates of Plug-In and Direct Predictors of DIs, Sorted by Sample Size.

Finally, in provinces with similar demographic and socioeconomic conditions, the distribution of men and women tends to show less divergence.

Figure 11 (left) shows that estimates of DIs are not close to zero. We note that the lowest DI value is around 0.3. In terms of labor market equality, these high predicted values reveal the magnitude of the problem: the labor market

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

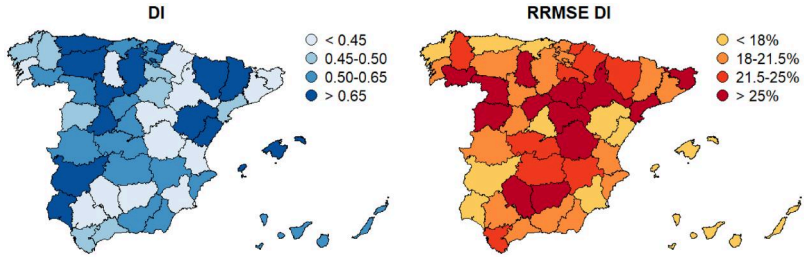

Figure 11. DI and RRMSE Estimates in SLFS2022.4.

disadvantages women and the occupational distribution is clearly not homogeneous. According to our results, public and private institutions should implement policies of labor equality and promote the inclusion of men and women in sectors where they are a minority.

To better understand where and how the differences between the men and women occupation sector distributions arise, we calculate for each occupational sector the average across provinces of its summands (contributions) in the formula of the plug-in predictors of DIs, that is

$$\hat{K}_r^{in} = \frac{1}{D}\sum_{d=1}^{D} \hat{K}_{dr}^{in}, \quad \hat{K}_{dr}^{in} = \hat{p}_{1dr}^{in} \log \frac{\hat{p}_{1dr}^{in}}{\hat{p}_{2dr}^{in}}, \quad r = 1, \ldots, 7,$$

and we estimate their corresponding RMSEs, $rmse(\hat{K}_r^{in})$, and CV in percent, $\text{cv}(\hat{K}_r^{in}) = 100 rmse(\hat{K}_r^{in})/|\hat{K}_r^{in}|$, by parametric bootstrap. Table 8 contains the sector contributions $\hat{K}_r^{in}$, $r = 1, \ldots, 7$ and the corresponding estimates of their RMSEs and CVs.

Table 8 shows that the contributions of occupational sectors 2, 4, 5, and 6 are negative for all provinces. This indicates that the presence of women in these sectors is higher than that of men. The contributions of sectors 1 and 6 are the closest to zero, showing that in these sectors the presence of men and women is homogeneous in all provinces. This sector 6 also has a very large CV, but this is because its mean is very close to zero and not because there is really a large dispersion. In any case, it does not have a notable influence on the estimates of the DIs. The remaining sectors show generally positive contributions to the DI, especially sector 7, with a mean of around 0.71 and RMSE close to 0.022 (both values much higher than the rest, which are comparatively quite homogeneous). This allows us to understand the results shown above: there is a clear majority of men employed in jobs related to industrial activities (mining, manufacturing, construction, etc.) and this difference greatly accentuates the divergence rate between provinces. The large presence of men in this sector is, we believe, due to historical and social reasons. As far as we can see,

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Table 8. Contributions to DI by Occupational Sectors

Contribution	OC1	OC2	OC3	OC4	OC5	OC6	OC7
$\hat{K}_r^{in}$	0.025	−0.071	0.045	−0.049	−0.103	−0.016	0.719
$rmse(\hat{K}_r^{in})$	0.001	0.003	0.004	0.001	0.003	0.005	0.022
$\text{cv}(\hat{K}_r^{in})$	4.674	4.194	8.739	2.454	3.614	33.641	3.108

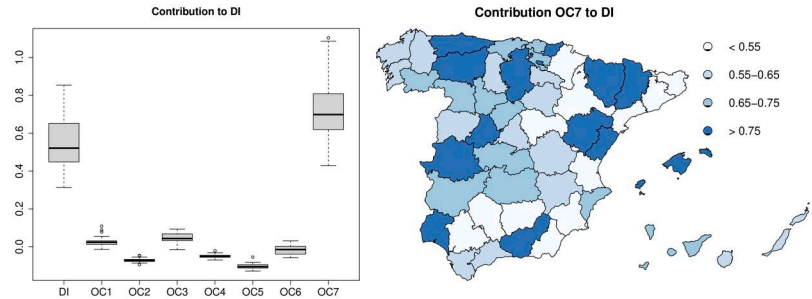

Figure 12. DI Predictions and Contributions by Each Occupation Sector (Left), and Contributions of Occupation Sector 7 by Province (Right).

the most influential sector in the DI predictor is sector 7. Among the other sectors, no occupational divergence by sex is observed.

Figure 12 (left) gives boxplots of the DI predictions $\hat{K}_d^{in}$ and of the contributions $\hat{K}_{dr}^{in}$ of each occupation sector. The impact of OC7 is remarkable. Figure 12 (right) shows the contributions $\hat{K}_{d7}^{in}$ of OC7 for each province. We can see that provinces in the west tend to have greater differences than those in the east. We note that the provinces with the highest DI predictions are not necessarily those with the longest tradition of industrial employment, such as Ávila (0.85), Huelva (0.83), Teruel (0.82), Lérida (0.79), or Baleares (0.73), but those where this difference is higher.

Figure 13 (left and right) presents maps of the Spanish provinces colored according to the Shannon entropy plug-in predictions for men (H_1) and women (H_2), respectively. The entropy predictions tend to be greater for men than for women through Spanish territory, this translates into a closer approach to the maximum entropy value (uniform distribution on occupational sectors) for men.

Figure 14 (left and right) presents maps of the estimated RRMSEs of the entropy plug-in predictors for men (RRMSE H_1) and women (RRMSE H_2), respectively. The estimates of RRMSEs are greater for women than for men, most of them do not exceed 5%. In both cases, the estimates are particularly low, this is because the variance is small for both sexes, which makes it

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

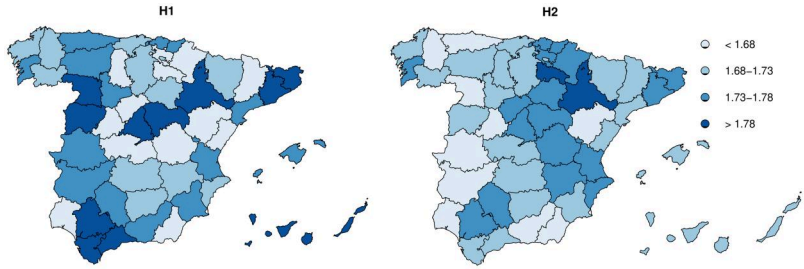

Figure 13. Shannon Entropy Predictions in SLFS2022.4, $k=1, 2$.

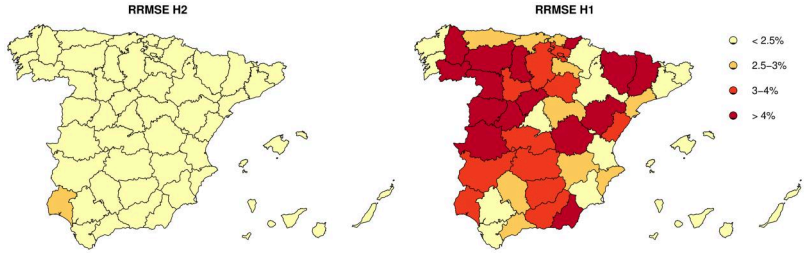

Figure 14. RRMSE of Shannon Entropy Predictors in SLFS2022.4, $k=1, 2$.

possible to predict with high accuracy ($\hat{\text{var}}(H_{1d}) \approx 0.03$ and $\hat{\text{var}}(H_{2d}) \approx 0.001$). The results obtained in figure 14 are consistent with those obtained in simulation 1. The RRMSEs from figures 11 (right) and 14 were obtained using the parametric bootstrap approach described in section 3.2 with $B = 500$ iterations.

6. CONCLUSIONS AND DISCUSSION

The DI is a relevant statistical indicator used in sociological studies to measure the "distance" between two groups to be compared with respect to a given variable of interest. The Shannon entropy measures how close the distributions of men and women in different occupation sectors are to the uniform distribution. If the sample sizes are large enough, then the DIs and the entropies can be estimated by direct methods. However, when we want to estimate DIs and entropies at disaggregation levels where the sample sizes are small, then we must resort to model-based estimation procedures. In this way, we can take advantage of the auxiliary information embedded in the models and, consequently, obtain more accurate predictions.

The analysis of compositional data using multivariate linear mixed models provides a flexible modeling strategy since it adapts to a hierarchically structured population according to sex, province, and occupation sector. In this way, a PMFH model can be fitted to logratio transformations of the direct

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

estimators of the proportions of men and women in each labor sector by provinces and subsequently incorporate auxiliary information to obtain more efficient predictors. For this, we have considered the alr, clr, and ilr transformations and we have introduced EBP, EBP-in, and plug-in predictors of entropies and DIs. To estimate MSEs, we applied a parametric bootstrap method and recommend using at least $B = 400$ iterations as a good compromise between accuracy and computational time. In the application to real data, we selected PMFH models for the three logratio transformations and observed that the clr-PMFH model had the best diagnostics. That is why we assumed that the model was "correct" and performed all subsequent statistical inferences based on it, including the estimation of the MSEs of the plug-in predictors based on the "incorrect" alr-PMFH and ilr-PMFH models.

However, our study has certain limitations. First, the auxiliary variables used in our model are extracted from surveys, not from censuses or administrative records. To mitigate this situation, we have taken the auxiliary variables from the last 20 SFLS to reduce the variability of the auxiliary information (compared to the target variables) and mitigate the effects of bias and increase in the MSE of the predictions. Second, the matrices of variances of the random effects are assumed to be diagonal. Considering full matrices would require estimating 42 variance parameters, 21 for each sex, which added to the beta coefficients of the model would make a total of 60 parameters. Our study takes $KD = 100$ domains, which means approximately two observations are required to estimate each variance. Third, there are no analytical estimators for MSEs concerning divergences. This is a general problem in the literature due to nonexistence of methodology of theoretical derivation for the MSEs of nonlinear univariate or multivariate indicators. Lastly, the model's fit and diagnosis could be improved, as it exhibits asymmetric residuals. Some of these drawbacks can be resolved through further research, proposed below. However, there are others that are due to the nonavailability of useful auxiliary variables from open-access statistical sources.

In light of these limitations, several avenues for future research emerge. First, it would be valuable to propose PMFH models that incorporate the auxiliary variables' measurement errors. Second, generalizing the model to matrices of random effects with nonzero off-diagonal elements would enhance its applicability. Third, addressing the challenge of analytical estimators for MSEs in multivariate nonlinear indicators is an important SAE research area. Fourth, extending the model to temporal data, considering other multivariable nonlinear indicators, and enhancing the model's fit and diagnosis would be extremely useful. Finally, considering non-linear multivariate area models for non-symmetric distributions deserves further exploration.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

REFERENCES

Aitchison, J. (1982), "The Statistical Analysis of Compositional Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, 44, 139–160.

Arima, S., Bell, W. R., Datta, G. S., Franco, C., and Liseo, B. (2017), "Multivariate Fay-Herriot Bayesian Estimation of Small Area Means under Functional Measurement Error," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180, 1191–1209.

Belov, D. I., and Armstrong, R. D. (2010), "Automatic Detection of Answer Copying via Kullback-Leibler Divergence and k-Index," *Applied Psychological Measurement*, 34, 379–392.

Benavent, R., and Morales, D. (2016), "Multivariate Fay-Herriot Models for Small Area Estimation," *Computational Statistics & Data Analysis*, 94, 372–390.

———. (2021), "Small Area Estimation under a Temporal Bivariate Area-Level Linear Mixed Model with Independent Time Effects," *Statistical Methods & Applications*, 30, 195–222.

Berg, E. J., and Fuller, W. A. (2012), "Estimators of Error Covariance Matrices for Small Area Prediction," *Computational Statistics & Data Analysis*, 56, 2949–2962.

Boubeta, M., Lombardía, M. J., and Morales, D. (2016), "Empirical Best Prediction under Area-Level Poisson Mixed Models," *Test*, 25, 548–569.

———. (2017), "Poisson Mixed Models for Studying the Poverty in Small Areas," *Computational Statistics & Data Analysis*, 107, 32–47.

Burgard, J. P., Esteban, M. D., Morales, D., and Pérez, A. (2020), "A Fay-Herriot Model When Auxiliary Variables Are Measured with Error," *Test*, 29, 166–195.

———. (2021), "Small Area Estimation under a Measurement Error Bivariate Fay-Herriot Model," *Statistical Methods & Applications*, 30, 79–108.

Burgard, J. P., Morales, D., and Wölwer, A.-L. (2022), "Small Area Estimation of Socioeconomic Indicators for Sampled and Unsampled Domains," *ASTA Advances in Statistical Analysis*, 106, 287–314.

Cabella, B. C. T., Sturzbecher, M. J., Tedeschi, W., Baffa, O., and Neves, U. P. C. (2008), "A Numerical Study of the Kullback-Leibler Distance in Functional Magnetic Resonance Imaging," *Brazilian Journal of Physics*, 38, 20–25.

Chambers, R., Salvati, N., and Tzavidis, N. (2016), "Semiparametric Small Area Estimation for Binary Outcomes with Application to Unemployment Estimation for Local Authorities in the UK," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 179, 453–479.

Cover, T. M. (1999), *Elements of Information Theory*, New Jersey, USA: John Wiley & Sons.

Datta, G., Lahiri, P., and Maiti, T. (2002), "Empirical Bayes Estimation of Median Income of Four-Person Families by State Using Time Series and Cross-Sectional Data," *Journal of Statistical Planning and Inference*, 102, 83–97.

Datta, G. S., Lahiri, P., Maiti, T., and Lu, K. L. (1999), "Hierarchical Bayes Estimation of Unemployment Rates for the States of the US," *Journal of the American Statistical Association*, 94, 1074–1082.

Egozcue, J. J., and Pawlowsky-Glahn, V. (2019), "Compositional Data: The Sample Space and Its Structure," *Test*, 28, 599–638.

Erciulescu, A. L., and Fuller, W. A. (2013), "Small Area Prediction of the Mean of a Binomial Random Variable," in *Joint Statistical Meeting 2013 Proceedings - Survey Research Methods Section, Session 37 Small-Area Estimation: Theory and Applications*, pp. 855–863. Available at http://www.asasrms.org/Proceedings/y2013f.html

———. (2019), "Bootstrap Prediction Intervals for Small Area Means from Unit-Level Nonlinear Models," *Journal of Survey Statistics and Methodology*, 7, 309–333.

Erciulescu, A. L., and Opsomer, J. D. (2022), "A Model-Based Approach to Predict Employee Compensation Components," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71, 1503–1520.

Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D., and Pérez, A. (2020), "Small Area Estimation of Proportions under Area-Level Compositional Mixed Models," *Test*, 29, 793–818.

Esteban, M. D., Morales, D., Pérez, A., and Santamaría, L. (2012), "Small Area Estimation of Poverty Proportions under Area-Level Time Models," *Computational Statistics & Data Analysis*, 56, 2840–2855.

Fabrizi, E., Ferrante, M. R., and Trivisano, C. (2016), "Bayesian Beta Regression Models for the Estimation of Poverty and Inequality Parameters in Small Areas," in *Analysis of Poverty Data by Small Area Methods*, ed. P. Monica, New York: Wiley.

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, 269–277.

Ferrante, M., and Trivisano, C, et al. (2010), "Small Area Estimation of the Number of Firms' Recruits by Using Multivariate Models for Count Data," *Survey Methodology*, 36, 171–180.

Franco, C., and Bell, W. R. (2015), "Borrowing Information over Time in Binomial/Logit Normal Models for Small Area Estimation," *Statistics in Transition New Series*, 16, 563–584.

Froud, H., Benslimane, R., Lachkar, A., and Ouatik, S. A. (2010), "Stemming and Similarity Measures for Arabic Documents Clustering," in *2010 5th International Symposium on I/V Communications and Mobile Network*, pp. 1–4, IEEE.

Ghosh, M., Nangia, N., and Kim, D. H. (1996), "Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach," *Journal of the American Statistical Association*, 91, 1423–1431.

González-Manteiga, W., Lombarda, M., Molina, I., Morales, D., and Santamaría, L. (2010), "Small Area Estimation under Fay–Herriot Models with Non-Parametric Estimation of Heteroscedasticity," *Statistical Modelling*, 10, 215–239.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2008), "Analytic and Bootstrap Approximations of Prediction Errors under a Multivariate Fay–Herriot Model," *Computational Statistics & Data Analysis*, 52, 5242–5252.

Guha, S., and Chandra, H. (2024), "Small Area Estimation under a Spatially Correlated Multivariate Area-Level Model," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187, 62–84.

Hall, P., and Maiti, T. (2006), "On Parametric Bootstrap Methods for Small Area Prediction," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68, 221–238.

Huang, E. T., and Bell, W. R. (2004), "An Empirical Study on using ACS Supplementary Survey Data in SAIPE State Poverty Models," in *2004 Proceedings of the American Statistical Association*, pp. 3677–3684, US Bureau of the Census, Washington DC.

Krause, J., Burgard, J. P., and Morales, D. (2022), "L2-Penalized Approximate Likelihood Inference in Logit Mixed Models for Regional Prevalence Estimation under Covariate Rank-Deficiency," *Metrika*, 85, 459–489.

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86.

Li, Y., and Wang, L. (2008), "Testing for Homogeneity in Mixture Using Weighted Relative Entropy," *Communications in StatisticsSimulation and Computation*, 37, 1981–1995.

Lin, X., Pittman, J., and Clarke, B. (2007), "Information Conversion, Effective Samples, and Parameter Size," *IEEE Transactions on Information Theory*, 53, 4438–4456.

López-Vizcaíno, E., Lombardía, M. J., and Morales, D. (2013), "Multinomial-Based Small Area Estimation of Labour Force Indicators," *Statistical Modelling*, 13, 153–178.

———. (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: Statistics in Society*, 178, 535–565.

Marhuenda, Y., Morales, D., and del Carmen Pardo, M. (2014), "Information Criteria for Fay–Herriot Model Selection," *Computational Statistics & Data Analysis*, 70, 268–280.

Marhuenda, Y., Morales, D., and Pardo, M. (2016), "Tests for the Variance Parameter in the Fay–Herriot Model," *Statistics*, 50, 27–42.

Menéndez, M., Morales, D., Pardo, L., and Salicrú, M. (1995a), "Asymptotic Behaviour and Statistical Applications of Divergence Measures in Multinomial Populations: A Unified Study," *Statistical Papers*, 36, 1–29.

Menéndez, M., Morales, D., Pardo, L., and Vajda, I. (1995b), "Divergence-Based Estimation and Testing of Statistical Models of Classification," *Journal of Multivariate Analysis*, 54, 329–354.

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

Molina, I., and Rao, J. N. (2010), "Small Area Estimation of Poverty Indicators," *Canadian Journal of Statistics*, 38, 369–385.

Molina, I., Saei, A., and José Lombardía, M. (2007), "Small Area Estimates of Labour Force Participation under a Multinomial Logit Mixed Model," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 170, 975–1000.

Morales, D., Esteban, M. D., Pérez, A., and Hobza, T. (2021), "A Course on Small Area Estimation and Mixed Models," in Methods, Theory and Applications in R, Switzerland: Springer.

Morales, D., Pagliarella, M. C., and Salvatore, R. (2015), "Small Area Estimation of Poverty Indicators under Partitioned Area-Level Time Models," *SORT*, 39, 19–34.

Morales, D., Pardo, L., Salicrú, M., and Menéndez, M. (1994), "Asymptotic Properties of Divergence Statistics in a Stratified Random Sampling and Its Applications to Test Statistical Hypotheses," *Journal of Statistical Planning and Inference*, 38, 201–221.

Morales, D., Pardo, L., and Vajda, I. (2000), "Rényi Statistics in Directed Families of Exponential Experiments," *Statistics: A Journal of Theoretical and Applied Statistics*, 34, 151–174.

Pardo, L. (2018), *Statistical Inference Based on Divergence Measures*, Florida, USA: Chapman and Hall/CRC.

Pardo, L., Morales, D., Salicrú, M., and Menéndez, M. (1997), "Large Sample Behavior of Entropy Measures When Parameters Are Estimated," *Communications in Statistics-Theory and Methods*, 26, 483–501.

Pawlowsky-Glahn, V., and Buccianti, A. (2011), *Compositional Data Analysis*, London: John Wiley & Sons.

Pfeffermann, D., and Burck, L. (1990), "Robust Small Area Estimation Combining Time Series and Cross-Sectional Data," *Survey Methodology*, 16, 217–237.

Porter, A. T., Wikle, C. K., and Holan, S. H. (2015), "Small Area Estimation via Multivariate Fay–Herriot Models with Latent Spatial Dependence," *Australian & New Zealand Journal of Statistics*, 57, 15–29.

Pratesi, M. (2016), *Analysis of Poverty Data by Small Area Estimation*, New York: John Wiley.

Rao, J. N., and Molina, I. (2015), *Small Area Estimation*, Hoboken: Wiley.

Rao, J. N., and Yu, M. (1994), "Small-Area Estimation by Combining Time-Series and Cross-Sectional Data," *Canadian Journal of Statistics*, 22, 511–528.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003), *Model Assisted Survey Sampling*, Berlin: Springer Science & Business Media.

Singh, B. B., Shukla, G. K., and Kundu, D. (2005), "Spatio-Temporal Models in Small Area Estimation," *Survey Methodology*, 31, 183.

Souza, D. F., and Moura, F. A. (2016), "Multivariate Beta Regression with Application in Small Area Estimation," *Journal of Official Statistics*, 32, 747–768.

Symeonaki, M. (2022), "A Relative Entropy Measure of Divergences in Labour Market Outcomes by Educational Attainment," in *Quantitative Methods in Demography: Methods and Related Applications in the Covid-19 Era*, New York: Springer International Publishing, pp. 351–358.

Ubaidillah, A., Notodiputro, K. A., Kurnia, A., and Mangku, I. W. (2019), "Multivariate Fay-Herriot Models for Small Area Estimation with Application to Household Consumption per Capita Expenditure in Indonesia," *Journal of Applied Statistics*, 46, 2845–2861.

Vilhena, D. A., Foster, J. G., Rosvall, M., West, J. D., Evans, J., and Bergstrom, C. T. (2014), "Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication," *Sociological Science*, 1, 221–238.

Volkau, I., Prakash, K. B., Ananthasubramaniam, A., Aziz, A., and Nowinski, W. L. (2006), "Extraction of the Midsagittal Plane from Morphological Neuroimages Using the Kullback–Leibler's Measure," *Medical Image Analysis*, 10, 863–874.

Ybarra, L. M., and Lohr, S. L. (2008), "Small Area Estimation When Auxiliary Information Is Measured with Error," *Biometrika*, 95, 919–931.

You, Y., and Rao, J. (2000), "Hierarchical Bayes Estimation of Small Area Means Using Multi-Level Models," *Survey Methodology*, 26, 173–181.

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

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Association for Public Opinion Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Journal of Survey Statistics and Methodology, 2024, 12, 1531–1566
https://doi.org/10.1093/jssam/smae023
Article

Downloaded from academic.oup.com/jssam/article/12/5/1531/7696003 by UNIVERSIDAD MIGUEL HERNANDEZ-BIBLIOTECA user on 21 September 2026