Mapping the Poverty Proportion in Small Areas under Random Regression Coefficient Poisson Models

Naomi Diz-Rosales, María José Lombardía, and Domingo Morales

Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Instituto Centro de Investigación Operativa IUCIO, Universitas Miguel Hernández, 03202 Elche, Spain
Correspondence: naomi.diz.rosales@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.18*

Abstract: In a complex socio-economic context, policy makers need highly disaggregated poverty indicators. In this work, we develop a methodology in small area estimation to derive predictors of poverty proportions under a random regression coefficient Poisson model, introducing bootstrap estimators of mean squared errors. Maximum likelihood estimators of model parameters and random effects mode predictors are calculated using a Laplace approximation algorithm. Simulation experiments are conducted to investigate the behaviour of the fitting algorithm, the predictors and the mean squared error estimator. The new statistical methodology is applied to data from the Spanish survey of living conditions to map poverty proportions by province and sex, developing a tool to support policy decision making.

1 Introduction

One out of every five people is at risk of poverty or social exclusion in the European Union (EU), with the figure rising to 26.4 % in Spain (Eurostat, 2022).

In order to reduce this number, the EU has set national targets for all its members. Among these measures, the SLC or Survey on living conditions, whose main objective is to provide comparable statistics on income distribution and social exclusion in order to support policy makers in the distribution of their policy packages, stands out. Thus, most European countries use the SLC to estimate poverty indicators. This survey in Spain, the Spanish survey of living conditions (SLCS), provides information on household income received during the year prior to the year of the interview, the domains foreseen being the Autonomous Communities (CCAA).

With these data, different tools can be developed, especially poverty maps, which are widely popularised as they provide a clear visual representation of the geographical distribution of poverty and the degree of inequality between territories in a country, becoming a key support for political decision-making. In fact, the estimation of these indicators and the use of the results for poverty mapping is of great international interest, initiated and sponsored in many cases by the United Nations and the World Bank, demanding their collection at increasingly lower levels of aggregation (see Molina and Rao (2010) for an exhaustive review).

This is a challenge in view of the small or even zero sample size, which can be assumed by Small Area Estimation (SAE) and which responds to the need to produce precise estimates on indicators of interest in areas or domains where the sample size is smaller than planned by the surveys from which the information on the target variable is extracted. Since its definition, with

the study by Fay III and Herriot (1979), research has followed one after the other, increasing significantly in recent years, driven by the Sustainable Development Goals (SDGs).

Thus, SAE methods are shown as potential alternatives, with the use of linear mixed models (LMMs) and generalised linear mixed models (GLMMs) standing out in recent years. However, all existing work so far in SAE considers the regression coefficients as fixed effects. This approach may be too rigid when the relationship between the target variable and the auxiliary variables is not constant across domains. This is common in the socio-economic context, covariates such as age, employment status, citizenship or education may have a different rate of influence on the poverty rate, depending on the region of study. In fact, we have as an example the Spanish case, where in terms of disparity between provinces, the different realities in the peninsular geography have been the subject of numerous investigations insinuating that the current structural inequalities were already established in the 1930s, with a pattern of increasing poverty from the north-east to the south-west of Spain (Tirado et al., 2016).

To provide more flexibility to the modelling process, Dempster et al. (1981), defined random slope mixed models for the first time in SAE. Despite an early initial development phase, with works such as those of Prasad and Rao (1990) to derive the empirical best linear unbiased predictor (EBLUP) and the analytical estimator of the mean squared error (MSE), research is limited, with the works of Hobza and Morales (2013) and Morales et al. (2021) standing out, and always in the context of LMMs. In fact, in SAE, to our current knowledge, GLMMs with random slope have not been defined.

Therefore, and due to the potential gain in flexibility of these approaches, in this work we present our current research, (Diz-Rosales et al., 2023, accepted for publication), aimed at developing new statistical methodology in SAE for poverty mapping by exploring the usefulness of area-level random slope Poisson GLMMs in the inference of poverty indicators. We introduce bootstrap estimators of the mean squared error. The maximum likelihood estimators of the parameters and the modal predictors of the random effects are calculated using a Laplace approximation algorithm. The behaviour of this fitting algorithm, as well as that of the predictors and the mean squared error estimators, is investigated by simulation experiments. Finally, the new statistical methodology is applied to the Spanish living conditions survey (SLCS) data. The objective is to estimate and map, by provinces, the proportions of women and men below the poverty threshold, thus developing a tool of social interest aimed at supporting social policy-making.

2 Data

The dataset selected for this study corresponds to the 2008 SLCS dataset, with a sample size of 35967. It should be noted that, despite its temporal distance with respect to the current year, the choice of this dataset lies in the fact that it has been widely used in other methodological approaches in SAE. In this way, it is possible to make comparisons of the results obtained using different procedures and, therefore, to make the necessary evaluations to improve the model for use with current data.

For the elaboration of the database, we constructed an aggregate data file with the 104 domains defined above. For each domain, the target variable of the Poisson model is the count of persons with an annual equivalised net incomes below a predetermined threshold established at 60 % of the median income per consumption unit, indicated in euros. The auxiliary variables are taken from the Spanish labour force survey (SLFS) of 2008, which provides information on the labour market participation of the population by relating it to characteristics of living conditions. Specifically, we consider the following categories where each category is the proportion of people who meet the defined condition:

- *Age*, with five categories: $\leqslant$15 years old (age0), 16 - 24 years old (age1), 25 - 49 years old (age2), 50 - 64 years old (age3), $\geqslant$ 65 years old (age4), with age0 as the reference category.

- *Education*, with four categories: $\leqslant$16 years old (edu0); illiterate persons with incomplete or complete primary education and/or lower secondary education (edu1); persons with complete secondary education and/or post-secondary education such as baccalaureate or vocational training (edu2); persons with university studies (edu3), with edu0 as the reference category.

- *Nationality*, with two categories: Spanish citizens including those with dual nationality (cit0); foreign citizenship (cit1), with cit0 as the reference category.

- *Labour status*, with four categories: $\leqslant$16 years (lab0), employed (lab1), unemployed (lab2), inactive (lab3), with lab0 as the reference category

For each factor at unit level, the sum of the proportions of its categories is one. Therefore, it is necessary to select a reference category and remove it from the dataset.

Finally, it is worth mentioning that, due to the socio-economic divergence across provinces (Tirado et al., 2016), we have created the *income group* variable, which classifies provinces into five categories ($K = 5$), $k = 1, \ldots, K$, based on an ascending order of the aggregate sum of the average income per household unit for men and women within each province. Thus, after an exploratory analysis of the data, we find that the relationship between the target variable and the covariates is different depending on the category k, and in the definition of the model, we introduce the random slopes by income groups of the provinces.

3 The area level random regression coefficient Poisson model

This section defines the basic principles of the methodology developed to define the area level random regression coefficient Poisson, which we refer to as the ARRCP model, and the derived predictors. A more in-depth development can be found in Diz-Rosales et al. (2023, accepted for publication).

Let us consider a count variable y_{ij} taking values on $\mathbb{N} \cup \{0\}$, where $i \in \mathbb{I} = \{1, \ldots, I\}$ and $j \in \mathbb{J} = \{1, \ldots, J\}$. Let $D = IJ$ be the total number of y-values. For example, y_{ij} could be the number of people living below the poverty line in a survey sample, the indexes i and j may represent province and sex, and D is the total number of domains defined by the crossings of the variables province and sex. In other words, we have a country partitioned into provinces and sexes. We further assume that each province can be grouped into one, and only one, of the K clusters, $\mathbb{I}_1, \ldots, \mathbb{I}_K$, of an income variable. Let $k(i)$ be the number of the category to which province i belongs, so that $k(i) \in \mathbb{K} = \{1, \ldots, K\}$. The number of provinces in the category $\mathbb{I}_K$ is $m_k = \#(\mathbb{I}_k)$, so that $D = J \sum_{k=1}^{K} m_k$.

We are dealing with area-level data for modelling and predicting the target variable y_{ij}. Let us assume that we have p explanatory variables with values $x_{\ell,ij}$, $\ell \in \mathbb{P} = \{1, \ldots, p\}$, $i \in \mathbb{I}$, $j \in \mathbb{J}$. For models with intercept, we take $x_{0,ij} = 1$ for all i and j. In what follows, we present the ARRCP model.

Let u_{ij}, $i \in \mathbb{I}$, $j \in \mathbb{J}$ be i.i.d. $N(0, 1)$ random variables. Let $\phi_\ell > 0$, $\ell \in \mathbb{P}$, be unknown standard deviation parameters. Let $\rho_{rs} \in (-1, 1)$, $r < s, r, s \in \mathbb{P}$, be unknown correlation parameters. Let $\boldsymbol{v}_k = (v_{1,k}, \ldots, v_{p,k})'$, $k \in \mathbb{K}$, be i.i.d. random vectors such that

$$\underset{1\leqslant\ell\leqslant p}{\text{diag}}\,(\phi_\ell)\boldsymbol{v}_k \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{vk}^{\phi\rho}), \quad \boldsymbol{V}_{vk}^{\phi\rho} = \begin{pmatrix} \phi_1^2 & \phi_1\phi_2\rho_{12} & \cdots & \phi_1\phi_p\rho_{1p} \\ \phi_2\phi_1\rho_{12} & \phi_2^2 & \cdots & \phi_2\phi_p\rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_p\phi_1\rho_{1p} & \phi_p\phi_2\rho_{2p} & \cdots & \phi_p^2 \end{pmatrix}.$$

Therefore, the variance of $\boldsymbol{v}_k$ is

$$\boldsymbol{V}_{vk} = \text{var}(\boldsymbol{v}_k) = \underset{1\leqslant \ell\leqslant p}{\text{diag}}\,(\phi_\ell^{-1})\boldsymbol{V}_{vk}^{\phi\rho}\underset{1\leqslant \ell\leqslant p}{\text{diag}}\,(\phi_\ell^{-1}) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{pmatrix}.$$

Let us define the vectors

$$\boldsymbol{u} = \underset{1\leqslant i\leqslant I}{\text{col}}(\underset{1\leqslant j\leqslant J}{\text{col}}(u_{ij})) \sim N_{IJ}(\boldsymbol{0}, \boldsymbol{I}_{IJ}), \quad \boldsymbol{v} = \underset{1\leqslant k\leqslant K}{\text{col}}(\boldsymbol{v}_k) \sim N_{pK}(\boldsymbol{0}, \boldsymbol{V}_v),$$

where $\boldsymbol{V}_v = \underset{1\leqslant k\leqslant K}{\text{diag}}\,(\boldsymbol{V}_{vk})$ and where diag and col are the diagonal and the column operator, respectively. We assume that $\boldsymbol{u}$ and $\boldsymbol{v}$ are independent. The distribution of the target variable y_{ij}, conditioned to the random effects u_{ij}, $v_{\ell,k(i)}$, $\ell \in \mathbb{P}$, is

$$y_{ij}|_{u_{ij},v_{1,k(i)},\ldots,v_{p,k(i)}} \sim \text{Poisson}(\nu_{ij}p_{ij}), \quad i \in \mathbb{I},\ j \in \mathbb{J},$$

where the offset (or size) parameters $\nu_{ij} > 0$ are known and correspond to the sample size when the model is applied to real data, and the binomial probability, p_{ij}, is the target parameter with range (0,1). For the natural parameters, we assume

$$\eta_{ij} = \log\mu_{ij} = \log\nu_{ij} + \log p_{ij} = \log\nu_{ij} + \sum_{\ell=1}^{p}\beta_\ell x_{\ell,ij} + \sigma u_{ij} + \sum_{\ell=1}^{p}\phi_\ell v_{\ell,k(i)}x_{\ell,ij}, \quad i \in \mathbb{I},\ j \in \mathbb{J}, \tag{18.1}$$

where $\mu_{ij} = E[y_{ij}|u_{ij}, v_{1,k(i)}, \ldots, v_{p,k(i)}]$. We may write $\boldsymbol{x}_{ij}\boldsymbol{\beta} = \sum_{\ell=1}^{p}\beta_\ell x_{\ell,ij}$, where $\boldsymbol{\beta} = \underset{1\leqslant \ell\leqslant p}{\text{col}}\,(\beta_\ell)$ is the column vector of regression parameters and $\boldsymbol{x}_{ij} = \underset{1\leqslant \ell\leqslant p}{\text{col}'}\,(x_{\ell,ij})$ is the row vector of known auxiliary variables. To finish the definition of the ARRCP model, we assume that the y_{ij}'s are independent conditioned to $\boldsymbol{u}$ and $\boldsymbol{v}$. The variance component parameters are $\sigma > 0$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)' \in \mathbb{R}_+^p$ and $\boldsymbol{\rho} = (\rho_{12}, \ldots, \rho_{1p}, \ldots, \rho_{p-1p})' \in (-1,1)^{p(p-1)/2}$, where $\mathbb{R}_+ = (0,\infty)$. The vector of model parameters is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma, \boldsymbol{\phi}', \boldsymbol{\rho}')'$. The total number of random effects is $H = IJ + pK$.

With the ARRCP model defined, we proceed to carry out the maximization, deriving the maximum likelihood estimators of the model parameters, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$, $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\rho}}$, and the mode predictors of the random effects, by means of a Laplace approximation algorithm (ML Laplace algorithm), using the R package lme4 1.1-33.

In addition, we define the best predictor (BP), the simplified best predictor (SP), the empirical best predictor (EBP), the empirical simplified best predictor (ESP) and the plug-in predictor (IN), to predict the porverty proportion by province and sex. The best predictor (BP) of p_{ij} is

$$\hat{p}_{ij}^{bp} = \hat{p}_{ij}^{bp}(\boldsymbol{\theta}) = E_\theta[p_{ij}|\boldsymbol{y}] = E_\theta[p_{ij}|\boldsymbol{y}_{jk(i)}]$$

The simplified best predictor (SP) of p_{ij} is

$$\hat{p}_{ij}^{sp} = \hat{p}_{ij}^{sp}(\boldsymbol{\theta}) = E_\theta[p_{ij}|\boldsymbol{y}] = E_\theta[p_{ij}|\boldsymbol{y}_{ij}]$$

The empirical best predictor (EBP) of p_{ij} is $\hat{p}_{ij}^{ebp} = \hat{p}_{ij}^{bp}(\hat{\boldsymbol{\theta}})$, and the empirical simplified best predictor (ESP) of p_{ij} is $\hat{p}_{ij}^{esp} = \hat{p}_{ij}^{sp}(\hat{\boldsymbol{\theta}})$. These predictor requires approximating a multivariate integral which we approximate by a Monte Carlo method.

The behaviour of these predictors is then evaluated by bootstrap simulation studies of at least 1000 iterations, comparing them with other predictors, such as the plug-in predictor (IN) of p_{ij} and μ_{ij} under the ARRCP model which has the form

$$\hat{p}_{ij}^{in} = \exp\{\boldsymbol{x}_{itj}\hat{\boldsymbol{\beta}} + \hat{\sigma}\hat{u}_{ij} + \sum_{\ell=1}^{p} \hat{\phi}_\ell \hat{v}_{\ell,k(i)} x_{\ell,ij}\},$$

where $\hat{u}_{ij}$ and $\hat{v}_{\ell,k(i)}$, $\ell = 1,\ldots,p$, are the mode predictors taken from the output of the ML Laplace algorithm, being $\hat{\mu}_{ij}^{in} = \nu_{ij}\hat{p}_{ij}^{in}$ the IN predictor of $\mu_{ij} = \nu_{ij}p_{ij}$.

As a result of the simulations, the IN predictor shows the best computational efficiency and accuracy trade-off. To understand these results, it should be noted that the estimation of the BP and EBP is a very complex multivariate integral approximation problem. In order to perform this Monte Carlo approximation, we have generated simulations with different configurations in the number of replications, until we have tested, for the moment, the approximation with the generation of 2500 independent random variables. However, as we can see from the results, this number is insufficient. The BP and EBP incorporate the Monte Carlo variance, which is high, and which has a variance underlying the estimation of the multivariate integral, which would require substantially more than 2500 for an optimal approximation. By contrast, these simulations are highly computationally expensive, increasing simulation times to the order of days.

On the other hand, SP and ESP also experience this problem, although to a lesser extent, since the integral to be approximated is considerably less complex than BP and EBP. The results in terms of bias are worse than those of the BP, EBP and IN predictors, although the estimation of the root mean squared error (RMSE) is notably better, being closer to the IN than the rest. This, together with an efficient computational time, makes it a candidate for use when few computational resources are available to simulate the EBP and/or the IN predictor cannot be used. Although, at the expense of obtaining more efficient results in terms of computational performance, and having verified the high performance of the IN predictor, we chose it as the starting predictor for this study.

Consequently, a second simulation study is performed to test the performance of the MSE estimator of this predictor based on the parametric bootstrap with and without bias correction. While the performance of the MSE estimators of both approaches was optimal, due to the substantial improvement in bias and the low computational cost, with virtually unbiased estimation, in the application to real data we use the parametric bootstrap estimation with bias correction.

4 Application to real data

During the model selection phase, we consider several criteria: a) significance of model parameters and socio-economic interpretability; b) convergence of the ML-Laplace approximation algorithm; c) validity of model assumptions; and d) lower conditional AIC.

As a result of the selection process, we define the ARRCP model, introduced in 18.1 in Section 3, with the following variables: y_{ij} is the sample count of people below the poverty threshold in province i and sex j, $\nu_{ij} = n_{ij}$ is the sample size, $x_{0,ij}$ is the intercept and $x_{1,ij}$, $x_{2,ij}$, $x_{3,ij}$ and $x_{4,ij}$ are the values of the auxiliary variables age3, edu1, cit1 and lab2 respectively. The selected model contains two random slopes for $x_{1,ij}$ and $x_{4,ij}$, so that the corresponding model parameters and random effects are $\phi_1, \phi_4, \rho_{14}, u_{ij} \sim N(0,1)$, $(v_{1,k}, v_{4,k})' \sim N_2(\mathbf{0}, \boldsymbol{V}_{14})$, $\mathbf{0} = (0,0)'$ and $\boldsymbol{V}_{14}(\rho_{14}) = \begin{pmatrix} 1 & \rho_{14} \\ \rho_{14} & 1 \end{pmatrix}$. The natural parameter is

$$\eta_{ij} = \log\mu_{ij} = \log n_{ij} + \sum_{\ell=1}^{5} \beta_\ell x_{\ell,ij} + \sigma u_{ij} + \phi_1 v_{1,k(i)} x_{1,ij} + \phi_4 v_{4,k(i)} x_{4,ij}.$$

The estimated model parameters are socioeconomically interpretable, with the auxiliary variables edu1 and lab2 having a protective effect on poverty, and age3 and cit1 helping to reduce it. In addition, we obtained the basic percentile bootstrap confidence intervals observing that, at 95 %, all are significant.

Once the model is selected, it undergoes the diagnostic phase starting with the evaluation of the Pearson residuals.

Having obtained good results, we proceed to further assess the performance in estimating the poverty ratio by province and sex. For this purpose, in Figure 1 (left) we plot the IN predictions and the classic Hájek estimates of the poverty proportions. This figure compares both types of estimators and analyses the effect of using this type of estimator in unplanned domains.

We can see that the direct estimator and the IN predictor diverge noticeably in domains with lower sample sizes, becoming closer as sample size increases. This trend is consistent in the RMSE estimate plotted on the right, where the error magnitude of the direct and IN estimator tend to decrease and equalise as the sample size increases. However, it is notable the smooth and decreasing behaviour of the RMSE of the IN predictor while in the case of the direct estimator the RMSE estimates are characterised by an abrupt trend with characteristic peaks.

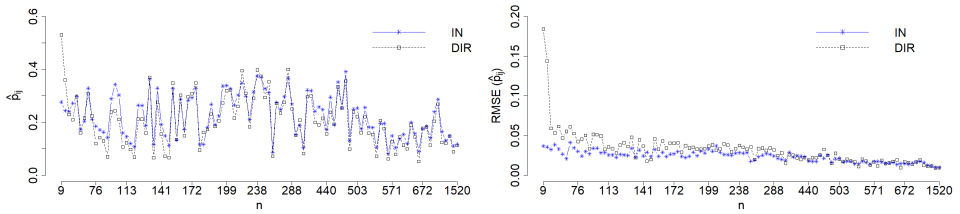

Figure 1: Poverty proportions estimates $\hat{p}_{ii}$ and associated $RMSE_{ii}$, ordered by sample size

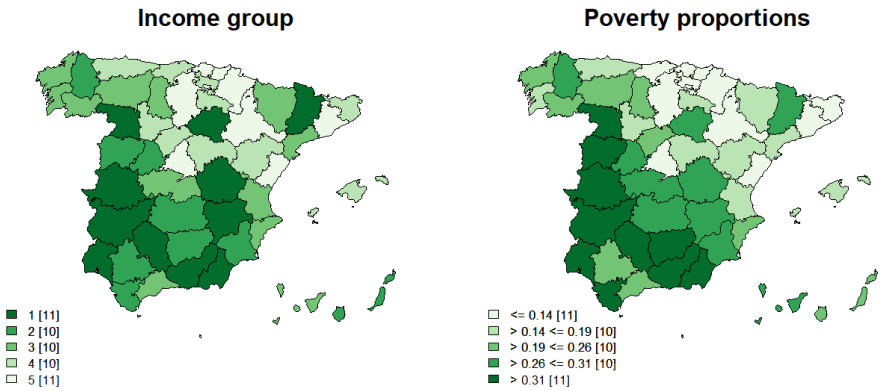

Figure 2: Estimated poverty proportions in Spanish provinces by income group

In order to improve visual comprehension, we represent the estimates of poverty proportions on maps widely used by socio-political powers. In particular, we illustrate in Figure 2 an example of poverty maps at the provincial level and compare it with a poverty map produced taking into account the existing K categories in the data. As can be seen, the intensity of colours is generally the same, while in those that diverge, they do so between contiguous groups, without a substantial difference, as in the provinces of Cádiz or Salamanca.

These maps allow us to evaluate at a glance the socio-economic state of the country, with a substantial percentage of areas with high levels of poverty, not being equally distributed, with a clear north-south, east-west pattern of increasing poverty rates. In addition, we have carried out a study of differences in poverty rates by sex using bootstrap basic percentile confidence interval, and although, at the 95 % and 90 % percentile, significant differences were only detected in 6 provinces, in line with other studies in the field, in all cases the poverty rate was higher among women.

We would like to complement the study with more indicators to overcome limitations such as the fact that the cost of living is not the same in all regions of the country, but the poverty line is the same. This does not detract from the conclusions, but it could help to have a more global vision in decision making.

Acknowledgments

This research is part of the grant PID2020-113578RB-I00, funded by MCIN/AEI/10.13039/501100011033/. It has also been supported by the Spanish grant PID2022-136878NB-I00, the Valencian grant Prometeo/2021/063, by the Xunta de Galicia (Competitive Reference ED431C-2020/14) and by CITIC that is supported by Xunta de Galicia, collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Sistema Universitario de Galicia (CIGUS). The first author was also sponsored by the Spanish Grant for Predoctoral Research Trainees RD 103/2019 being this work part of grant PRE2021-100857, funded by MCIN/AEI/10.13039/501100011033/ and ESF+.

Bibliography

A. P. Dempster, D. B. Rubin, and R. K. Tsutakawa. Estimation in covariance components models. *Journal of the American Statistical Association*, 76(374):341–353, 1981.

N. Diz-Rosales, M. J. Lombardía, and D. Morales. Poverty mapping under area-level random regression coefficient poisson models. *Journal of Survey Statistics and Methodology*, 2023, accepted for publication.

Eurostat. Persons at risk of poverty or social exclusion by age and sex - eu 2020 strategy. *https://ec.europa.eu/eurostat/databrowser/view/ilc_peps01/default/table?lang=en*, 2022. [Online; accessed 4-June-2022].

R. E. Fay III and R. A. Herriot. Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

T. Hobza and D. Morales. Small area estimation under random regression coefficient models. *Journal of Statistical Computation and Simulation*, 83(11):2160–2177, 2013.

I. Molina and J. Rao. Small area estimation of poverty indicators. *Canadian Journal of statistics*, 38(3):369–385, 2010.

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

N. N. Prasad and J. N. Rao. The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, 85(409):163–171, 1990.

D. A. Tirado, A. Díez-Minguela, and J. Martinez-Galarraga. Regional inequality and economic development in spain, 1860–2010. *Journal of Historical Geography*, 54:87–98, 2016.