SORT 48 (1) January-June 2024, 1-28 DOI: 10.57645/20.8080.02.16

# Small area estimation of the proportion of single-person households: Application to the Spanish Household Budget Survey

María Bugallo Porto[1,*], Domingo Morales González[1]
and María Dolores Esteban Lefler[1]

**Abstract**

Household composition reveals vital aspects of the socioeconomic situation and major changes in developed countries for decision-making and mapping the distribution of single-person households is highly relevant and useful. Driven by the Spanish Household Budget Survey data, we propose a new statistical methodology for small area estimation of proportions and total counts of single-person households. Estimation domains are defined as crosses of province, sex and age group of the main breadwinner of the household. Predictors are based on area-level zero-inflated Poisson mixed models. Model parameters are estimated by maximum likelihood and mean squared errors by parametric bootstrap. Several simulation experiments are carried out to empirically investigate the properties of these estimators and predictors. Finally, the paper concludes with an application to real data from 2016.

***MSC:*** *62J12, 62P25, 62D05.*

***Keywords:*** *Small area estimation, zero-inflated Poisson mixed model, area-level data, Household Budget Survey, single-person household.*

## 1. Introduction

National statistical offices plan surveys to provide a cost effective way of obtaining accurate estimates at a certain level of aggregation. Nonetheless, disaggregated statistics can facilitate more effective targeting of decision-making, but obviously require more information to adequately represent population subgroups. If domain sample sizes are

* *Corresponding author:* mbugallo@umh.es

[1] Operations Research Center, Miguel Hernández University of Elche (Spain). Address: Edificio Torretamarit - Avda. de la Universidad s/n, 03202 Elche (Alicante).

Received: January 2023
Accepted: November 2023

large enough, we can accurately estimate domain characteristics using direct estimators, such as the Hájek estimator (Hájek, 1971). The term "small areas" is commonly used to describe domains with too small sample sizes to obtain precise direct estimates. In these cases, indirect estimation techniques, relying on statistical modelling, will have to be used. Small area estimation (SAE) addresses this challenge by borrowing strength from auxiliary variables, data from other domains and underlying dependency structures.

Given the topic of our research, in recent decades, most developed countries have faced major demographic changes that directly affect household composition (Cohen, 2021), with new forms of cohabitation replacing the traditional concept of "*two-parent family with children*" (Lesthaeghe, 2014). In a context of social transformation, living alone has become a sign of individual autonomy and freedom (Fritsch, Riederer and Seewann, 2023), even if it is sometimes still stereotyped (Greitemeyer, 2009). Meanwhile, loneliness and its impact on physical and mental health are an increasingly widespread problem (Snell, 2017), accentuating the symptoms of cognitive diseases (Lee and Lee, 2021; Park et al., 2016). Especially, among elderly single-person households, the need for medical care is expected to be high, and even more so compared to other age groups. Hence the natural need for research aimed at curbing these problems.

Among the main indicators of loneliness, we can mention the proportion and total count of single-person households by domains defined by territorial and demographic features. Indeed, the disaggregated mapping of these indicators provides valuable information for governments to implement social and health policies aimed at improving the well-being of people suffering from loneliness. Hence, more specific studies are needed. In addition, the number and size of households in the coming years is closely related to demographic projections (Ortiz-Ospina, 2019) and their distribution across provinces, sex and age groups is therefore of particular interest (Cho et al., 2019). For that purpose, this paper develops a new statistical methodology and illustrates its use with an application to the Spanish Household Budget Survey (SHBS), where the aim is to estimate proportions of single-person households by Spanish province, sex and age group. However, it can be applied to other contexts where the same problem holds.

The following is an overview of the state of the art. SAE uses linear mixed models (LMMs) and generalized linear mixed models (GLMMs) that can be fitted to either unit or area-level data. Area-level models have the advantage of easily incorporate auxiliary variables from statistical sources other than the sample. Namely, Torabi and Rao (2014) and Cai and Rao (2022) use subarea-models to deal with hierarchically structured data. Zhang and Chambers (2004) develops log-linear structural models suitable to estimate small area cross-classified counts based on survey data. Esteban et al. (2012), Marhuenda, Molina and Morales (2013); Marhuenda, Morales and Pardo (2014) and Morales, Pagliarella and Salvatore (2015) estimate poverty proportions based on LMMs. For GLMMs, binomial and multinomial mixed models are applied to estimate proportions by Molina, Saei and Lombardía (2007), Ghosh et al. (2009), Chandra and Chambers (2011), Chen and Lahiri (2012), Chambers, Salvati and Tzavidis (2012), López-Vizcaíno, Lombardía and Morales (2013, 2015), Militino, Ugarte and Goicoa

(2015), Chambers, Salvati and Tzavidis (2016), Hobza and Morales (2016), Liu and Lahiri (2017), as well as Hobza, Morales and Santamaría (2018). Poisson (PO) and Negative Binomial (NB) mixed models are employed to estimate counts and proportions by Dreassi, Petrucci and Rocco (2014), Tzavidis et al. (2015), Boubeta, Lombardía and Morales (2016, 2017) and Morales, Krause and Burgard (2022), among others. As for the computational limitations of PO-GLMMs, but with a unit-level approach, the conjugate form of the Gamma-PO model allows for computationally light estimation and prediction procedures (Berg, 2022). However, none of the above cited papers deal with data with excess zeros.

In scientific and technical studies it is common to find count data with many zeros (Zuur et al., 2009; Michael and Thomas, 2016). This is the case for our target variable, the count of single-person households by domains. A possible solution is to fit a Fay-Herriot (FH) model after a transformation, and apply the methodology of Berg and Fuller (2012) to obtain a non-zero variance estimate if the observed value is zero. Another approach is to consider models in which the probability of the null count is modified with respect to that which would correspond to a given probability distribution. Because of their flexibility, zero-inflated models play a relevant role. Without wishing to be exhaustive, we cite some papers where these models are used in SAE. Pfeffermann, Terryn and Moura (2008) consider situations where the value of the target variable is zero or an observation from a continuous distribution. They analyse the assessment of literacy proficiency with the possible outcome of zero, indicating illiteracy, or a positive score measuring the literacy level. Chandra, Bathla and Sud (2010) and Chandra and Sud (2012) introduce unit-level mixtures between zero and a LMM. They estimate domain means of continuous variables when the census vector contains a substantial proportion of zeros. Chandra and Chambers (2011) generalize their previous proposal by modelling logarithms. Anggreyani, Indahwati and Kurnia (2015) estimate infant mortality using plug-in predictors based on area-level mixed effects zero-inflated PO models. Krieg, Boonstra and Smeets (2016) and Sadik, Anisa and Aqmaliyah (2019) have carried out simulation experiments for unit-level mixtures between zero and a nested error regression model under a Bayesian approach. Hartono, Kurnia and Indahwati (2017) deal with area-level zero-inflated binomial models, with an application to unemployment data in Indonesia. Datta and Mandal (2015) and Sugasawa, Kubokawa and Ogasawara (2017) propose uncertain random effects, which are expressed as mixtures of a normal distribution and a one-point-at-zero distribution. Bugallo et al. (2023) model the number of fires in small areas using a zero-inflated NB mixed model.

Currently, there are no published studies that address the estimation of proportions of single-person households in small areas. However, it is essential for a more accurate implementation of social policies, as well as for clarifying certain economic aspects related to the housing sector and the private consumption of basic resources. Because of this challenge, we introduce a zero-inflated PO mixed model, that is, a mixture model with a logistic mixed model on a latent variable that indicates whether we count zero or according to a PO mixed model. Based on that model, we construct predictors of domain-

level counts and proportions. To estimate mean squared errors (MSE) of small area predictors, we lay out a parametric bootstrap method following González-Manteiga et al. (2007) and González-Manteiga et al. (2008). For assessment and illustration, several simulation experiments and a detailed application to real data are included. Regarding the latter, special attention has been reserved for the excess of zeros and the conciliation between the area-level model-based approach and the traditional survey sampling design-based approach. Further comparisons are also made with a FH model and an area-level zero-inflated NB mixed model.

The main document is organized as follows. Section 2 describes the data and SAE problem. Section 3 introduces the area-level zero-inflated PO mixed model. Section 4 provides model-based predictors of domain counts and proportions. Section 5 presents bootstrap-based confidence intervals (CI) of model parameters and MSE estimators of the predictors. Section 6 addresses the case study. Section 7 summarizes some conclusions. The paper includes supplementary material organized in four appendices. Appendix A describes the Laplace approximation to the model log-likelihood and the algorithm to calculate the maximum likelihood (ML) estimators of model parameters and obtain modal predictors of random effects. Appendix B empirically investigates the behaviour of the fitting algorithm, predictors and MSE estimators. Appendix C gives some additional simulation results. Appendix D maps relative root mean squared error (RRMSE) estimates for the application to real data.

## 2. Data and problem of interest

This paper presents and applies a new SAE methodology, based on an area-level zero-inflated PO mixed model, to estimate proportions of single-person households in small areas. The script has been approached from an applied point of view, in order to provide a reference text for future research on zero-inflated data in SAE. As far as the dataset is concerned, we use the 2016 SHBS (SHBS2016). The anonymized data file can be downloaded from the Spanish Statistical Office (INE) website. Regarding sample sizes, the SHBS2016 is designed to calculate precise direct estimators at NUTS 2 level, but it does not publish results at a lower level of aggregation. Below that level, sample sizes are quite small and direct estimators lose precision. In our research, we consider $D = 416$ domains defined at NUTS 3 level by Spanish province ($I = 52$) crossed by sex ($J = 2$) and age group ($K = 4$). Given the sample sizes of SHBS2016, we are faced with an SAE problem. In fact, the quartiles of the small area sample sizes are $q_0 = 1$, $q_{0.25} = 17$, $q_{0.5} = 34$, $q_{0.75} = 72$ and $q_1 = 367$, respectively. Therefore, it is desirable to use more sophisticated prediction methods rather than direct estimators. In terms of methodology, Section 2.1 describes our research framework; Section 2.2 introduces the explanatory variables of the case study and Section 2.3 focuses on the zero inflation problem.

### 2.1. Count and size variables

Further notation is introduced below. Formally, the finite population of Spanish households, $U$, can be partitioned in subpopulations $U_{ijk}$, $i \in \mathbb{I} = \{1,\ldots,I\}$, $j \in \mathbb{J} = \{1,\ldots,J\}$, $k \in \mathbb{K} = \{1,\ldots,K\}$, defined by province, sex (*sex1*: men, *sex2*: women) and age group (*age1*: less than 45 years; *age2*: between 46 and 55 years; *age3*: between 56 and 64 years; *age4*: 65 years or older) of the main breadwinner. This is to say, each $U_{ijk}$ is disjoint and $U = \bigcup_{i=1}^{I}\bigcup_{j=1}^{J}\bigcup_{k=1}^{K} U_{ijk}$. Let $N$ and $N_{ijk}$ be the sizes of populations $U$ and $U_{ijk}$, respectively.

At unit-level, the variable of interest is dichotomic, i.e. $y_{ijkl} = 1$ if the household $u_{ijkl} \in U_{ijk}$ is single-person and $y_{ijkl} = 0$, otherwise. Let $s = \bigcup_{i=1}^{I}\bigcup_{j=1}^{J}\bigcup_{k=1}^{K} s_{ijk}$ be a SHBS sample extracted from $U$, so that $s_{ijk} \subset U_{ijk}$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$. Let $n$ and $n_{ijk}$ be the sample sizes of $s$ and $s_{ijk}$, respectively. For ease of exposition, we write $l = 1,\ldots,n_{ijk}$ for the households in $s_{ijk}$ and $l = n_{ijk}+1,\ldots,N_{ijk}$ for the households in $U_{ijk}\backslash s_{ijk}$.

The domain parameters of interest are the total count and proportion of single-person households in $U_{ijk}$, i.e.

$$Y_{ijk} = \sum_{l=1}^{N_{ijk}} y_{ijkl}, \quad \overline{Y}_{ijk} = \frac{Y_{ijk}}{N_{ijk}}, \quad i \in \mathbb{I},\, j \in \mathbb{J},\, k \in \mathbb{K}. \tag{2.1}$$

Let $w_{ijkl}$ be the household sampling weight of $u_{ijkl} \in U_{ijk}$. The sample count and the Hájek estimator of $Y_{ijk}$ and $N_{ijk}$ are

$$y_{ijk\cdot} = \sum_{l=1}^{n_{ijk}} y_{ijkl}, \quad \hat{Y}_{ijk}^{dir} = \sum_{l=1}^{n_{ijk}} w_{ijkl}\, y_{ijkl}, \quad \hat{N}_{ijk}^{dir} = \sum_{l=1}^{n_{ijk}} w_{ijkl}, \quad i \in \mathbb{I},\, j \in \mathbb{J},\, k \in \mathbb{K}.$$

The sample proportion and the Hájek estimator of $\overline{Y}_{ijk}$ are

$$\overline{y}_{ijk\cdot} = \frac{y_{ijk\cdot}}{n_{ijk}}, \quad \hat{\overline{Y}}_{ijk}^{dir} = \frac{\hat{Y}_{ijk}^{dir}}{\hat{N}_{ijk}^{dir}}, \quad i \in \mathbb{I},\, j \in \mathbb{J},\, k \in \mathbb{K}. \tag{2.2}$$

Once the count and size variables have been presented, it is important to be aware of the following scheme. Section 3 details the area-level zero-inflated PO mixed model and Section 4 proposes model-based predictors of the domain parameters defined in (2.1). Nevertheless, this requires an external file with auxiliary variables aggregated at domain level. In any case, it must contain the dependent variable of the area-level model, $y_{ijk}$, the size variable (offset), $m_{ijk}$, and a vector of domain-level auxiliary variables, $\boldsymbol{x}_{ijk}$ (see Section 2.2). As far as $y_{ijk}$ and $m_{ijk}$ are concerned, two options can be considered:
*Option 1.* Take $y_{ijk} = \lfloor \hat{Y}_{ijk}^{dir} \rfloor$ and $m_{ijk} = \lfloor N_{ijk} \rfloor$, where $\lfloor\cdot\rfloor$ is the closest integer operator. Let $\hat{\mu}_{yijk}$ be a model-based predictor of the expected value of $y_{ijk}$. The predictors of $\overline{Y}_{ijk}$ and $Y_{ijk}$ are

$$\hat{\overline{Y}}_{ijk} = \frac{\hat{\mu}_{yijk}}{m_{ijk}}, \quad \hat{Y}_{ijk} = \hat{\mu}_{yijk}, \quad i \in \mathbb{I},\, j \in \mathbb{J},\, k \in \mathbb{K}.$$

By taking the direct estimators of domain totals as the dependent variable of the area-level model, Option 1 reconciles the area-level model-based approach and the sample design approach to inference in finite populations. This is an important argument in favour of Option 1. On the other hand, the fitting algorithm or the calculation of predictors may become unstable when the values of the dependent variable are large, which require more refined programming.

*Option 2*. Take $y_{ijk} = y_{ijk\cdot}$ and $m_{ijk} = n_{ijk}$. Let $\hat{\mu}_{yijk}$ be a model-based predictor of the expected value of $y_{ijk}$. The predictors of $\overline{Y}_{ijk}$ and $Y_{ijk}$ are

$$\hat{\overline{Y}}_{ijk} = \frac{\hat{\mu}_{yijk}}{m_{ijk}}, \quad \hat{Y}_{ijk} = \hat{N}^{dir}_{ijk}\hat{\overline{Y}}_{ijk}, \quad i \in \mathbb{I},\, j \in \mathbb{J},\, k \in \mathbb{K}.$$

Boubeta, Lombardía and Morales (2016) applies Option 2 for area-level PO mixed models, as it is computationally more robust, but it does not include the sampling weights into the model. As omission of sampling weights is an important problem with Option 2, because it can lead to biased predictors, our choice of Option 1 is properly justified, even if it makes programming more difficult.

### 2.2. Domain-level auxiliary information

Population sizes and domain-level auxiliary variables have been estimated from the 2016 Spanish Labour Force Survey (SLFS). The SLFS is published quarterly, includes nearly 65,000 dwellings, equivalent to approximately 160,000 people, and collects data on the labour force and its various categories, as well as on the population outside the labour market. The anonymized data file can be downloaded from the INE website. The sample size of each quarterly SLFS is larger than three times the size of an annual SHBS. From the first to the last quarter of 2016, there are about $4 \cdot 160,000$ respondents. As there are $D = 416$ estimation domains, it is expected an average of 1538 respondents per domain. In order to improve our results, we jointly use data from the four quarters of 2016 and apply (2.2). In this way the effects of the variances of the covariate means on the properties of the prediction procedure are considered negligible.

The set of domain-level auxiliary variables is calculated by estimating the proportion of people in the following factor categories: *Citizenship*: Spanish (cit1) and foreign (cit2); *Education*: primary or less (edu1), basic secondary education (edu2), advanced secondary education (edu3) and higher education, such as university (edu4); *Labour situation*: employed (lab1), unemployed (lab2) and inactive (lab3); *Civil status*: unmarried (civ1), married (civ2), widower (civ3) and separated or divorced (civ4); *Dwelling mobility*: more than a year in the same dwelling (dwe1) and the opposite (dwe2). The above-mentioned auxiliary variables are proportions, bounded in the interval [0, 1], i.e. they are continuous variables, not binary indicators. Since the sum of proportions in the categories of each factor is one, and based on their socio-economic meaning, we omit one category from each factor. Namely, we have deleted cit2, edu2, lab3, civ1, dwe2.

### 2.3. Zero inflation

So far we have discussed the necessity of auxiliary information, but we have not addressed the problem of excess zeros, nor even demonstrated its occurrence. Nevertheless, it is important to assess the presence of false zeros to model counts of single-person households by province, sex and age group. The reason lies in the low number of respondents at some crosses and thus the difficulty of detecting single-person households. Throughout the paper, it will be shown why the incorporation of zero-inflated structures is more appropriate for the case study. As we assume that $y_{ijk}$ counts the number of single-person households in $U_{ijk}$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, it can be described by an area-level PO mixed model with offset parameter $m_{ijk}$ and some explanatory variables. However, the target variable is aggregated by province, sex and age group, so that the number of households in $s_{ijk}$ may be too small. Moreover, there are 28 domains with zero single-person households in SHBS2016. As the number of zeros seems to be too large, it has been decided to fit an area-level zero-inflated PO mixed model to $(y_{ijk}, m_{ijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$.

**Table 2.1.** *Distribution of domains where single-person households are not observed in SHBS2016, by sex and age group of the main breadwinner.*

| | sex | | |
|---|---|---|---|
| age group | *sex1* | *sex2* | Total |
| *age1* | 3 | 8 | 11 |
| *age2* | 2 | 8 | 10 |
| *age3* | 2 | 2 | 4 |
| *age4* | 3 | 0 | 3 |
| Total | 10 | 18 | 28 |

Table 2.1 presents the distribution of zeros by sex and age group in SHBS2016. It is shown that the 28 zeros are mainly concentrated in certain sex-age group categories. In fact, it can be suggested that single-person households inhabited by young and middle-age women are likely to be more difficult to capture in the count, i.e. their expected proportion is lower. The opposite is true at older ages. In any case, the number of zeros appears to be too large for what would be expected under a PO distribution. This motivates that a zero-inflated PO mixed model will have a better performance. Section 6 and Appendix B analytically justify the importance of incorporate the zero-inflated structure, both in terms of significance and goodness-of-fit. The area-level PO mixed model and the area-level zero-inflated PO mixed model will be compared and the latter will be chosen because it will give better results.

In order to test the dependence between the count of zeros/non zeros and provinces, sex and age groups, we have applied the Pearson's Chi-Squared test in $2 \times I$, $2 \times J$ and $2 \times K$ contingency tables, calculating p-values by Monte Carlo (MC). As a result, p-values close to 0.06 are reached for province and age group as inputs, increasing to 0.18

for sex. Based on Table 2.1 and the results of the above tests, we have decided to consider only age-group randomness to model zero-inflated probabilities. Furthermore, applying the same tests to assess the dependence between the count of single-person households (less/greater than 1, 2 or 3) and provinces, sex and age groups, only the randomness of the age group is significant. Guided by the promise of finding a good, simple model, Section 3 presents our methodological proposal.

## 3. Area-level zero-inflated Poisson mixed model

This section describes the area-level zero-inflated PO mixed model proposed as a basis to derive small area predictors of the proportion of single-person households by domains. All mathematical steps are detailed, justifying the soundness of what is presented. The formulation of the model is given in an orderly fashion, followed by the description of the fitting algorithm in Appendix A, the ML-Laplace approximation. Although the model is proposed in a general form, it is adapted for application to real data where appropriate. In fact, the model description is based on the stratification used in the real data example in Section 6, because it is in these domains that the need to incorporate a zero-inflated structure to model the response variable has been assessed. Even so, it is easily adaptable to other situations involving the general zero inflation problem.

Let us consider a count variable $y_{ijk}$ taking values on $\mathbb{N}\cup\{0\}$, $i\in\mathbb{I}$, $j\in\mathbb{J}$, $k\in\mathbb{K}$. Let $D=IJK$ be the total number of $y$-values. As a particular case, a country divided into provinces, sex and age groups can be modelled as follows. Let $z_{ijk}$, $\boldsymbol{x}_{1,ijk}=(x_{1,ijk1},\ldots,x_{1,ijkq_1})$ and $\boldsymbol{x}_{2,ijk}=(x_{2,ijk1},\ldots,x_{2,ijkq_2})$ be latent (non observable) variables and $1\times q_1$ and $1\times q_2$ row vectors containing area-level explanatory variables, respectively. Define the vectors and matrices $\boldsymbol{y}_{ij}=\underset{1\le k\le K}{\text{col}}(y_{ijk})$, $\boldsymbol{z}_{ij}=\underset{1\le k\le K}{\text{col}}(z_{ijk})$, $\boldsymbol{X}_{1,ij}=\underset{1\le k\le K}{\text{col}}(\boldsymbol{x}_{1,ijk})$. $\boldsymbol{X}_{2,ij}=\underset{1\le k\le K}{\text{col}}(\boldsymbol{x}_{2,ijk})$, $\boldsymbol{y}=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\boldsymbol{y}_{ij}))$, $\boldsymbol{z}=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\boldsymbol{z}_{ij}))$, $\boldsymbol{X}_1=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\boldsymbol{X}_{1,ij}))$ and $\boldsymbol{X}_2=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\boldsymbol{X}_{2,ij}))$. In order to understand how the data are stacked according to the col(.) operator, we rely on the application to the SHBS2016 data as a useful example. In this dataset, the $D=416$ domains are sorted by age group and, within each age group, the Spanish provinces are concatenated, first for males and then for females.

Let $u_{1,k}$, $u_{2,ijk}$ be independent $N(0,1)$ random effects, $\boldsymbol{u}_1=\underset{1\le k\le K}{\text{col}}(u_{1,k})\sim N_K(\boldsymbol{0},\boldsymbol{I})$, $\boldsymbol{u}_2=\underset{1\le i\le I}{\text{col}}(\underset{1\le j\le J}{\text{col}}(\underset{1\le k\le K}{\text{col}}(u_{2,ijk})))\sim N_{IJK}(\boldsymbol{0},\boldsymbol{I})$, $\boldsymbol{u}=(\boldsymbol{u}_1^\intercal,\boldsymbol{u}_2^\intercal)^\intercal$. The bivariate vector $(y_{ijk},z_{ijk})$ follow an area-level zero-inflated PO (aZIP13) mixed model if

$$z_{ijk}\overset{ind}{\sim}\text{BE}(p_{ijk}),\ P(y_{ijk}=0/z_{ijk}=1)=1,\ P(y_{ijk}=t/z_{ijk}=0)=\frac{e^{-\mu_{ijk}}\mu_{ijk}^t}{t!},\ t\in\{0\}\cup\mathbb{N},$$

where $0<p_{ijk}<1$, $\mu_{ijk}=m_{ijk}\lambda_{ijk}$, $m_{ijk}\in\mathbb{N}$ is known, $\lambda_{ijk}>0$ and $p_{ijk}$ and $\lambda_{ijk}$ depend on the explanatory variables $\boldsymbol{x}_{1,ijk}$ and $\boldsymbol{x}_{2,ijk}$, on the regression parameters $\boldsymbol{\beta}_1=(\beta_{11},\ldots,\beta_{1q_1})^\intercal$ and $\boldsymbol{\beta}_2=(\beta_{21},\ldots,\beta_{2q_2})^\intercal$, and on the standard deviation parameters $\phi_1>$

0 and $\phi_2 > 0$ by means of the link functions

$$\text{logit}(p_{ijk}) = \log \frac{p_{ijk}}{1-p_{ijk}} = \boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k} = \sum_{\ell=1}^{q_1} x_{1,ijk\ell}\beta_{1\ell} + \phi_1 u_{1,k}$$

$$\log(\lambda_{ijk}) = \boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk} = \sum_{\ell=1}^{q_2} x_{2,ijk\ell}\beta_{2\ell} + \phi_2 u_{2,ijk}, \quad i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}.$$

Inverting the above functions, it follows that

$$p_{ijk} = \frac{\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}}, \quad \lambda_{ijk} = \exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}, \quad i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}. \tag{3.1}$$

In short, the proposed model is a mixture of two mixed submodels. First, the BE submodel drives the mixture and incorporates the information derived from the excess of zeros. Subsequently, the PO submodel deals with the modelling of count variables. To complete its definition, it is assumed that $(y_{ijk}, z_{ijk})^\intercal$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, are independent conditioned to $\boldsymbol{u}$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_2^\intercal, \phi_1, \phi_2)^\intercal$ be the vector of model parameters and define $\xi_{ijk} = I_{\{0\}}(y_{ijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$. This is to say, $\xi_{ijk} = 1$ if $y_{ijk} = 0$ and $\xi_{ijk} = 1$, otherwise. It holds that

$$\begin{aligned}
&P(y_{ijk}|u_{1,k}, u_{2,ijk}; \boldsymbol{\theta}) = \xi_{ijk}\left[p_{ijk} + (1-p_{ijk})e^{-\mu_{ijk}}\right] + (1-\xi_{ijk})\left[(1-p_{ijk})\frac{e^{-\mu_{ijk}}\mu_{ijk}^{y_{ijk}}}{y_{ijk}!}\right] \\
&= \left(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}\right)^{-1} \\
&\cdot \left\{\xi_{ijk}\left[\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\} + \exp\left\{-m_{ijk}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}\right\}\right]\right. \\
&+ (1-\xi_{ijk})\exp\left\{y_{ijk}(\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}) - m_{ijk}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}\right. \\
&\left.\left. + y_{ijk}\log m_{ijk} - \log y_{ijk}!\right\}\right\}, \ i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}.
\end{aligned}$$

By the independence assumptions, we have that

$$P(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\theta}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} P(y_{ijk}|u_{1,k}, u_{2,ijk}; \boldsymbol{\theta}).$$

Therefore, the likelihood function of the aZIP13 mixed model is

$$\begin{aligned}
P(\boldsymbol{y}; \boldsymbol{\theta}) &= \int_{\mathbb{R}^{K(1+IJ)}} P(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\theta}) f_u(\boldsymbol{u})\, d\boldsymbol{u} \qquad (3.2) \\
&= \prod_{k=1}^{K} \int_{\mathbb{R}^{1+IJ}} \Big(\prod_{i=1}^{I}\prod_{j=1}^{J} P(y_{ijk}|u_{1,k}, u_{2,ijk}; \boldsymbol{\theta}) f_{N(0,1)}(u_{2,ijk})\, du_{2,ijk}\Big) f_{N(0,1)}(u_{1,k})\, du_{1,k},
\end{aligned}$$

and the respective log-likelihood function is

$$\ell(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{k=1}^{K} \log \int_{\mathbb{R}^{1+IJ}} \Big( \prod_{i=1}^{I} \prod_{j=1}^{J} P(y_{ijk}|u_{1,k}, u_{2,ijk};\boldsymbol{\theta}) f_{N(0,1)}(u_{2,ijk})\, du_{2,ijk} \Big) f_{N(0,1)}(u_{1,k})\, du_{1,k}.$$

Given $\boldsymbol{y}$, the ML estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\, \ell(\boldsymbol{\theta};\boldsymbol{y})$, where $\Theta = \mathbb{R}^{q_1+q_2} \times \mathbb{R}_+^2$ and $\mathbb{R}_+ = (0,\infty)$. The expression of $\ell(\boldsymbol{\theta};\boldsymbol{y})$ contains integrals in $\mathbb{R}^{1+IJ}$. To maximize it, two functions can be applied sequentially. The first one would compute the integral on $\mathbb{R}^{1+IJ}$ and the second one would perform the maximization on $\boldsymbol{\theta}$. As this approach is not efficient, Appendix A describes the ML-Laplace approximation as an alternative maximization method.

## 4. Prediction of totals and proportions

Under the assumption that $y_{ijk}$, $i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}$, follows the proposed aZIP13 mixed model, this section is devoted to the development of new small area predictors. Typical of the literature, the inference is focused on the expected values

$$\mu_{yijk} \triangleq E[y_{ijk}|\boldsymbol{u}_{ijk}] = m_{ijk}(1-p_{ijk})\lambda_{ijk},\ i \in \mathbb{I},\ j \in \mathbb{J},\ k \in \mathbb{K}, \tag{4.1}$$

where $p_{ijk} = p_{ijk}(u_{1,k})$ and $\lambda_{ijk} = \lambda_{ijk}(u_{2,ijk})$ are defined in (3.1). In an orderly fashion, first the plug-in predictor is introduced. Subsequently, the best predictor and its empirical version are derived (see Molina, Saei and Lombardía (2007) for futher details). At the expense of the theoretical properties, simpler alternatives are finally proposed looking for a better computational performance. Under a scenario based on SHBS2016, they will be compared in simulation experiments in Appendix B so as to justify the application to real data in Section 6.

Firstly, by plugging ML estimators and modal predictors, the population-based quantities given by (4.1) can be predicted using the plug-in (IN) predictor, defined as

$$\hat{\mu}_{yijk}^{in} = m_{ijk}\big(1+\exp\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1\hat{u}_{1,k}\}\big)^{-1} \exp\{\boldsymbol{x}_{2,ijk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2\hat{u}_{2,ijk}\}.$$

Among the different predictors that can be mentioned, this is the simplest approach to understand and the easiest to calculate. Indeed, its ease of interpretation and calculation, as well as its computational performance and execution times, are unsurpassed. Nevertheless, there are other potentially competitive alternatives. Let us define $\boldsymbol{y}_k = \underset{1\le i\le I}{\text{col}}\big(\underset{1\le j\le J}{\text{col}}(y_{ijk})\big)$, $\boldsymbol{u}_{2,k} = \underset{1\le i\le I}{\text{col}}\big(\underset{1\le j\le J}{\text{col}}(u_{2,ijk})\big)$, $\boldsymbol{v}_k = (u_{1,k}, \boldsymbol{u}_{2,k}^{\intercal})^{\intercal}$. The best predictor (BP) of (4.1) is $\hat{\mu}_{yijk}^{bp}(\boldsymbol{\theta}) = m_{ijk}E[(1-p_{ijk})\lambda_{ijk}|\boldsymbol{y}_k]$. The conditional expectation $E_{ijk} = E[(1-p_{ijk})\lambda_{ijk}|\boldsymbol{y}_k]$ is

$$E_{ijk} = \frac{\int_{\mathbb{R}^{1+IJ}} \big(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}\big)^{-1} \exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\} P(\boldsymbol{y}_k|\boldsymbol{v}_k) f(\boldsymbol{v}_k)\, d\boldsymbol{v}_k}{\int_{\mathbb{R}^{1+IJ}} P(\boldsymbol{y}_k|\boldsymbol{v}_k) f(\boldsymbol{v}_k)\, d\boldsymbol{v}_k}.$$

Denote the numerator and denominator of $E_{ijk}$ by $A_{ijk} = A_{ijk}(\boldsymbol{y}_k, \boldsymbol{\theta})$ and $B_k = B_k(\boldsymbol{y}_k, \boldsymbol{\theta})$, respectively. Define $\xi_{rtk} = I_{\{0\}}(y_{rtk})$, $r \in \mathbb{I}$, $t \in \mathbb{J}$, $k \in \mathbb{K}$. It holds that

$$A_{ijk} = \int_{\mathbb{R}^{1+IJ}} \frac{\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2 + \phi_2 u_{2,ijk}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\}} \prod_{r=1}^{I}\prod_{t=1}^{J} \omega_{rtk} f_{N(0,1)}(u_{1,k}) f_{N(0,1)}(u_{2,rtk})\, du_{1,k}\, du_{2,rtk},$$

$$B_k = \int_{\mathbb{R}^{1+IJ}} \prod_{r=1}^{I}\prod_{t=1}^{J} \omega_{rtk} f_{N(0,1)}(u_{1,k}) f_{N(0,1)}(u_{2,rtk})\, du_{1,k}\, du_{2,rtk},$$

$$\begin{aligned}\omega_{rtk} = (1+\exp\{\boldsymbol{x}_{1,rtk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\})^{-1} \Bigg\{ \xi_{rtk} \Bigg[ \exp\{\boldsymbol{x}_{1,rtk}\boldsymbol{\beta}_1 + \phi_1 u_{1,k}\} \\ + \exp\Big\{ -m_{rtk}\exp\{\boldsymbol{x}_{2,rtk}\boldsymbol{\beta}_2 + \phi_2 u_{2,rtk}\}\Big\}\Bigg] + (1-\xi_{rtk})\exp\Big\{ y_{rtk}(\boldsymbol{x}_{2,rtk}\boldsymbol{\beta}_2 + \phi_2 u_{2,rtk}) \\ - m_{rtk}\exp\{\boldsymbol{x}_{2,rtk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u_{2,rtk}\} + y_{rtk}\log m_{rtk} - \sum_{a=1}^{y_{rtk}} \log a\Big\}\Bigg\}.\end{aligned}$$

The empirical best predictor (EBP) is $\hat{\mu}^{ebp}_{yijk} = \hat{\mu}^{bp}_{yijk}(\hat{\boldsymbol{\theta}})$ and can be calculated by a MC method using antithetic variables to reduce variability Hobza and Morales (2016). The outline is as follows:

1. Calculate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^{\mathsf{T}}, \hat{\boldsymbol{\beta}}_2^{\mathsf{T}}, \hat{\phi}_1, \hat{\phi}_2)^{\mathsf{T}}$.
2. For $s = 1, \ldots, S$, generate $u^{(s)}_{1,k}$, $u^{(s)}_{2,rtk}$ i.i.d. $N(0,1)$, $u^{(S+s)}_{1,k} = -u^{(s)}_{1,k}$, $u^{(S+s)}_{2,rtk} = -u^{(s)}_{2,rtk}$.
3. Calculate $\hat{\mu}^{ebp}_{yijk} = m_{ijk}\hat{A}_{ijk}/\hat{B}_k$, where

$$\hat{A}_{ijk} = \frac{1}{2S}\sum_{s=1}^{2S} \frac{\exp\{\boldsymbol{x}_{2,ijk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u^{(s)}_{2,ijk}\}}{1+\exp\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u^{(s)}_{1,k}\}} \prod_{r=1}^{I}\prod_{t=1}^{J} \hat{\omega}_{rtk},\ \hat{B}_k = \frac{1}{2S}\sum_{s=1}^{2S}\prod_{r=1}^{I}\prod_{t=1}^{J}\hat{\omega}_{rtk}, \tag{4.2}$$

$$\begin{aligned}\hat{\omega}_{rtk} = \frac{1}{1+\exp\{\boldsymbol{x}_{1,rtk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u^{(s)}_{1,k}\}} \Bigg\{ \xi_{rtk}\Bigg[ \exp\{\boldsymbol{x}_{1,rtk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u^{(s)}_{1,k}\} \\ + \exp\Big\{ -m_{rtk}\exp\{\boldsymbol{x}_{2,rtk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u^{(s)}_{2,rtk}\}\Big\}\Bigg] + (1-\xi_{rtk})\exp\Big\{ y_{rtk}(\boldsymbol{x}_{2,rtk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u^{(s)}_{2,rtk}) \\ - m_{rtk}\exp\{\boldsymbol{x}_{2,rtk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u^{(s)}_{2,rtk}\} + y_{rtk}\log m_{rtk} - \sum_{a=1}^{y_{rtk}}\log a\Big\}\Bigg\},\ \xi_{rtk} = I_{\{0\}}(y_{rtk}).\end{aligned}$$

It has been noted that (4.2) contains products with $IJ$ terms. Given the nature of our problem, these products are close to zero under Option 1, leading to numerical precision problems in Section 6 and Appendix B. Facing this challenge, we have introduced a simplified version of the BP by conditioning to $y_{ijk}$ instead of $\boldsymbol{y}_k$. This simplified predictor (SP) is $\hat{\mu}^{sp}_{yijk}(\boldsymbol{\theta}) = m_{ijk}E[(1-p_{ijk})\lambda_{ijk}|y_{ijk}]$. The conditional expectation

$E_{ijk}^{sp} = E[(1-p_{ijk})\lambda_{ijk}|y_{ijk}]$ is

$$E_{ijk}^{sp} = \frac{\int_{\mathbb{R}^2}\left(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1+\phi_1 u_{1,k}\}\right)^{-1}\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2+\phi_2 u_{2,ijk}\}P(y_{ijk}|\boldsymbol{u}_{ijk})f(\boldsymbol{u}_{ijk})\,d\boldsymbol{u}_{ijk}}{\int_{\mathbb{R}^2}P(y_{ijk}|\boldsymbol{u}_{ijk})f(\boldsymbol{u}_{ijk})\,d\boldsymbol{u}_{ijk}},$$

Denote the numerator and denominator of $E_{ijk}^{sp}$ by $A_{ijk}^{sp} = A_{ijk}^{sp}(y_{ijk},\boldsymbol{\theta})$ and $B_{ijk}^{sp} = B_{ijk}^{sp}(y_{ijk},\boldsymbol{\theta})$, respectively. It holds that

$$A_{ijk}^{sp} = \int_{\mathbb{R}^2}\frac{\exp\{\boldsymbol{x}_{2,ijk}\boldsymbol{\beta}_2+\phi_2 u_{2,ijk}\}}{(1+\exp\{\boldsymbol{x}_{1,ijk}\boldsymbol{\beta}_1+\phi_1 u_{1,k}\})}\,\omega_{ijk}\,f_{N(0,1)}(u_{1,k})f_{N(0,1)}(u_{2,rtk})\,du_{1,k}\,du_{2,rtk},$$

and

$$B_{ijk}^{sp} = \int_{\mathbb{R}^2}\omega_{ijk}\,f_{N(0,1)}(u_{1,k})f_{N(0,1)}(u_{2,rtk})\,du_{1,k}\,du_{2,rtk}.$$

The empirical simplified predictor (ESP) is $\hat{\mu}_{yijk}^{esp} = \hat{\mu}_{yijk}^{sp}(\hat{\boldsymbol{\theta}})$ and can be approximated by numerical approximation of integrals. However, the following antithetical MC algorithm is applied:

1. Calculate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^{\intercal}, \hat{\boldsymbol{\beta}}_2^{\intercal}, \hat{\phi}_1, \hat{\phi}_2)^{\intercal}$.
2. For $s = 1,\ldots,S$, generate $\boldsymbol{u}_{ij}^{(s)} = \left(u_{1,k}^{(s)}, u_{2,ijk}^{(s)}\right)^{\intercal}$ i.i.d. $N_2(\boldsymbol{0},\boldsymbol{I}_2)$, $\boldsymbol{u}_{ij}^{(S+s)} = -\boldsymbol{u}_{ij}^{(s)}$.
3. Calculate $\hat{\mu}_{yijk}^{esp} = m_{ijk}\hat{A}_{ijk}^{sp}/\hat{B}_{ijk}^{sp}$, where

$$\hat{A}_{ijk}^{sp} = \frac{1}{2S}\sum_{s=1}^{2S}\frac{\exp\{\boldsymbol{x}_{2,ijk}\hat{\boldsymbol{\beta}}_2+\hat{\phi}_2 u_{2,ijk}^{(s)}\}}{(1+\exp\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1+\hat{\phi}_1 u_{1,k}^{(s)}\})}\hat{\omega}_{ijk} \quad \text{and} \quad \hat{B}_{ijk}^{sp} = \frac{1}{2S}\sum_{s=1}^{2S}\hat{\omega}_{ijk}.$$

Because of the numerical precision of R, calculating exponential functions to predict $\mu_{yijk}$ may result in negative values that are too small. Consequently, $\omega_{ijk}$ would be close to zero. These overflow problems were detected by Boubeta, Lombardía and Morales (2016) and motivated these authors to choose Option 2, more computationally stable. In our case, as defined, the ESP allows us to solve them. Therefore, we assume Option 1, which is more convenient, as it reconciles to some extent the design-based and model-based approaches. Consequently, in simulations experiments in Appendix B and the case study in Section 6, the ESP will be used and the EBP will be omitted.

## 5. Bootstrap inference

This section presents bootstrap-based CIs for the model parameters and estimators of the MSEs of the predictors. For the latter, we adapt the procedures used by González-Manteiga et al. (2007, 2008).

### 5.1. Confidence intervals for model parameters

Let $\theta_\ell$ be a component of the vector of model parameters $\boldsymbol{\theta}$. Let $\alpha \in (0,1)$. The following procedure calculates a $(1-\alpha)\%$ percentile bootstrap CI for $\theta_\ell$.

1. Fit the model to the sample and calculate the ML estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top, \hat{\phi}_1, \hat{\phi}_2)^\top$.
2. Repeat $B$ times ($b = 1, \ldots, B$):
   (a) For $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, generate $u_{1,k}^{*(b)} \sim N(0,1)$, $u_{2,ijk}^{*(b)} \sim N(0,1)$ and calculate
   $$p_{ijk}^{*(b)} = \exp\left\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u_{1,k}^{*(b)}\right\}\left(1+\exp\left\{\boldsymbol{x}_{1,ijk}\hat{\boldsymbol{\beta}}_1 + \hat{\phi}_1 u_{1,k}^{*(b)}\right\}\right)^{-1}, \quad (5.1)$$
   $$\lambda_{ijk}^{*(b)} = \exp\left\{\boldsymbol{x}_{2,ijk}\hat{\boldsymbol{\beta}}_2 + \hat{\phi}_2 u_{2,ijk}^{*(b)}\right\}.$$
   (b) Generate $z_{ijk}^{*(b)} \sim \text{BE}(p_{ijk}^{*(b)})$. If $z_{ijk}^{*(b)} = 1$, do $y_{ijk}^{*(b)} = 0$. If $z_{ijk}^{*(b)} = 0$, generate $y_{ijk}^{*(b)} \sim \text{PO}(m_{ijk}\lambda_{ijk}^{*(b)})$.
   (c) On the basis of the bootstrap sample $(y_{ijk}^{*(b)}, m_{ijk}, \boldsymbol{x}_{ijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, calculate the ML estimate $\hat{\theta}_\ell^{*(b)}$.
3. Sort the values $\hat{\theta}_\ell^{*(b)}$, $b = 1, \ldots, B$, from smallest to largest. They are $\hat{\theta}_{\ell(1)}^* \leq \ldots \leq \hat{\theta}_{\ell(B)}^*$. A $(1-\alpha)\%$ percentile bootstrap CI for $\theta_\ell$ is $\left(\hat{\theta}_{\ell(\lfloor(\alpha/2)B\rfloor)}^*, \hat{\theta}_{\ell(\lfloor(1-\alpha/2)B\rfloor)}^*\right)$.

### 5.2. Mean squared error estimation

The model-based MSE of the EBP, ESP or IN predictor, $\hat{\mu}_{yijk}$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, can be estimated using a resampling method. The following procedure calculates a parametric bootstrap estimator of $MSE(\hat{\mu}_{yijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$.

1. Fit the model to the sample and calculate the ML estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top, \hat{\phi}_1, \hat{\phi}_2)^\top$.
2. Repeat B times ($b = 1, \ldots, B$):
   (a) Run Steps (a) and (b) of the algorithm detailed in Section 5.1.
   (b) For $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, calculate $\mu_{yijk}^{*(b)} = m_{ijk}(1-p_{ijk}^{*(b)})\lambda_{ijk}^{*(b)}$.
   (c) On the basis of the bootstrap sample $(y_{ijk}^{*(b)}, m_{ijk}, \boldsymbol{x}_{ijk})$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, calculate the ML estimate $\hat{\boldsymbol{\theta}}^{*(b)}$ and the predictor $\hat{\mu}_{yijk}^{*(b)}$.
3. Output: $mse^*(\hat{\mu}_{yijk}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\mu}_{yijk}^{*(b)} - \mu_{yijk}^{*(b)}\right)^2$, $i \in \mathbb{I}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$.
4. An estimator of the model-based MSE of the Hájek estimator is
$$mse^*(\hat{\bar{Y}}_{ijk}) = \frac{1}{B}\sum_{b=1}^{B}\left(y_{ijk}^{*(b)} - \mu_{yijk}^{*(b)}\right)^2, \quad i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}.$$

## 6. Application to real data

As a starting point, some considerations are presented to place the application in context and to encourage us in the work we are about to undertake. Regarding 2016, the Spanish Household Projection 2016–2031 addresses demographic trends and social patterns currently observed in Spain in terms of the number of households. Its authorship is attributed to the INE. It shows that households will increase by 4.9%, despite the decrease in the number of inhabitants, because of a reduction in the expected number of residents per dwelling, from 2.50 in 2016 to 2.35 in 2031. Related to this, between 2016 and 2031 the smallest households (one or two people in a shared dwelling) would continue to grow, while the largest ones would decrease, with a relative increase of 19.6% of single-person households. As a result, there will be more than 5.5 million single-person households (28.6%), with 12% of the Spanish population living alone.

For methodological purposes, this section applies the aZIP13 mixed model to the SHBS2016 data so as to estimate proportions of single-person households in small areas. Regarding the SHBS, it is published annually by the INE to study the nature and destination of consumer spending and the living conditions of households. The SHBS2016 includes around 22,000 dwellings, selected by means of a two-stage stratified random sampling carried out independently in each Autonomous Community (NUTS 2 level). Broadly speaking, the first stage units are territories with around 2,000 dwellings, called census sections. The second stage units are dwellings, interviewing all individuals over 16 years of age who reside in them. In each NUTS 2 region, the first stage units are stratified following a geographical criterion, which assigns the stratum according to the size of the municipality to which the section belongs. Sections are selected within each stratum with probability proportional to their population size. Dwellings are selected, within each section, with equal probability by means of systematic sampling with random start. The target variable $y_{ijk}$ is the direct estimate of the number of single-person households in a domain where $i$, $j$ and $k$ represent the province of residence, sex and age group of the main breadwinner, respectively. Furthermore, direct estimates of population sizes and area-level auxiliary variables have been obtained from the four 2016SLFS microdata. What is more, they have been considered as true population values because of the precision derived from the acceptable sample sizes of the 2016SLFS surveys.

Table 6.1 shows the ML estimates of the regression parameters (RP) $\boldsymbol{\beta}_1$, $\phi_1$ (BE submodel), $\boldsymbol{\beta}_2$ and $\phi_2$ (PO submodel), the p-values to test $H_0 : \beta_{t\ell} = 0$, $t = 1,2$, $\ell = 1,...,q_t$, and $H_0 : \phi_t = 0$, $t = 1,2$, and the normal-asymptotic and bootstrap CIs at a 95% confidence level. For convenience, their lower (LB) and upper (UB) bounds are provided. Normal-asymptotic CIs are discussed in Appendix A and bootstrap CIs in Section 5.

The final model incorporates only those variables that are significant at 5%. The flexibility achieved by making the random effects of the count model domain-dependent allows us to reduce the importance of the set of domain-level variables and incorporate

only those that actually add relevant knowledge. In order, edu4, ci1, edu1, tm1, ec4, lab1, lab2 are removed. The BE submodel contains one auxiliary variable, $x_{1,1}$ = intercept, and the PO submodel four: $x_{2,1}$ = intercept, $x_{2,2}$ = edu3, $x_{2,3}$ = civ2 and $x_{2,4}$ = civ3.

The only parameter of the BE submodel, $\beta_{11}$, is significantly non-zero and its CIs have an acceptable short length, which guarantees some precision in its estimation. Actually, the latter provides strong evidence in favour of the zero-inflated structure. According to Table 6.1, none of the area-level auxiliary variables is relevant to explain the zero-inflated probabilities, supporting our contribution. Null counts are caused by the difficulty of detecting single-person households in domains with small sample sizes. The basic zero-inflated probability is $p_0(\hat{\beta}_{11}) = 0.063$, which implies that the basic probability of obtaining an observation from the PO submodel is 0.937. However, it has already been proven that it is also important to take into account the age-group randomness. Here, it is confirmed that the asymptotic and bootstrap 95% CIs for $\phi_1$ do not contain the zero.

**Table 6.1.** *Regression parameters of the final aZIP13 mixed model.*

| | | BE submodel | | PO submodel | | | | |
|---|---|---|---|---|---|---|---|---|
| | RP | $\beta_{11}$ | $\phi_1$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{23}$ | $\beta_{24}$ | $\phi_2$ |
| | Estimate | -2.696 | 0.398 | -1.857 | 2.138 | -0.649 | 3.881 | 0.517 |
| | p-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Asymp. | LB 95% | -3.270 | 0.091 | -2.319 | 1.007 | -1.057 | 3.207 | 0.482 |
| | UB 95% | -2.121 | 1.752 | -1.395 | 3.269 | -0.242 | 4.554 | 0.555 |
| Boot. | LB 95% | -3.317 | 0.0002 | -2.312 | 1.051 | -1.016 | 3.215 | 0.480 |
| | UB 95% | -2.162 | 0.859 | -1.432 | 3.270 | -0.222 | 4.577 | 0.554 |

For the PO submodel, it could be suggested that a medium-high level of education ($\beta_{22}$), as well as being widower ($\beta_{24}$), contribute to increase the count of single-person households by domains, because their signs are significantly positive. On the other hand, an increase in the proportion of people who are married ($\beta_{23}$) implies a decrease in the number of single-person households, assuming that the other auxiliary variables are fixed. Given the group effect, it can be inferred that Spanish citizenship, employment status and dwelling mobility are not relevant to model the count of single-person households. The proportion of inhabitants with primary or university education and the proportion of separated or divorced people are also irrelevant. Last but not least, the asymptotic and bootstrap 95% CIs for $\phi_2$ do not contain the zero, confirming the necessity of modelling the counts with a random-effect model.

Back to the modelling of the zero-inflated structure, recall that Table 2.1 suggested that the number of zeros appears to be too large for what would be expected under a PO distribution. This statement is confirmed by comparing the number of zeros found in $B =$ 1000 bootstrap resamples under the PO mixed model and the proposed aZIP13 mixed

model. Indeed, for the PO mixed model there are no null counts in any resample, with 38 single-person households in the lowest case. The number of zeros in the data exceeds the number of zeros that could plausibly be generated by the fitted PO distribution. For the aZIP13 mixed model, each resample contains an average of 28 zeros, closely mimicking the structure of Table 2.1 and, thus, the behaviour of the target variable.

Hereafter, we assume the aZIP13 mixed model that Table 6.1 presents. To have more confidence in this model as a true generating model, Section 6.1 addresses its validation and Appendix B performs some simulation experiments under the SHBS2016 scenario. Importantly, they support the use of the IN predictor.

### 6.1. Model validation

Residual analysis is used to validate a model as well as to detect potential underlying dependency relationships. As the aZIP13 mixed model is an area-level model, model diagnosis is also performed at that level of aggregation. Besides, we are interested in the conciliation of the model-based approach and the design-based approach to SAE. Further notation is introduced below. Let us define the raw residuals (RR) as $e_{ijk} = y_{ijk} - \mu^{in}_{yijk},\ i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}$. Under Option 1, $y_{ijk} = \lfloor \hat{Y}^{dir}_{ijk} \rfloor$ and $e_{ijk} = \lfloor \hat{Y}^{dir}_{ijk} \rfloor - \mu^{in}_{yijk}$. The standardized residuals (SR) are defined by dividing the RRs by its standard deviation.

In what follows, validation results are shown for a better interpretation of the application to real data. To start with, Figure 6.1 plots the SRs of the aZIP13 mixed model versus domain indexes (left) and predicted values of the proportion of single-person households in original (center) and log scale (right). In dotted red, the line $y = 0$ is added.

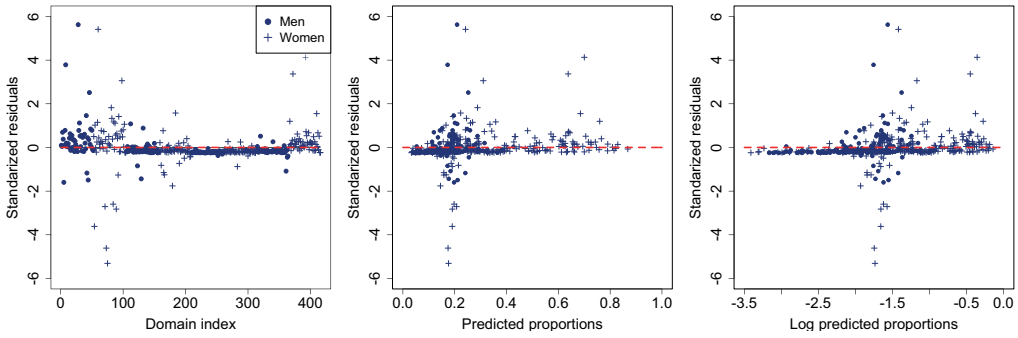


**Figure 6.1.** *SRs versus domain indexes (left) and predicted values of the proportion of single-person households in original (center) and log scale (right).*

As general conclusions drawn from Figure 6.1, it can be seen that SRs have a pattern of symmetry around zero and are mainly found in $[-3,3]$. The central plot has a low percentage of domains with large predicted probabilities, which exceed the threshold of 0.7, and correspond to domains with predominantly single-person households, i.e. inhabited by elderly women. Regarding the right plot, plotting SRs against log predicted

probabilities allows us to detect a conical pattern in the scatterplot. That is, as the log predicted probabilities increases, so does the variability of SRs. This phenomenon is in agreement with the theoretical dispersion of the aZIP13 mixed model.

As expected, SRs are highly variable between provinces, with different sample sizes and socioeconomic conditions. Namely, there are 11 areas with absolute SRs greater than 3, which represents 2.650% of the domains. Directly related to housing prices, the most affected are Madrid and Barcelona. Related to age group and sex, changes are minor. However, *age4* contains the largest amount of outliers.

### *6.1.1. Zero inflation validation*

Area-level models do not aim to chase the scatterplot, but to smooth it and provided more accurate results. It is therefore crucial to understand the importance of zero-inflated probabilities, as they solve the problems of overfitting of the PO mixed model to the Hájek estimates. Indeed, Figure 6.2 shows this improvement in domains with null counts of single-person households (left) and with less than 5 counts (center). All observations are sorted according to the domain index. The line charts plot the Hájek estimates, the IN predictions and those relative to the IN predictor of the PO mixed model with the same set of area-level auxiliary variables as the aZIP13 mixed model, denoted as IN0. The advantage of the IN predictor over the IN0 predictor also applies when comparing with the IN predictor that uses a constant zero-inflated probability $p_{ijk} = p$, denoted as IN1 in Appendix B, although it is not included for ease of exposition.

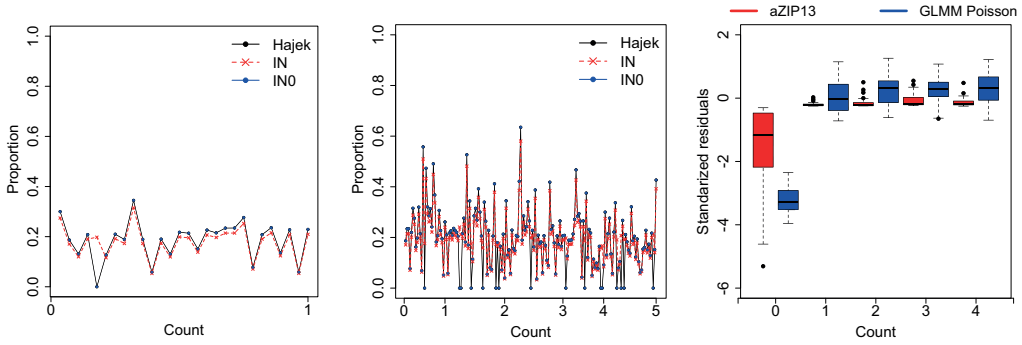


**Figure 6.2.** *Predicted proportions of single-person households in domains with null counts (left) and less than 5 counts (center) and boxplots for the respective SRs (right).*

On balance, the IN predictor of the aZIP13 mixed model is the one that smoothes the results the most. In addition, a challenge encountered in modelling direct estimators is that, in areas with tiny sample sizes, some households in the sample represent too many households in the population. The main concern is to find area-level outliers. Figure 6.2 (right) shows boxplots of the SRs from the aZIP13 mixed model and the PO mixed model in domains with less than 5 counts, grouped according to the observed single-person household counts. When single-person households are not observed, the

PO mixed model is clearly worse and, for low counts, its performance does not improve either: the variability of the boxes is higher. It is concluded that the aZIP13 mixed model performs satisfactorily, both in terms of the significance level of the RPs, the validation via SRs and the fit of zero outcomes. This is a great support for the proposed methodology.

### 6.2. Predictions and error measures

This section provides Hájek estimates and IN predictions of the proportion of single-person households by province, sex and age group. Figure 6.3 shows line charts of these values sorted by domain index (left) and sample size (center), as well as a comparison of both (right). Among the most noteworthy findings, model-based predictors correct the excessively large Hájek estimates, especially for elderly women in Madrid and Barcelona. Even more, it is inferred that the IN predictor smoothes the results of the Hájek estimator, although it still presents problems when dealing with extreme proportions. On the other hand, if single-person households are not observed, the Hájek estimator has no margin of error, although the model never comes to such a low proportion. The same is true for values close to one. This can be seen in Figure 6.3 (left). As it is unlikely, our research is a methodological improvement. In addition, it can be observed that household composition does not affect all domains equally: as the age group increases, the proportion of single-person households also increases.

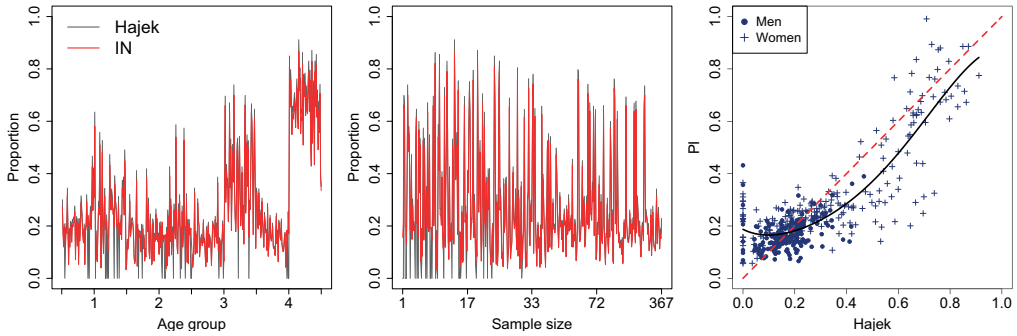


**Figure 6.3.** *IN proportions of single-person households sorted by domain (left) and sample size (center), and Hájek estimates versus IN proportions (right).*

According to Figure 6.3 (center), the IN predictor gets closer to the Hájek estimator as the sample size increases, which is one of the most convincing aspects of the data analysis. Eventually, Figure 6.3 (right) plots the Hájek estimates versus the IN proportions. It can be seen that the dots are evenly distributed around $y = x$. To support this statement, a local polynomial regression of degree 3, with an appropriate bandwidth, is plotted to smoothly represent the relationship between ordinates and abscissas. Consequently, we can underline a crucial advantage of our approach: the theoretical properties of the Hájek estimator, such as asymptotic design-based unbiasedness, are, to some extent, inherited by the IN predictor based on the aZIP13 mixed model.

Table 6.2 (a) reports IN proportions of single-person households by sex and age group. Predominantly, it is the population of *age2* that is least likely to live in single-person households, followed by *age1*. The current trend projects an increase in the proportion of single-person households, with the number of households inhabited by elderly women skyrocketing. This phenomenon is associated with the ageing process, which progressively involves the emancipation of children and widowhood.

**Table 6.2.** *Tabular results for the IN predictor and RRMSEs (%) of the proportion of single-person households by sex and age group of the main breadwinner.*

| | sex | | |
|---|---|---|---|
| age group | *sex1* | *sex2* | Total |
| *age1* | 0.1988 | 0.2389 | 0.2187 |
| *age2* | 0.1596 | 0.1838 | 0.1736 |
| *age3* | 0.1468 | 0.3694 | 0.2612 |
| *age4* | 0.1707 | 0.6479 | 0.4394 |
| Total | 0.1830 | 0.3371 | 0.2621 |

**(a)** *IN proportions aggregated by province.*

| | sex | | |
|---|---|---|---|
| age group | *sex1* | *sex2* | Total |
| *age1* | 20.8360 | 19.8250 | 20.3350 |
| *age2* | 20.3787 | 22.0940 | 21.2451 |
| *age3* | 20.9409 | 12.6800 | 16.6921 |
| *age4* | 20.1576 | 18.1149 | 19.0072 |
| Total | 20.6300 | 19.3125 | 19.9537 |

**(b)** *IN RRMSEs (%) aggregated by province.*

As for the error measures, we calculate the parametric bootstrap estimator of the MSE of $\mu^{in}_{yijk}$, $i \in \mathbb{I}, j \in \mathbb{J}, k \in \mathbb{K}$, following Section 5. $B = 2000$ resamples are used. To avoid scale dependencies, and as usual, the script should be focused on RRMSEs. However, the non-relative version, the root-MSE (RMSE), is preferable because it allows a better understanding of what happens with null counts. Accordingly, Figure 6.4 plots model-based estimates of RMSEs for the IN predictor versus design-based standard deviations (RVAR) for the Hájek estimator (left) and versus model-based estimates of RMSEs for the Hájek estimator (right). See Morales et al. (2021) (Section 2.5) for further details about the RVARs of the Hajék estimator.

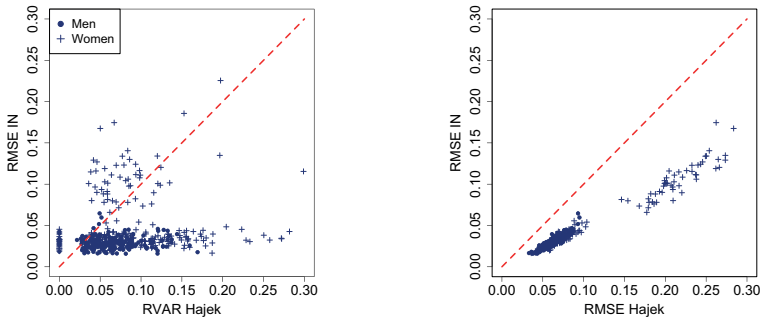


**Figure 6.4.** *Model-based estimates of RMSEs for the IN predictor versus design-based RVARs (left) and model-based estimates of RMSEs (right) for the Hájek estimator.*

Broadly speaking, both plots in Figure 6.4 show that the IN predictor has lower RMSE in all domains and, in most of them, it is also lower than the design-based RVAR of the Hájek estimator. The reduction of the model-based RMSE is therefore prominent when we use the IN predictor instead of the Hájek estimator. Nevertheless, the Hájek estimator has an estimated variance of zero for an observed zero and the RMSE of the IN predictor is always greater than zero. As there are 28 zeros in SHBS2016, this implies 28 aligned points in the lower left corner of Figure 6.4 (left). Clearly, we have already reported that these are false zeros.

In terms of magnitude, the RMSE is higher for elderly women, and it is attributable to the high predicted and/or estimated proportions for these domains. Therefore, it is also useful to provide summary measures of the RRMSE, expressing the error in percentage terms. Table 6.2 (b) contains the bootstrap estimates of the RRMSE (in %) for the IN predictor by sex and age group. As a general conclusion, all values are around 20%, with a slightly lower average for women and especially for *age3*. Hence, the IN predictions of the proposed aZIP13 mixed model have low RRMSEs, as expected in SAE. Appendix D of Supplementary Material maps these relative errors by province, sex and age group.

## 6.3. Mapping proportions of single-person households

The case study concludes by analysing the socioeconomic findings drawn from the area-level predictions. In this sense, the proposed methodology offers the opportunity to analytically read the appreciable differences by Spanish province, sex and age group. Figures 6.5–6.8 map the provincial distribution of single-person households for men (left) and women (right) according to the age group of the main breadwinner.

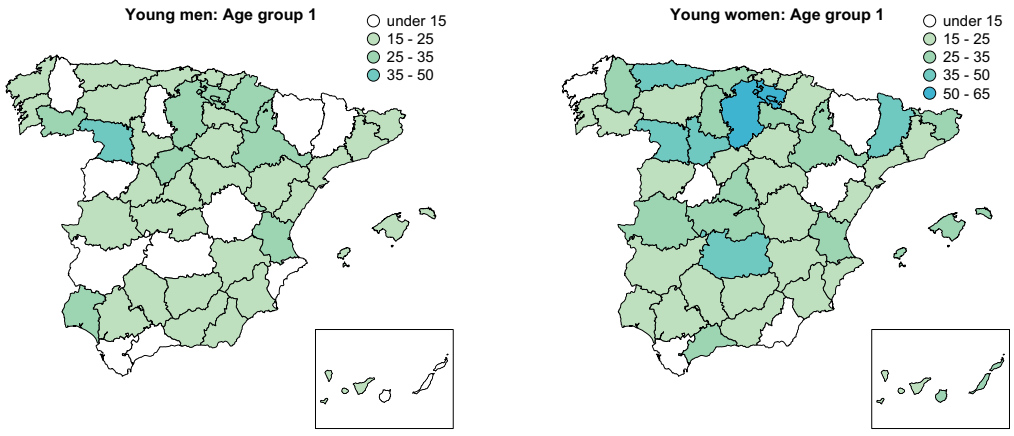


**Figure 6.5.** *Percentages of single-person households for young men (left) and women (right).*

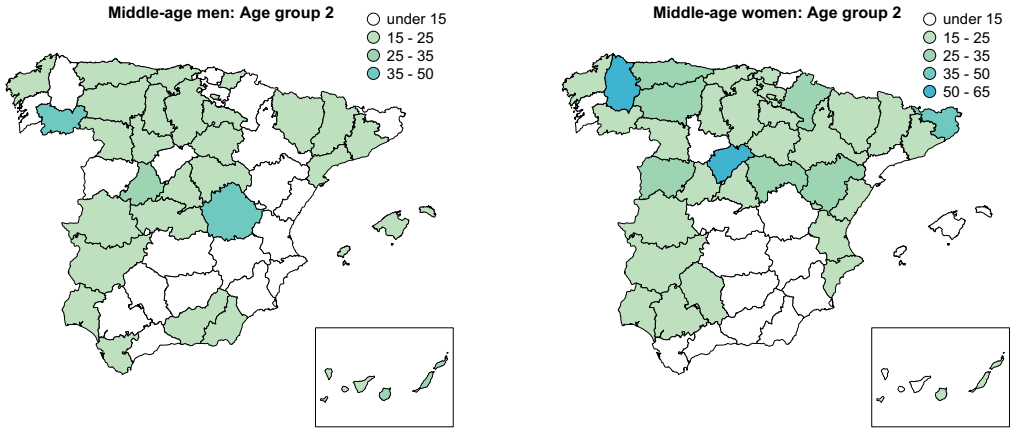


**Figure 6.6.** *Percentages of single-person households for middle-age men (left) and women (right).*

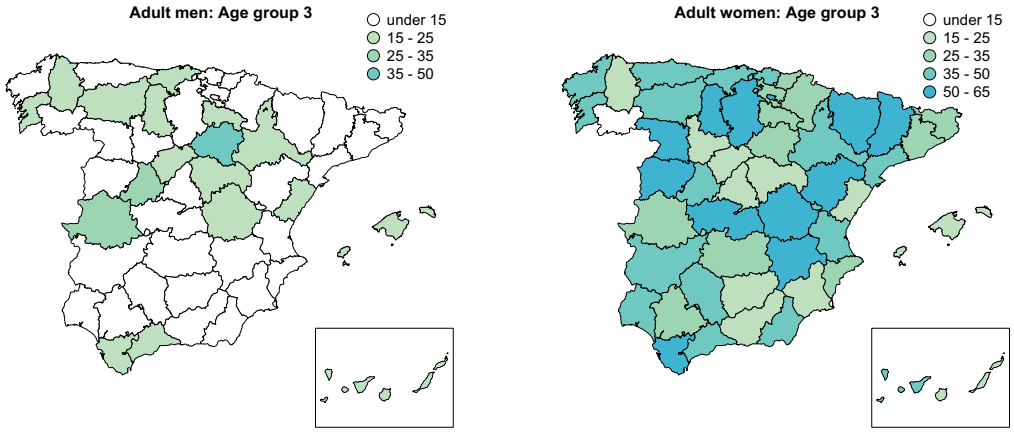


**Figure 6.7.** *Percentages of single-person households for adult men (left) and women (right).*

On the one hand, they show that the highest proportions of single-person households are found in the centre and north-west of Spain, with lower rates in the south and Canary Islands. As expected, the distribution between neighboring provinces, or between those whose demographic and socioeconomic conditions are similar, is generally homogeneous. This fact justifies how model-based predictors lead to smoother results (and closer to reality) than direct estimators. In addition, an interesting spatial pattern emerges, as it can be observed an inverse relationship between house prices and the proportion of single-person households. Thus, lower proportions are estimated for the Catalan Coast, Madrid, Balearic Islands and Málaga. In other words, the Spanish provinces with the highest average prices.

On the other hand, over the course of a person's life, their lifestyle can be expected to change, with the age group directly affecting the composition of households. Most

notably, old age is linked to another factor that alters household composition: mortality. So sex and *age4* are crucial here. Moreover, the increase in quality of life implies not only an increase in life expectancy but also in the autonomy of the elderly, which results in an increase in the number of single-person households inhabited by retired people. Most men live with their partners until their death. In contrast, women have a longer life expectancy (implying a greater accumulation at the top of the demographic pyramid) and the average age of their partners is higher, so they will live alone to a greater extent. Accordingly, Figures 6.5–6.8 map a significant difference between men and women, with clearly higher proportions of dwellings inhabited only by women.

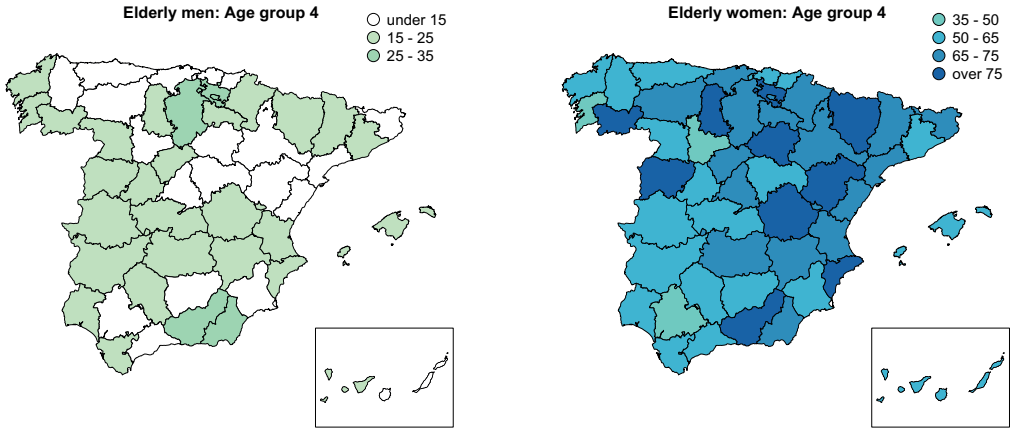


**Figure 6.8.** *Percentages of single-person households for elderly men (left) and women (right).*

## 7. Conclusions

Households are a key unit in a country's socioeconomic decision-making. Therefore, statistical studies of household composition in small and disaggregated areas are of great interest. Against this background, this paper addresses the prediction of total counts and proportions of single-person households by province, sex and age group of the main breadwinner. Given the difficulty of detecting single-person households from survey data, it is also important to model the disaggregated probabilities of false zeros. To do so, it has been taken into account that area-level zero-inflated PO mixed models are quite flexible to predict and explain count variables. In addition, they successfully model zero-inflated outcomes and have been applied in many fields of research. Consequently, the paper deals with an important, common but rather underestimated issue in SAE, which is the problem of zero inflation data.

To fit the model, we have calculated ML estimators of the model parameters and modal predictors of random effects by applying the ML-Laplace approximation. Then, we have considered the EBP, ESP and IN predictors. In theory, the EBP is very attractive because of its properties of approximately null bias and small RMSE. However, its formula contains double products of exponentials and integrals in $\mathbb{R}^{1+IJ}$. The evaluation

of exponential functions usually cause overflow problems when the observed counts are large, which is quite common under Option 1. This produces computational instability problems, especially when applying bootstrap resampling procedures, which made it necessary to omit the EBP from our simulation experiments. Finally, we have investigated the behaviour of the remaining predictors by generating the target variable from the same model as the one selected in the application to real data. Ultimately, we found that the ESP seems very attractive as it has a very low bias, but the IN predictor seems more interesting, as it has a small RMSE and lower computational cost. That is why we have decided to use the IN predictor in the case study. Regarding MSE estimation, we propose a parametric bootstrap procedure and recommend to use $B = 600$ iterations as a good tradeoff between accuracy and computational time.

Simulations also empirically investigated what happens if excess zeros are ignored in the prediction. Namely, if the excess of zeros is large, predictions based on the PO mixed model are rather inefficient. According to our results, the same applies if constant zero-inflated probabilities are considered, so that age-group randomness is required.

Section 6 presents an application to the 2016SHBS and illustrates how to use the proposed methodology. It has been concluded that living alone is a common residential choice across all age groups, influenced by marital separations, emancipation of children, cohabiting relationships and lifestyle in general. Declining fertility and increasing life expectancy are leading to an ageing population. Therefore an overwhelming increase in the proportion of single-person households is expected. What is more, differences in household composition for men and women are more pronounced among the elderly. In addition, RRMSE estimates are below 30% in most domains, which is a fairly good accuracy for a SAE problem.

## Acknowledgements

**Funding Body:** This work has been supported by the Spanish Ministry of Universities, through the project PGC2018-096840-B-I00 and by the Valencian Government, through the project PROMETEO-2021-063. It has also benefited from a study grant from the Manuel Ventura Figueroa Foundation.

## References

Anggreyani, A., Indahwati, I. and Kurnia, A. (2015). Small area estimation for estimating the number of infant mortality using a mixed effects zero inflated Poisson model. *Indonesian Journal of Statistics*, 20, 2, 108-115.

Berg, E.J. (2022). Empirical best prediction of small area means based on a unit-level Gamma-Poisson model. *Journal of Survey Statistics and Methodology*, 11, 4, 873-894.

Berg, E.J. and Fuller, W.A. (2012). Estimators of error covariance matrices for small area prediction. *Computational Statistics and Data Analysis*, 56, 10, 2949-2962.

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

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

Bugallo, M., Esteban, M.D., Marey-Pérez, M.F. and Morales, D. (2023). Wildfire prediction using zero-inflated negative binomial mixed models: Application to Spain. *Journal of Environmental Management*, 328, 116788.

Cai, S. and 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. and Tzavidis, N. (2012). *M-quantile regression for binary data with application to small area estimation.* Centre for Statistical and Survey Methodology, University of Wollongong.

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*, 179, 2, 453-479.

Chandra, H., Bathla, H.V.L. and Sud, U.C. (2010). Small area estimation under a mixture model. *Statistics in Transition*, 11, 3, 503-516.

Chandra, H. and Chambers, R. (2011). Small area estimation for skewed data in presence of zeros, *Calcutta Statistical Association Bulletin*, 63, 249-252.

Chandra, H. and Sud, U.C. (2012). Small area estimation for zero inflated data. *Communications in Statistics-Simulation and Computation*, 41, 632-643.

Chen, S. and Lahiri, P. (2012). Inferences on small area proportions. *Journal of the Indian Society of Agricultural Statistics*, 66, 121-124.

Cho, Y.K., Shim, K.W., Suk, H.W., Lee, H.S., Lee, S.W., Byun, A.R. and Lee, H.N. (2019). Differences between one-person and multi-person households on socioeconomic status, health behaviour, and metabolic syndrome across gender and age groups. *Korean Journal of Family Practice*, 9, 373-82.

Cohen, P.N. (2021). The rise of one-Person households. *Socius*, 7.

Datta, G.S. and Mandal, A. (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110, 512, 1735-1744.

Dreassi, E., Petrucci, A. and Rocco, E. (2014). Small area estimation for semicontinuous skewed spatial data: An application to the grape wine production in Tuscany. *Biometrical Journal*, 56, 1, 141-156.

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

Fritsch, N.S., Riederer, B. and Seewann, L. (2023). Living alone in the city: Differentials in subjective well-being among single households 1995-2018. *Applied Research in Quality of Life*, 18, 2065-2087.

Ghosh, M., Kim, D., Sinha, K., Maiti, T., Katzoff, M. and Parsons, V.L. (2009). Hierarchical and empirical Bayes small domain estimation and proportion of persons without health insurance for minority subpopulations. *Survey Methodology*, 35, 53-66.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. and Santamaría, L. (2007). 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. and Santamaría, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics and Data Analysis*, 52, 5242-5252.

Greitemeyer, T. (2009). Stereotypes of singles: Are singles what we think? *European Journal of Social Psychology*, 39, 3, 368-383.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling".

Hartono, B., Kurnia, A. and Indahwati, I. (2017). Zero inflated binomial models in small area estimation with application to unemployment data in Indonesia. *International Journal of Computer Science and Network*, 6, 6, 746-752.

Hobza, T. and Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics*, 32, 3, 661-69.

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

Krieg, S., Boonstra, H.J. and Smeets, M. (2016). Small area estimation with zero-inflated data: A simulation study. *Journal of Official Statistics*, 32, 4, 963-986.

Lee, S.J. and Lee, S.H. (2021). Comparative analysis of health behaviours, health status, and medical needs among one-person and multi-person household groups: Focused on the ageing population of 60 or more. *Korean Journal of Family Medicine*, 42, 2, 73-83.

Lesthaeghe, R. (2014). The second demographic transition: A concise overview of its development. *Proc. of the National Academy of Sciences*. 111, 51, 18112-18115.

Liu, B. and Lahiri, P. (2017). Adaptive hierarchical Bayes estimation of small area proportions. *Calcutta Statistical Association Bulletin*, 69, 2, 150-164.

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

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

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

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

Michael, F. and Thomas, D. (2016). *Discrete data analysis with R: Visualization and modeling techniques for categorical and count data*. Chapman and Hall.

Militino, A.F., Ugarte, M.D. and Goicoa, T. (2015). Deriving small area estimates from information technology business surveys. *Journal of the Royal Statistical Society, Series A*, 178, 4, 1051-1067.

Molina, I., Saei, A. and Lombardía, M.J. (2007). Small area estimates of labour force participation under multinomial logit mixed model. *The Journal of the Royal Statistical Society, Series A*, 170, 975-1000.

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

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

Morales, D., Krause, J. and Burgard, J.P. (2022). On the use of aggregate survey data for estimating regional major depressive disorder prevalence. *Psychometrika*, 87, 4.

Ortiz-Ospina, E. (2019). The rise of living alone: How one-person households are becoming increasingly common around the world. *Our World in Data*.

Park, B.Y., Kwon, H.J., Ha, M.N. and Burm, E.A. (2016). A comparative study on mental health between elderly living alone and elderly couples: Focus on gender and demographic characteristics. *Journal of Korean Public Health Nursing*, 20, 195-20.

Pfeffermann, D., Terryn, B. and Moura, F.A.S. (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology*, 34, 2, 235-249.

Sadik, K., Anisa, R. and Aqmaliyah, E. (2019). Small area estimation on zero-inflated data using frequentist and Bayesian approach. *Journal of Modern Applied Statistical Methods*, 18, 1, eP2677.

Snell, K.D.M. (2017). The rise of living alone and loneliness in history. *Social History*, 42, 1, 2-28.

Sugasawa, S., Kubokawa, T. and Ogasawara, K. (2017). Empirical uncertain bayes methods in area-level models. *Scandinavian Journal of Statistics*, 44, 3, 684-706.

Torabi, M. and Rao, J.N.K. (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, 127, 36-55.

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

Zhang, L.C. and Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B*, 66, 2, 479-496.

Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A. and Smith, G.M. (2009). *Mixed effects models and extensions in ecology with R*. Chapter 11. Springer.