Social Indicators Research
https://doi.org/10.1007/s11205-024-03393-w

ORIGINAL RESEARCH

Model-Based Estimation of Small Area Dissimilarity Indexes: An Application to Sex Occupational Segregation in Spain

María Bugallo[1] · Domingo Morales[1] · María Dolores Esteban[1] · Maria Chiara Pagliarella[2]

Accepted: 1 July 2024

Abstract
This paper introduces a new statistical methodology for estimating Duncan dissimilarity indexes of occupational segregation by sex in administrative areas and time periods. Given that direct estimators of the proportion of men (or women) in the group of employed people for each occupational sector are not accurate enough in the considered estimation domains, we fit to them a three-fold Fay–Herriot model with random effects at three hierarchical levels. Based on the fitted area-level model, empirical best predictors of the cited proportions and Duncan segregation indexes are derived. A parametric bootstrap algorithm is implemented to estimate the mean squared error. Some simulation studies are included to show how the proposed predictors have a good balance between bias and mean squared error. Data from the Spanish Labour Force Survey are used to illustrate the performance of the new statistical methodology and to give some light about the current state of sex occupational segregation by province in Spain. Research claims that there is a sex gap that persists despite advances in the inclusion of women in the labour market in recent years and that is related to the unequal sharing of family responsabilities and the stigmas still present in modern societies.

Keywords Small area estimation · Fay–Herriot model · Labour force survey · Duncan segregation index · Sex occupational segregation

Mathematics Subject Classification 62D05 · 62P25 · 62J05

1 Introduction

The Duncan segregation index (DSI) is a dissimilarity index proposed by Duncan and Duncan (1955) to measure segregation by location and, because of its properties and ease of calculation, is widely used in sociological studies. The DSI is a measure of segregation

✉ María Bugallo
mbugallo@umh.es

[1] Operations Research Center, Miguel Hernández University of Elche, Elche, Spain

[2] Dipartimento di Economia e Giurisprudenza, Universita degli Studi di Cassino e del LazioMeridionale, Cassino, Italy

Published online: 17 July 2024

that can be applied to individuals differentiated by a dichotomous classification variable in groups defined by sex, race, origin, religion or culture, among others. Locations should be interpreted in a broad sense. Examples of locations are residential areas, education levels or occupation sectors. The current research explores the use of the DSI to measure occupational segregation by sex, where the group variable is sex and the location variable is occupation sector.

In this paper, the DSI measures the dissimilarity between the higher-than-expected presence of men (women) over women (men) in different sectors of labour occupation. To do so, it compares the percentage of men and women employed in each occupation sector and provides a numerical value that is lower the closer the occupational distribution is to equality. If all job sectors have equal proportions of employed men and women, the DSI becomes zero. On the opposite side, the DSI becomes one and segregation reaches its maximum.

Many authors have explored the properties of the DSI or have applied it in sociological studies. Among the most important contributions, Taeuber and Taeuber (1965) analyzed the segregation of the Afro-American population in Chicago neighborhoods; Reardon and Firebaugh (2002) used the DSI to measure inequalities and interpreted it as a relative mean deviation; Reardon and O'Sullivan (2004) reviewed the properties of the DSI and proposed alternative indicators; Alonso-Villar and Del Río (2010) presented data on occupational segregation by sex in Spain; Roberto (2016) analyzed data on race segregation in US cities; Salardi (2016) investigated the evolution of gender and racial occupational segregation across labour markets in Brazil; and Das and Kotikula (2019) discussed the factors that drive gender-based employment segregation.

As a general feature, the studies cited above assume that the available information is completely reliable. In practice, data may come from surveys and are, therefore, subject to sampling errors. If data come from administrative registers or surveys with large sample sizes, the DSI calculation is straightforward. Nevertheless, direct estimation techniques may be unreliable if the sample sizes are small, and this is a problem that merits methodological research.

Obtaining DSI estimates for small areas, and over time, can be done using labor force survey data. In order to do so, we can calculate direct estimators of the proportions of men and women among the employed population in each sector and then plug these estimators into the DSI formula. However, if domain sample sizes are small, direct estimators will not be sufficiently precise. Consequently, they will not produce good DSI estimates. In fact, direct estimators are calculated using only data from the domain of the dependent variable and the corresponding sampling weights. On the other hand, they have good bias and variance properties under the sampling design distribution, when the sample sizes are large enough. Otherwise, they are not precise and it is desirable to use model-based estimators to reduce variability by introducing auxiliary information in the inferential process.

Small area estimation (SAE) methodologies provide more reliable granular level estimates by fitting statistical models to unit-level or area-level data and obtaining predictors based on them. This is the usual way to incorporate additional information from other domains, auxiliary variables and hierarchical, spatial or temporal data dependency structures. As a result, indiscriminate increases in survey sample sizes are avoided. The term "*small area*" is commonly used to refer to a small geographic area or a subgroup of the population defined according to some combination of socio-demographic characteristics but where, in any case, direct estimation is not accurate enough due to the smallness of the sample size. For instance, if a survey is designed to obtain precise direct estimates at national level and results disaggregated by region, or for a certain minority group, are of interest, these unplanned estimation domains are called small areas. Budget and lack of planning in the initial design are the most common causes of small or even zero sample sizes, prompting the need for further statistical research.

Rao and Molina (2015), Pratesi (2016) and Morales et al. (2021) provide an introduction to SAE and, in particular, to the area-level model-based approach derived from the seminal paper of Fay and Herriot (FH) (1979). In this regard, area-level linear mixed models, adapted to hierarchical structures or temporal correlations, offer a solution to this problem. A baseline example is the estimation of domain totals, means and proportions, using empirical best linear unbiased predictors (EBLUP) based on the FH model.

Regarding the generalizations of the FH model, some temporal extensions were given by Pfeffermann and Burck (1990), Rao and Yu (1994), Ghosh et al. (1996), Datta et al. (1999, 2002), You and Rao (2000) and Singh et al. (2005). Marhuenda et al. (2016) propose tests for the variance parameter in the FH model. In addition, some extensions of the FH model assuring estimates in the interval [0, 1] have been proposed in the literature. For example LMMs with suitable transformations (González-Manteiga et al., 2002) and Beta regression models (Janicki, 2020). Concerning the estimation of the proportions, Esteban et al. (2012), Marhuenda et al. (2013, 2014) and Morales et al. (2015) have proposed predictors based on variants of the FH model. Multivariate FH models were studied by Huang and Bell (2004), Porter et al. (2015), González-Manteiga et al. (2008) and Benavent and Morales (2016, 2021). Under area-level Poisson, binomial or multinomial regression mixed models, predictors of counts and proportions were introduced by Boubeta et al. (2016, 2017), Burgard et al. (2021, 2022), Krause et al. (2022), Morales et al. (2022), López-Vizcaíno et al. (2013, 2015) and Chambers et al. (2016). All in all, the above non exhaustive collection of relevant papers have in common that they have introduced area-level SAE methodology for predicting domain proportions and totals. This is a partial step that this paper also deals with.

Nevertheless, the scientific literature also presents numerous contributions to SAE based on unit-level models. Molina and Rao (2010) have proposed empirical best predictors (EBP) based on a nested error regression model (NER). Marhuenda et al. (2017), Guadarrama et al. (2022) and Esteban et al. (2022) have extended the EBP approach to two-fold, temporal and multivariate NERs. Hobza and Morales (2016) and Hobza et al. (2018) have fitted a logit mixed model to unit-level poverty data and derived EBPs for poverty proportions. Tzavidis et al. (2008), Marchetti et al. (2012) and Tzavidis et al. (2015) have introduced robust estimators by using a quantile regression approach. Against this background, the EBP approach and the quantile regression procedures can be applied to predict domain indicators defined by non-linear transformations of means, totals and proportions. Thus, they represent alternative methodologies to the one proposed in this manuscript based on area-level models.

We have considered the extension of the two-fold Fay–Herriot (FH2) model to three levels of hierarchy. At this regard, the FH2 model has been introduced by Rao and Yu (1994) and studied by Esteban et al. (2012), Marhuenda et al. (2013) and Morales et al. (2015), among others. It is a model adapted to area-level data indexed by areas and subareas. The three-fold Fay–Herriot (FH3) model, recently proposed by Marcis et al. (2023), can further describe data structured in areas, subareas and time periods or subsubareas. This is the case of the employment data used to estimate sex segregation by province, occupation sector and time period. The proposed area-level methodology introduces sample weights into the inferential process, taking into account the sampling distribution. Under the FH3 model, Krenzke et al. (2020) have estimated adult literacy of US counties and Cai and Rao (2022) have studied some variable selection methods. Based on the FH3 model, we have obtained

EBLUPs of male and female domain proportions in the set of employed for each occupational sector and derived DSI predictors. We have not used a specific model for proportions because we are interested in using a three-fold nested model that allows the population to be hierarchised in provinces, occupational sectors and time periods. Indeed, our contribution focuses on the estimation of DSIs. So as to estimate the corresponding mean squared errors (MSE), we have introduced a parametric bootstrap algorithm by following Hall and Maiti (2006) and González-Manteiga et al. (2008, 2010). The bootstrap procedure also allows studying the loss of precision of model-based predictors when using added auxiliary variables obtained with sampling errors.

As for the issue at hand, modern societies promote the fair treatment and legal protection of women and minority groups, and governments look for places where systemic discrimination occurs. This motivates research into statistical methodologies for mapping segregation at different levels of aggregation. In this sense, we introduce a new statistical methodology for mapping DSIs and present an application to data from the Spanish Labour Force Survey (SLFS). This can be of great help to policy makers to decide where to implement specific equality policies. According to the EUROSTAT 2016 NUTS classification, the SLFS is designed to obtain reliable direct estimators in NUTS 3 regions (provinces). However, SLFS sample sizes are rather small in provinces crossed by occupational sectors. At this regard, our study aims to estimate DSIs by province from the last quarter of 2020 to the last quarter of 2021, both included. Other papers also modelling SLFS data include Baíllo and Molina (2009) and Herrador et al. (2011).

The rest of the paper is organized as follows. Section 2 introduces the data, the DSI and the SAE problem. Section 3 presents the FH3 model, the residual maximum likelihood estimators (REML) of the model parameters, the EBLUPs of the domain proportions of employed men and women, the DSI predictors and the MSE estimators. In a simulation study we may be interested in: (1) analysing the effect of increasing the sample size or the number of domains (among other elements), or (2) testing the behaviour of estimators and predictors in scenarios close to that of the application to real data. While we have addressed both issues, we have moved the latter to Supplementary Material. Thus, in order to learn from the simulations, the usual section order has been reversed as follows. Section 4 includes some simulation experiments to investigate the performance of the DSI predictors and MSE estimators. Sect. 5 deals with the application to real data. Last but not least, Sect. 6 provides some relevant conclusions. The paper contains Supplementary Material organised in two sections. Section A describes the steps of the simulation experiments and presents additional results based on artificial data. Section B provides tables with supplementary information about the application to real data.

2 Data and Dissimilarity Indexes

This paper introduces and applies the SAE methodology needed to estimate sex occupational segregation by Spanish province and time period. It uses data from the SLFS, which is a quarterly survey that follows a two-stage stratified random sampling approach to draw samples from each province. Primary sampling units are census sections, which are geographical areas with around 500 dwellings or approximately 3000 people. Census sections are grouped into strata according to the size of the municipality to which they belong. Secondary sampling units are dwellings, and all individuals aged 16 or over in the selected dwelling are interviewed. Here, it is important to bear in mind that we are only interested in the subpopulation of employed respondents. Furthermore, as we have included the autonomous cities Ceuta and Melilla in the set of provinces, $D = 52$. Finally, we have chosen $T = 5$ consecutive quarterly periods: 2020.4, 2021.1, 2021.2, 2021.3 and 2021.4. The occupation sector (OC) variable has been derived from the Spanish National Classification of Occupations (CNO2011). Three categories have, however, been aggregated due to the smallness of the sample sizes.

Table 1 describes the encoding of the OC variable, which has $R = 7$ mutually exclusive categories. The aggregated occupations are denoted with an .*, grouping the most similar ones based on their definition. The main reason is to avoid zero or very small sample sizes, where even suitable statistical models do not provide reliable results. All in all, the final set of categories covers a broad spectrum of jobs, achieving an accurate level of knowledge about the respondents in terms of their main occupation.

In the following, we will introduce some mathematical notation to define the DSI across provinces and time periods. Let U_{drt} be a subset (estimation domain) of the population, relative to time period t and conformed by N_{drt} employed people aged 16 or over, resident in province d and working in sector r. Let y_{drt1j} be a dichotomic variable such that $y_{drt1j} = 1$ if the individual j of U_{drt} is male, and $y_{drt1j} = 0$, otherwise. Let $y_{drt2j} = 1 - y_{drt1j}$ be the analogous variable for females. The population means of these variables are

$$\overline{Y}_{drt1} = \frac{1}{N_{drt}} \sum_{j=1}^{N_{drt}} y_{drt1j}, \quad \overline{Y}_{drt2} = \frac{1}{N_{drt}} \sum_{j=1}^{N_{drt}} y_{drt2j}. \tag{2.1}$$

For $d = 1, \dots, D$, $t = 1, \dots, T$, the DSI of province d at time period t is

$$S_{d.t} = \frac{1}{2} \sum_{r=1}^{R} S_{drt}, \quad S_{drt} = \left| \frac{N_{drt}\overline{Y}_{drt1}}{\sum_{i=1}^{R} N_{dit}\overline{Y}_{dit1}} - \frac{N_{drt}\overline{Y}_{drt2}}{\sum_{i=1}^{R} N_{dit}\overline{Y}_{dit2}} \right|, \quad r = 1, \dots, R. \tag{2.2}$$

In our research, the DSI quantifies sex occupational segregation and measures how evenly (or unevenly) the population of both sexes is distributed in each occupational sector. Segregation is measured as the degree to which the spatial distribution of the female group deviates from that of the male one. If men and women are distributed in equal proportions across occupation sectors, there is no segregation. Therefore, the DSI has a straightforward interpretation: it corresponds to the proportion of women (or men) who would have to move to another occupational sector to balance the distribution. Accordingly, movements

Table 1 Encoding of OC. The .* means that the category has been aggregated

Code	Description
OC1	Directors and managers. Senior public and private figures
OC2	Scientists and intellectual technicians and professionals
OC3*	(i) Military occupations. (ii) Technicians and support staff
OC4	Accounting, administrative and other office employees
OC5	Catering, protection and commercial staff
OC6*	(i) Unskilled workers. (ii) Primary sector workers
OC7*	(i) Plant and machinery operators. (ii) Craftsmen and skilled workers in the manufacturing and construction industries.

Table 2 Percentiles of the sample sizes and sampling fractions in the 2020.4−2021.4 SLFS data

	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
SS	6	34	56	82	104	121	143	175	218	297	1,013
SF (in %)	0.091	0.202	0.249	0.290	0.350	0.407	0.469	0.538	0.652	0.811	1.779

Table 3 Employed men and women by occupation sector in the SLFS of 2021.4

Occupation sector		OC1	OC2	OC3	OC4	OC5	OC6	OC7
Men	Total	11,023	31,816	29,101	12,720	31,845	27,798	63,984
	Proportion	0.6981	0.3952	0.6311	0.3172	0.3803	0.5083	0.9058
Women	Total	5,191	44,518	17,947	28,116	48,398	29,021	6,694
	Proportion	0.3019	0.6048	0.3689	0.6828	0.6197	0.4917	0.0942

would have to occur from occupations in which the group is overrepresented to occupations in which it is underrepresented.

In practice, the theoretical DSI values defined in (2.2) should be estimated by using SLFS sample data. However, our estimation domains are not planned in the SLFS so we first investigate whether the sample sizes are large enough to provide accurate direct estimates of the dissimilarities S_{drt}'s. For this sake, we will introduce some additional notation. Let n_{drt} and $\hat{N}_{drt}^{dir}$ be the sample size and the estimated population size (sum of the sampling weights) of U_{drt}. Let $n = \sum_{d=1}^{D} \sum_{r=1}^{R} \sum_{t=1}^{T} n_{drt}$ be the global sample size. The estimated sampling fractions (in %), are defined as relative sample sizes as

$$f_{drt} = 100 \frac{n_{drt}}{\hat{N}_{drt}^{dir}}, \quad d = 1, \dots, D,\ r = 1, \dots, R,\ t = 1, \dots, T, \tag{2.3}$$

and are not uniformly distributed in U_{drt}. The latter is shown in Table 2, which presents the deciles of the sample sizes (SS) and the estimated sampling fractions (SF).

It can be observed in Table 2 that 20% (50%) of the U_{drt}'s have samples sizes smaller than 56 (121) and that the average sample size, equal to 149, is between $q_{0.6} = 143$ and $q_{0.7} = 175$, indicating that the sample size distribution is positively skewed. Furthermore, sampling fractions allow us to know the percentage of individuals of the subsets U_{drt} who actually belong to the sample. As they are all lower than 1.779, the representativeness of the samples in the crosses is quite small. Consequently, this is a SAE problem and direct estimators, such as the Hájek estimator, are not accurate enough. Therefore, the inference problem requires the incorporation of more sophisticated prediction methods.

Table 3 contains the total and proportion of men and women in the subset of employees, by main occupation, for the SLFS2021.4 data. Analytically studying occupational segregation by sex, categories OC7 and, to a lesser extent, OC1 stand out.

The Hájek estimators of $\overline{Y}_{drt1}$ and $\overline{Y}_{drt2}$ are direct estimators that are calculated by using only data of the SLFS sample s_{drt} of the subset U_{drt} and the sampling weights w_{drt} 's. They are therefore ratios between two quantities, $\hat{Y}_{drt\kappa}^{dir}$ and $\hat{N}_{drt}^{dir}$, given by

$$\hat{\overline{Y}}^{dir}_{drt\kappa} = \frac{\hat{Y}^{dir}_{drt\kappa}}{\hat{N}^{dir}_{drt}} = \frac{\sum_{j\in s_{drt}} w_{drtj} y_{drt\kappa j}}{\sum_{j\in s_{drt}} w_{drtj}}, \quad \kappa = 1, 2. \tag{2.4}$$

Fig. 1 (left) plots the Hájek estimates of the proportion of men who were employed for the SLFS of 2021.4. The estimated proportions are sorted by occupation sector and province, so that the aggregated data file, at time period $t = 5$, is organized into 7 occupation sectors and 52 provinces. The dotted line $y = 0.5$, that corresponds to the equal distribution between employed men and women, is added. It can be seen that there are main occupations for which a higher proportion of employed men are expected to be found. Again, categories OC1 and OC7 stand out. Figure 1 (center) confirms it definitively. Compared to Table 3, these estimates follow the same trend as the sample data. However, they are too inaccurate in terms of a SAE problem. In fact, the coefficients of variation (CV) of the Hájek estimator take really large values when the sample size is small, decreasing as it increases, as can be seen in Fig. 1 (right). Moreover, they have been calculated assuming unbiasedness (see Morales et al. (2021), Chap. 3), which gives them an additional advantage. The proposed SAE methodology will allow, without assuming unbiasedness in the error estimation, to reduce the variability of the Hájek estimates. Indeed, we will provide comparable error measures that will be lower than the CVs of the Hájek estimator for small and moderate sample sizes.

As an extreme value in terms of the CV of the Hájek estimates, we should mention Melilla and the OC6 category, with an estimated male ratio of 0.10 and a standard deviation close to 0.095. However, its sample size is 10 and, therefore, not large enough to provide accurate direct estimates. This observation has been removed from Fig. 1 (right) for aesthetics.

To overcome the lack of precision of the Hájek estimator, we will consider auxiliary information, hierarchical structures and model-based predictors, which are the ones that drive our research. It must be said that national statistical offices have censuses and/or administrative files, so they are able to use auxiliary variables measured without error. Nevertheless, the access is often restricted, so the scientific community is forced to fit measurement error models or to use estimates for the aggregated auxiliary information as population values. As for the former, measurement error models are rather sophisticated because they are not LMMs and their study should be investigated elsewhere. For more information, Hariyanto et al. (2018) provides a comprehensive and up-to-date account of these models in the context of SAE.

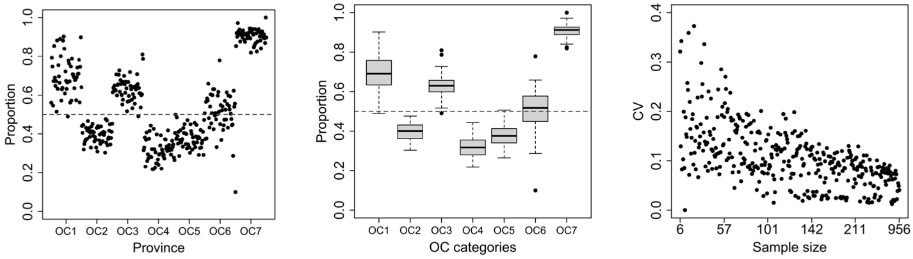


Fig. 1 Hájek estimates of the proportion of employed men (left–center) by OC category and CVs (right) sorted by sample size. Data from the SLFS of 2021.4

In our research, the selected auxiliary variables are Hájek estimates of the proportion of individuals in U_{drt} that belong to the categories of the following factors:

Age group, with 3 categories: between 16 and 30 years (age3–1), between 30 and 50 years (age3–2) and over 50 years (age3–3).
Citizenship, with 2 categories: Spanish (cit1) and not Spanish (cit2).
Education, with 4 categories: primary or less (edu1), basic secondary education (edu2), advanced secondary education (edu3) and higher education, such as university (edu4).
Working hours, with 2 categories: full-time (work1) and part-time work (work2).
Professional status, with 5 categories: self-employed (st1), cooperative or family business (st2), public (st3) and private (st4) sector salaried employee and others (st5).

The set of categories of each factor is exhaustive, so the estimated proportions sum one. Based on their socioeconomic meaning, we have limited to 11 auxiliary variables defined at the level of the U_{drt} subsets. Particularly, we have removed age3–2, cit2, edu2, work2 and st5. First of all, cit1 and work1 are complementary to cit2 and work2, respectively, so the choice of the former or the latter is of little interest. As for age3–2, it represents the intermediate category, so we consider it more informative to include the age variables that account for the two edge groups, which to some extent also applies to edu2. Finally, we dropped st5, defined as "*others*", for being the most ambiguous variable to account for professional status.

For the sake of accuracy, we jointly use data from the last five SLFSs to estimate the covariates for each quarter and the population sizes N_{drt} used to calculate the DSI values in (2.2). Therefore, the effects of the variances of the covariate means and population sizes in the properties of the prediction procedure are considered negligible. This allows for an approximate 5-fold increase in available data and reduces temporal variability. In simulation experiments in Sect. 4 we empirically verify that this does not lead to underestimating the final variability. As an example, the vector of covariates for $t = 1$ (last quarter of 2020) is estimated using the SLFS data from 2019.4 to 2020.4, both surveys included. Table 4 compares the quartiles of the variances of the Hájek estimates of the selected auxiliary variables with those of the response variable.

All area-level variables being proportions, it is safe to say that the variability of the covariates is significantly lower than that of $\hat{\overline{Y}}^{dir}_{drt1}$ and close to zero. In addition, the elevation factors are the inverses of the inclusion probabilities, which are deterministic, after a calibration process whose randomness is minimal. Therefore, the population sizes estimated as sums of elevation factors have negligible variability. In order to take advantage of these auxiliary data to refine the estimation of the proportion of men and women who are employed in each occupation sector, and to obtain the DSI predictions by province and time period, Sect. 3 details the FH3 model.

3 FH3 Model-Based Statistical Methodology

3.1 Model

This section describes a three-fold Fay–Herriot (FH3) model. It is worth recall that the FH3 model for the SAE of domain linear indicators has recently been introduced by Marcis et al.

Table 4 Quartiles of the variances of the Hájek estimates of the selected auxiliary variables and the response variable. Data from the 2020.4−2021.4 SLFS

	age3–1	age3–3	cit1	edu1	edu3	edu4	work1	st1	st2	st3	st4	$\hat{\overline{Y}}^{dir}_{drt1}$
$q_{0.25}$	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0012
$q_{0.5}$	0.0008	0.0007	0.0004	0.0001	0.0008	0.0007	0.0005	0.0005	0.0000	0.0005	0.0009	0.0022
$q_{0.75}$	0.0013	0.0012	0.0008	0.0003	0.0014	0.0016	0.0008	0.0009	0.0000	0.0011	0.0016	0.0036

(2023). Therefore, we have adapted their methodology, being our contribution the prediction of DSIs and related inference issues. Here, the response variable is the Hájek estimator of the proportion of men in each estimation domain U_{drt}, defined in (2.1). The FH3 model is defined in two steps, with the simplified notation $y_{drt} = \hat{\overline{Y}}^{dir}_{drt1}$ and $\mu_{drt} = \overline{Y}_{drt1}$. The first step starts from the sampling model, indicating that y_{drt} is an unbiased estimator of μ_{drt}, i.e.

$$y_{drt} = \mu_{drt} + e_{drt},\ \ e_{drt} \sim N(0, \sigma^2_{drt}),\ \ \sigma^2_{drt} > 0,\ \ d = 1, \dots,\ r = 1, \dots, R,\ t = 1, \dots, T, \tag{3.1}$$

where the error variances are assumed to be known.

The selection of σ^2_{drt} is worthy of comment. In practice, we use the generalized variance function (GVF) method to calculate σ^2_{drt}. For this purpose, a regression model is fitted to the direct estimates of the design-based variance of y_{drt}, $\hat{\sigma}^{dir,2}_{drt}$, obtained in advance from the unit-level survey data. See e.g. Remark 2.3 in Morales et al. (2021). Following Section 16.4 in Morales et al. (2021), we define the log-linear model

$$\log(\hat{\sigma}^{dir,2}_{drt}) = b_0 + b_1 y_{drt} + b_2 n_{drt} + \varepsilon_{drt}, \tag{3.2}$$

where the ε_{drt}'s are i.i.d. $N(0, \sigma^2_A)$ and $\sigma^2_A > 0$. Intuitively, b_1 is expected to be positive and b_2 negative. The final σ^2_{drt} equals the variance values predicted by the GVF model (3.2), i.e.

$$\sigma^2_{drt} = \exp(\hat{\sigma}^2_A/2) \cdot \exp\big(\hat{b}_0 + \hat{b}_1 y_{drt} + \hat{b}_2 n_{drt}\big), \tag{3.3}$$

where the factor $\exp(\hat{\sigma}^2_A/2)$ is the usual bias correction term in a log-linear regression analysis to prevent underestimation. This allows for the smoothing of the direct estimates $\hat{\sigma}^{dir,2}_{drt}$.

In a second step, a linking model is constructed assuming a hierarchical linear relationship between μ_{drt} and a row vector $\boldsymbol{x}_{drt}$ of p auxiliary variables, i.e.

$$\mu_{drt} = \boldsymbol{x}_{drt}\boldsymbol{\beta} + u_{1,d} + u_{2,dr} + u_{3,drt},\ \ d = 1, \dots,\ r = 1, \dots, R,\ t = 1, \dots, T, \tag{3.4}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ row vector of regression parameters, $u_{1,d} \sim N(0, \sigma^2_1)$, $u_{2,dr} \sim N(0, \sigma^2_2)$, $u_{3,drt} \sim N(0, \sigma^2_3)$ and $\sigma^2_1,\ \sigma^2_2,\ \sigma^2_3 > 0$ are variance parameters. We further assume independence between errors and random effects.

The FH3 model is a linear mixed model that can be expressed in the single form

$$y_{drt} = \boldsymbol{x}_{drt}\boldsymbol{\beta} + u_{1,d} + u_{2,dr} + u_{3,drt} + e_{drt},\quad d = 1, \dots, D,\ \ r = 1, \dots, R,\ \ t = 1, \dots, T. \tag{3.5}$$

For $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (\sigma^2_1, \sigma^2_2, \sigma^2_3)$, the REML log-likelihood function is

$$l_{reml}(\boldsymbol{\theta}) = -\frac{DRT - p}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{X}| - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y},$$

where the column and diagonal operators define the vectors and matrices

$$\begin{aligned}
\boldsymbol{X} &= \underset{1\le d\le D}{\mathrm{col}}\Big(\underset{1\le r\le R}{\mathrm{col}}\Big(\underset{1\le t\le T}{\mathrm{col}}(\boldsymbol{x}_{drt})\Big)\Big), \quad \boldsymbol{V}_e = \underset{1\le d\le D}{\mathrm{diag}}\Big(\underset{1\le r\le R}{\mathrm{diag}}\Big(\underset{1\le r\le R}{\mathrm{diag}}(\sigma^2_{drt})\Big)\Big),\\
\boldsymbol{V} &= \sigma^2_1 \underset{1\le d\le D}{\mathrm{diag}}(\boldsymbol{1}_{RT}\boldsymbol{1}'_{RT}) + \sigma^2_2 \underset{1\le d\le D}{\mathrm{diag}}\Big(\underset{1\le r\le R}{\mathrm{diag}}(\boldsymbol{1}_T\boldsymbol{1}'_T)\Big) + \sigma^2_3\boldsymbol{I}_{DRT} + \boldsymbol{V}_e,\\
\boldsymbol{y} &= \underset{1\le d\le D}{\mathrm{col}}\Big(\underset{1\le r\le R}{\mathrm{col}}\Big(\underset{1\le t\le T}{\mathrm{col}}(y_{drt})\Big)\Big), \quad \boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1},
\end{aligned}$$

and $\mathbf{1}_m$ and $\boldsymbol{I}_m$ denote the $m \times 1$ vector of ones and the $m \times m$ identity matrix, respectively. The REML estimators of the variance components, $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$, are obtained by maximizing $l_{reml}(\boldsymbol{\theta})$. We apply the Fisher-scoring algorithm with updating equation

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \boldsymbol{F}^{-1}(\boldsymbol{\theta}^{(k)})\boldsymbol{S}(\boldsymbol{\theta}^{(k)}), \tag{3.6}$$

where $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{\theta}) = (S_1, S_2, S_3)'$ is the score vector and $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{\theta}) = \big(F_{ab}\big)_{a,b=1,2,3}$ is the Fisher information matrix. For $a, b = 1, 2, 3$, the components of $\boldsymbol{S}$ and $\boldsymbol{F}$ are

$$S_a = \frac{\partial l_{reml}}{\partial \theta_a} = -\frac{1}{2}\mathrm{tr}(\boldsymbol{P}\boldsymbol{V}_a) + \frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{V}_a\boldsymbol{P}\boldsymbol{y}, \quad F_{ab} = \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\boldsymbol{V}_a\boldsymbol{P}\boldsymbol{V}_b),$$

where

$$\boldsymbol{V}_1 = \frac{\partial \boldsymbol{V}}{\partial \theta_1} = \underset{1\le d\le D}{\mathrm{diag}}\,(\mathbf{1}_{RT}\mathbf{1}'_{RT}),\;\; \boldsymbol{V}_2 = \frac{\partial \boldsymbol{V}}{\partial \theta_2} = \underset{1\le d\le D}{\mathrm{diag}}\Big(\underset{1\le r\le R}{\mathrm{diag}}\,(\mathbf{1}_T\mathbf{1}'_T)\Big),\;\; \boldsymbol{V}_3 = \frac{\partial \boldsymbol{V}}{\partial \theta_3} = \boldsymbol{I}_{DRT}.$$

To estimate $\boldsymbol{\beta}$ and to predict $\boldsymbol{u} = (\boldsymbol{u}'_1, \boldsymbol{u}'_2, \boldsymbol{u}'_3)'$, where

$$\boldsymbol{u}_1 = \underset{1\le d\le D}{\mathrm{col}}(u_{1,d}), \quad \boldsymbol{u}_2 = \underset{1\le d\le D}{\mathrm{col}}\Big(\underset{1\le r\le R}{\mathrm{col}}(u_{2,dr})\Big), \quad \boldsymbol{u}_3 = \underset{1\le d\le D}{\mathrm{col}}\Big(\underset{1\le r\le R}{\mathrm{col}}\Big(\underset{1\le t\le T}{\mathrm{col}}(u_{3,drt})\Big)\Big),$$

we use the REML estimator of $\boldsymbol{\beta}$ and the REML-EBLUP of $\boldsymbol{u}$, i.e.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}, \quad \hat{\boldsymbol{u}} = \hat{\boldsymbol{V}}_u\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}\big(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\big), \tag{3.7}$$

where $\hat{\boldsymbol{V}}$ is obtained by plugging, $\hat{\boldsymbol{V}}_u =$ diag $(\hat{\sigma}_1^2\boldsymbol{I}_D, \hat{\sigma}_2^2\boldsymbol{I}_{DR}, \hat{\sigma}_3^2\boldsymbol{I}_{DRT})$, $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3)$, and

$$\boldsymbol{Z}_1 = \underset{1\le d\le D}{\mathrm{diag}}\big(\mathbf{1}_{RT}\big), \quad \boldsymbol{Z}_2 = \underset{1\le d\le D}{\mathrm{diag}}\Big(\underset{1\le r\le R}{\mathrm{diag}}\,(\mathbf{1}_T)\Big), \quad \boldsymbol{Z}_3 = \boldsymbol{I}_{DRT}.$$

The EBLUP of μ_{drt} is $\hat{\mu}_{drt} = \boldsymbol{x}_{drt}\hat{\boldsymbol{\beta}} + \hat{u}_{1,d} + \hat{u}_{2,dr} + \hat{u}_{3,drt}$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ are given in (3.7). Consequently, each $\hat{\mu}_{drt}$ contains area-level auxiliary information that will reduce the variance of the Hájek estimates $\hat{Y}^{dir}_{drt1}$ without needing to increase the sample sizes.

3.2 Prediction of the Duncan Segregation Index

Let us assume that y_{drt} follows the FH3 model (3.5) and define $\boldsymbol{u}_{drt} = (u_{1,d}, u_{2,dr}, u_{3,drt})'$, so that $\boldsymbol{u}_{drt} \sim N_K(\mathbf{0}, \boldsymbol{V}_{u,drt})$, $\boldsymbol{V}_{u,drt} =$ diag $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$, $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \sigma_3^2,)$ and $K = 1 + R + RT$. Let us consider the domain target parameters

$$S_{drt} = \left| \frac{N_{drt}\mu_{drt}}{\sum_{i=1}^{R} N_{dit}\mu_{dit}} - \frac{N_{drt}(1 - \mu_{drt})}{\sum_{i=1}^{R} N_{dit}(1 - \mu_{dit})} \right|. \tag{3.8}$$

For $d = 1, \dots, D$, $r = 1, \dots, R$, $t = 1, \dots, T$, the plug-in predictors of S_{drt} and $S_{d.t}$ are

$$\hat{S}^{in}_{d.t} = \frac{1}{2}\sum_{r=1}^{R} \hat{S}^{in}_{drt}, \quad \hat{S}^{in}_{drt} = \left| \frac{N_{drt}\hat{\mu}_{drt}}{\sum_{i=1}^{R} N_{dit}\hat{\mu}_{dit}} - \frac{N_{drt}(1 - \hat{\mu}_{drt})}{\sum_{i=1}^{R} N_{dit}(1 - \hat{\mu}_{dit})} \right|. \tag{3.9}$$

For $d = 1, \dots, D$, $r = 1, \dots, R$, $t = 1, \dots, T$, the marginal predictor (MP) of S_{drt} is

$$\hat{S}_{drt}^{mp} = E[S_{drt}|y_{drt}] = \frac{\int_{\mathbb{R}^3} S_{drt}(\boldsymbol{u}_{drt}, \boldsymbol{\beta}) f(y_{drt}|\boldsymbol{u}_{drt}) f(\boldsymbol{u}_{drt})\, d\boldsymbol{u}_{drt}}{\int_{\mathbb{R}^3} f(y_{drt}|\boldsymbol{u}_{drt}) f(\boldsymbol{u}_{drt})\, d\boldsymbol{u}_{drt}} = \frac{A_{drt}(y_{drt}, \boldsymbol{\beta}, \boldsymbol{\theta})}{B_d(y_{drt}, \boldsymbol{\beta}, \boldsymbol{\theta})},$$

where

$$A_{drt}(y_{drt}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\mathbb{R}^3} S_{drt}(\boldsymbol{u}_{drt}, \boldsymbol{\beta}) \exp\Big\{ - \frac{1}{2\sigma_{drt}^2} e_{drt}^2 \Big\} f(\boldsymbol{u}_{drt}) d\boldsymbol{u}_{drt},$$
$$B_d(y_{drt}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\mathbb{R}^3} \exp\Big\{ - \frac{1}{2\sigma_{drt}^2} e_{drt}^2 \Big\} f(\boldsymbol{u}_{drt}) d\boldsymbol{u}_{drt}$$

and $e_{drt} = y_{drt} - \boldsymbol{\mu}_{drt} = y_{drt} - \boldsymbol{x}_{drt}\boldsymbol{\beta} - u_{1,d} - u_{2,dr} - u_{3,drt}$.

The empirical marginal predictor (EMP) of S_{drt} is

$$\hat{S}_{drt}^{emp} = \frac{A_{drt}(y_{drt}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}{B_d(y_{drt}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}, \quad d = 1, \dots, D, \; r = 1, \dots, R, \; t = 1, \dots, T. \tag{3.10}$$

The following algorithm gives a Monte Carlo approximation of $\hat{S}_{drt}^{emp}$.

1. Fit the model to the data $(y_{drt}, \boldsymbol{x}_{drt})$ and obtain the REML estimates of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$.
2. For $\ell = 1, \dots, L$, do
   (a) Draw $u_{1,d}^{(\ell)} \sim N(0, \hat{\sigma}_1^2)$, $u_{2,dr}^{(\ell)} \sim N(0, \hat{\sigma}_2^2)$, $u_{3,drt}^{(\ell)} \sim N(0, \hat{\sigma}_3^2)$, $\boldsymbol{u}_{drt}^{(\ell)} = (u_{1,d}^{(\ell)}, u_{2,dr}^{(\ell)\prime}, u_{3,drt}^{(\ell)\prime})'$ and set $\boldsymbol{u}_{drt}^{(L+\ell)} = -\boldsymbol{u}_{drt}^{(\ell)}$, $d = 1, \dots, D$, $r = 1, \dots, R$, $t = 1, \dots, T$.
   (b) For $d = 1, \dots, D$, $r = 1, \dots, R$, $t = 1, \dots, T$, calculate $\hat{S}_{drt}^{emp} = \hat{A}_{drt}/\hat{B}_d$, where

$$\hat{A}_{drt} = \frac{1}{2L} \sum_{\ell=1}^{2L} S_{drt}(\boldsymbol{u}_{drt}^{(\ell)}, \hat{\boldsymbol{\beta}}) \exp\Big\{ - \frac{e_{drt}^2}{2\sigma_{drt}^2} \Big\}, \quad \hat{B}_d = \frac{1}{2L} \sum_{\ell=1}^{2L} \exp\Big\{ - \frac{e_{drt}^2}{2\sigma_{drt}^2} \Big\}$$

and $e_{drt}^{(\ell)} = y_{drt} - \boldsymbol{x}_{drt}\hat{\boldsymbol{\beta}} - u_{1,d}^{(\ell)} - u_{2,dr}^{(\ell)} - u_{3,drt}^{(\ell)}$.

The EMP of $S_{d.t}$ is

$$\hat{S}_{d.t}^{emp} = \frac{1}{2} \sum_{r=1}^{R} \hat{S}_{drt}^{emp}, \quad d = 1, \dots, D, \; r = 1, \dots, R, \; t = 1, \dots, T. \tag{3.11}$$

The best predictor of S_{drt}, $\hat{S}_{drt}^{bp} = E[S_{drt}|\boldsymbol{y}]$, is also a potentially attractive alternative: theoretically it has a minimum MSE within the class of unbiased predictors. However, its computation requires to approximate an integral in $\mathbb{R}^K$, with $K = 43$ in the application to real data. This is computationally intensive and the main reason why we consider that the EBP approach, under the proposed FH3 model, is not an useful alternative for predicting DSIs in academia or in the production of public statistics.

Sect. 4 includes results from several simulation experiments that have been carried out to investigate and compare the behaviour of the plug-in predictor and the EMP in real data based scenarios. Section A of Supplementary Material describes the steps of the simulations and presents additional results based on artificial data.

3.3 Parametric Bootstrap Estimation of the MSE of the Small Area Estimates

Under the FH3 model, we have adapted the parametric bootstrap procedure proposed by Marcis et al. (2023) to estimate the MSE of $\hat{\mu}_{drt}$ and $\hat{S}_{d.t} \in \{\hat{S}^{in}_{d.t}, \hat{S}^{emp}_{d.t}\}$. The steps of our algorithm are described below.

1. Fit the FH3 model to the data $(y_{drt}, \boldsymbol{x}_{drt})$ and obtain the REML estimates of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$.
2. Repeat B times ($b = 1, \dots, B$):
   (a) For $d = 1, \dots, D$, generate $u^{*(b)}_{1,d} \sim N(0, \hat{\sigma}^2_1)$. Construct the vector $\boldsymbol{u}^{*(b)}_1 = \underset{1 \le d \le D}{\text{col}}(u^{*(b)}_{1,d})$.
   (b) For $d = 1, \dots, D$, $r = 1, \dots, R$, generate $u^{*(b)}_{2,dr} \sim N(0, \hat{\sigma}^2_2)$. Construct the vector $\boldsymbol{u}^{*(b)}_2 = \underset{1 \le d \le D}{\text{col}}(\underset{1 \le r \le R}{\text{col}}(u^{*(b)}_{2,dr}))$.
   (c) For $d = 1, \dots, D$, $r = 1, \dots, R$, $t = 1, \dots, T$, generate $u^{*(b)}_{3,drt} \sim N(0, \hat{\sigma}^2_3)$. Construct the vector $\boldsymbol{u}^{*(b)}_3 = \underset{1 \le d \le D}{\text{col}}(\underset{1 \le r \le R}{\text{col}}(\underset{1 \le t \le T}{\text{col}}(u^{*(b)}_{3,drt})))$.
   (d) For $d = 1, \dots, D$, $r = 1, \dots, R$, $t = 1, \dots, T$, generate $e^{*(b)}_{drt} \sim N(0, \sigma^2_{drt})$. Construct the vector $\boldsymbol{e}^{*(b)} = \underset{1 \le d \le D}{\text{col}}(\underset{1 \le r \le R}{\text{col}}(\underset{1 \le t \le T}{\text{col}}(e^{*(b)}_{drt})))$.
   (e) Calculate the bootstrap vectors
   $$\boldsymbol{y}^{*(b)} = \boldsymbol{\mu}^{*(b)} + \boldsymbol{e}^{*(b)}, \quad \boldsymbol{\mu}^{*(b)} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_1\boldsymbol{u}^{*(b)}_1 + \boldsymbol{Z}_2\boldsymbol{u}^{*(b)}_2 + \boldsymbol{Z}_3\boldsymbol{u}^{*(b)}_3.$$
   (f) For $d = 1, \dots, D$, calculate the bootstrap quantities
   $$S^{*(b)}_{d.t} = \frac{1}{2}\sum_{r=1}^{R} S^{*(b)}_{drt}, \quad S^{*(b)}_{drt} = \left| \frac{N_{drt}\mu^{*(b)}_{drt}}{\sum_{i=1}^{R} N_{dit}\mu^{*(b)}_{dit}} - \frac{N_{drt}(1 - \mu^{*(b)}_{drt})}{\sum_{i=1}^{R} N_{dit}(1 - \mu^{*(b)}_{dit})} \right|.$$
   (g) Fit the FH3 model to the bootstrap vector $\boldsymbol{y}^{*(b)}$. Calculate $\hat{\boldsymbol{\theta}}^{*(b)}$, $\hat{\boldsymbol{\beta}}^{*(b)}$, the EBLUP $\hat{\boldsymbol{\mu}}^{*(b)}$, with components $\hat{\mu}^{*(b)}_{drt}$, and the predictors $\hat{S}^{*(b)}_{d.t}$, $d = 1, \dots, D$, $t = 1, \dots, T$.
3. Output: For $d = 1, \dots, D$, $r = 1, \dots, R$, $t = 1, \dots, T$, calculate
$$mse^*(\hat{\mu}_{drt}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\mu}^{*(b)}_{drt} - \mu^{*(b)}_{drt}\right)^2, \quad mse^*(\hat{S}_{d.t}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{S}^{*(b)}_{d.t} - S^{*(b)}_{d.t}\right)^2. \quad (3.12)$$

Remark 3.1 The auxiliary variables of the FH3 model must be known at domain level, from censuses or administrative records, as they must be free of sampling errors to reduce the variability of the small area estimates. In practice, however, this is not the norm, leading researchers to resort to strategies that allow estimating such area-level variables with low variability. A common technique is to use data from many consecutive surveys to increase sample sizes in the direct estimation of the auxiliary information.

If the explanatory variables have non-negligible sampling errors, the algorithm described above could lead to underestimates of the actual MSE of $\hat{\mu}_{drt}$ and $\hat{S}_{d.t} \in \{\hat{S}^{in}_{d.t}, \hat{S}^{emp}_{d.t}\}$. As a solution to this potential problem, we propose to modify Step 2 (e) so as to include the potencial non-negligible variability of $\boldsymbol{x}_{drt}$. Having said that, it should be clarified that the proposed modification does assume the uncorrelation between the columns of $\boldsymbol{x}_{drt}$ and between $\boldsymbol{x}_{drt}$ and y_{drt}. Nonetheless, correlation relationships are expected to be even lower. In addition, if the first column of $\boldsymbol{x}_{drt}$ represents the intercept, the modification does not apply.

Let us rewrite Step 2 (e) as follows:

2. (e) For $d = 1, \dots, D, r = 1, \dots, R, t = 1, \dots, T, k = 1, \dots, p$, generate $v_{drtk}^{*(b)} \backsim N(0, \sigma_{drtk}^2)$, where σ_{drtk}^2 is the design-based variance of the k-th component of $\boldsymbol{x}_{drt} = (x_{drt1}, \dots, x_{drtp})$ and p is the dimension of $\boldsymbol{x}_{drt}$. This must be skipped for the intercept. If we use Hájek estimates, σ_{drtk}^2 can be replaced by the direct estimate of the design-based variance of x_{drtk}, $d = 1, \dots, D, r = 1, \dots, R, t = 1, \dots, T, k = 1, \dots, p$. Now, we calculate the modified bootstrap vectors
$$\boldsymbol{y}^{*(b)} = \boldsymbol{\mu}^{*(b)} + \boldsymbol{e}^{*(b)}, \quad \boldsymbol{\mu}^{*(b)} = (\boldsymbol{X} + \boldsymbol{v}^{*(b)})\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_1 \boldsymbol{u}_1^{*(b)} + \boldsymbol{Z}_2 \boldsymbol{u}_2^{*(b)} + \boldsymbol{Z}_3 \boldsymbol{u}_3^{*(b)},$$
where $\boldsymbol{v}^{*(b)} = \underset{1 \le d \le D}{\text{col}} (\underset{1 \le r \le R}{\text{col}} (\underset{1 \le t \le T}{\text{col}} (\boldsymbol{v}_{drt}^{*(b)})))$ and $\boldsymbol{v}_{drt}^{*(b)} = (v_{drt1}^{*(b)}, \dots, v_{drtp}^{*(b)}) \in \mathbb{R}^p$.

4 Real Data Simulation Experiments

Based on the application to real data, two simulation experiments have been performed. The real set of area-level auxiliary variables, the variance of the direct estimator and the fitted model, described around Table 9, have been used to simulate the target variable, y_{drt}, $d = 1, \dots, D, r = 1, \dots, R, t = 1, \dots, T$. At this regard, y_{drt} represents the direct estimator of the proportion of employed men in province d, occupation sector r and time period t. The main reason to present the simulation results first is to learn for the application to real data.

Simulation 1 investigates the performance of the Fisher-Scoring algorithm (3.6) and studies the behaviour of the DSI predictors. Simulation 2 deals with the MSE bootstrap estimation and provides a recommendation on the number of replicates to be used. Thus, the behaviour of the estimators and predictors is studied under the assumption that the fitted model is the true one. In this context, it is important to bear in mind that there are $D = 52$ provinces, $R = 7$ occupation sectors and $T = 5$ time periods. So there are 1820 estimation domains defined by the crosses of province, occupation sector and time period. The purpose is to predict sex segregation in provinces and time periods, which is equivalent to 260 DSI-domains. To complete the empirical analysis, Section A of Supplementary Material describes the steps of the simulation processes and presents additional results based on artificial data.

4.1 Simulation 1

The goal of Simulation 1 is to investigate the behaviour of the fitting algorithm and the performance of the predictors of S_{drt} and $S_{d.t}$, $d = 1, \dots, D, r = 1, \dots, R, t = 1, \dots, T$. We run Simulation 1 with $I = 10^3$ iterations and assume the same scenario as in the application to the real data. The bias and error measures that we will calculate are defined below. For a model parameter $\hat{\tau} = \hat{\beta}_k$, $k = 0, 1, \dots, 7$ or $\hat{\tau} = \sigma_l^2$, $l = 1, 2, 3$, we calculate

Table 5 Performance of REML estimators of $\boldsymbol{\beta}$ assuming that the explanatory variables are deterministic (top) and taking into account their sampling errors (bottom)

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
True	−0.3272	0.1424	0.0894	−0.3043	0.8887	0.2054	0.6203	0.1346
BIAS	−0.0013	−0.0004	0.0013	−0.0006	0.0014	−0.0003	0.0138	−0.0002
RMSE	0.0437	0.0181	0.0384	0.0189	0.0341	0.0292	0.1860	0.0208
RBIAS	−0.3868	−0.3052	1.5000	−0.2087	0.1572	−0.1289	2.2300	−0.1662
RRMSE	13.3463	12.6809	42.9240	6.2156	3.8364	14.2279	29.9924	15.4232
	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
BIAS	−0.0015	0.0008	0.0041	0.0016	−0.0002	0.0008	0.0009	0.0006
RMSE	0.0423	0.0190	0.0385	0.0181	0.0345	0.0279	0.1938	0.0204
RBIAS	−0.4502	0.5937	4.6093	0.5370	−0.0188	0.3713	0.1466	0.4299
RRMSE	12.9333	13.3777	43.1294	5.9561	3.8827	13.5834	31.2484	15.1388

Table 6 Performance of REML estimators of the variances assuming that the explanatory variables are deterministic (left) and taking into account their sampling errors (right)

	σ_1^2	σ_2^2	σ_3^2	σ_1^2	σ_2^2	σ_3^2
True	0.0116	0.0022	0.0011	0.0116	0.0022	0.0011
BIAS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
RMSE	0.0024	0.0002	0.0001	0.0024	0.0002	0.0001
RBIAS	−0.0076	0.1650	−0.0237	−0.0657	−0.2959	0.0196
RRMSE	20.792	10.701	11.360	20.3584	10.6609	11.5938

$$BIAS(\hat{\tau}) = \frac{1}{I}\sum_{i=1}^{I}(\hat{\tau}^{(i)} - \tau), \quad RMSE(\hat{\tau}) = \left(\frac{1}{I}\sum_{i=1}^{I}(\hat{\tau}^{(i)} - \tau)^2\right)^{1/2},$$

and for a predictor $\hat{S}_{d.t} \in \{\hat{S}^{in}_{d.t}, \hat{S}^{emp}_{d.t}\}$, $d = 1, \dots, D$, $t = 1, \dots, T$, we calculate

$$ABIAS = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}\left|\frac{1}{I}\sum_{i=1}^{I}(\hat{S}^{(i)}_{d.t} - S^{(i)}_{d.t})\right|, \quad RMSE = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}\left(\frac{1}{I}\sum_{i=1}^{I}(\hat{S}^{(i)}_{d.t} - S^{(i)}_{d.t})^2\right)^{1/2}.$$

The corresponding relative performance measures (in %) are

$$RBIAS(\hat{\tau}) = 100\frac{BIAS(\hat{\tau})}{\tau}, \quad RRMSE(\hat{\tau}) = 100\frac{RMSE(\hat{\tau})}{\tau},$$

$$RBIAS_{dt} = 100\frac{BIAS_{dt}}{S_{d.t}}, \; RRMSE_{dt} = 100\frac{RMSE_{dt}}{S_{d.t}}, \; S_{d.t} = \frac{1}{I}\sum_{i=1}^{I}S^{(i)}_{d.t},$$

$$ARBIAS = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}|RBIAS_{dt}|, \quad RRMSE = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}RRMSE_{dt}.$$

Table 5 (top) and Table 6 (left) present the results for the model parameter estimators. As can be noticed, for the $\boldsymbol{\beta}$ coefficients the biases are small but the root-MSEs (RMSE) are not, implying that the variance is the main component of the MSE. Such variability is probably attributable to the relationship between the number of estimation domains and the number of parameters estimated by the model, $DRT/(8+3) = 165.45$, which is not large enough to activate the asymptotic properties of the ML estimators. For the estimators of the variances, the RBIAS is small and the RRMSE does not present notably large values either, with the worst result being the one corresponding to $\hat{\sigma}_1^2$.

Table 7 (left) provides the absolute and relative performance measures for the EMPs and the plug-in predictors of the DSI values. We use $L = 500$ iterations in the integral approximation.

For the plug-in predictor, the average across DSI-domains of the absolute relative bias (ARBIAS) is close to 11% and the RRMSE average (RRMSE) does not exceed 28%, which is quite satisfactory. We therefore use the plug-in predictor in the application to real data. In the case of the EMP, the ARBIAS is greater than 56% and the RRMSE is close to 80%. At this regard, it should be noted that the EMP is not obtained exactly, only approximately, because the integrals that appear in its expression cannot be calculated analytically. Approximations are generated by the antithetic Monte Carlo method and calculations are subject to the number of iterations, partly justifying its poor results. Moreover, good theoretical properties are attributed to the best predictor, not to marginal or empirical versions.

Up to this point, we have assumed that the area-level auxiliary variables are deterministic. This assumption leads to the results in Table 5 (top) and Tables 6-7 (left). However, it has already been mentioned in Remark 3.1 that if the auxiliary information does not come from censuses or administrative registers, but from estimates, it is potentially likely to add more variability to the small domain estimates. For this reason, we have also considered in the real data simulations the scenario in which the area-level auxiliary variables have non-negligible sampling errors. For each iteration $i = 1, \dots, I$, the new area-level explanatory variables are generated as follows:

$$\boldsymbol{X} + \boldsymbol{v}^{*(i)}, \ \text{ where } \boldsymbol{v}^{*(i)} = \underset{1\le d\le D}{\text{col}}\Big(\underset{1\le r\le R}{\text{col}}\Big(\underset{1\le t\le T}{\text{col}}(\boldsymbol{v}^{*(i)}_{drt})\Big)\Big), \ \boldsymbol{v}^{*(i)}_{drt} = (v^{*(i)}_{drt1}, \dots, v^{*(b)}_{drtp}) \in \mathbb{R}^p, \tag{4.1}$$

$v^{*(i)}_{drtk} \backsim N(0, \sigma^2_{drtk})$, $d = 1, \dots, D$, $r = 1, \dots, R$, $t = 1, \dots, T$, $k = 1, \dots, p$, and σ^2_{drtk} is the design-based variance of the k-th component of $\boldsymbol{x}_{drt} = (x_{drt1}, \dots, x_{drtp})$.

Table 5 (bottom) and Table 6 (right) present the results for the model parameter estimators under scenario (4.1). At this point, to compare differences between both scenarios (the deterministic scenario and scenario (4.1)), the error measures must be interpreted in absolute terms to avoid small variations caused by changes in the denominators when relativizing. Having said that, it is concluded that there are practically no changes in the performance of the Fisher-Scoring algorithm when adding the random terms $v^{*(i)}_{drtk}$. This is further evidence for the validity of our methodology. In addition, and as can be seen in Table 7 (right), generate the area-level explanatory variables according to scenario (4.1) leads to virtually no changes in the performance measures of the predictors $S_{d.t}$. This is another argument in favour of the fact that the variability they add by estimating them with

Table 7 Performance of predictors of $S_{d.t}$ assuming that the explanatory variables are deterministic (left) and taking into account their sampling errors (right)

	Plug-in	EMP	Plug-in	EMP
ABIAS	0.0509	0.3242	0.0504	0.3423
RMSE	0.0997	0.3491	0.1004	0.3503
ARBIAS	11.1862	56.4842	11.7298	57.284
RRMSE	27.6593	79.3712	27.9553	78.6337

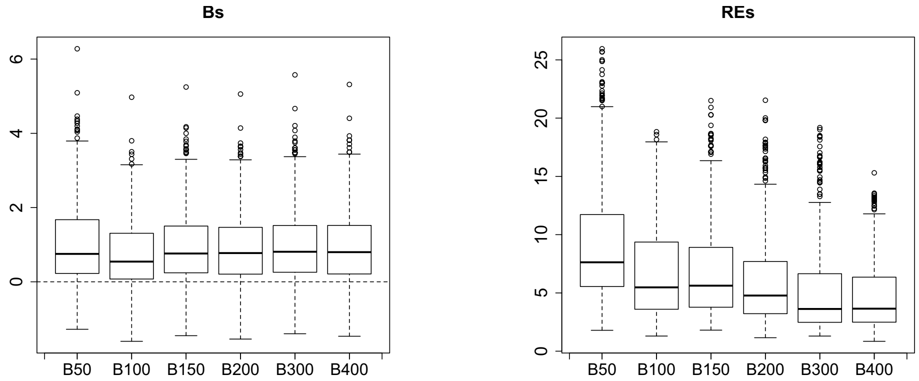


Fig. 2 Boxplots of RB_{dt}'s (left) and RRE_{dt}'s (right) for $B = 50, 100, 150, 200, 300, 400$

Table 8 Performance measures for $B = 50, 100, 150, 200, 300, 400$

B	50	100	150	200	300	400
ARB	1.2294	0.9633	1.1634	1.1067	1.1799	1.1213
RRE	9.5056	6.9331	7.1849	6.5066	5.6235	4.8295

5 consecutive periods of the SLFS is minimal. Taking all of the above into consideration, we deduce that it is not necessary to propose a measurement error model for the problem at hand, i.e., the estimation of DSIs in small areas based on explanatory variables estimated with five consecutive SLFSs.

4.2 Simulation 2

Simulation 2 studies the behavior of the parametric bootstrap estimator of the MSE of the plug-in predictor of $\hat{S}^{in}_{d.t}$, denoted by mse_{dt}, $d = 1, \dots, D$, $t = 1, \dots, T$. The real MSE of $\hat{S}^{in}_{d.t}$ is taken from Simulation 1 and denoted by MSE_{dt}, $d = 1, \dots, D$, $t = 1, \dots, T$. It is assumed that the area-level auxiliary variables are deterministic. Since it is computationally more demanding, we run Simulation 2 with $I = 500$ iterations. Moreover, as absolute measures are more difficult to interpret, we focus our study on relative measures. To do so, we first calculate

$$B_{dt} = \frac{1}{I}\sum_{i=1}^{I}\left(mse^{*(i)}_{dt} - MSE_{dt}\right),\; RE_{dt} = \left(\frac{1}{I}\sum_{i=1}^{I}\left(mse^{*(i)}_{dt} - MSE_{dt}\right)^2\right)^{1/2},$$

$d = 1, \dots, D, t = 1, \dots, T$. Then we define the relative performance measures (in %)

$$RB_{dt} = 100\frac{B_{dt}}{MSE_{dt}},\; RRE_{dt} = 100\frac{RE_{d}}{MSE_{dt}},\; d = 1, \dots, D,\; t = 1, \dots, T;$$

$$ARB = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}|RB_{dt}|,\; RRE = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}RRE_{dt}.$$

Figure 2 plots five boxplots of the relative biases, RB_{dt}, and the relative root-MSEs, RRE_{dt}, $d = 1, \dots, D$, $t = 1, \dots, T$, for $B = 50, 100, 150, 200, 300, 400$. The left boxplots show that the relative biases do not decrease as the size of B increases, showing a slight positive bias around 1.2%. The right boxplots show that the relative root-MSEs are lower than 20% and decrease as B increases, achieving good results for values greater than or equal to 300 resamples. Table 8 confirms it, with the averages of the absolute relative biases (ARB) stabilized around 1.2% and the averages of the relative root-MSEs (RRE) decreasing as B increases, but suggesting some stabilization around $B = 300$ iterations.

5 Prediction of Sex Occupational Segregation by Spanish Province

This section applies the developed SAE methodology to the SLFS data from 2020.4 to 2021.4, both surveys included. We fit the FH3 model to all data, but we mainly focus on the statistical results for the latest available period to offer conclusions. The main reason is the proximity in time, which allows us to analyze the results closer to the present day, but also to value brevity. We recall that the response variable, y_{drt}, is the Hájek estimator of the proportion of men in the subset of employed people of province d, occupation sector r and time period t. The error variances σ^2_{drt} have previously been predicted according to the GVF method (3.3), after fitting model (3.2). The precision of the estimates of the auxiliary information has been improved by using a time window of five quarters, which includes the current quarter and the previous four quarters. By virtue of Table 4 and the results of the model-based simulations in Sect. 4, the sampling error of their estimation is assumed to be negligible. Even so, Remark 3.1 has been taken into account to provide bootstrap estimates of the MSE of the DSI plug-in predictions.

5.1 Model Selection and Model Parameter Estimation

In order to fit the FH3 model to each y_{drt}, those auxiliary variables that were not significant at 5% were recursively removed. At this regard, our results and conclusions are subject to the available information and therefore, with other territorial divisions, occupational sectors or time periods, the final set of explanatory variables may vary. Nevertheless, the main objective here is to illustrate how to fit a simple but informative model to reduce the variability of the DSI estimates. As a result, age3–1, age3–3, edu3 and st3 were eliminated because its REML estimate was not significantly different from zero at 5%. The failure to consider age groups suggests that sex segregation is persistent over time, despite the age of the worker.

Table 9 Regression parameters of the final FH3 model

	β_0	β_1^{cit1}	β_2^{edu1}	β_3^{edu4}	β_4^{work1}	β_5^{st1}	β_6^{st2}	β_7^{st4}
Estimate	−0.3272	0.1424	0.0894	−0.3043	0.8887	0.2054	0.6203	0.1346
SE	0.0441	0.0191	0.0384	0.0181	0.0350	0.0280	0.1925	0.0205
p-value	0.0000	0.0000	0.0201	0.0000	0.0000	0.0000	0.0013	0.0000
LB 95%	−0.4136	0.1050	0.0140	−0.3398	0.8201	0.1505	0.2430	0.0945
UB 95%	−0.2408	0.1798	0.1647	−0.2689	0.9574	0.2603	0.9976	0.1747

Table 9 presents the REML estimate of $\boldsymbol{\beta}$, together with the asymptotic standard errors (SE) and p-values to test $H_0 : \beta_k = 0$, $k = 0, 1, \dots, 7$. It also contains the lower (LB) and upper (UB) bounds of the 95% asymptotic confidence intervals (CI).

The effect of the explanatory variables derived from Table 9 is consistent with a socio-economic interpretation. Once the rest of the variables are fixed, their sign indicates their contribution (positive or negative) to estimate the proportion of employed men by estimation domain.

Regarding the model variances, we obtain $\hat{\sigma}_1^2 = 0.0117$, $\hat{\sigma}_2^2 = 0.0022$ and $\hat{\sigma}_3^2 = 0.0011$. At a 95% confidence level, the asymptotic CIs for the variances are

$$CI_{\sigma_1^2} = (0.0070, 0.0164), \quad CI_{\sigma_2^2} = (0.0017, 0.0026), \quad CI_{\sigma_3^2} = (0.0008, 0.0013).$$

As they do not contain zero, it is justified to make further inferences based on the FH3 model.

5.2 Model Validation

For the diagnosis analysis of the FH3 model, we consider the raw residuals, defined as

$$\hat{e}_{drt} = y_{drt} - \hat{\mu}_{drt}, \; d = 1, \dots, D, \; r = 1, \dots, R, \; t = 1, \dots, T, \tag{5.1}$$

and the standardized residuals, defined by dividing by their own standard deviation. To analyse outliers, even though all standardized residuals move in a suitably short range close to 0, three boxplots are included in Fig. 3. From left to right, they are grouped according to time period, province and occupation sector. The last two boxplots use only data from the SLFS of 2021.4. We observe that: (1) the standardized residuals present an homogeneous pattern in terms of time period, (2) provinces have more notable influence, although none of them has a particularly anomalous behaviour, and (3) the Hájek estimates tend to overestimate occupational categories OC1, OC2, OC3 and OC7 because their boxes fall mostly in the positive half-plane. Another important result that can be inferred is the adequacy of the standardized residuals in terms of rank: they take values from − 3 to 2, with a single outlier, located in Melilla.

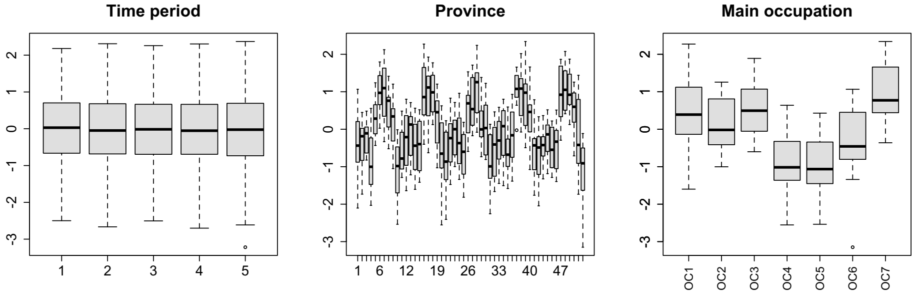


Fig. 3 Boxplot of the standardized residuals

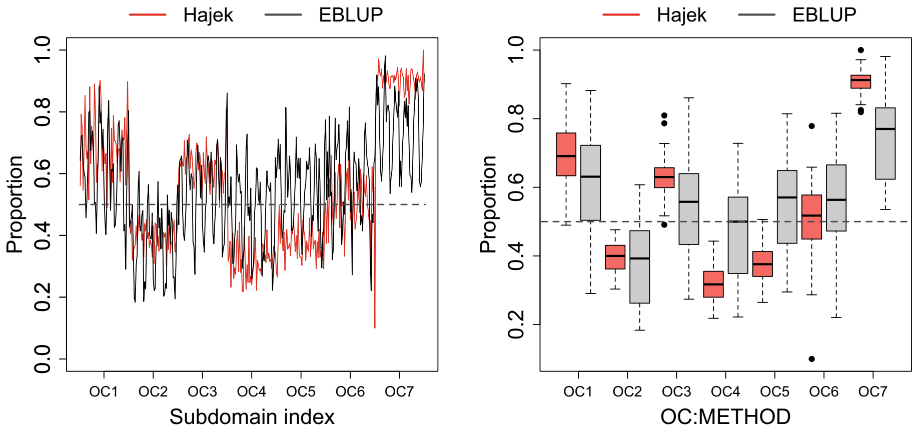


Fig. 4 On the left, EBLUPs and direct estimates of the proportion of employed men. On the right, boxplots of the same quantities. Data from the SLFS of 2021.4

5.3 Prediction of the Proportions of Employed Men

This section provides some results and maps to show the predicted proportions of men who are employed in each occupation sector by province and time period. To interpret them, let us recall that the aim of the area-level model-based predictors is not to reproduce the trend of the direct estimators, but to smooth them and provide more accurate results, relying on auxiliary information, complex correlation structures and data from other domains.

Figure 4 (left) plots the EBLUPs and the Hájek direct estimates of the proportion of employed men in the last quarter of 2021. The dotted line $y = 0.5$ is included to compare the distance between both approaches and the balanced distribution of the population. As desired, it can be seen that model-based predictions smooth the behaviour of the Hájek estimates, with atypically high and low proportions, and show a better predictive performance. It is observed that the EBLUPs and the direct estimates follow the same trend, although the first ones are closer to $y = 0.5$. Figure 4 (right) includes some boxplots of the EBLUPs and the Hájek direct estimates of the proportion of employed men, for each occupation sector and the lastest time period SLFS2021.4. The boxes of the EBLUPs and the direct estimates follow the same pattern, although they are not completely identical.

To make a fair comparison of the relative error measures, we estimate the RRMSE of $\hat{\mu}_{drt}$ by dividing the squared root of the bootstrap estimate $mse^*(\hat{\mu}_{drt})$, defined in (3.12), by the Hájek estimate y_{drt}. Next, we run the bootstrap algorithm with $B = 2000$ resamples, taking into account Remark 3.1. More concretely, we estimate the RRMSEs of the EBLUP as follows:

$$\text{RRMSE}(\hat{\mu}_{drt}) = \frac{\sqrt{mse^*(\hat{\mu}_{drt})}}{y_{drt}}, \quad d = 1, \dots, D,\ r = 1, \dots, R,\ t = 1, \dots, T. \tag{5.2}$$

Figure 5 (left) shows that, in a large percentage of estimation domains, the EBLUP has lower RRMSE than the design-based CV of the Hájek estimator.

Table 10 contains the deciles of the model-based estimates of the RRMSEs of the EBLUP proportions of employed men and CVs of the Hájek estimator for the 2021.4 SLFS

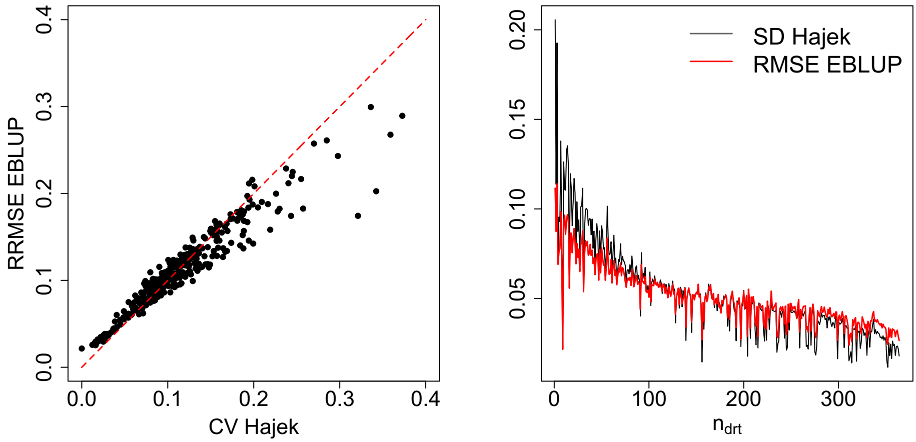


Fig. 5 On the left, RRMSEs and design-based CVs. On the right, RMSEs and design-based standard deviations (SD) sorted by sample size. Data from the SLFS of 2021.4

Table 10 Percentiles of sample sizes, RRMSEs of the EBLUP proportions of employed men and CVs of the Hájek estimator. Data from the SLFS of 2021.4

	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_{drt}	6	30	57	81	101	118	142	172	211	294	956
RRMSE	0.0269	0.0466	0.0662	0.0779	0.0876	0.0974	0.1077	0.1285	0.1496	0.1868	0.3153
CV	0.0000	0.0291	0.0629	0.0773	0.0876	0.0985	0.1128	0.1267	0.1530	0.1872	0.3727

data. It is obtained that the percentiles of the CVs prior to the median are lower, as they correspond to estimation domains with higher sample sizes, where direct estimates report reliable results. However, after the median, the CVs have higher percentiles than those of the RRMSEs of the EBLUP. The reason, again, is the sample size.

Similarly, the consistency of the EBLUP proportions of employed men is empirically checked in Fig. 5 (right) by plotting the estimated RMSE of the EBLUPs against the domain sample sizes. The design-based standard deviations of the Hájek estimator are also included to confirm what happens with the magnitude of the sample sizes. Since the sample sizes are highly variable in our estimation domains for the SLFS data, it is advisable to use model-based predictors instead of direct estimators. Under the model-based approach, the EBLUP also has some theoretical good properties, such as asymptotic unbiasedness. Overall, the proposed model performs satisfactorily, both in terms of the significance level of the estimated parameters and in the reduction of the CVs of the Hájek estimator when the sample sizes are small.

5.4 Prediction of Duncan Segregation Indexes

This section calculates the DSI plug-in predictions by province from 2020.4 to 2021.4. With a view to get an idea of the distribution of segregation by main occupation, Fig. 6 plots the disaggregated DSI predictions for the SLFS of 2021.4. Based on these results,

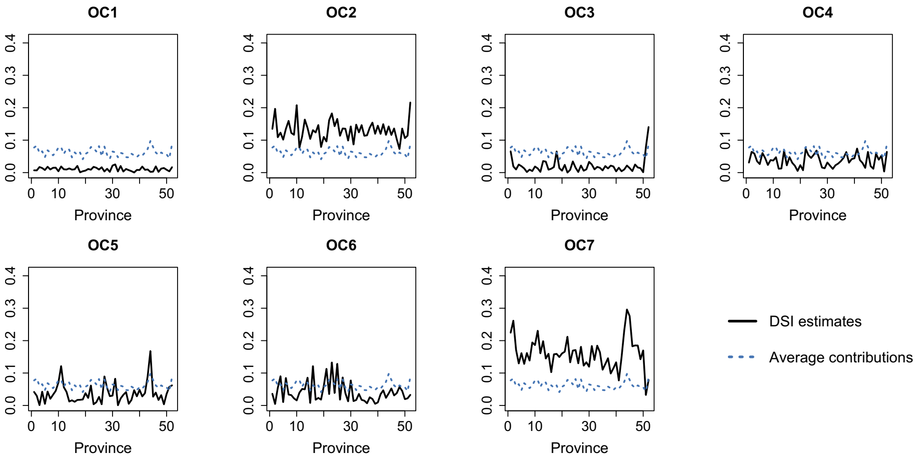


Fig. 6 DSI estimates and average contributions. Data from the SLFS2S021.4

categories OC2, OC6 and OC7 are the main contributors to the DSI. Thus, segregation is concentrated in two main groups: high-skilled scientific and intellectual jobs and traditionally manual or low-skilled jobs. On the opposite direction are categories OC1 and OC4. For instance, directors and managers of public and private institutions and, in general, accountants, administrative and other office employees work in less sex-segregated jobs.

Given a province d and a time period t, if all main occupations contribute equally to the DSI values, there is a common average value s_{dt} such that $S_{drt} = s_{dt}$, $r = 1, \dots, R$, and

$$S_{d.t} = \frac{1}{2}\sum_{r=1}^{R} S_{drt} = \frac{1}{2}Rs_{dt}, \quad s_{dt} = \frac{2S_{d.t}}{R}. \tag{5.3}$$

By comparing s_{dt} with each S_{drt}, $r = 1, \dots, R$, we can illustrate the contribution of each main occupation to the provincial indicators. Since we have focused on 2021.4, it follows that $t = 5$ and the graphical analysis is restricted to this last quarter. In Fig. 6, the values $\hat{s}^{in}_{d5}$, calculated by replacing $S_{d.5}$ by $\hat{S}^{in}_{d.5}$, are added in dashed blue. In terms of interpretation, an employment sector r with $\hat{S}^{in}_{dr5}$ far from the average contribution $\hat{s}^{in}_{d5}$ adds up a lot when calculating the plug-in predictor $\hat{S}^{in}_{d.t}$ of province d. The latter is true for sectors OC2, OC6 and OC7 and the opposite applies to OC1 and OC4. Results are, therefore, consistent with the previous findings.

Table 11 presents the provincial averages of the DSI plug-in predictions for $t = 5$, i.e.

$$\hat{S}^{in}_{.r5} = \frac{1}{D}\sum_{d=1}^{D} \hat{S}^{in}_{dr5}, \quad r = 1, \dots, R. \tag{5.4}$$

Table 11 Provincial average DSI values by main occupation

	OC1	OC2	OC3	OC4	OC5	OC6	OC7
$\hat{S}^{in}_{.r5}$	0.0110	0.1311	0.0194	0.0362	0.0396	0.0396	0.1596

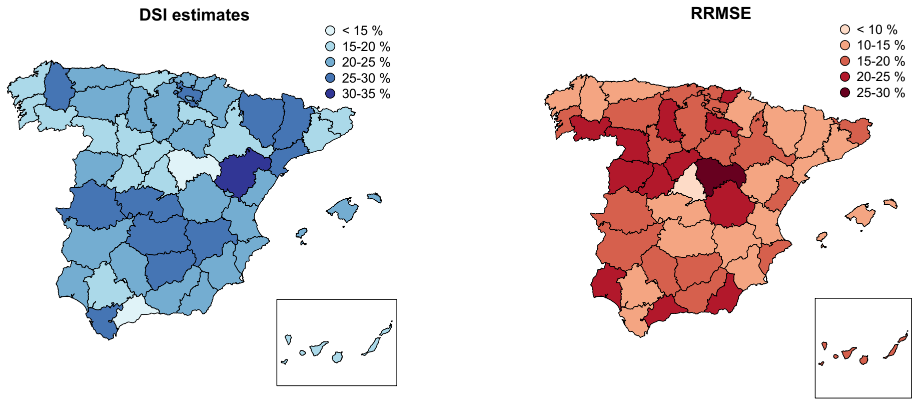


Fig. 7 DSI estimates (left) and RRMSE (right). Data from the SLFS of 2021.4

Among the main occupations with highest DSIs, OC2 and OC7 stand out. On the opposite side, we have found OC1. However, the average measures presented therein overshadow the provincial variability of segregation. This can also be seen in Fig. 6.

Figure 7 (left) colours Spain according to the DSI predictions for the fourth quarter of 2021. So, it allows us to analyse how sex segregation differs across provinces. We observe that the largest discrepancies are found in Teruel, Albacete and Álava, among others. Indeed, between 30 and 35% of the employed population of Teruel would have to change their occupational sector to achieve a uniform distribution by province. The cause of the high sex segregation in Álava is due to the mining industry. Historically, male labour has always been more predominant in this sector, including plant and machinery operators and assemblers, as well as the construction and mining industries. In the other highlighted provinces, sex segregation mainly occurs in the category OC2, which covers highly skilled scientific and intellectual jobs.

There is no clear spatial pattern in the sense that it is not possible to say that certain larger regions of the Iberian Peninsula are more prone to sex segregation than others. Furthermore, the distribution among provinces with similar demographic and socioeconomic conditions is, in general, homogeneous. In terms of labour equality, the high predicted values for many provinces reveal the magnitude of the problem: the labour market disadvantages women and the occupational distribution is clearly non-homogeneous. According to our research, public and private institutions should implement measures of work equality and promote the inclusion of men and women in those sectors in which their presence is minority.

As for the error measures, we calculate the parametric bootstrap estimator of $\hat{S}^{in}_{d.t}$, $mse^{*}(\hat{S}^{in}_{d.t})$, given by (3.12). We generate $B = 2000$ bootstrap resamples, taking into account Remark 3.1. The estimated RRMSE of $\hat{S}^{in}_{d.t}$ is obtained by dividing the RMSE by the DSI estimates, i.e.

$$\mathrm{RRMSE}(\hat{S}^{in}_{d.t}) = \frac{\sqrt{mse^{*}(\hat{S}^{in}_{d.t})}}{\hat{S}^{in}_{d.t}}, \quad d = 1, \ldots, D,\ t = 1, \ldots, T. \tag{5.5}$$

Fig. 7 (right) shows the bootstrap estimates of the RRMSE for the DSI predictions, which enables us to visually quantify the precision of our results. It can be concluded that most

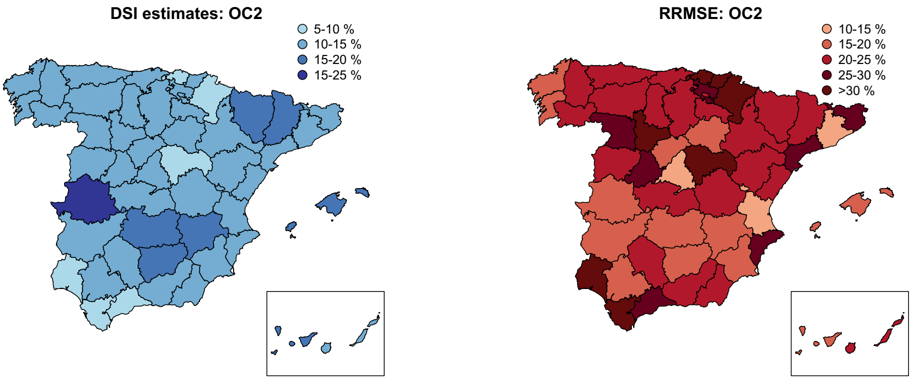


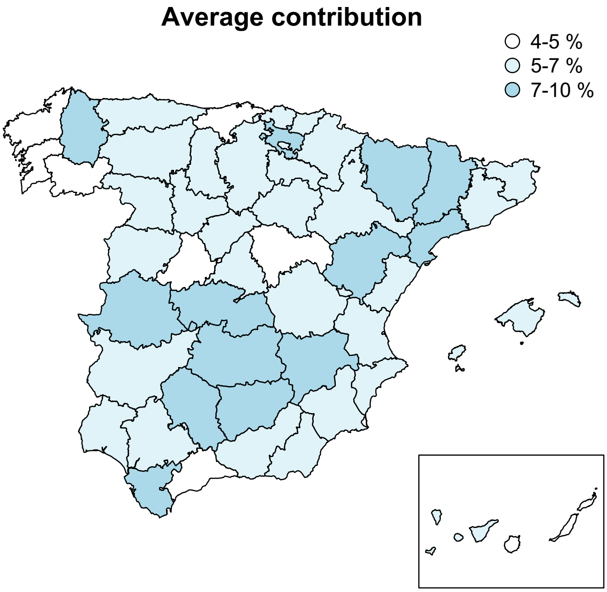
Fig. 8 DSI estimates (left) and RRMSE (right) for OC2. Data from the SLFS of 2021.4

provinces are accompanied by RRMSEs below 25%, which is quite acceptable in the SAE setup. Most RRMSEs are lower than 20% and even 10% in several domains. For more information, the DSI preditions and corresponding RRMSE estimates for all DSI-domains, i.e., by province and time period, are included in Section B of Supplementary Material.

Recalling the findings in Figs. 6, 8 (left) shows the DSI estimates for the OC2 category. Looking further into this main occupation, the province with the highest contribution to the DSI is Cáceres. As expected, the left-hand plots in Figs. 7 and 8 follow a similar pattern and the OC2 category is crucial for the Spanish sex segregation.

Figure 8 (right) maps the bootstrap estimates of the RRMSE for the OC2 category, using data from the SLFS2021.4. At first glance, we conclude that our predictions are sufficiently accurate in terms of an SAE problem, with RRMSE below 30% in most provinces, exceeding it in isolated cases. Comparing the right-hand plots in Figs. 7 and 8, higher RRMSEs are achieved if the target is to predict segregation in the OC2 category. This is in line with

Fig. 9 Average contributions. Data from the SLFS of 2021.4



the theoretical findings, as the variability involved in predicting an average is expected to be smaller than the one involved in predicting each of its summands, potentially increasing the prediction error of a particular category. For more details, Section B of Supplementary Material contains the results of all DSI preditions for the OC2 category, together with the RRMSE estimates.

Figure 9 plots the average contributions $\hat{s}^{in}_{d5}$, $d = 1, ..., D$. Consequently, it is possible to identify the provincial average trend, highlighting those provinces whose performance is positive. Specifically, 29 provinces have an average contribution of between 5 and 7% and 13 contribute between 7 and 10%. The remaining ones report values below 5%. This leads to the conclusion that, on average, its contribution to the Spanish sex segregation is lower than that of the other provinces and, therefore, inequality is lower.

Taking advantage of the available temporal information, Fig. 10 (left) shows the map of the DSI differences between the last quarter 2021.4 and the first quarter 2020.4, i.e. $\hat{S}^{in}_{d.5} - \hat{S}^{in}_{d.1}$, $d = 1, \dots, D$. We have observed that segregation shows appreciable changes over the observation period, with a maximum decrease close to 7% units and a maximum increase bordering on 10%. However, several provinces in the center of Spain do not seem to be affected by any change. In absolute terms, the situation has worsened in 17 provinces, improved in 7 and remained stable in 26 (between −0.01 and 0.01). In Madrid and Barcelona, which are the most populated regions, no significant changes have been observed.

Figure 10 (right) shows the differences $\hat{S}^{in}_{d25} - \hat{S}^{in}_{d21}$, $d = 1, \dots, D$. As far as OC2 category is concerned, the evolution of segregation moves in the range mentioned above for the average case, but the situation varies slightly for many provinces. Even so, the changes observed over 2021 do not refer to a sufficiently long period of time to capture the results of potentially applicable policy decisions, and therefore they are not statistically significant. Nevertheless, our model provides relevant advances in the study of the temporal evolution of sex segregation in SAE situations. Consequently, the proposed methodology could be applied in other studies with data from longer time periods, such as years or decades.

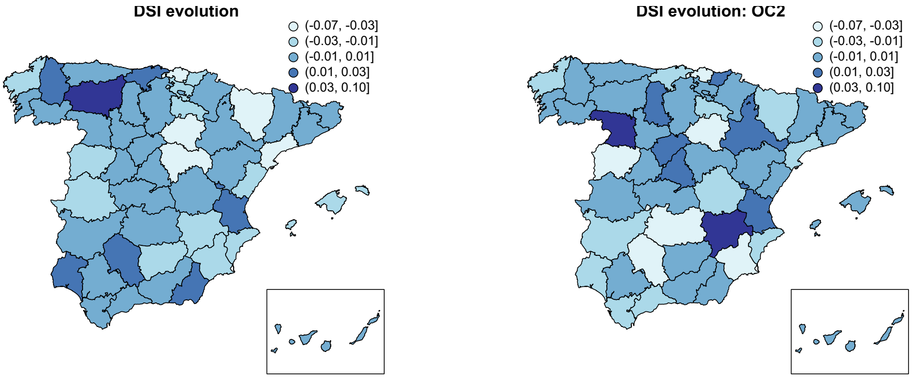


Fig. 10 Evolution of the DSI values over the horizon 2020.4–2021.4

6 Conclusions

The DSI is a relevant statistical indicator used in many sociological studies to measure segregation. It can be calculated based on data from administrative records or surveys. In the latter case, we can estimate it directly, which perform adequately if the sample size in the target population is large enough. Nevertheless, in a more comprehensive study, it should be calculated at a more disaggregated demographic and territorial level. In such cases, direct estimators are no longer reliable and we have to resort to indirect model-based approaches. In this way, we can take advantage of the auxiliary information embedded in the models and, as a consequence, obtain more accurate predictions.

At this point, multilevel linear mixed models are quite flexible. Namely, the advantages of the FH3 model over the existing literature are, on the one hand, being an area model, which facilitates the availability of auxiliary information. On the other hand, its level of hierarchy is adapted to the nature of our data. Nested error regression models may also be appropriate, but the lack of census data and administrative registers would limit their predictive capability to that of ANOVA-type models. We very much doubt that such models would perform better. In addition, the current model has smoothed the results of the direct estimator and provided estimates of the proportions in the interval [0, 1]. Having said that, there is room for improvement and it would be interesting to develop fitting algorithms for new FH models, nested at three hierarchical levels and adapted to the modelling of proportions.

The FH3 model is adapted to a population hierarchically structured in provinces, occupation sectors and time periods. Thus, we can model the direct estimator of the proportion of men (or women) in each labour sector and subsequently incorporate auxiliary information to obtain more efficient predictors. We have considered the plug-in predictor and the EMP. We have not implemented the EBP because it requires to approximate integrals in $\mathbb{R}^{43}$, which is computationally unfeasible. The plug-in predictor is easily calculable, but the EMP requires to approximate integrals on $\mathbb{R}^3$, so it is less efficient. We have investigated the behaviour of these predictors in simulation studies based on real and artificial data. As a result, the plug-in predictor seems more interesting, as it also has a small RMSE. In addition, we have considered scenarios in which the auxiliary variables have small sampling errors and found that the changes in the final results are minimal. Therefore, we do not recommend fitting measurement error models in our case study. To estimate MSEs, we apply a parametric bootstrap method and advise to use $B = 300$ iterations as a good compromise between accuracy and computational time.

In the application to the SLFS of 2020.4–2021.4, we have only used the plug-in predictor. We also present results for the DSI components and, in particular, for the OC2 category, related to scientific and intellectual technicians and professionals. The plug-in predictors have lower MSE than the direct Hájek estimators. We have mapped the Spanish DSIs and most of the predictions have estimated RRMSE below 25%, which is a fairly good accuracy for an SAE problem. Finally, we have found which provinces and sectors are most affected by sex segregation and exemplified how it can be monitored over time.

To sum up, we believe that this work can be a valuable starting point for promoting SAE in sociological studies of current interest. As of today, this is particularly useful for addressing points 4, 5, 8, 9 and 10 of the Sustainable Development Goals (SDGs) set by the United Nations General Assembly (UNGA), designed in 2015 and to be achieved until 2030.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11205-024-03393-w.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work has been supported by the Spanish Ministry of Science,Innovation and Universities, through the Project PGC2018-096840-B-I00 and by the Valencian RegionalMinistry of Innovation, Universities, Science and Digital Society, through the ProjectPROMETEO-2021-063. In addition, it has also benefited from a study grant from the ManuelVentura Figueroa Foundation.

Data Availability Statement In the GitHub repository https://github.com/small-area-estimation/FH3DUNCAN, you will find our dataset and computer code and a detailed description of its contents

References

Alonso-Villar, O., & Del Río, C. (2010). Segregation of female and male workers in Spain: Occupations and industries. *Revista de Economía Pública, 194*(3), 91–121.

Baíllo, A., & Molina, I. (2009). Mean-squared errors of small-area estimators under a unit-level multivariate model. *Statistics, 43*(6), 553–569.

Benavent, R., & Morales, D. (2016). Multivariate Fay–Herriot models for small area estimation. *Computational Statistics and Data Analysis, 94*, 372–390.

Benavent, R., & Morales, D. (2021). Small area estimation under a temporal bivariate area-level linear mixed model with independent time effects. *Statistical Methods and Applications, 30*(1), 195–222.

Boubeta, M., Lombardía, M. J., & Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. *TEST, 25*, 548–569.

Boubeta, M., Lombardía, M. J., & Morales, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics and Data Analysis, 107*, 32–47.

Burgard, J. P., Krause, P., Münnich, R., & Morales, D. (2021). L2-penalized temporal logit mixed models for the estimation of regional obesity prevalence over time. *Statistical Methods in Medical Research, 30*(7), 1744–1768.

Burgard, J. P., Krause, P., & Morales, D. (2022). A measurement error Rao-Yu model for regional prevalence estimation over time using uncertain data obtained from dependent survey estimates. *TEST, 31*(1), 204–234.

Cai, S., & Rao, J. N. K. (2022). Selection of auxiliary variables for three-fold linking models in small area estimation: A simple and effective method. *Stats, 5*(1), 128–138.

Chambers, R., Salvati, N., & 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, 179*(2), 453–479.

Das, S., Kotikula, A. (2019). Gender-based employment segregation: Understanding causes and policy interventions. Jobs working paper, issue 26. The World Bank Group.

Datta, G. S., Lahiri, P., Maiti, T., & Lu, K. L. (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *Journal of the American Statistical Association, 94*, 1074–1082.

Datta, G. S., Lahiri, P., & 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.

Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review., 20*(2), 210–217.

Esteban, M. D., Morales, D., Pérez, A., & Santamaría, L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis, 56*, 2840–2855.

Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D., & Pérez, A. (2022). Empirical best prediction of small area bivariate parameters. *Scandinavian Journal of Statistics, 49*, 1699–1727.

Fay, R. E., & 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.

Ghosh, M., Nangia, N., & Kim, D. (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., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2002). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics and Data Analysis, 51*, 2720–2733.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model. *Computational Statistics and Data Analysis, 52*, 5242–5252.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2010). Small area estimation under Fay–Herriot models with nonparametric estimation of heteroscedasticity. *Statistical Modelling, 10*(2), 215–239.

Guadarrama, M., Morales, D., & Molina, I. (2022). Time stable empirical best predictors under a unit-level model. *Computational Statistics and Data Analysis., 160*, 107226.

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small-area prediction. *Journal of the Royal Statistical Society, B, 68*, 221–238.

Herrador, M., Esteban, M. D., Hobza, T., & Morales, D. (2011). A modified nested-error regression model for small area estimation. *Statistics, 47*(2), 258–273.

Hariyanto, S., Notodiputro, K., Kurnia, A., & Sadik, K. (2018). Measurement error in small area estimation: A literature review. *IOP Conference Series: Earth and Environmental Science, 187*, 012034.

Hobza, T., & Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of official statistics, 32*(3), 661–69.

Hobza, T., Morales, D., & Santamaría, L. (2018). Small Area Estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST, 27*(2), 270–294.

Huang, E. & Bell, W. (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). U.S. Bureau of the Census.

Janicki, R. (2020). Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. *Communications in Statistics - Theory and Methods, 49*(9), 2264–2284.

Krause, J., Burgard, J. P., & Morales, D. (2022). L2-penalized approximate likelihood inference in logit mixed models for regional prevalence estimation under covariate rank-deficiency. *Metrika, 85*, 459–489.

Krenzke, T., Mohadjer, L., Li, J., Erciulescu, A., Fay, R. E., Ren, W., VanDeKerckhove, W., Li, L., & Rao, J. N. K. (2020). *Program for the international assessment of adult competencies (PIAAC): State and county estimation methodology report; technical report*. Washington: Institute of Education Sciences, National Center for Education Statistics.

López-Vizcaíno, E., Lombardía, M. J., & Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling, 13*(2), 153–178.

López-Vizcaíno, E., Lombardía, M. J., & Morales, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Association, Series A, 178*(3), 535–565.

Marhuenda, Y., Molina, I., & Morales, D. (2013). Small area estimation with spatio-temporal Fay–Herriot models. *Computational Statistics and Data Analysis, 58*, 308–325.

Marhuenda, Y., Morales, D., & Pardo, M. C. (2014). Information criteria for Fay–Herriot model selection. *Computational Statistics and Data Analysis, 70*, 268–280.

Marhuenda, Y., Morales, D., & Pardo, M. C. (2016). Tests for the variance parameter in the Fay–Herriot model. *Statistics, 50*(1), 27–42.

Marhuenda, Y., Molina, I., Morales, D., & Rao, J. N. K. (2017). Poverty mapping in small areas under a two-fold nested error regression model. *Journal of the Royal Statistical Society, series A, 180*(4), 1111–1136.

Marchetti, S., Tzavidis, N., & Pratesi, M. (2012). Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. *Computational Statistics and Data Analysis, 56*, 2889–2902.

Marcis, L., Morales, D., Pagliarella, M. C., & Salvatore, R. (2023). Three-fold Fay–Herriot model for small area estimation and its diagnostics. *Statistical Methods and Applications, 32*, 1563–1609.

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics, 38*, 369–385.

Morales, D., Pagliarella, M. C., & Salvatore, R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. *SORT-Statistics and Operations Research Transactions, 39*(1), 19–34.

Morales, D., Esteban, M. D., Pérez, A., & Hobza, T. (2021). *A course on small area estimation and mixed models*. Springer.

Morales, D., Krause, J., & Burgard, J. P. (2022). On the use of aggregate survey data for estimating regional major depressive disorder prevalence. *Psychometrika, 87*(1), 344–368.

Pfeffermann, D., & Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology, 16*, 217–237.

Pratesi, M. (2016). *Analysis of poverty data by small area estimation*. Wiley.

Porter, A. T., Wikle, C. K., & Holan, S. H. (2015). Small area estimation via multivariate Fay–Herriot models with latent spatial dependence. *Australian and New Zealand Journal of Statistics, 57*(1), 15–29.

Rao, J. N. K., & Yu, M. (1994). Small area estimation by combining time series and cross sectional data. *Canadian Journal of Statistics, 22*, 511–528.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). Wiley.

Reardon, S. F., & Firebaugh, G. (2002). Measures of multigroup segregation. *Sociological Methodology, 32*, 33–67.

Reardon, S. F., & O'Sullivan, D. (2004). Measures of spatial segregation. *Sociological Methodology, 34*, 121–62.

Roberto, E. (2016). The divergence Index: A decomposable measure of segregation and inequality. arXiv:1508.01167v2 [stat.ME] 5 Dec 2016.

Salardi, P. (2016). The evolution of gender and racial occupational segregation across formal and non-formal labor markets in Brazil, 1987 to 2006. *Review of Income and Wealth, 62*(S1), 68–89.

Singh, B., Shukla, G., & Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology, 31*, 183–195.

Taeuber, K. E., & Taeuber, A. F. (1965). *Negroes in cities: Residential segregation and neighborhood change*. Aldine Pub.

"The 17 Goals. Sustainable Development" (2015). Dept of Economic and Social Affairs. New York City.

Tzavidis, N., Salvati, N., Pratesi, M., & Chambers, R. (2008). M-quantile models with application to poverty mapping. *Statistical Methods and Applications, 17*, 393–411.

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E., & Chambers, R. (2015). Robust small area prediction for counts. *Statistical Methods in Medical Research, 24*(3), 373–395.

You, Y., & Rao, J. N. K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology, 26*, 173–181.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.