Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 16 No. 1, 2018 (New Series), pp 269-288

Selection of Small Area Estimators

María José Lombardía[1] **, Esther López-Vizcaíno**[2] **and Cristina Rueda**[3]

[1] *Universidade da Coruña, Spain,*
[2] *Instituto Galego de Estatística, Spain,*
[3] *Universidad de Valladolid, Spain*

Received: January 23, 2018; Reviewed: Feb 21, 2018; Accepted: March 11, 2018

Abstract

In this paper we consider small area estimators using a selection approach based on a Mixed Generalized Akaike Information Criterion statistic. The estimators are model-based that use auxiliary information represented by a set of explicative variables. The candidate models are defined using different types of relationship between the response and the explicatives, going from a simple univariate linear to a multivariate and non-parametric one.

Numerical results show that the procedure selects the estimators with the smallest mean squared error from a set of candidates. The good performance of the selection procedure is also compared with alternative selection approaches. In addition, the important practical advantages in a real application are shown.

*K*ey words: Akaike Information Criterion, Bootstrap, Fay-Herriot model, Generalized Degree of Freedom, monotone model, spline regression, small area estimation.

1 Introduction

In recent years, the demand for small area estimation (SAE) has increased considerably due to its use in the formulation of policies and programs, in the distribution of government funds and in regional plans. Demand from the private sector has also increased significantly related with business decisions. As a result, we found related to the subject: working groups, conferences (the International SAE Conference, starting in 2005, that continues to be celebrated, among others) and research projects (such European projects as EURAREA, SAMPLE and AMELI, among others), as well as important papers published in international prestigious statistical and sampling journals. One of the figures that has contributed to this growth is J.N.K. Rao, whose contributions have been key in the methodological development in this field. It has also been a source of inspiration for all our work in the field of small areas and in particular in this paper, in which we refer to some of his

Supported by the MINECO grants (MTM2017-82724-R, MTM2015-71217-R, MTM2014-52876-R and MTM2013-41383-P, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.
Corresponding author: María José Lombardía
Email id: maria.jose.lombardia@udc.es

papers and books.

In what follows, we present an estimation selection approach based on a Mixed Generalized Akaike Information Criterion statistic, accompanied by an important application with data from the survey on income and living conditions in Galicia (north-west Spain) in 2015. A fundamental aspect to determine the degree of social cohesion in a population is to know the distribution of income. Being aware of this, at the European Summit in Laeken (Brussels) in 2001, the heads of the European Union Member States set a series of indicators in order to quantify the progress in the objectives of the fight against poverty and social exclusion. These indicators have since been published by the Statistical Office of the European Union (Eurostat).

The region of Galicia has been characterized by a marked duality that is materialized in the existence of areas that have managed to successfully incorporate industrialization processes, access to international trade or the development of a service sector of importance, as opposed to others that have remained with more traditional productive models linked to primary, family farms. The most visible consequence of these differences was the massive emigration in these last zones, with the consequent depopulation.

Galicia is divided into four provinces: Coruña, Lugo, Ourense and Pontevedra. Each province is hierarchically partitioned into counties and municipalities. Until now, the income differences between households in Galicia only could be analyzed at the province or area level. In any case, the provincial division hides important characteristics to which attention should be paid. For example, in the province of Pontevedra, coastal areas are very different from the eastern areas, which are more similar to their neighboring areas in the province of Ourense. Therefore, it is fundamental to disaggregate the provincial analysis. There is an increasing interest in obtaining income data at the smallest possible geographical level from local authorities, academics, commercial organizations and independent researchers. These data are essential for the identification of deprived and disadvantaged communities, provision of information to practitioners and for the profiling of geographical areas.

The objective of this paper is to estimate the distribution of the mean income in the galician counties. This will allow us to quantify social cohesion in order to assess the differences in the citizens' standard of living. The survey on income and living conditions (ILC) provides information about the distribution of income and social exclusion. The sampling design of the ILC is stratified, with strata defined by the size of the municipalities. Most of the municipalities are not represented in the sample, so the direct estimates at the municipal or county level have a low accuracy. In this context, estimating the mean income in the counties is an SAE problem, whose techniques give more accurate estimators for individual areas, by borrowing strengh from other areas. Some of the most relevant references in SAE are the monographs of Rao (2003) and Rao and Molina (2015), and the reviews of Ghosh and Rao (1994), Rao (1999), Pfeffermann (2002, 2013) and Jiang and Lahiri (2006). Some of the most important papers about income in SAE are Fay and Herriot (1979), who obtain small area estimators of median income in small places of the U.S.A. Ghosh et al. (1996) propose an extension of the basic Fay-Herriot model to handle time series and cross sectional data to estimate the median income for four-person families for the fifty American states and the district of Columbia. With these latter data Datta et al. (2002) propose a hierarchical

time series model to estimate the median income. Molina and Rao (2010) propose an empirical Bayes method, based on a nested error model to estimate poverty incidences and poverty gaps in Spanish provinces by gender.

Small area estimators, based on area level models, are likely to achieve high precision when the model is correctly specified. In this point, the question of model selection plays an important role. We talk about estimation selection instead of model selection, in a small area context, as the model is only an artifact to derive the area estimators. Even if a model is far from being true, the area estimators derived from that model could be more accurate to estimate the parameter of interest rather than using the direct estimator. In fact, the working model (the model used to analyze the data and derive the estimators) is often simpler than the data-generating model (the unknown model that generates the data). Moreover, working models with different types of relationship between the response and the explicatives are considered in the selection process: nonparametric models based on P-splines, models with monotone assumptions and linear models.

The general approach to the problem is as follows, a response vector $(Y_1, \ldots, Y_D)$ is considered, where D is the number of domains of study and $Y_d \sim N(m_d, V_d)$. A working model M is defined as a mapping from $\mathbf{R}^D$ to $\mathbf{R}^D$, which produces a vector of fitted values $(\hat{m}_1, \ldots, \hat{m}_D)$. The first goal is to select estimators with the smallest Mean Squared Error (MSE), defined as $MSE = \sum_{d=1}^{D} E(\hat{m}_d - m_d)^2$.

In SAE, the main interests are the totals, means and proportions in each area. The formulations of these in terms of parameters depends on the working model and usually, when that model includes random effects, the researchers consider the conditional means under the random terms as the parameters of interest. A good selection of the working model from a set of candidate models is very important because, if the model chosen to fit the data is a bad one, far from the generating model, the corresponding estimators can lead to erroneous inferences.

In this paper, we propose to perfom the selection step using an AIC statistic. In general terms, the value of AIC (Akaike (1973)) for a model M is defined as $AIC(M) = -2log(l(M)) + 2P$, where $l(M)$ is the model likelihood and P is a penalty term. The model M with the lowest AIC is selected. When mixed models are among the candidate models, different versions for the penalty term, and either conditional or marginal log-likelihoods, have been considered in the literature. Some references of interest are Vaida and Blanchard (2005), Muller et al. (2013), You et al. (2016), and Lombardía et al. (2017) among others. From these papers we highlight three GAIC statistics useful in small area problems: cGAIC, yGAIC and xGAIC, which will be defined in Section 2.

We show below that the xGAIC, proposed in Lombardía et al. (2017), almost always selects the estimator with the smallest MSE. The candidate models are derived using different explicatives and functional forms relating the explicatives with the response. Moreover, this paper can be viewed as an extension of Lombardía et al. (2017).

We organize the remainder of the paper as follows. The background defining parametric and nonparametric models to fit the data, and the approach to estimate the MSEs and the GAIC statistics, are given in Section 2. In Section 3, numerical studies, conducted to compare the MSE of

the proposed strategy and others are considered. Finally, in Section 4, the real data are analyzed. Section 5 contains the conclusions and discussion.

2 Background

A complete revision and discussion of the estimators in SAE can be seen in the monographs of Rao (2003) and Rao and Molina (2015).

2.1 Area model-based estimators

Let Y be the response variable, in our case, the direct estimator of area mean log-income, and $d = 1, ..., D$ the domains of interest. In this section, we introduce the candidates for working models:

$$Y_d = \mu_d + e_d = \theta_d + u_d + e_d, \qquad d = 1, ..., D;$$

where $\mu_d = \theta_d + u_d$ and $\theta_d = f(x_{1d}, \ldots, x_{pd})$, with p auxiliary variables; u_d independent and identically distributed as $N(0, \sigma_u^2)$, with σ_u^2 unknown and $e_d \sim N(0, \sigma_d^2)$ with σ_d^2 assumed known.

Model-based estimators are defined as $\hat{m}_d = \hat{\mu}_d = \hat{\theta}_d + \hat{u}_d$, where $\hat{\theta}_d$ is the estimator of the fixed part of the model and $\hat{u}_d$ the predictor of the random effect u_d. Depending on the functional form of θ, several small area models can be defined. We focus on three types of models: linear, monotone and spline.

First, we consider θ_d as a linear function of the auxiliary variables, $\theta_d = \mathbf{x}_d\boldsymbol{\beta}$, with $\mathbf{x}_d = (x_{1d}, \ldots, x_{pd})$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. The resulting working model is an area level model well-known in SAE literature (Fay and Herriot (1979)):

$$\text{(LM)}: \quad Y_d = \mathbf{x}_d\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \ldots, D.$$

To fit the model, Maximun Likelihood Estimation and the `sae` package in R language (Molina and Marhuenda (2015)) are considered obtaining:

$$\hat{\mu}_d = \mathbf{x}_d\hat{\boldsymbol{\beta}} + \hat{u}_d, \quad d = 1, \ldots, D.$$

Here $\hat{\boldsymbol{\beta}}$ is the empirical best linear unbiased estimator of $\boldsymbol{\beta}$ and $\hat{u}_d$ is the empirical best linear unbiased predictor of u_d. More details about the estimation of the model parameters and variance components can be seen in the monograph of Rao (2003), Rao and Molina (2015) and the paper Lombardía et al. (2017), among others.

Secondly, we consider a semiparametric monotone model for θ_d:

$$\theta_d = \sum_{j=1}^{p_1} \beta_j x_{jd} + \sum_{j=p_1+1}^{p} h_j(x_{jd}),$$

and the working model is:

$$\text{(MM)}: \quad Y_d = \sum_{j=1}^{p_1} \beta_j x_{jd} + \sum_{j=p_1+1}^{p} h_j(x_{jd}) + u_d + e_d, \quad d = 1, \ldots, D;$$

with $h_j()$ monotone functions. To obtain the Maximun Likelihood estimators for the area parameters and the estimator for the variance of the random effects, we use the methodology proposed in Rueda and Lombardía (2012) and construct the model-based estimator as:

$$\hat{\mu}_d = \left(1 - \frac{\sigma_u^2}{\sigma_d^2 + \sigma_u^2}\right) \widehat{\theta}_d + \frac{\sigma_u^2}{\sigma_d^2 + \sigma_u^2} Y_d, \quad d = 1, \ldots, D;$$

where $\widehat{\boldsymbol{\theta}}$ is the projection of $\mathbf{Y}$ onto $\mathbf{K}$, $P(\mathbf{Y}|\mathbf{K})$, and $\mathbf{K} = \mathbf{L}_0 + \mathbf{S}_1 + \ldots + \mathbf{S}_{p_2}$ is a convex region in R^D defined by the restrictions imposed. $\mathbf{L}_0$ is the linear subspace of dimension p_1 spanned by columns in matrix $(\mathbf{x}_1, \ldots, \mathbf{x}_{p_1})$ and, for $j > p_1$, each $\mathbf{S}_j$ is the order cone associated to $\mathbf{x}_j$, $\mathbf{S}_j = \{u \in R^D / u_d \leq u_{d'} \Leftrightarrow x_{jd} \leq x_{jd'}\}$. $P(\mathbf{Y}|\mathbf{K})$ is obtained using a cyclic pool adjacent algorithm (CPAVA) similar to the backfitting procedure built around the PAVA (Robertson et al. (1988)). To obtain $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}_u^2$, we propose an iterative procedure following the ideas of Rueda et al. (2010) and Lombardía et al. (2017).

Finally,

$$\theta_d = \sum_{j=1}^{p_1} \beta_j x_{jd} + \sum_{j=p_1+1}^{p} f_j(x_{jd}), \quad d = 1, \ldots, D;$$

where $f_j()$ are any smooth functions to be estimated using penalized spline regression.

So, we can write the working model as the following mixed effects model

$$\text{(SM)}: \quad \mathbf{Y} = \boldsymbol{\theta} + \mathbf{u} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{u} + \mathbf{e},$$

where $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$ represents the spline function. According to the base used for P-splines, $\mathbf{X}$ and $\mathbf{Z}$ have different forms. In particular, in this work we use B-Splines. Being $\mathbf{X} = [\mathbf{1}, \mathbf{x}, \ldots, \mathbf{x}^{(d-1)})]$, with d the order of the differences in the penalty matrix, and $\mathbf{Z} = \mathbf{B}\mathbf{R}\boldsymbol{\Sigma}^{-1/2}$, with $\mathbf{B}$, is the matrix of the spline basis obtained from the covariate $\mathbf{X}$, while $\mathbf{R}$ and $\boldsymbol{\Sigma}$ are matrices that form part of the decomposition in singular values of the penalty matrix. Having described the base, the connection with a mixed model is immediate. Some applications of this type of model in SAE are Opsomer et al. (2008) and Ugarte et al. (2009), among others.

In order to fit the model, it is suitable to treat $\mathbf{Z}\mathbf{v}$ as a random effect term, with $\mathbf{v} \sim N(0, \boldsymbol{\Sigma}_v = \sigma_v^2 \boldsymbol{I}_{c-2})$, where c is the number of columns in the original base $\mathbf{B}$. Using Maximun Likelihood estimation we obtain:

$$\hat{\mu}_d = \mathbf{x}_d \hat{\boldsymbol{\beta}} + \hat{v}_d + \hat{u}_d, \quad d = 1, \ldots, D;$$

where $\hat{\boldsymbol{\theta}}_d = \mathbf{x}_d \hat{\boldsymbol{\beta}} + \hat{v}_d$. More details of the estimation process can be seen in Lombardía et al. (2017).

2.2 Mean Squared Error

An important problem in SAE is how to assess the prediction error: $MSE(\hat{m}_d) = E\left[(\hat{m}_d - m_d)^2\right]$. In practice, this measure cannot be estimated because the unknown model generating the data is not usually the same as the working model. Nevertheless, under the working model, the MSE of the predictor of the area characteristic of interest, $MSE(\hat{\mu}_d)$, can be estimated and is used as an estimator of $MSE(\hat{m}_d)$.

It is well known that the estimation of the prediction MSE is complicated because of the added variability induced by the estimation of the model hyper-parameters. There are several works published on this problem, taking the LM introduced in the previous section as the working model. Prasad and Rao (1990) developed MSE estimators with a bias of order o(1/D), Datta and Lahiri (2000) extended the estimation of Prasad and Rao to the more general mixed linear model and Das et al. (2004) extended the estimation of Datta and Lahiri (2000) by relaxing some independent assumptions. See Datta (2009) for an extensive review of methods for estimating the MSE of the empirical best linear unbiased predictor under LM. As an alternative, the resampling methods appear. We can cite among others: Lohr and Rao (2009) who give a jackknife estimator, and Hall and Maiti (2006), González-Manteiga et al. (2008a) and González-Manteiga et al. (2008b), who propose bootstrap estimators.

In this paper, we will use a parametric bootstrap to estimate the MSE. This allows us to give a predictor based on any of the previous working models. The steps are as follows:

1. Calculate the estimators of the model parameters, which are used to generate the model in the next step: $\hat{\theta}_d = \widehat{f}(x_{1d}, \ldots, x_{qd})$, $\hat{u}_d$ and $\hat{\sigma}_u^2$ from LM, MM or SM.
2. Repeat B times $(b = 1, \ldots, B)$
   (a) Generate the random part of the model, $u_d^{*(b)}$ and $e_d^{*(b)}$ as independents $N(0, \widehat{\sigma}_u^2)$ and $N(0, \sigma_d^2)$ respectively, $d = 1, \ldots, D$. Now, construct the bootstrap model $y_d^{*(b)} = \mu_d^{*(b)} + e_d^{*(b)} = \hat{\theta}_d + u_d^{*(b)} + e_d^{*(b)}$.
   (b) From each bootstrap sample $\{y_d^{*(b)}, \mathbf{x}_d\}$, calculate $\widehat{\mu}_d^{*(b)} = \hat{\theta}_d^{*(b)} + \hat{u}_d^{*(b)}$, with $\hat{\theta}_d^{*(b)} = \widehat{f}^{*(b)}(x_{1d}, \ldots, x_{pd})$ and $\hat{u}_d^{*(b)}$ calculated as in step 1 from the b-th bootstrap sample.
3. Approximate the MSE by Monte Carlo,

$$\widehat{MSE}(\hat{\mu}_d) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\mu}_d^{*(b)} - \mu_d^{*(b)})^2. \tag{2.1}$$

2.3 GAIC statistics

A key element for the construction of estimators based on the model is to have a working model that fits the data well. At this point, the selection methods become a tool of great interest, allowing us to choose the best model under which the estimator will be built.

The model selection has received much attention in the literature, one of the most popular approaches to model selection is to use the Akaike information criterion AIC (Akaike (1973)). In

general terms, the value of the AIC for a model M is defined as $AIC(M) = -2log(l(M)) + 2P$, where $l(M)$ is the model likelihood and P is a penalty term. The model M with the lowest AIC is selected.

The penalty term was equal, originally, to the number of parameters in the model (p) or to the Degrees of Freedom (DF), which coincides with p for simple models. However, for more complex models, such as the lasso or shrinkage estimation, the concept of DF is not so simple (Kato (2009) and Tibshirani and Taylor (2012)). A lot of work has been done over the recent years in deriving measures of the complexity of models in such cases, in particular, to be used as a penalty term in the AIC. Several related concepts have been used, the concepts of Divergence and Effective Degrees of Freedom, for example, by Rueda (2013) or Hansen and Sokol (2014). Other authors have used the concept of Generalized Degrees of Freedom (GDF), originally defined in Ye (1998) for normal models and also considered in different models by Shen and Huang (2006), Gao and Fang (2011), Zhang et al. (2012), among others.

In the particular case of models with random effects, the problem of model selection is that it considers different versions for the penalty terms and either conditional or marginal loglikelihoods. In the context of SAE, Pfeffermann (2013) proposes to follow the ideas of Vaida and Blanchard (2005), explaining that in linear mixed model selection the marginal likelihood should be used when the interest is the population parameters and the conditional when the interest is the clusters or domains. Rao and Molina (2015), following this idea, say that a conditional AIC is more relevant when the focus is on the estimation of the realized random effects and the regression parameters. Han (2013) also gives a conditional AIC for the Fay-Herriot model. The most recent contributions to the subject are the new definition of GDF proposed by You et al. (2016), who also derive AIC ($yAIC$) measures using the new GDF as the penalty, and Lombardía et al. (2017), who deal with the issue of model selection in SAE problems when no linear models are considered and where the new statistic, $xGAIC$, is also derived.

In this work, we compare the behavior of three AICs suitable for the selection of models in SAE: $cAIC$ built from a purely conditional approach; $yAIC$, which conbines the conditional loglikelihood with the GDF proposed in You et al. (2016); and $xAIC$, which combines a quasi-log-likelihood with a GDF estimator as proposed in Lombardía et al. (2017).

For the construction of these estimators, we follow the general AIC statistic for model M, $AIC(M) = -2log(l(M)) + 2P$. To calculate the loglikelihood, we consider the conditional focus:

$$\log(l_c(M)) = -\frac{1}{2}D\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_{y|u}| - \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{V}_{y|u}^{-1}(\mathbf{Y}-\boldsymbol{\mu}),$$

where $\mathbf{V}_{y|u} = Var(\mathbf{Y}|\mathbf{u})$ and $\boldsymbol{\mu} = E(\mathbf{Y}|\mathbf{u}) = \boldsymbol{\theta} + \mathbf{u}$, assuming that $\mathbf{Y}|\mathbf{u} \sim N(\boldsymbol{\mu}, \mathbf{V}_{y|u})$. We also consider the quasi-likelihood, as in Lombardía et al. (2017), which considers the focus in the random effect and the total variability as follows:

$$\log(l_x(M)) = -\frac{1}{2}D\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_{y}| - \frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{V}_{y}^{-1}(\mathbf{Y}-\boldsymbol{\mu}),$$

where $\mathbf{V}_y = Var(\mathbf{Y})$ is the marginal variance.

For the penalty term, we consider the GDF, which is defined as a measure of the sensibility of each fitted value $\hat{m}_d$, to the perturbation in the corresponding observed value m_d, for $d = 1, \ldots, D$, as follows:

$$GDF = \sum_{d=1}^{D} \frac{\partial E(\hat{m}_d)}{\partial m_d}.$$

To estimate GDF, different expectations have been defined in the literature, the most recent being due to You et al. (2016) and Lombardía et al. (2017), where it has been calculated from the conditional mean estimator $\hat{\mu}_d$ and the marginal expectation $E_Y(Y_d) = \theta_d$:

$$xGDF = \sum_{d=1}^{D} \frac{\partial E_Y(\hat{\mu}_d)}{\partial \theta_d}. \tag{2.2}$$

Assuming that the interest is the domain or area, we consider the version calculated from the conditional mean estimator and conditional expectation (Vaida and Blanchard (2005), Lombardía et al. (2017)):

$$cGDF = \sum_{d=1}^{D} \frac{\partial E_{Y|u}(\hat{\mu}_d)}{\partial \mu_d}. \tag{2.3}$$

Finally, we select the model with the smallest value of $GAIC = -2\log(l(\hat{M})) + \widehat{GDF}$, where $l(\hat{M})$ depends on the fitted model, and the GDF is a known quantity in simple models, but unknown in complex modeling procedures. You et al. (2016) and Lombardía et al. (2017) propose estimating (2.2) and (2.3) by bootstrap, so we denote the corresponding estimators by $\widehat{xGDF}$ and $\widehat{cGDF}$, respectively.

Considering the above mentioned, we can take a purely conditional measure

$$cGAIC(M) \;=\; -2\log(l_c(\hat{M})) + \widehat{cGDF};$$

the proposal of You et al. (2016), which combines the conditional loglikelihood with the $xGDF$

$$yGAIC(M) \;=\; -2\log(l_c(\hat{M})) + \widehat{xGDF};$$

and the mixed $GAIC$, $xGAIC$, introduced by Lombardía et al. (2017)

$$xGAIC(M) = -2\log(l_x(\hat{M})) + \widehat{xGDF}.$$

These measures will be studied and compared in the next section.

3 Simulation study

The objective of this simulation is to study the behavior of the estimators selected by the $xGAIC$, $yGAIC$ and $cGAIC$. We consider the MSE as a predictive measure of performance.

The data-generating models use auxiliary variables and other parameter values from the real case analyzed in the next section.

The simulated data are generated as $y_d = m_d + e_d$ for each domain $d = 1, ..., D$, where $D = 53$ and $e_d \sim N(0, \sigma_d^2)$. The values of σ_d^2 are those from the real case. Also, 10 times those values are defined to account for scenarios with different variability.

To derive m_d, we consider six models, the first three are standard area level models in SAE, a linear model (LM), a monotone but non linear mixed model (MM) and a non-monotone mixed model (NM), as follows,

- (LM): $m_d = \alpha + \beta x_d + u_d$;
- (MM): $m_d = \alpha - \frac{\beta}{30*(x_d)^3} + u_d$;
- (NM): $m_d = \alpha + \sin(\pi \frac{x_d - \min x_d}{\max x_d - \min x_d}) + u_d$;

where x_d is the variable $65age$ in the real case, α and β are the intercept and slope estimates from fitting a linear model to the real data, and $u_d \sim N(0, \sigma_u^2)$ with $\sigma_u^2 = 0, 0.2$.

The other three models, (G1),(G2) and (G3), are defined in a similar way to those above, but shifting the intercept term and the random effect for nine areas that have been randomly selected. The models are given as follows,

- (G1): $m_d = \alpha + a_d + \beta x_d + v_d$
- (G2): $m_d = \alpha + 10 * a_d - \frac{\beta}{30*(x_d)^3} + v_d$
- (G3): $m_d = \alpha + 2 * a_d + \sin(\pi \frac{x_d - \min x_d}{\max x_d - \min x_d}) + v_d$

where, $v_d \sim N(0, \sigma_{v_d}^2)$, $\sigma_{v_d}^2 = \sigma_u^2 + c_d$. $a_d \in \{0.12, 0.10, -0.15, 0.10, -0.15, -0.15, -0.12, 0.12, 0.12\}$ and $c_d \in \{0.22, 0.11, 0.33, 0.11, 0.33, 0.33, 0.52, 0.22, 0.22\}$ respectively for the nine areas selected. $a_d = c_d = 0$ for the other 44 areas. Also, we consider $\sigma_u^2 = 0, 0.2$.

For each scenario, we fit the linear, monotone and P-spline models, and analyze data as follows:

- Repeat $I = 500$ times ($i = 1, \ldots, 500$)
  - Generate samples (y_d, x_d), $d = 1, \ldots, D$ under different models and record m_d.
  - Fit the models and calculate for each: $\hat{\mu}_d$, $xGAIC$, $cGAIC$, $yGAIC$
  - Record the estimator selected by using the minimum $xGAIC$, $cGAIC$ and $yGAIC$.
- Derive global statistics:
  - Correct functional form classification rates from the $xGAIC$, $cGAIC$ and $yGAIC$ measures.

– Empirical MSE:

$$EMSE(\hat{\mu}_d) = \frac{1}{I}\sum_{i=1}^{I}(\hat{\mu}_d^{(i)} - m_d^{(i)})^2$$

with $\hat{\mu}_d$ calculated as in Section 2.1.

Next, we present the results of the simulation. Tables 1 and 2 show the results when data are generated from the models LM, MM and NM, while Table 3 shows the results when models G1,G2 and G3 are considered.

Table 1 shows the percentage of times that the Fay-Herriot, Monotone or P-spline models are selected by $xGAIC$, $cGAIC$ or $yGAIC$. Correct classification rates are the numbers in the diagonals. $xGAIC$ almost always selects the real model, whereas $cGAIC$ and $yGAIC$ do not. The difference in the correct classification rates between the three methods are significant in the nine scenarios considered.

Table 2, shows the $EMSE$ for different estimators of m_d, the Direct estimator (which is the generated observation), the model-based estimators from Fay-Herriot, Monotone and P-Spline models, and the estimators selected by the $xGAIC$, $cGAIC$ and $yGAIC$. It is shown that the estimators selected by the $xGAIC$ have a smaller $EMSE$ compared with the estimators selected by $cGAIC$ and $yGAIC$, in the nine scenarios, even being up to 10 times smaller, in some cases. In addition, the $EMSE$ of the estimator selected by the $xGAIC$ is smaller than the $EMSE$ of the Fay-Herriot model-based estimator, which is the most popular choice in SAE applications.

Finally, in Table 3, we give the $EMSE$ for G1, G2 and G3. In these cases, we use a generator model that does not match any working model. Also, in this case, the estimator selected by the $xGAIC$ is the one with the smallest $EMSE$.

Table 1: **Classification rates (%) using $xGAIC$, $cGAIC$ and $yGAIC$.**

	xGAIC			cGAIC			yGAIC		
Model	FH	Monotone	P-Spline	FH	Monotone	P-Spline	FH	Monotone	P-Spline
$\sigma_u^2 = 0, \sigma_d^2$									
LM	95.5	0.0	4.5	44.0	53.0	3.0	76.7	18.8	4.5
MM	0.0	100.0	0.0	9.3	90.7	0.0	10.6	89.4	0.0
NM	0.0	0.0	100.0	0.0	96.7	3.3	24.1	52.8	23.1
$\sigma_u^2 = 0.2, \sigma_d^2$									
LM	85.8	3.5	10.8	53.0	39.1	7.9	66.8	24.7	8.5
MM	0.0	95.0	5.0	60.8	39.2	0.0	51.6	48.4	0.0
NM	1.2	25.5	73.3	38.4	55.7	5.9	45.7	49.8	4.5
$\sigma_u^2 = 0.2, \sigma_d^2 * 10$									
LM	78.4	6.6	15.0	52.9	40.0	7.2	74.5	13.8	11.7
MM	0.0	92.2	7.8	0.8	99.2	0.0	24.2	75.8	0.0
NM	7.5	27.1	65.4	11.4	84.0	4.6	26.9	62.7	10.3

4 Application to real data

In this application, we deal with data from the ILC in Galicia in 2015. The statistical population in this survey is composed of all persons living in private households and the sample uses a two-stage sampling with primary sampling unit stratification. The first stage is formed by census sections grouped into strata, in agreement with the size of the municipality to which they belong. The second stage is formed by main family dwellings. No sub-sampling is carried out within this sample, investigating all dwellings that are their usual residence. The sample includes $9,216$ dwellings distributed in 512 census sections.

This survey is designed to obtain precise estimates at province and area level. The problem is to get reliable estimates for unplanned small domains. Sample sizes in the unplanned domains may be too small to obtain reliable direct estimates. Our domains of interest are the 53 counties and our goal is to estimate the mean income in the private households in the counties of Galicia in 2015. For this data set, the minimum sample size in the counties is 18, the first quartile is 49 and the median 108. Therefore, obtaining reliable estimates for target domains is a small area estimation problem and borrowing strength from auxiliary data is recommended.

The ILC does not produce official estimates at the domain level, but the analogous direct estimates of the total Y_d, the mean $\bar{Y}_d = Y_d/N_d$ and the size N_d are

$$\hat{Y}_d^{dir} = \sum_{j \in S_d} w_j \, y_j, \;\; \hat{\bar{Y}}_d^{dir} = \hat{Y}_d^{dir} / \hat{N}_d^{dir}, \;\; \hat{N}_d^{dir} = \sum_{j \in S_d} w_j,$$

where S_d is the sample in domain d and w_j is the official calibrated sampling weight. Considering that the sample weights w_j correspond to the inverse of the probability of selecting the individual j, π_j, then we have that

Table 2: $EMSE$ **for Direct estimator (Dir), the estimator from Fay-Herriot, monotone and the P-spline models, and for the estimators selected by the** $xGAIC$**,** $cGAIC$ **and** $yGAIC$**.**

Model	Dir	FH	Monotone	P-Spline	xGAIC	cGAIC	yGAIC
$\sigma_u^2 = 0, \sigma_d^2$							
LM	0.0148	0.0004	0.0068	0.0004	0.0004	0.0038	0.0017
MM	0.0148	0.0143	0.0057	0.0023	0.0057	0.0068	0.0069
NM	0.0148	0.0164	0.0160	0.0014	0.0014	0.0155	0.0126
$\sigma_u^2 = 0.2, \sigma_d^2$							
LM	0.0148	0.0133	0.0135	0.0136	0.0133	0.0134	0.0134
MM	0.0150	0.0146	0.0136	0.0145	0.0136	0.0142	0.0140
NM	0.0148	0.0141	0.0139	0.0140	0.0136	0.0140	0.0140
$\sigma_u^2 = 0.2, \sigma_d^2 * 10$							
LM	0.1494	0.0765	0.0821	0.0826	0.0778	0.0792	0.0778
MM	0.1520	0.1260	0.0873	0.0842	0.0871	0.0878	0.0975
NM	0.1499	0.0992	0.0911	0.1407	0.0883	0.0960	0.0970

Table 3: $EMSE$ **for Direct estimator (Dir), the estimator from Fay-Herriot, monotone and the P-spline models, and for the estimators selected by the** $xGAIC$**,** $cGAIC$ **and** $yGAIC$**.**

Model	Dir	FH	Monotone	P-Spline	xGAIC	cGAIC	yGAIC
$\sigma_{v_d}^2 = c_d, \sigma_d^2$							
G1	0.0152	0.0098	0.0104	0.0107	0.0096	0.0098	0.0096
G2	0.0152	0.0147	0.0143	0.0136	0.0138	0.0145	0.0143
G3	0.0166	0.0148	0.0136	0.0111	0.0105	0.0137	0.0141
$\sigma_{v_d}^2 = 0.2 + c_d, \sigma_d^2$							
G1	0.0148	0.0136	0.0137	0.0136	0.0136	0.0136	0.0136
G2	0.0152	0.0149	0.0145	0.0143	0.0144	0.0148	0.0147
G3	0.0150	0.0143	0.0141	0.0139	0.0140	0.0142	0.0142
$\sigma_{v_d}^2 = 0.2 + c_d, \sigma_d^2 * 10$							
G1	0.1471	0.0809	0.0866	0.0839	0.0821	0.0846	0.0830
G2	0.1551	0.1327	0.1137	0.1033	0.1121	0.1147	0.1195
G3	0.1492	0.0995	0.0937	0.0875	0.0881	0.0940	0.0947

$$\hat{Y}_d^{dir} = \sum_{j \in S_d} w_j \, y_j = \sum_{j \in S_d} \frac{1}{\pi_j} \, y_j, \;\; \hat{\bar{Y}}_d^{dir} = \hat{Y}_d^{dir} / \hat{N}_d^{dir} = \frac{\sum_{j \in S_d} w_j \, y_j}{\hat{N}_d^{dir}}.$$

If $\pi_j \neq 0$, the variance estimator of the direct estimator would be:

$$\hat{V}_\pi(\hat{\bar{Y}}_d^{dir}) \approxeq \frac{1}{\hat{N}_d^2} \sum_{j \in S_d} w_j (w_j - 1) \left(y_j - \hat{\bar{Y}}_d^{dir} \right)^2. \tag{4.1}$$

This formula is obtained with the simplifications: $w_j = \frac{1}{\pi_j}$, $\pi_{jj} = \pi_j$ and $\pi_{ij} = \pi_i \pi_j$ for $i \neq j$ in the second order inclusion probabilities. Using this, we calculate the coefficient of variation of the direct estimator for all the counties. We obtain five counties with the coefficient of variation higher than 20%. At this point, it is desirable to take into account the fact, that, in the statistics of labour

force, the Office for National Statistics (ONS) in the UK considers that an estimate is publishable and therefore official, if the coefficient of variation is less than 20%.

The models are formulated using the log-income by private household as the response variable, $\log(\widehat{\bar{Y}}^{dir})$, where $\widehat{\bar{Y}}^{dir}$ is the direct estimator obtained from the ILC. In our dataset, we also have these auxiliary variables:

- *Age*: the percentage of population under 20 years ($20age$) and 65 years and over ($65age$).
- *Population density*: the population divided by total land area for each county ($density$).
- *Education level*: the proportion of people with low education (low) and higher education ($higher.educ$).

Table 4 shows the correlations between the auxiliary variable and the response variable. From this table, it can be seen that $higher.educ$ is the most positively correlated variable with the log-income. As expected, $65age$ and low are the most negatively correlated variables with the log-income, that is, the regions with the oldest population and low education have lower incomes.

Table 4: **Correlations between the predictors and the response.**

	density	$20age$	$65age$	low	$higher.educ$
log-income	0.30	0.35	-0.55	-0.50	0.59

In addition, non significative auxiliary variables have been discarded after fitting a Fay-Herriot model, with the exception of *Population density*, selected by the experts. Figure 1 shows the scatterplots from the auxiliary variables against the response variable for the chosen variables. The assumption of a monotone relationship between auxiliary and response is reasonable.

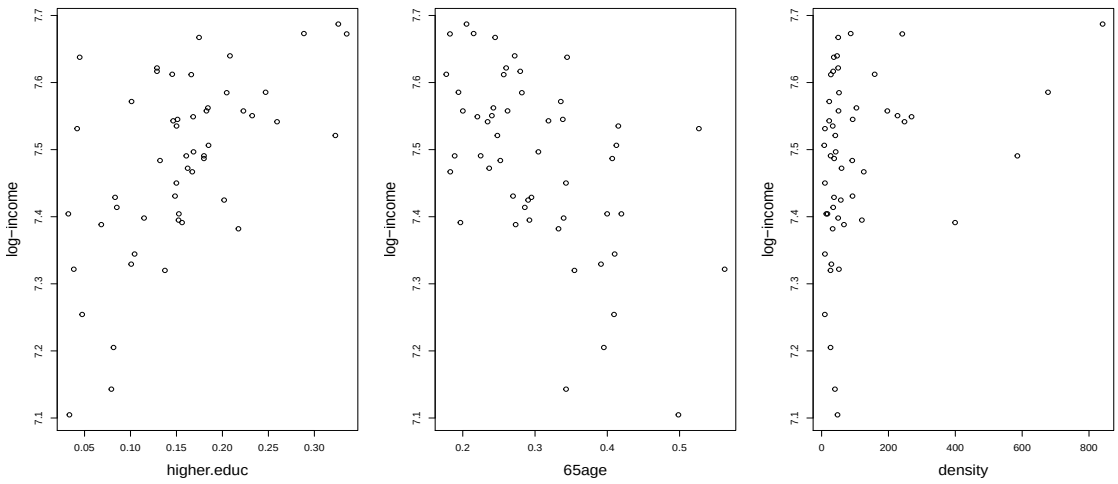

Figure 1: Relation between the auxiliary variables and the response variable.

Several models have been fitted to the data set, taking two or three predictors. The problem of the selection of auxiliary variables is solved simultaneously with the problem of the specification of their functional form. The candidate models are formulated as those in Section 2: linear (LM), monotone (MM) or non-monotone (SM), with different predictors. Relevant statistics are shown in Table 5. From the numbers in this table, it can be seen that the $xGAIC$ selects model M6, which is defined by using the linear function for $higher.educ$ and the non monotone function for $65age$. In contrast, $cGAIC$ and $yGAIC$ select model M7, which is defined using a non monotone function for both $higher.educ$ and $65age$, and also with a greater estimated variance of the random effect than M6.

Table 5: $xGAIC$, $cGAIC$, $yGAIC$ and $\hat{\sigma}_u^2$ for the fitted models.

Model label	Predictors	xGAIC	cGAIC	yGAIC	$\hat{\sigma}_u^2$
M1	higher.educ/65age (LM)	-92.28	-91.66	-92.28	0
M2	higher.educ/65age/density (LM)	-93.89	-94.96	-93.62	0
M3	higher.educ (LM) 65age (MM)	-87.78	-88.97	-88.21	0.00047
M4	higher.educ/65age (MM)	-94.72	-94.19	-94.72	0
M5	higher.educ (LM) 65age/density (MM)	-84.3	-84.9	-84.9	0
M6	higher.educ (LM) 65age (SM)	-96.91	-98.12	-97.75	0.0003
M7	higher.educ/65age (SM)	-81.51	-110.73	-103.73	0.0023
M8	higher.educ (LM) 65age/density (SM)	-81.78	-108.82	-102.69	0.0022

Figure 2 plots the residuals for the two models selected, M6 and M7. The residuals are randomly distributed above and below zero and no rare pattern is observed.

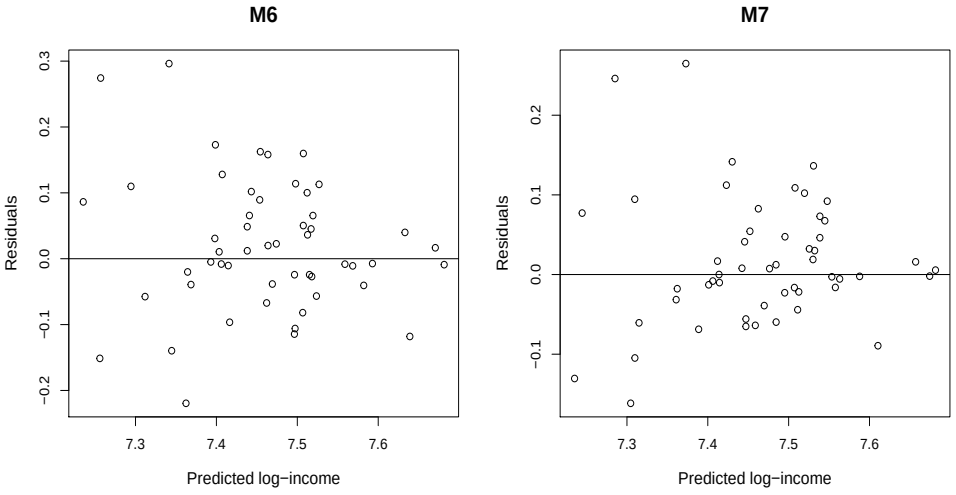

Figure 2: Residuals for models M6 and M7.

Figure 3 plots the model-based versus the direct estimates of the log-income in the counties of Galicia for M6 and M7. We observe that the model-based estimators take values rather different from the direct estimates. We also observe that the model-based estimates are in general smaller than the direct ones.

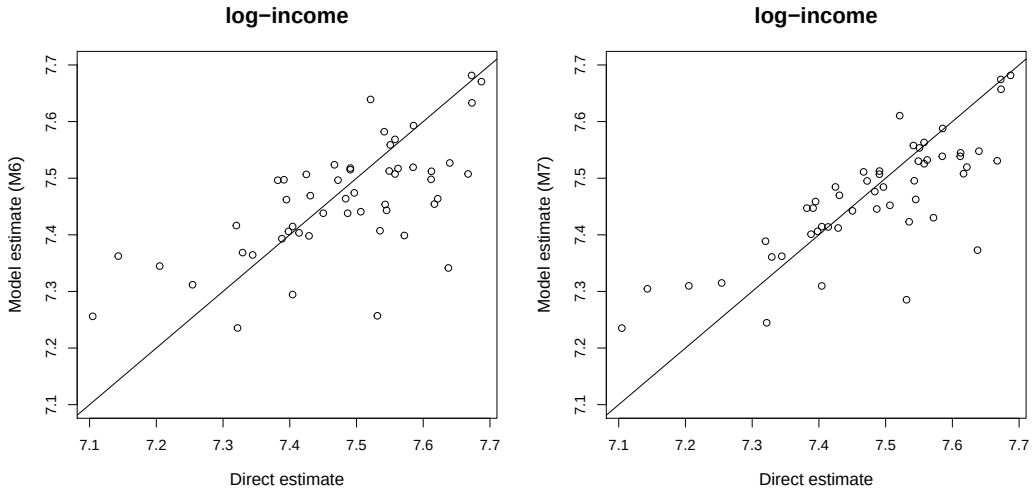

Figure 3: Model-based versus direct estimates of the log-income.

Figure 4 plots the parametric bootstrap estimates of the root mean squared errors ($\widehat{RMSE}$) of the model-based estimators for models M6 and M7, and the $\widehat{RMSE}$ for the corresponding direct estimates from (4.1). It can be seen that the $\widehat{RMSE}$ of the direct estimator is much greater than the model-based estimators under M6 and M7. Also, that the estimator of the model M6 gives the best results in terms of $\widehat{RMSE}$. Again, it is clear that the $xGAIC$ shows the best behavior if we compare it with $cGAIC$ and $yGAIC$, which select a more complicated estimator and with a higher $\widehat{RMSE}$. These results are in line with those obtained in the simulation study, see Table 2.

Finally, and from the previous results, Figure 5 maps the mean income in the counties for the model-based estimator under M6. The colors are darker in areas with a higher level of income. Note that the counties with the highest income are in the west, the north coast and the cities of Lugo and Ourense. On the other hand, it can be observed that the provinces that are in the southeast of Galicia, in general terms, have the lowest income values. The exception in this latter situation is the dynamic county of Valdeorras, in the southeast of Galicia. Also, Figure 6 shows the relative $\widehat{RMSE}$ ($\widehat{RRMSE}$) of the mean income in the Galician counties estimated using parametric bootstrap as follows from eq. (2.1) :

$$\widehat{RRMSE}(\hat{\mu}_d) = \frac{\sqrt{\widehat{MSE}(\hat{\mu}_d)}}{\hat{\mu}_d}.$$

As can be seen in Figure 4, we can observe that all the counties have an $\widehat{RRMSE}$ lower than 5.27% and all the values are much lower than the $\widehat{RRMSE}$ of the direct estimator.

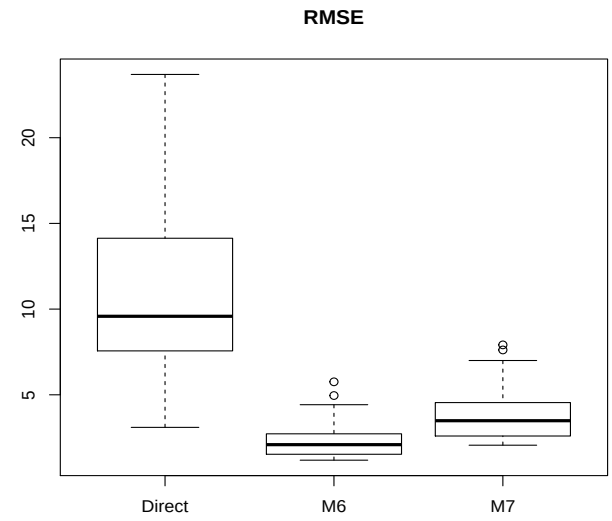

Figure 4: $\widehat{RMSE}$ of direct and model based estimators of log-income.

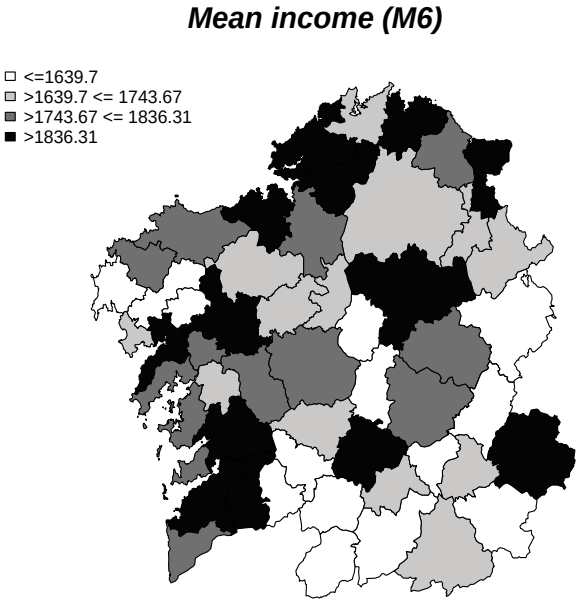

Figure 5: Map of the mean income in the Galician counties.

5 Conclusions

The first, and most important contribution, from the theoretical point of view, is the good performance of the $xGAIC$, against $cGAIC$ and $yGAIC$, to select the best model in terms of

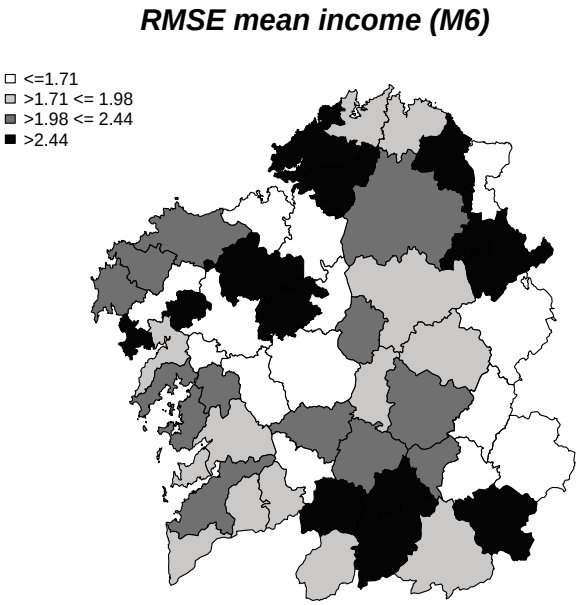

Figure 6: Map of the $\widehat{RRMSE}$ of the mean income in the Galician counties (%).

MSE. We verify this in the simulations and in the results obtained in the real case, where the $xGAIC$ selects a model with the lowest MSE. In fact, this paper can be considered an extension of Lombardia et al (2017), which shows how $xGAIC$ performs remarkably better than $cGAIC$ with a smaller missclassification rate. We have also shown that $xGAIC$ is better than the alternatives in terms of MSE.

The second most important conclusion of this paper is that the common practice of comparing estimators by comparing the estimators of the mean-squared errors is not the most desirable, as the MSE estimator is ussually model dependent. In practice, estimators are proposed that may not be better than direct estimators because an incorrect model is being used. It is necessary to take this fact into account, so it is important to carry out a wide model selection process that should include parametric and non parametric formulations and different sets of auxiliaries.

As for the income results in Galicia, we can conclude that the northwest coast, in general terms, is the most dynamic part with a higher level of income. There is a big problem in the southeast, which is essentially rural. Fixing the population in rural areas, with decent living and income levels, is a basic requirement to ensure territorial balance and environmental sustainability. Galicia is aging, and this is more accentuated in rural areas, due to the emigration of young people and the fall in the birth rate. This reality, with obvious economic and sociological consequences, demands answers from social policies.

References

Akaike, H. (1973). Information theory and the maximum likelihood principle. In *International Symposium on Information Theory*, pages 267 – 281, Budapest. Akademiai Kiado.

Das, K., Jiang, J., and Rao, J. (2004). Mean squared error of empirical predictor. *Annals of Statistics*, **32**, 818–840.

Datta, G. (2009). Model-based approach to small area estimation, ind d.pfeffermann and c.r. rao (eds.). *Sample Surveys: Inference and Analysis, Handbook of Statistics*, 29B, 251–288.

Datta, G. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, **10**, 613–627.

Datta, G. S., Lahiri, P., and Maiti, T. (2002). Empirical bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, **102**, 83–97.

Fay, R. and Herriot, R. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, **70**, 311–319.

Gao, X. and Fang, Y. (2011). A note on the generalized degrees of freedom under L1 loss function. *J. of Statistical Planning and Inference*, **141**, 677–686.

Ghosh, M., Nangia, N., and Kim, D. (1996). Estimation of median income of four-person families: A bayesian time series approach. *Journal of the American Statistical Association*, **91**, 1423–1431.

Ghosh, M. and Rao, J. (1994). Small area estimation: An appraisal. *Statistical Science*, **9**, 55–93.

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

González-Manteiga, W., Lombardía, M., Molina, I., Morales, D., and Santamaría, L. (2008b). Boostrap mean squared error os a small-area eblup. *Journal of Statistical Computation and Simulation*, **78**, 443–462.

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

Han, B. (2013). Conditional akaike information criterion in the fay-herriot model. *Statistical Methodology*, **11**, 53–67.

Hansen, N. and Sokol, A. (2014). Degrees of freedom for nonlinear least squares estimation. *Submitted*.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15(1)**, 1–96.

Kato, K. (2009). On the degrees of freedom in shrinkage estimation. *J. of Multivariate Analysis*, **100**, 1138–1352.

Lohr, S. and Rao, J. (2009). Jackknife estimation of mean squared error of small area predictors in non-linear mixed models. *Biometrika*, **96**, 457–468.

Lombardía, M., López-Vizcaíno, E., and Rueda, C. (2017). Mixed generalized akaike information (xgaic) for small area models. *Journal of Royal Statistical Society Series A*, **180**, 1229–1252.

Molina, I. and Marhuenda, Y. (2015). sae: An r package for small area estimation. *The R Journal*, **7(1)**, 81–98.

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

Muller, S., Scealy, J., and Welsh, A. (2013). Model selection in linear mixed models. *Statistical Science*, **28(2)**, 135–167.

Opsomer, J., Claeskens, G., Ranalli, M., Kauermann, G., and Breidt, F. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of Royal Statistical Society Series B*, **70**, 265–286.

Pfeffermann, D. (2002). Small area estimation. new developments and directions. *International Statistical Review*, **70**, 125–143.

Pfeffermann, D. (2013). New important developments in small area estimation. statistical science. *Statistical Science*, pages 40–68.

Prasad, N. and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 67–72.

Rao, J. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, **25**, 175–186.

Rao, J. (2003). *Small Area Estimation*. Wiley, New York.

Rao, J. and Molina, I. (2015). *Small Area Estimation*. Wiley, New York.

Robertson, T., Wright, F., and Dykstra, R. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, New York.

Rueda, C. (2013). Degrees of freedom and model selection in semiparametric additive monotone regression. *J. of Multivariate Analysis*, **117**, 88–99.

Rueda, C. and Lombardía, M. (2012). Small area semiparametric additive isotone models. *Statistical Modelling*, **12**, 503–525.

Rueda, C., Menéndez, J., and Gómez, F. (2010). Small area estimators based on restricted mixed models. *TEST*, **19**, 558–568.

Shen, X. and Huang, H.-C. (2006). Optimal model assessment, selection, and combination. *Journal of the American Statistical Association*, **101(474)**, 554 – 568.

Tibshirani, R. T. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, **40(2)**, 1198–1232.

Ugarte, M., Goicoa, T., Militino, A., and Durbán, M. (2009). Spline smoothing in small area trend estimation and forecasting. *Computational Statistics and Data Analysis*, **53**, 3616–3629.

Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.*, **93**, 120–131.

You, C., Muller, S., and Ormerod, J. (2016). On generalized degrees of freedom with application in linear mixed models selection. *Statistics and Computing*, **26**, 199–210.

Zhang, B., Shenb, X., and Mumford, S. (2012). Generalized degrees of freedom and adaptive model selection in linear mixed-effects models. *Computational Statistics and Data Analysis*, **56**, 574–586.