ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

*J. R. Statist. Soc.* A (2017)
**180**, *Part* 4, *pp.* 1229–1252

# Mixed generalized Akaike information criterion for small area models

María José Lombardía,

*Universidade da Coruña, Spain*

Esther López-Vizcaíno

*Instituto Galego de Estatística, Santiago de Compostela, Spain*

and Cristina Rueda

*Universidad de Valladolid, Spain*

[Received April 2016. Final revision May 2017]

**Summary.** A mixed generalized Akaike information criterion *xGAIC* is introduced and validated. It is derived from a quasi-log-likelihood that focuses on the random effect and the variability between the areas, and from a generalized degree-of-freedom measure, as a model complexity penalty, which is calculated by the bootstrap. To study the performance of xGAIC, we consider three popular mixed models in small area inference: a Fay–Herriot model, a monotone model and a penalized spline model. A simulation study shows the good performance of xGAIC. Besides, we show its relevance in practice, with two real applications: the estimation of employed people by economic activity and the prevalence of smokers in Galician counties. In the second case, where it is unclear which explanatory variables should be included in the model, the problem of selection between these explanatory variables is solved simultaneously with the problem of the specification of the functional form between the linear, monotone or spline options.

*Keywords*: Akaike information criterion; Bootstrap; Fay–Herriot model; Generalized degree of freedom; Monotone model; Small area estimation; Spline regression

## 1. Introduction

The question of model selection has received much attention in the literature in the past (starting with the well-known paper by Akaike (1973)), and also in recent years due, among other reasons, to the increasing complexity of modelling approaches. In particular, the question has received considerable attention in the context of linear mixed models; for a comprehensive review of various approaches we refer the reader to Muller *et al.* (2013). However, in small area estimation (SAE), this is still a problem that has been only tentatively studied. One of the most popular approaches to model selection is to use the Akaike information criterion AIC, which is the objective of this paper.

In general terms, the value of AIC for a model $M$ is defined as $\mathrm{AIC}(M) = -2\log\{l(M)\} + 2D$, where $l(M)$ is the model likelihood and $D$ is a penalty term, which was originally equal to the number of parameters in the model, $p$ (Akaike, 1973). The model with the lowest value of AIC is then selected. It is also usual to use the degrees of freedom DF instead of $p$; DF coincides

*Address for correspondence*: María José Lombardía, Departamento de Matemáticas, Universidade de Coruña, Campus Elviña, Faculade de Informática, A Coruña 15071, Spain.
E-mail: maria.jose.lombardia@udc.es

 0964–1998/17/1801229

with $p$ for simple models like the normal linear regression model and with the number of *free parameters* or the parameters of the final model, in other cases. However, this is not always so simple for more complex models such as the lasso or shrinkage estimation; see among others Kato (2009) and Tibshirani and Taylor (2012). Much work has been done over the last few years in deriving measures of the complexity of models in such cases, to be used, particularly, as a penalty term in 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 researchers have used the concept of generalized degrees of freedom (GDF), which was originally defined in Ye (1998) for normal models and also considered in different models by Shen and Huang (2006), Gao and Fang (2011) or Zhang *et al.* (2012), among others.

In the particular case of models with random effects, the most interesting references are, chronologically ordered, Hodges and Sargent (2001), Vaida and Blanchard (2005), Greven and Kneib (2010), Yu and Yau (2012), Zhang *et al.* (2012), Muller *et al.* (2013), Overholser and Xu (2014) and You *et al.* (2016). All of them use AIC-measures to address the problem of model selection but consider different versions for the penalty terms and either conditional or marginal log-likelihoods. Moreover, some of them, using similar solutions for the penalty and the same log-likelihood, propose different estimation approaches. The most recent contribution to the subject is the new definition of GDF that was proposed by You *et al.* (2016), who also derived new conditional AIC, cAIC, and marginal AIC, mAIC, measures using the new GDF as the penalty. They reported the ability of these criteria to select the data-generating models, with simulations, in various settings.

In the context of SAE, Pfeffermann (2013), in a review about new important developments in SAE, thought about the problem of model selection and followed the ideas of Vaida and Blanchard (2005), who explained in detail the advantage of cAIC over mAIC in many applications of SAE. They argued that, in linear mixed model selection, the marginal likelihood should be used when the interest is the population parameters and the conditional likelihood when the interest is the clusters or domains. Rao and Molina (2015), following this idea, said that cAIC is more relevant when the focus is on estimation of the realized random effects and the regression parameters. Han (2013) studied cAIC in the Fay–Herriot model and affirmed that cAIC is suitable for measuring the prediction performance of a working model in SAE. Marhuenda *et al.* (2014) studied the bias corrections to AIC for the Fay–Herriot model. Model selection in SAE problems, when $P$-splines models are the candidate models, was considered by Jiang *et al.* (2010), who proposed a fence method; but, as far as we know, there are no other references dealing with this issue for a general model formulation such as we consider here, despite the fact that there has been a larger number of SAE applications, in recent years, using estimators based on non-linear models (Opsomer *et al.* (2008), Jiang *et al.* (2010), Salvati *et al.* (2010), Sperlich and Lombardía (2010), Rueda and Lombardía (2012) or Torabi and Shokoohi (2015), among others). In fact, analytical values for GDF are known only when the fitted model is linear (Han (2013) and references therein).

In this work, we deal with the selection model issue in the more realistic setting where the functional form of the predictors cannot, or should not, be assumed to be linear. Non-parametric models based on $P$-splines and monotone assumptions will be the opponents to linear models in the selection process. These models have been considered in some of the above references.

With that goal in mind, we propose a new AIC, which we refer to as xGAIC. The novelty of the proposal is twofold: on the one hand, xGAIC is derived by using a quasi-log-likelihood that focuses on the random effect and the variability between the areas, and, on the other hand, the penalty is a GDF-measure, inspired by that of You *et al.* (2016) and Ye (1998). In addition, as the focus in SAE is the domains, xGAIC is compared with an alternative derived by using

a conditional likelihood, cGAIC. A bootstrap approach to estimate GDF is considered that distinguishes the mixed, xGDF, or the conditional, cGDF, focus. Moreover, both proposals will be compared with the conditional AIC proposed by Vaida and Blanchard (2005), vAIC, which has been proposed by several researchers for model selection in SAE applications (see Pfeffermann (2013)). We shall show, using simulation results and two real cases, the good performance of xGAIC, compared with those of cGAIC and vAIC, for selecting the predictors and its functional form in SAE problems.

We organize the remainder of the paper as follows. Section 2 introduces the linear, monotone and $P$-spline mixed models. In Section 3, the new information criteria statistics, xGAIC and cGAIC, are derived. Also in Section 3, an estimation approach, based on the bootstrap, to calculate the penalties is described. The behaviour of the new methodology is illustrated in Section 4, with simulation studies and in Section 5, with two real applications: one of socio-economic interest and the other in the field of health. Section 6 discusses some issues related to the application of the proposed methodology to real data and gives the main conclusions and a brief discussion of future research. Finally, Appendix A includes methodological developments of interest.

## 2. Model description

In this section, we consider some linear and non-linear models of interest to study two real applications in the field of socio-economy and health. In the following subsections, we briefly detail how they are used in SAE. We take $D$ as the number of domains or small areas of interest and $p$ auxiliary variables $(X_1, \ldots, X_p)$. Let $\mathbf{X}$ be the matrix of auxiliary information with dimension $D \times p$, and $\mathbf{x}_d$ be a vector containing the aggregated (population) values of $p$ auxiliary variables for domain $d$.

The model is composed of two stages. In the first stage, a model called the sampling model is used to represent the sampling error of direct estimators. Let $\mu_d$ be the characteristic of interest in the $d$th area and $y_d$ be a direct estimator of $\mu_d$. The sampling model indicates that the direct estimator $y_d$ is unbiased and can be expressed as

$$y_d = \mu_d + e_d, \qquad d = 1, \ldots, D,$$

where $e_d \sim N(0, \sigma_d^2)$ are independent with $\sigma_d^2$ known. In practice, we take the design-based variance of direct estimator $y_d$. In the second stage, the domain characteristics $\mu_d \sim N(\theta_d, \sigma_u^2)$, where $\theta_d = f(x_{1d}, \ldots, x_{pd})$ is a linear or non-linear function, depending on the model that is considered. Hereinafter we use $\theta_d$ or $f(x_{1d}, \ldots, x_{pd})$ interchangeably, and the variance $\sigma_u^2$ is unknown, with the random effect $u_d \sim N(0, \sigma_u^2)$. So, the final model can be expressed as a single model in the form

$$y_d = \theta_d + u_d + e_d, \qquad d = 1, \ldots, D.$$

### 2.1. Fay–Herriot model

The Fay–Herriot model has been widely used in the literature of SAE (Fay and Herriot, 1979). Under this model,

$$\theta_d = f(x_{1d}, \ldots, x_{pd}) = \mathbf{x}_d \boldsymbol{\beta}, \qquad d = 1, \ldots, D,$$

where $\boldsymbol{\beta}$ is the vector of the regression coefficients. So, the Fay–Herriot model is

$$y_d = \mathbf{x}_d \boldsymbol{\beta} + u_d + e_d, \qquad d = 1, \ldots, D.$$

In matrix notation,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e},$$

where $\mathbf{u} \sim N(0, \boldsymbol{\Sigma}_u = \sigma_u^2 \mathbf{I}_D)$ is the small area random effect which is independent of the model error $\mathbf{e} \sim N(0, \boldsymbol{\Sigma}_e)$, and $\mathbf{I}_D$ is the identity matrix with dimension $D$. Note that the variability of $\mathbf{e}$ is known and different in each area: $\boldsymbol{\Sigma}_e = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$. Then, the covariance matrix of the response variable $\mathbf{Y}$ is given by $\text{var}(\mathbf{Y}) = \mathbf{V}_y = \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e$.

To fit the model, we use maximun likelihood (ML) estimation and the functions that are available in package `sae` in the R language (Molina and Marhuenda, 2015). If the variance components are known, the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) of the random part are obtained as

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{Y}$$

and

$$\tilde{\mathbf{u}} = \boldsymbol{\Sigma}_u \mathbf{V}_y^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}});$$

then $\tilde{\boldsymbol{\mu}} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{u}}$. But, in practice, the variance components $\sigma_u^2$ are unknown, so well-known methods, such as ML or restricted maximum likelihood (REML) can be used to estimate them. Details of the calculation can be seen in Appendix A. The ML update equation of $\sigma_u^2$ is

$$\hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{D - (1/\hat{\sigma}_u^2)\text{tr}(\hat{\mathbf{T}}_u)}$$

where $\hat{\mathbf{T}}_u = (\hat{\boldsymbol{\Sigma}}_e^{-1} + \hat{\boldsymbol{\Sigma}}_u^{-1})^{-1}$. These equations can be solved numerically from an initial value $\sigma_{u0}^2$ of $\hat{\sigma}_u^2$, which is replaced in $\tilde{\boldsymbol{\beta}}$ to obtain $\boldsymbol{\beta}_0$. These values, $\sigma_{u0}^2$ and $\boldsymbol{\beta}_0$, are replaced in $\tilde{\mathbf{u}}$ and $\hat{\sigma}_u^2$ and the process is iterated until convergence. The empirical BLUE $\hat{\boldsymbol{\beta}}$ and the empirical BLUP $\hat{\mathbf{u}}$ are obtained by replacing, in the above expressions, the variance components for their estimates. So we have $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$ and $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

### 2.2. Monotone model

Under the monotone model,

$$\theta_d = f(x_{1d}, \ldots, x_{pd}) = \sum_{j=1}^{p_1} \beta_j x_{jd} + \sum_{j=p_1+1}^{p} h_j(x_{jd}), \qquad d = 1, \ldots, D,$$

where $h_j(\cdot)$ are monotone functions. To obtain the ML estimators for the area parameters and the estimator for the variance of the random effects we use the methodology that was proposed in Rueda and Lombardía (2012), which is now briefly summarized.

In the simple case where $\sigma_u^2$ is known, the weight matrix $\mathbf{W} = \mathbf{V}_y^{-1}$ is also known, and $\hat{\theta}_d$ is the projection of $\mathbf{Y}$ with weight $\mathbf{W}$ onto a cone, as follows:

$$\hat{\theta}_d = \sum_{j=1}^{p_1} \hat{\beta}_j x_{jd} + \sum_{j=p_1+1}^{p} \hat{h}_j(x_{jd}) = P_W(\mathbf{Y}|\mathbf{K}).$$

$\mathbf{K} = \mathbf{L}_0 + \mathbf{S}_1 + \ldots + \mathbf{S}_{p_2}$ is a convex region in $R^n$ defined by the restrictions that are 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 with $\mathbf{x}_j$, $S_j = \{u \in R^n / u_d \leqslant u_{d'} \Leftrightarrow x_{jd} \leqslant x_{jd'}\}$. $P_W(\mathbf{Y}|\mathbf{K})$ is obtained by using a cyclic pool adjacent algorithm in a style of a backfitting procedure built

around the pool adjacent algorithm PAVA, where PAVA is a popular algorithm (Robertson *et al.*, 1988) for solving univariate monotone regressions.

From $\hat{\theta}_d$, an empirical maximum likelihood predictor for the area means is easily derived by

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

In the case where $\sigma_u^2$ is unknown, we propose an iterative procedure to obtain $\hat{\boldsymbol{\theta}} = P_W(\mathbf{Y}|\mathbf{K})$ and $\hat{\sigma}_u^2$. The procedure is based on the next equality proposed in Rueda *et al.* (2010):

$$E_{\boldsymbol{\theta}}[h(\sigma_u^2)|D_K(\mathbf{Y}) = l] = D - l$$

where $h(\sigma_u^2) = \|\mathbf{Y} - P_W(\mathbf{Y}|\mathbf{K})\|_W^2$ and $D_K(\mathbf{Y})$ measures the degrees of freedom of the model, which is obtained from case B of Rueda (2013). The estimators are obtained by solving the equation $h(\sigma_u^2) = D - l$ iteratively, where $l$ is the dimension of subspace $\mathbf{L}_K$, such that $P_W(\mathbf{Y}|\mathbf{K}) = P_W(\mathbf{Y}|\mathbf{L}_K)$ and letting $\sigma_u^2 = 0$, when no positive solution exists. The procedure includes a control variable $c_0$ to assure that $l$ does not vary in the next iteration once $l$ has attained the same value in two successive iterations. The $\hat{\sigma}_u^2$ from the Fay–Herriot model can be used as the initial values.

### 2.3. Penalized spline model

Under the penalized spline model, we take

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

where $p = p_1 + p_2$ is the number of area auxiliary variables, and $f_j(\cdot)$ are any smooth functions to be estimated by using penalized spline regression.

Using $P$-splines, we can write the model as the mixed effects model

$$\mathbf{Y} = \boldsymbol{\theta} + \mathbf{u} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{u} + \mathbf{e},$$

where $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$ represents the spline function. According to the base that is used for $P$-splines, $\mathbf{X}$ and $\mathbf{Z}$ have different forms.

(a) Truncated polynomial spline basis: $\mathbf{X} = (\mathbf{1}, \mathbf{x}, \ldots, \mathbf{x}^p)$ and $\mathbf{Z} = ((x_i - k_K)_+^p)$, where $p$ is the degree of spline, $(x)_+^p$ denotes the function $x^p I_{x>0}$ and $k_1 < \ldots < k_K$ is a set of fixed knots.

(b) $B$-splines: $\mathbf{X} = (\mathbf{1}, \mathbf{x}, \ldots, \mathbf{x}^{(d-1)})$, where $d$ is the order of the differences in the penalty matrix, and $\mathbf{Z} = \mathbf{B}\mathbf{R}\boldsymbol{\Sigma}^{-1/2}$, with $\mathbf{B}$ the matrix of the spline basis obtained from the covariate $\mathbf{X}$, whereas $\mathbf{R}$ 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. To fit the model, it is suitable to treat $\mathbf{Z}\mathbf{v}$ as a random-effect term, with $\mathbf{v} \sim N(0, \boldsymbol{\Sigma}_v = \sigma_v^2 \mathbf{I}_{c-2})$, where $c$ is the number of columns in the original base $\mathbf{B}$. Then, the covariance matrix of the variable $\mathbf{Y}$ is given by $\text{var}(\mathbf{Y}) = \mathbf{V}_y = \mathbf{Z}\boldsymbol{\Sigma}_v\mathbf{Z}' + \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e$, adding an additional term if we compare it with the Fay–Herriot model. Some examples of the use of $P$-splines in small areas can be seen in Opsomer *et al.* (2008) and Ugarte *et al.* (2009), among others.

If the variance components are known, the BLUE of $\boldsymbol{\beta}$ is obtained as

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{Y}$$

and the BLUP of the random part is

$$\tilde{\mathbf{v}} = \boldsymbol{\Sigma}_v \mathbf{Z}' \mathbf{V}_y^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$
$$\tilde{\mathbf{u}} = \boldsymbol{\Sigma}_u \mathbf{V}_y^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

In practice, the variance components $(\sigma_v^2, \sigma_u^2)$ are unknown; then well-known methods, such as ML or REML, can be used to estimate them, as in the Fay–Herriot model (see details of the calculation in Appendix A). Then, the ML update estimates of $\sigma_v^2$ and $\sigma_u^2$ are respectively

$$\hat{\sigma}_v^2 = \frac{\hat{\mathbf{v}}'\hat{\mathbf{v}}}{D - (1/\hat{\sigma}_v^2)\mathrm{tr}(\hat{\mathbf{T}}_v)}$$

and

$$\hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{D - (1/\hat{\sigma}_u^2)\mathrm{tr}(\hat{\mathbf{T}}_u)},$$

where $\hat{\mathbf{T}}_v$ and $\hat{\mathbf{T}}_u$ are the empirical versions of $\mathbf{T}_v$ and $\mathbf{T}_u$ respectively; see the expression in Section 2.1. Replacing the true variances with their estimates $(\hat{\sigma}_v^2, \hat{\sigma}_u^2)$, we obtain the empirical BLUE $\hat{\boldsymbol{\beta}}$ and the empirical BLUPs $\hat{\mathbf{v}}$ and $\hat{\mathbf{u}}$. The fitted model can be obtained in R with various packages: `mgcv`, `SemiPar` or `nlme`, among others (Wood, 2006; Pinheiro *et al.*, 2016; R Core Team, 2015), but not directly, because, in our case, $\boldsymbol{\Sigma}_e$ is not constant between areas and is assumed to be known. To solve these problems, we propose to use the following algorithm.

*Step 1*: estimate $\mathbf{X}$ and $\mathbf{Z}$ by using the function `gamm` of package `mgcv` in R.
*Step 2*: estimate the initial values of $\hat{\sigma}_{u,0}^2$ and $\hat{\sigma}_{v,0}^2$ from the output of the function `gamm` of R and calculate $\hat{\mathbf{V}}_0 = \mathbf{Z}\hat{\boldsymbol{\Sigma}}_{v,0}\mathbf{Z}' + \hat{\boldsymbol{\Sigma}}_{u,0} + \boldsymbol{\Sigma}_e$.
*Step 3*: for step $k$, calculate

$$\hat{\beta}_k = (\mathbf{X}'\hat{\mathbf{V}}_{y,k-1}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_{k-1}^{-1}\mathbf{Y},$$

$$\hat{v}_k = \hat{\boldsymbol{\Sigma}}_v \hat{\mathbf{V}}_{y,k-1}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

$$\hat{u}_k = \hat{\boldsymbol{\Sigma}}_{u,k-1} \hat{\mathbf{V}}_{y,k-1}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

$$\hat{\sigma}_{v,k}^2 = \frac{\hat{\mathbf{v}}_k'\hat{\mathbf{v}}_k}{D - (1/\hat{\sigma}_{v,k-1}^2)\mathrm{tr}(\hat{\mathbf{T}}_{v,k-1})},$$

$$\hat{\sigma}_{u,k}^2 = \frac{\hat{\mathbf{u}}_k'\hat{\mathbf{u}}_k}{D - (1/\hat{\sigma}_{u,k-1}^2)\mathrm{tr}(\hat{\mathbf{T}}_{u,k-1})},$$

$$\hat{\mathbf{T}}_{v,k} = (\mathbf{Z}'\boldsymbol{\Sigma}_e^{-1}\mathbf{Z} + \hat{\boldsymbol{\Sigma}}_{v,k}^{-1})^{-1},$$

$$\hat{\mathbf{T}}_{u,k} = (\boldsymbol{\Sigma}_e^{-1} + \hat{\boldsymbol{\Sigma}}_{u,k}^{-1})^{-1}.$$

*Step 4*: stop when $\max\{|\hat{\sigma}_{u,k}^2 - \hat{\sigma}_{u,k-1}^2|, |\hat{\sigma}_{v,k}^2 - \hat{\sigma}_{v,k-1}^2|, \|\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_{k-1}\|\} \leqslant \varepsilon$.

Finally, we have $\hat{\mu} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}} + \hat{\mathbf{u}}$ and $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}}$.

## 3. Akaike information criterion

As we discussed in Section 1, various AIC-statistics have been defined in the literature which

are derived from two versions of the likelihood, conditional or marginal, and different versions of the penalty term. Although there is no clear consensus on whether to use one or another in general problems, conditional likelihood is the most popular choice in SAE.

In this section, we introduce a new statistic, xGAIC, which does not follow a marginal or a conditional approach. In contrast, it is derived by using quasi-log-likelihood and a naturally linked GDF-measure.

xGAIC is partially inspired in the proposal by You *et al.* (2016), as we also define the GDF-measure xGDF as the marginal expected estimated value of $Y_d$ with respect to the corresponding underlying true means. However, the solution for the estimation of xGDF, which is intractable analytically for non-linear models, such as those considered in this paper; and, more importantly, the use of a quasi-log-likelihood, which is a measure combining marginal focus and conditional focus, are novel contributions of this paper. In addition, we also introduce in this section, for comparison, a conditional version: cGAIC. The mixed and conditional likelihood are presented in Section 3.1, whereas, in Section 3.2, we discuss the proposed approach to estimate xGDF and the conditional GDF. A pure marginal approach is not considered in this paper, as it is more suitable for model selection without random effects and the focus here is a mixed models selection for SAE applications. Finally, in Section 3.3, the expressions for xGAIC and cGAIC are included, along with that of vAIC, exhibiting the differences and similarities between them.

### *3.1. The calculation of log-likelihood*

Let the general model be

$$\mathbf{Y} = \boldsymbol{\theta} + \mathbf{u} + \mathbf{e},$$

where the marginal likelihood approach assumes that $\mathbf{Y} \sim N(\boldsymbol{\theta}, \mathbf{V}_y)$ and $\mathbf{V}_y = \text{var}(\mathbf{Y})$. Following the calculations of Section 2, $\mathbf{V}_y = \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e$ for the Fay–Herriot model and the monotone model, and for the $P$-spline model $\mathbf{V}_y = \mathbf{Z}\boldsymbol{\Sigma}_v\mathbf{Z}' + \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_e$.

In contrast, the conditional likelihood approach assumes that $\mathbf{Y}|\mathbf{u} \sim N(\boldsymbol{\mu}, \mathbf{V}_{y|u})$ with $\boldsymbol{\mu} = E(\mathbf{Y}|\mathbf{u}) = \boldsymbol{\theta} + \mathbf{u}$ and $\mathbf{V}_{y|u} = \text{var}(\mathbf{Y}|\mathbf{u})$. In the previous examples, $\mathbf{V}_{y|u} = \boldsymbol{\Sigma}_e$ for the Fay–Herriot model and the monotone model, and $\mathbf{V}_{y|u} = \mathbf{Z}\boldsymbol{\Sigma}_v\mathbf{Z}' + \boldsymbol{\Sigma}_e$ for the $P$-spline model.

Then, the marginal log-likelihood is calculated as

$$\log\{l_{\text{m}}(M)\} = -\tfrac{1}{2}D\log(2\pi) - \tfrac{1}{2}\log|\mathbf{V}_y| - \tfrac{1}{2}(\mathbf{Y}-\boldsymbol{\theta})'\mathbf{V}_y^{-1}(\mathbf{Y}-\boldsymbol{\theta})$$

and the conditional log-likelihood is calculated as

$$\log\{l_{\text{c}}(M)\} = -\tfrac{1}{2}D\log(2\pi) - \tfrac{1}{2}\log|\mathbf{V}_{y|u}| - \tfrac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{V}_{y|u}^{-1}(\mathbf{Y}-\boldsymbol{\mu}).$$

As an alternative, we propose a quasi-log-likelihood, which considers the focus on the random effect and the total variability, combining the conditional and marginal approach, as follows:

$$\log\{l_x(M)\} = -\tfrac{1}{2}D\log(2\pi) - \tfrac{1}{2}\log|\mathbf{V}_y| - \tfrac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{V}_y^{-1}(\mathbf{Y}-\boldsymbol{\mu}).$$

### *3.2. The calculation of generalized degrees of freedom*

Ye (1998) proposed a GDF-concept that motivated the proposal in You *et al.* (2016), which in turn motivated our proposal. Our first proposal is an extension of the GDF of You *et al.* (2016) from the unit level model to the area level model, and the second considers only the conditional distribution of $Y$.

We denote by $E_Y(\cdot)$ and $\text{cov}_Y(\cdot)$ the expectation and covariance with respect to the marginal distribution of $Y$ respectively:

$$
\begin{aligned}
\text{xGDF} &= \sum_{d=1}^{D} \frac{\partial E_Y(\hat{\mu}_d)}{\partial \theta_d} \\
&= \sum_{d=1}^{D} \frac{\partial}{\partial \theta_d} \left[ (2\pi)^{-D/2} |\mathbf{V}_y|^{-1/2} \int \hat{\mu}_d(\mathbf{y}) \exp\left\{ -\frac{1}{2} (\mathbf{y}-\boldsymbol{\theta})' \mathbf{V}_y^{-1} (\mathbf{y}-\boldsymbol{\theta}) \right\} d\mathbf{y} \right] \\
&= \sum_{d=1}^{D} (2\pi)^{-D/2} |\mathbf{V}_y|^{-1/2} \int \left\{ -\frac{1}{2} \frac{\partial (\mathbf{y}-\boldsymbol{\theta})' \mathbf{V}_y^{-1} (\mathbf{y}-\boldsymbol{\theta})}{\partial \theta_d} \right\} \hat{\mu}_d(\mathbf{y}) \exp\left\{ -\frac{1}{2} (\mathbf{y}-\boldsymbol{\theta})' \mathbf{V}_y^{-1} (\mathbf{y}-\boldsymbol{\theta}) \right\} d\mathbf{y} \\
&= \sum_{d=1}^{D} (2\pi)^{-D/2} |\mathbf{V}_y|^{-1/2} \int \sum_{i=1}^{D} V_y^{di} (y_i - \theta_i) \hat{\mu}_d(\mathbf{y}) \exp\left\{ -\frac{1}{2} (\mathbf{y}-\boldsymbol{\theta})' \mathbf{V}_y^{-1} (\mathbf{y}-\boldsymbol{\theta}) \right\} d\mathbf{y} \\
&= \sum_{d=1}^{D} \sum_{i=1}^{D} V_y^{di} \int (y_i - \theta_i) \hat{\mu}_d(\mathbf{y}) \Phi(\mathbf{y}) d\mathbf{y} \\
&= \sum_{d=1}^{D} \sum_{i=1}^{D} V_y^{di} E_Y \{ \hat{\mu}_d (y_i - \theta_i) \} \\
&= \sum_{d=1}^{D} \sum_{i=1}^{D} V_y^{di} \text{cov}_Y (\hat{\mu}_d, y_i),
\end{aligned}
$$

where $\Phi(\cdot)$ denotes the probability distribution function of $\mathbf{y}$, $\hat{\mu}_d$ is the estimator of $\mu_d$ (see Section 2) and $V_y^{di}$ is the $di$-element of the matrix $\mathbf{V}_y^{-1}$. We must implicity make a choice between a marginal and a conditional expectation and we choose the marginal expectation. This GDF is a measure of the sensitivity of the expected estimate of the response with respect to the corresponding underlying means.

For comparison, we also propose a conditional GDF based on the conditional distribution of $Y$. We denote by $E_{Y|u}(\cdot)$ and $\text{cov}_{Y|u}(\cdot)$ the expectation and covariance with respect to the conditional approach respectively. Again, we obtain an empirical measure of the model complexity, but in this case the underlying true mean is $\boldsymbol{\mu} = E(\mathbf{Y}|\mathbf{u}) = \boldsymbol{\theta} + \mathbf{u}$:

$$
\begin{aligned}
\text{cGDF} &= \sum_{d=1}^{D} \frac{\partial E_{Y|u}(\hat{\mu}_d)}{\partial \mu_d} \\
&= \sum_{d=1}^{D} \frac{\partial}{\partial \mu_d} \left[ (2\pi)^{-D/2} |\mathbf{V}_{y|u}|^{-1/2} \int \hat{\mu}_d(\mathbf{y}) \exp\left\{ -\frac{1}{2} (\mathbf{y}-\boldsymbol{\mu})' \mathbf{V}_{y|u}^{-1} (\mathbf{y}-\boldsymbol{\mu}) \right\} d\mathbf{y} \right] \\
&= \sum_{d=1}^{D} (2\pi)^{-D/2} |\mathbf{V}_{y|u}|^{-1/2} \int \left\{ -\frac{1}{2} \frac{\partial (\mathbf{y}-\boldsymbol{\mu})' \mathbf{V}_{y|u}^{-1} (\mathbf{y}-\boldsymbol{\mu})}{\partial \mu_d} \right\} \\
&\qquad \times \hat{\mu}_d(\mathbf{y}) \exp\left\{ -\frac{1}{2} (\mathbf{y}-\boldsymbol{\mu})' \mathbf{V}_{y|u}^{-1} (\mathbf{y}-\boldsymbol{\mu}) \right\} d\mathbf{y} \\
&= \sum_{d=1}^{D} (2\pi)^{-D/2} |\mathbf{V}_{y|u}|^{-1/2} \int \sum_{i=1}^{D} V_{y|u}^{di} (y_i - \mu_i) \hat{\mu}_d(\mathbf{y}) \exp\left\{ -\frac{1}{2} (\mathbf{y}-\boldsymbol{\mu})' \mathbf{V}_{y|u}^{-1} (\mathbf{y}-\boldsymbol{\mu}) \right\} d\mathbf{y} \\
&= \sum_{d=1}^{D} \sum_{i=1}^{D} V_{y|u}^{di} \int (y_i - \mu_i) \hat{\mu}_d(\mathbf{y}) \Phi(\mathbf{y}|\mathbf{u}) d\mathbf{y} \\
&= \sum_{d=1}^{D} \sum_{i=1}^{D} V_{y|u}^{di} E_{Y|u} [\hat{\mu}_d (y_i - \mu_i)] \\
&= \sum_{d=1}^{D} \sum_{i=1}^{D} V_{y|u}^{di} \text{cov}_{Y|u} (\hat{\mu}_d, y_i).
\end{aligned}
$$

where $V^{di}_{(y|u)}$ is the $di$-element of the matrix $\mathbf{V}^{-1}_{y|u}$. This expression depends on $\mathbf{u}$, which is substituted by $\hat{\mathbf{u}}$.

The analytic values of xGDF and cGDF are difficult to obtain for most non-linear mixed models. We propose the parametric bootstrap as an alternative, following the idea of You *et al.* (2016). We give a solution to approximate xGDF and cGDF, proposing the following plug-in type of estimators that use bootstrap resampling to facilitate their calculation. For this, we use the full model REML estimates to resample $Y$, according to the consistency property of REML estimates (Jiang, 1996).

### *3.2.1. Parametric bootstrap*

#### *3.2.1.1. Mixed approach.*

*Step 1*: calculate the estimators of the model parameters (as in Section 2), which are used to generate the model in the next step—$\hat{\theta}_d = \hat{f}(x_{1d}, \ldots, x_{qd})$, the fitted model following the Fay–Herriot, monotone or $P$-spline model, $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_e$.
*Step 2*: repeat the following steps $B$ times ($b = 1, \ldots, B$).

(a) Generate the random part of the model $u^*_d$ and $e^*_d$ as independents $N(0, \hat{\sigma}^2_u)$ and $N(0, \sigma^2_d)$ distributions respectively, $d = 1, \ldots, D$. Now, construct the bootstrap model $y^{*(b)}_d = \mu^{*(b)}_d + e^{*(b)}_d$, with $\mu^{*(b)}_d = \hat{\theta}_d + u^{*(b)}_d$, and the variance–covariance matrix $\hat{\mathbf{V}}_y$.
(b) From each bootstrap sample $\{y^{*(b)}_d, \mathbf{x}_d\}$, calculate $\hat{\mu}^{*(b)}_d = \hat{\mu}_d(y^{*(b)}_d, \mathbf{x}_d) = \hat{\theta}^{*(b)}_d + \hat{u}^{*(b)}_d$, with $\hat{\theta}^{*(b)}_d = \hat{f}^{*(b)}(x_{1d}, \ldots, x_{pd})$ and $\hat{u}^{*(b)}_d$ calculated as $\hat{f}(x_{1d}, \ldots, x_{pd})$ and $\hat{u}_d$ from the $b$th bootstrap sample respectively.

*Step 3*: approximate xGDF by Monte Carlo sampling,

$$\widehat{\text{xGDF}} = \sum_{d=1}^{D} \sum_{i=1}^{D} \frac{1}{B-1} \sum_{b=1}^{B} V^{*(b),di}_y (\hat{\mu}^{*(b)}_d - \bar{\hat{\mu}}^*_d)(y^{*(b)}_i - \bar{y}^*_i)$$

where $\bar{\hat{\mu}}^*_i = (1/B)\Sigma_{b=1}^{B} \hat{\mu}^{*(b)}_i$ and $\bar{y}^*_i = (1/B)\Sigma_{b=1}^{B} y^{*(b)}_i$, and $V^{*(b),di}_y$ is the $di$-element of the inverse of the $\mathbf{V}^{*(b)}_y$-matrix.

#### *3.2.1.2. Conditional approach.*

*Step 1*: we calculate the estimators of the model parameters as in Section 2—$\hat{\theta}_d = \hat{f}(x_{1d}, \ldots, x_{qd})$ and $\hat{u}_d$.
*Step 2*: repeat the following steps $B$ times ($b = 1, \ldots, B$).

(a) Generate the random part of the model $e^*_d$ as $N(0, \sigma^2_d)$, $d = 1, \ldots, D$. Construct the bootstrap model $y^{*(b)}_d = \hat{\mu}_d + e^{*(b)}_d$, with $\hat{\mu}_d = \hat{\theta}_d + \hat{u}_d$ and the variance–covariance matrix $\hat{\mathbf{V}}_{y|u}$.
(b) From each bootstrap sample $\{y^{*(b)}_d, \mathbf{x}_d\}$, calculate $\hat{\mu}^{*(b)}_d = \hat{\mu}_d(y^{*(b)}_d, \mathbf{x}_d) = \hat{\theta}^{*(b)}_d + \hat{u}^{*(b)}_d$, calculated as $\hat{\theta}_d$ and $\hat{u}_d$ from the $b$th bootstrap sample respectively.

*Step 3*: approximate cGDF by Monte Carlo sampling,

$$\widehat{\text{cGDF}} = \sum_{d=1}^{D} \sum_{i=1}^{D} \frac{1}{B-1} \sum_{b=1}^{B} V^{*(b),di}_{y|u} (\hat{\mu}^{*(b)}_d - \bar{\hat{\mu}}^*_d)(y^{*(b)}_i - \bar{y}^*_i)$$

where $V^{*(b),di}_{y|u}$ is the $di$-element of the inverse of the $\mathbf{V}^{*(b)}_{y|u}$-matrix.

In this way, we finally compute $\widehat{\text{xGDF}}$ or the $\widehat{\text{cGDF}}$ as required.

### *3.3. Generalized Akaike information criterion statistics*

We define the generalized Akaike information criterion GAIC for a small area model $M$,

$$y_d = f(x_{1d}, \ldots, x_{pd}) + u_d + e_d, \qquad d = 1, \ldots, D,$$

as follows:

$$\mathrm{GAIC}(M) = -2\log\{l(M)\} + \mathrm{GDF}.$$

Here $l(M)$ may be $l_x(M)$ or $l_c(M)$ (or a marginal version) and GDF is estimated by $\widehat{\mathrm{xGDF}}$ or $\widehat{\mathrm{cGDF}}$ (or other candidates). Here, we propose to combine $l_x(M)$ and $\widehat{\mathrm{xGDF}}$ by considering the random effect and the variability between areas to define

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

And, for comparison, we also define

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

In fact, we could obtain several definitions of GAIC by taking other combinations. You *et al*. (2016) suggested using the same GDF-estimator combined with $l_c(M)$ and also with a marginal likelihood, whereas Pfeffermann (2013), in a review about SAE, proposed the use of the conditional AIC of Vaida and Blanchard (2015) for linear mixed models, which is defined as

$$\mathrm{vAIC} = -2\log\{l_c(\hat{M})\} + P^*$$

where

$$P^* = \frac{n(n-k-1)(\rho+1) + n(k+1)}{(n-k)(n-k-2)},$$

where $k$ is the number of covariates and $\rho = \mathrm{tr}(H)$, and $H$ defines the matrix mapping the observed vector $y$ onto the fitted vector $\hat{y} = X\hat{\beta} + \hat{u}$ such that $\hat{y} = Hy$. As far as we know, the matrix $H$ has not been derived in the literature for monotone models, so we can only calculate the expression of vAIC for the Fay–Herriot and $P$-spline models. The difference between vAIC and cGAIC is in the penalty term. In the next section, we compare the values for cGDF and $P^*$ that were obtained in the simulations and the performance of the three criteria when the Fay–Herriot and $P$-spline models are considered, showing that cGAIC provides similar results to those of vAIC.

The way that xGAIC is calculated, it can be used for selection from a range of models, going from a simple linear model to the most complex non-parametric model. In SAE applications, particularly in those which are included in this paper, we propose to use the mixed estimators for the small areas resulting from the model with the lowest value of xGAIC. cGAIC-selection is also recorded for comparison. Surprisingly, quite different models can be selected by these two criteria, as can be seen in the following sections where real and simulated data sets are analysed.

## 4. Simulation study

### 4.1. Simulation experiment 1

The goal of this first simulation experiment is to analyse the behaviour of the proposed methodology for the selection between linear, monotone and spline models. To generate the data, we considered the Labour Force Survey (LFS) example that is analysed in Section 5.1 as a reference, and normality assumptions. Then, the simulated model is $y_d = f(x_d) + u_d + e_d$ $(d = 1, \ldots, D)$, where $D = 77$, $x_d \sim U(0, 1)$, $u_d \sim N(0, \sigma_u^2)$ and $e_d \sim N(0, \sigma_d^2)$. Various scenarios are designed, based on different definitions for $f(\cdot)$, and various $\sigma_u$- and $\sigma_d$-values, $d = 1, \ldots, D,$.

Regarding the definition of $f(\cdot)$, three models are considered: a linear model LM,

$$f(x_d) = \beta_0 + \beta_1 x_d,$$

a monotone but non-linear model MM,

$$f(x_d) = \beta_0 + \log(x_d),$$

and a non-monotone model NM,

$$f(x_d) = \beta_0 + \sin(\pi x_d).$$

For the variance parameters, we consider $\sigma_d^2 = \sigma_{d,\mathrm{LFS}}^2$ $(d = 1, \ldots, 77)$ values in the interval [0,0.09] and $\sigma_u^2 = \sigma_{u,\mathrm{LFS}}^2 = 0.21$, where $\sigma_{d,\mathrm{LFS}}^2$ and $\sigma_{u,\mathrm{LFS}}^2$ are the values of $\sigma_d^2$ and $\hat{\sigma}_u^2$ in example 1 of the real applications, and the four combinations resulting from also using $\sigma_{d,\mathrm{LFS}}^2 \times 10$ and $\sigma_{u,\mathrm{LFS}}^2/10$:

(a) Var1, $\sigma_{u,\mathrm{LFS}}^2$ and $\sigma_{d,\mathrm{LFS}}^2$;
(b) Var2, $\sigma_{u,\mathrm{LFS}}^2/10$ and $\sigma_{d,\mathrm{LFS}}^2$;
(c) Var3, $\sigma_{u,\mathrm{LFS}}^2$ and $\sigma_{d,\mathrm{LFS}}^2 \times 10$;
(d) Var4, $\sigma_{u,\mathrm{LFS}}^2/10$ and $\sigma_{d,\mathrm{LFS}}^2 \times 10$.

In short, we consider a total of 12 scenarios, corresponding to the three model specifications and the four variance parameter configurations. For each scenario, we generate and analyse data as follows.

(a) Repeat $I = 500$ times $(i = 1, \ldots, 500)$.
  (i) Generate samples $(y_d, x_d)$, $d = 1, \ldots, D$ under LM, MM and NM.
  (ii) Fit the Fay–Herriot model FH, the monotone model and the *P*-spline model and calculate, for each, $\hat{\sigma}_u^2$, $\widehat{\mathrm{cGDF}}$, $\widehat{\mathrm{xGDF}}$, cGAIC, xGAIC and $P^*$.
  (iii) Record the model selected, by using the minimum cGAIC or xGAIC, between the Fay–Herriot, monotone or *P*-spline model and also record whether the cGAIC- or xGAIC-selected model agrees with the generated model.
  (iv) Record $\hat{\sigma}_u^2$ corresponding to the model selected, by using the minimum cGAIC or xGAIC.
(b) Derive global statistics:
  (i) average values of $\hat{\sigma}_u^2$ , $\widehat{\mathrm{cGDF}}$ and $\widehat{\mathrm{xGDF}}$ for the Fay–Herriot, monotone and *P*-spline models;
  (ii) average values of $P^*$ for the Fay–Herriot and *P*-spline models;
  (iii) correct classification rates from using xGAIC or cGAIC;
  (iv) relative root-mean-squared errors RRMSE for $\hat{\sigma}_u^2$, corresponding to the model that is selected by cGAIC or xGAIC,

$$\mathrm{RRMSE}(\hat{\sigma}_u^2) = \frac{\sqrt{\left\{ (1/I) \sum_{i=1}^{I} (\hat{\sigma}_u^{2(i)} - \sigma_u^2)^2 \right\}}}{\sigma_u^2}.$$

Tables 1–3 include the main statistics from the simulation results for the 12 scenarios: mean values of penalties, percentages of correct classification and RRMSEs of $\hat{\sigma}_u^2$ for the selected model respectively.

Firstly, in Table 1, mean values of $\widehat{\mathrm{xGDF}}$, $\widehat{\mathrm{cGDF}}$ and $P^*$ are shown. Note that $P^*$-values are not given for the monotone models because, as far as we know, they have not been derived in this context. Numbers in Table 1 show that the values of $\widehat{\mathrm{xGDF}}$, $\widehat{\mathrm{cGDF}}$ and $P^*$ are quite similar: a little higher for $P^*$ than for $\widehat{\mathrm{xGDF}}$ or $\widehat{\mathrm{cGDF}}$ but in the same range. In fact, vAIC and cGAIC

**Table 1.** Average values of $\widehat{\text{xGDF}}$, $\widehat{\text{cGDF}}$ and $P^*$ under various simulated scenarios

| *Scenario* | *Results for* $\widehat{xGDF}$ | | | *Results for* $\widehat{cGDF}$ | | | *Results for* $P^*$ | |
|---|---|---|---|---|---|---|---|---|
| | *FH* | *Monotone* | *P-spline* | *FH* | *Monotone* | *P-spline* | *FH* | *P-spline* |
| *Var1* | | | | | | | | |
| LM | 74.83 | 75.05 | 74.86 | 74.89 | 74.99 | 74.86 | 76.05 | 76.05 |
| MM | 76.29 | 74.98 | 75.02 | 76.30 | 75.11 | 75.10 | 77.47 | 76.26 |
| NM | 75.44 | 75.45 | 74.98 | 75.47 | 75.41 | 74.96 | 76.60 | 76.17 |
| *Var2* | | | | | | | | |
| LM | 64.92 | 65.69 | 64.69 | 64.49 | 65.33 | 64.26 | 66.23 | 66.22 |
| MM | 76.12 | 67.15 | 69.42 | 76.12 | 67.23 | 69.95 | 77.31 | 70.10 |
| NM | 74.02 | 73.62 | 64.28 | 74.01 | 73.60 | 64.25 | 74.60 | 66.13 |
| *Var3* | | | | | | | | |
| LM | 65.09 | 65.80 | 64.79 | 64.70 | 65.25 | 64.37 | 66.20 | 66.18 |
| MM | 72.47 | 65.74 | 65.75 | 72.54 | 65.47 | 65.70 | 73.69 | 67.40 |
| NM | 67.61 | 68.27 | 65.22 | 67.28 | 67.60 | 64.64 | 69.04 | 66.37 |
| *Var4* | | | | | | | | |
| LM | 40.17 | 42.79 | 39.39 | 38.75 | 40.36 | 38.16 | 41.10 | 41.10 |
| MM | 71.57 | 45.59 | 49.24 | 71.62 | 43.88 | 50.03 | 72.74 | 52.70 |
| NM | 60.14 | 58.36 | 38.26 | 59.60 | 57.22 | 37.70 | 60.23 | 40.70 |

perform in much the same way when they are used as model selection criteria, as the difference between these statistics is only due to the penalty term. However, both statistics perform in a different way from xGDF, as the former differs from the latter, mostly, in the likelihood term. This fact is shown with the simulated data sets in Section 4.2. Therefore, the values of vAIC and cGAIC are also in the same range as the difference between them is that we see due to the penalty term which results, as we show below, in quite similar selections using both criteria. However, the behaviour of xGAIC is quite different, as we shall show below.

From Table 2, it can be seen that xGAIC almost always selects a valid model, whereas cGAIC does not (note that model LM is a particular case of MM or NM). When an LM model, in particular, is generated, xGAIC selects LM less frequently (rates ranging from 22% to 37%) than cGAIC (rates ranging from 25% to 47%), but the model selected provides $\sigma_u^2$-estimators with a similar efficiency to those provided by cGAIC (see Table 3). Moreover, in this case, the MM and the *P*-spline model, which are also correct models, have similar degrees of complexity to the linear model, as shown by the GDF-values in Table 1, resulting in quite similar area estimators.

When model MM is generated, the correct classification rates from xGAIC are equal to 100% in the four cases considered, whereas cGAIC has rates ranging from 30% to 57%. Moreover, cGAIC incorrectly selected a Fay–Herriot model 7–37% of the time. The consequences of choosing a wrong model, the Fay–Herriot model, are more serious in this case. First, the $\sigma_u^2$-estimators are much worse and, also, the complexity of the model selected compared with the correct model increases in this case, as shown by the relatively high values of GDF when a Fay–Herriot model is fitted. These comments come from the numbers in Table 1 and Table 3, and are more important in the case $\sigma_{d,\text{LFS}}^2 \times 10$ and $\sigma_{u,\text{LFS}}^2/10$. When model NM is generated, again, xGAIC performs much better than cGAIC; the model selected, by the latter measure, is markedly worse in terms of complexity and efficiency of $\hat{\sigma}_u^2$.

Furthermore, some researchers have dealt with the interesting issue of 'double dipping' (e.g.

**Table 2.** Percentage of times that the Fay–Herriot, monotone or *P*-spline models are selected by xGAIC and cGAIC under various simulated scenarios

| *Scenario* | *Results for xGAIC* | | | *Results for cGAIC* | | |
|---|---|---|---|---|---|---|
| | *FH* | *Monotone* | *P-spline* | *FH* | *Monotone* | *P-spline* |
| *Var1* | | | | | | |
| LM | 36.98 | 61.85 | 1.17 | 47.04 | 44.07 | 8.89 |
| MM | 0 | 100 | 0 | 36.8 | 30.2 | 33 |
| NM | 1.35 | 14.35 | 84.3 | 34.98 | 33.63 | 31.39 |
| *Var2* | | | | | | |
| LM | 24.37 | 73.11 | 2.52 | 34.18 | 60.5 | 5.32 |
| MM | 0 | 100 | 0 | 13.6 | 46.4 | 40 |
| NM | 0 | 0 | 100 | 15.72 | 24.93 | 59.35 |
| *Var3* | | | | | | |
| LM | 36.57 | 56.12 | 7.31 | 38.10 | 47.96 | 13.95 |
| MM | 0 | 100 | 0 | 35.27 | 39.83 | 24.90 |
| NM | 3.60 | 2.80 | 93.60 | 10.80 | 54.80 | 34.40 |
| *Var4* | | | | | | |
| LM | 22.35 | 63.22 | 12.36 | 24.48 | 66.12 | 9.48 |
| MM | 0 | 100 | 0 | 7.00 | 56.60 | 36.40 |
| NM | 0 | 0 | 100 | 8.80 | 59.03 | 32.18 |

Jiang *et al*. (2015), page 1137) when some quantity is evaluated not necessarily under the same model as the true model. We consider that the effect of this issue dealing with the GDF-estimators included in the GAIC-measures, is minimal; first, as numbers in Table 1 show the dependence of GDF-estimators on the underlying model and, second, and mainly, because the good empirical performance of xGAIC implies that the estimator works, at least for using xGAIC as the selection model measure.

### *4.2. Simulation experiment 2*

The goal of simulation 2 is to study the performance of the method proposed as the variable selection criterion. We shall use the Fay–Herriot model with the objective of comparing our proposal with vAIC. We simulate three scenarios with one, two and three variables. In this simulation, we have considered the auxiliary information and variance parameters from the second real application in the next section (the health survey) to generate the data: scenario S1,

$$y_d = \beta_0 + \beta_1 X_1 + u_d + e_d;$$

scenario S2,

$$y_d = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_d + e_d;$$

scenario S3,

$$y_d = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u_d + e_d.$$

Here $d = 1, \ldots, D$, $D = 41$, $u_d \sim N(0, \sigma_u^2)$ and $e_d \sim N(0, \sigma_d^2)$, with $\sigma_u^2 = 0.4$ and $\sigma_d^2$ in the interval [0.02,0.64] and $X_1 = 65\text{age}$, $X_2 = \text{unemp}$ and $X_3 = \text{low}$ from the Women's Health Survey. Then, we fit the models considering one ($X_1$), two ($X_1$,$X_2$) and three variables ($X_1$, $X_2$, $X_3$) and calculate the percentage of times where the real model is chosen with xGAIC, cGAIC and vAIC.

**Table 3.** RRMSE of $\hat{\sigma}_u^2$ by using the model selected by xGAIC and cGAIC under various simulated scenarios

| *Scenario* | *Results for xGAIC* | *Results for cGAIC* |
|---|---|---|
| *Var1* | | |
| LM | 0.09 | 0.08 |
| MM | 0.12 | 0.59 |
| NM | 0.11 | 0.20 |
| *Var2* | | |
| LM | 0.12 | 0.11 |
| MM | 0.18 | 1.60 |
| NM | 0.10 | 0.91 |
| *Var3* | | |
| LM | 0.10 | 0.10 |
| MM | 0.13 | 0.57 |
| NM | 0.11 | 0.20 |
| *Var4* | | |
| LM | 0.19 | 0.16 |
| MM | 0.36 | 1.27 |
| NM | 0.15 | 0.97 |

We present the results in Table 4. It can be seen that xGAIC has much better classification rates than cGAIC and vAIC. In the case of S1, when the real model has only one variable, the correct classification rate for xGAIC is 55%, whereas cGAIC and vAIC have rates of 35.6% and 43.2% respectively. When scenario S3 is generated, the correct classification rate for xGAIC is 100%, whereas cGAIC and vAIC have rates of 73.6% and 55% respectively. Moreover, when the real model has three variables, cGAIC selects 36.4% of the times a model with fewer variables than it needs, and vAIC 44% of the times. In this case, the consequences of choosing a wrong model are more important.

## 5. Real applications

We illustrate the behaviour of the new methodology with two real applications: one of socio-economic interest and the other in the field of health. The objective in the first application is the estimation of the total number of employed people by economic activity in Galicia (a region in the north-west of Spain). Knowing the number of employed people in each of the economic activities is important for the government authorities to make decisions about economic sectors in decline or to encourage potentially emerging sectors. In the second application, the objective is to estimate the prevalence of smokers in the counties of Galicia. Tobacco consumption is still a major public health problem in Spain, where one in seven deaths in the population aged 35 years and over is attributable to it. The sociodemographic characteristics of the Galician population have significant territorial differences, which suggests that the prevalence of smokers is also different from one area to another. Knowing these differences is useful for designing specific prevention and intervention programmes to control smoking, as well as serving to evaluate its result, or to prioritize the allocation of resources.

In both cases, the direct estimators are not reliable estimators for domains with small sample sizes. In addition, the auxiliary information is available, so we propose considering model-based estimators that borrow strength across domains through regression models on the auxiliaries.

**Table 4.** Correct classification rates from using xGAIC, cGAIC and vGAIC under various simulated scenarios

| *Scenario* | *Rates (%) for 1 variable* | *Rates (%) for 2 variables* | *Rates (%) for 3 variables* |
|---|---|---|---|
| *S1* | | | |
| xGAIC | 55.0 | 25.2 | 19.8 |
| cGAIC | 35.6 | 34.2 | 30.2 |
| vAIC | 43.2 | 22.6 | 34.2 |
| *S2* | | | |
| xGAIC | 0 | 70.01 | 29.99 |
| cGAIC | 14.2 | 46.0 | 39.8 |
| vAIC | 39.8 | 34.8 | 25.7 |
| *S3* | | | |
| xGAIC | 0 | 0 | 100 |
| cGAIC | 11.6 | 14.8 | 73.6 |
| vAIC | 11.2 | 33.8 | 55.0 |

We consider, as candidate models, those defined in Section 2, which include the Fay–Herriot and additive models with linear and monotone or $P$-spline components and a random effect. In the first case, only one auxiliary is considered, so only three models are compared. In the second case, where several auxiliary variables are available, up to 15 candidate models are compared and the problem of selection between the more informative auxiliary variables is solved simultaneously with the problem of the specification of their functional form.

Different models are selected after applying xGAIC- or cGAIC-criteria in both applications and some evidence suggests that a better choice is provided, in both cases, by xGAIC. In particular, whereas the xGAIC-model selection agrees with the smallest $\hat{\sigma}_u^2$ derived from fitting the different candidate models, in both cases, cGAIC does not. $\sigma_u^2$ is an important parameter in SAE applications, because it is a measure of model accuracy, as it quantifies the variability that is not explained by the model, and it is also the main factor in deriving the small area model-based estimators.

### *5.1. Application to Labour Force Survey*

We deal with data from the LFS of Galicia in the fourth quarter of 2013. The domains in this case are economic activity and the response variable is the total number of employed people in each domain $d$, $Y_d$, which is the number of people currently employed in the activity. Denote by $P_d$ the population in economic activity $d$. Our goal is to estimate

$$Y_d = \sum_{j \in P_d} y_j,$$

where $y_j = 1$ if the $j$th person in the domain $d$ is employed, and $y_j = 0$ otherwise.

The LFS 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

$$\hat{Y}_d^{\text{dir}} = \sum_{j \in S_d} w_j y_j,$$

$$\hat{\bar{Y}}_d^{\text{dir}} = \hat{Y}_d^{\text{dir}} / \hat{N}_d^{\text{dir}},$$

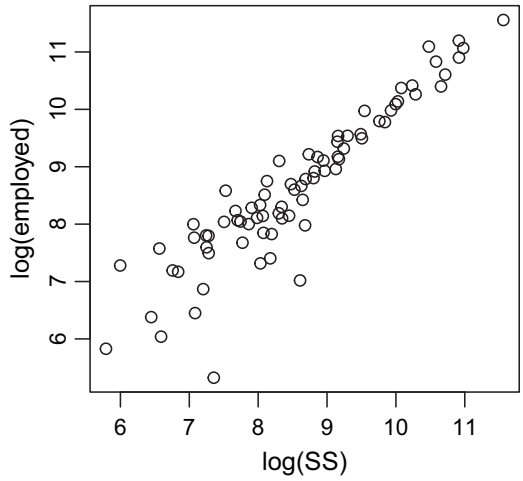


**Fig. 1.** Scatter plot between the target variable and the auxiliary variable

$$\hat{N}_d^{\mathrm{dir}} = \sum_{j \in S_d} w_j,$$

where $S_d$ is the sample domain and $w_j$s are the official calibrated sampling weights.

The LFS is designed to obtain precise estimates in the activity sector ('*Nomenclature statistique des activités économiques dans la Communauté européenne*', one digit). The problem of the LFS is that, when the domains are below the planned level, we find very small sample sizes of domains and therefore very high sampling errors. For the fourth quarter of 2013, the minimum sample size in the domains is 1, the first quartile is 12 and the median is 31; therefore, for some domains with the direct estimator, it is not possible to obtain a reliable estimate of our objective.

We consider 77 domains after discarding seven domains with a very low number of employed people. In our data set, we have four variables.

(a) cnae: this variable indicates economic activity, e.g. agriculture, forestry and the food industry.
(b) $Y$ is the direct estimator of the total number of employed people in each economic activity.
(c) SS is the number of people registered in the social security system in each economic activity.
(d) $\sigma^2_{d,\mathrm{LFS}}$ is the variance of the direct estimator of $Y$.

The models are formulated by using the log-transform of the response and auxiliary, to fit the normality error assumption better. In Fig. 1, we present the scatter plot of the two variables, log(SS) against $\log(Y_d) = \log(\text{employed})$, by economic activity. The relationship between the auxiliary variable and the response variable is apparently monotone, and almost linear.

We consider three additive mixed regression models, corresponding to those defined in Section 2 for only one auxiliary. Table 5 shows the main statistics that are related to the calculation of GAICs. From numbers in Table 5, we see that xGAIC selected a Fay–Herriot model whereas cGAIC selected a monotone model, being $\hat{\sigma}^2_u$ higher in the latter. In this case, where the assumption of linearity is fair, the more simple Fay–Herriot model seems a better choice, since the differences between the models are so small, in view of the similarity between the $\widehat{\mathrm{GDF}}$-values; all three values were very close to $D = 77$.

**Table 5.** $\widehat{\text{cGDF}}$ and cGAIC versus $\widehat{\text{xGDF}}$ and xGAIC for the Fay–Herriot, monotone and *P*-spline models and the corresponding values of $\hat{\sigma}_u^2$

| *Model* | $\widehat{cGDF}$ | $cGAIC$ | $\widehat{xGDF}$ | $xGAIC$ | $\hat{\sigma}_u^2$ |
|---|---|---|---|---|---|
| FH | 74.8 | −296.2 | 74.5 | 99.8 | 0.21 |
| Monotone | 76.0 | −297.5 | 75.8 | 109.0 | 0.24 |
| *P*-spline | 73.9 | −296.1 | 74.7 | 100.1 | 0.21 |

### *5.2. Application to health data*

In this application, we deal with data from the surveys from the behavioural risk factors information system in Galicia in 2010–2011. The sample design of this survey is stratified random sampling with equal allocation by sex and age group. Our domains of interest are 41 areas obtained from the 53 counties of Galicia. Our goal is to estimate the prevalence of smokers by sex among the population aged 16 years and over, in the 41 areas of Galicia in the period 2010–2011.

This survey is designed to obtain precise estimates at province level. The problem is to obtain reliable estimates for domains below the planned level because of small sample sizes. For 2010–2011, the minimum sample size in the domains for men is 44, the first quartile is 69 and the median 93.

We use the logarithm of smokers as the response, $\log(\hat{Y}^{\text{dir}})$, where $\hat{Y}^{\text{dir}}$ is the direct estimator obtained from the survey. In our data set we also have the following auxiliary variables:

(a) age, the percentage of the population under 15 years, 15age, from 15 to 24, years, 15–24, from 25 to 44 years, 25–44, from 45 to 64 years, 45–64, and 65 years and over, 65age;
(b) degree of urbanization, the percentage of the population that live in a densely populated area, zdp, an intermediate area, zip, and a thinly populated area, zpp;
(c) activity, the proportion of employed, emp, unemployed, unemp, and inactive people, inac;
(d) education level, the proportion of people with low education, low, secondary education, sec, and higher education, higher.educ.

As several studies have revealed, some sociodemographic characteristics of the population, such as sex, age, level of education, employment status or degree of urbanization, are associated with tobacco consumption (Li *et al.*, 2009; Srebotnjak *et al.*, 2010), and these characteristics vary from one region to another. In Galicia, 28% of men and 18% of women aged 16 years and over currently smoke, according to data from the survey of behavioural risk factors in 2011. In this study, we separate men and women, as in the study of Srebotnjak *et al.* (2010), because the behaviour and pattern of tobacco consumption are very different between the sexes. In Fig. 2, we present the prevalence of smokers and some summaries of the main variables used in this study by sex. It can be seen that the sociodemographic characteristics for men and women are quite different.

Table 6 shows the correlations between the auxiliary variables and the response variable. From Table 6 it can be seen that low, higher.educ, emp and sec are the most correlated variables for men and zdp, zpp, 15age and 65age for women.

In addition non-significant auxiliary variables were discarded after fitting a Fay–Herriot model, with the exception of emp and unemp selected by the experts. Fig. 3 shows the scatter plots from the four auxiliary variables selected against the response variable (smokers by area)

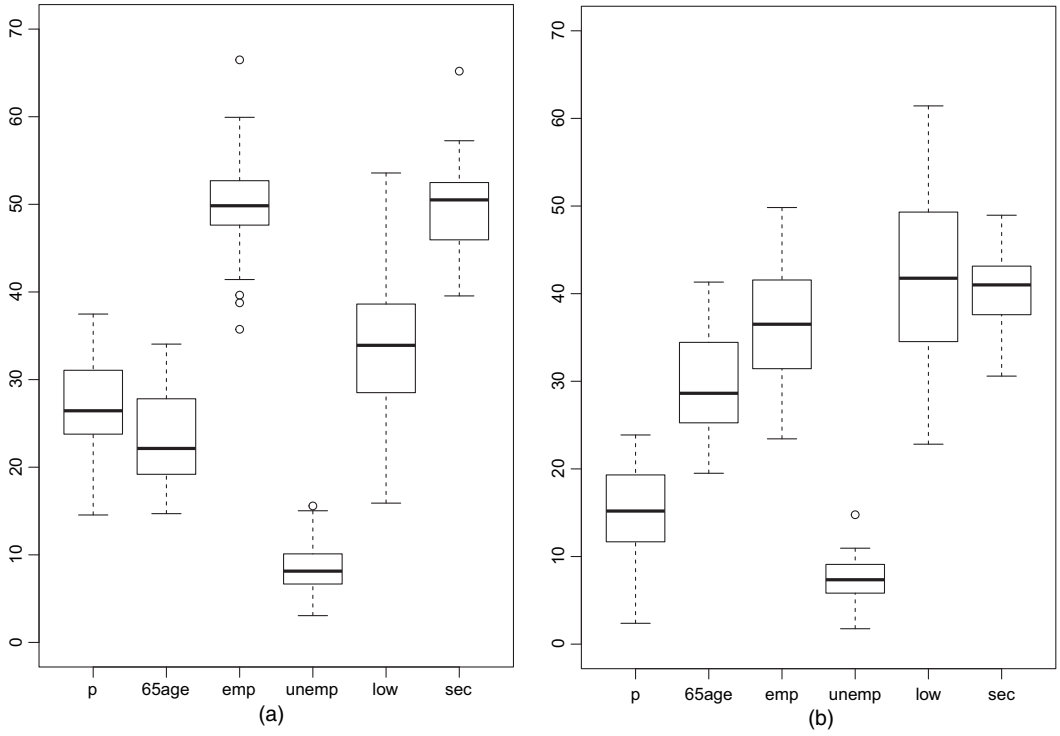


**Fig. 2.** Boxplots of the main variables and the prevalence of smokers $p$ for (a) men and (b) women

**Table 6.** Correlations between the predictors and the response

| *Predictor* | | *Results for men* | *Results for women* |
|---|---|---|---|
| zdp | $X_1$ | 0.27 | 0.56 |
| zip | $X_2$ | −0.14 | −0.19 |
| zpp | $X_3$ | −0.21 | −0.54 |
| 15age | $X_4$ | 0.22 | 0.54 |
| 15–24 | $X_5$ | 0.10 | 0.33 |
| 25–44 | $X_6$ | 0.26 | 0.50 |
| 45–64 | $X_7$ | 0.02 | 0.19 |
| 65age | $X_8$ | −0.27 | −0.52 |
| emp | $X_9$ | 0.29 | 0.17 |
| unemp | $X_{10}$ | −0.08 | 0.43 |
| inactive | $X_{11}$ | −0.22 | −0.32 |
| low | $X_{12}$ | −0.38 | −0.40 |
| sec | $X_{13}$ | 0.28 | 0.31 |
| higher.educ | $X_{14}$ | 0.29 | 0.36 |

for men. The assumption of a monotone relationship between auxiliary and response is sensible in all four cases.

Up to 15 additive mixed models, defined from two, three or four predictors, have been fitted to the data set. The models are formulated as those in Section 2 and differ from each other in which predictors are modelled as linear, monotone or $P$-spline. We have included the relevant

**Table 7.** Models fitted to the men's data: $\widehat{\text{GDF}}$ and GAIC conditional and mixed values, and $\hat{\sigma}_u^2$

| *Model label* | *Linear predictors* | *Monotone predictors* | *P-spline predictors* | $\widehat{cGDF}$ | *cGAIC* | $\widehat{xGDF}$ | *xGAIC* | $\hat{\sigma}_u^2$ |
|---|---|---|---|---|---|---|---|---|
| M1 | $X_{12}, X_8$ | | | 37.2 | −18.5 | 37.1 | 79.5 | 0.40 |
| M2 | $X_{12}$ | $X_8$ | | 40.9 | −14.7 | 41.1 | 78.8 | 0.35 |
| M3 | $X_{12}$ | | $X_8$ | 36.5 | −18.6 | 36.4 | 75.9 | 0.37 |
| M4 | $X_{12}, X_8, X_{13}$ | | | 37.1 | −18.4 | 37.4 | 77.9 | 0.38 |
| M5 | $X_{12}, X_{13}$ | $X_8$ | | 41.0 | −14.4 | 40.8 | 78.5 | 0.35 |
| M6 | $X_{12}, X_{13}$ | | $X_8$ | 36.7 | −18.4 | 35.4 | 71.7 | 0.34 |
| M7 | $X_{12}$ | $X_8, X_{13}$ | | 40.9 | −14.9 | 40.8 | 67.2 | 0.26 |
| M8 | $X_{12}$ | | $X_8, X_{13}$ | 36.2 | −17.8 | 34.2 | 70.6 | 0.30 |
| M9 | $X_{12}, X_8, X_{13}, X_9$ | | | 37.4 | −18.1 | 36.7 | 76.1 | 0.38 |
| M10 | $X_{12}, X_{13}, X_9$ | $X_8$ | | 41.0 | −12.9 | 40.4 | 69.9 | 0.28 |
| M11 | $X_{12}, X_{13}, X_9$ | | $X_8$ | 36.9 | −17.5 | 34.6 | 68.9 | 0.31 |
| M12 | $X_{12}, X_{13}$ | $X_8, X_9$ | | 41.0 | −14.1 | 40.8 | 78.4 | 0.34 |
| M13 | $X_{12}, X_{13}$ | | $X_8, X_9$ | 36.6 | −20.0 | 35.6 | 68.5 | 0.31 |
| M14 | $X_{12}$ | $X_8, X_9, X_{13}$ | | 40.8 | −13.0 | 40.4 | 64.9 | 0.24 |
| M15 | $X_{12}$ | | $X_8, X_9, X_{13}$ | 36.2 | −17.3 | 34.0 | 68.5 | 0.26 |

**Table 8.** Models fitted to the women's data: $\widehat{\text{GDF}}$ and GAIC conditional and mixed values, and $\hat{\sigma}_u^2$

| *Model label* | *Linear predictors* | *Monotone predictors* | *P-spline predictors* | $\widehat{cGDF}$ | *cGAIC* | $\widehat{xGDF}$ | *xGAIC* | $\hat{\sigma}_u^2$ |
|---|---|---|---|---|---|---|---|---|
| W1 | $X_{12}, X_8$ | | | 34.7 | 10.5 | 35.2 | 89.5 | 0.48 |
| W2 | $X_{12}$ | $X_8$ | | 40.2 | 15.4 | 40.2 | 83.5 | 0.35 |
| W3 | $X_{12}$ | | $X_8$ | 32.9 | 10.5 | 31.5 | 74.7 | 0.31 |
| W4 | $X_{12}, X_8, X_{13}$ | | | 34.4 | 10.6 | 34.3 | 86.1 | 0.46 |
| W5 | $X_{12}, X_{13}$ | $X_8$ | | 40.0 | 15.7 | 39.8 | 82.5 | 0.35 |
| W6 | $X_{12}, X_{13}$ | | $X_8$ | 32.3 | 10.6 | 30.8 | 75.7 | 0.29 |
| W7 | $X_{12}$ | $X_8, X_{13}$ | | 39.8 | 15.6 | 40.0 | 78.7 | 0.30 |
| W8 | $X_{12}$ | | $X_8, X_{13}$ | 31.8 | 10.0 | 29.8 | 70.5 | 0.29 |
| W9 | $X_{12}, X_8, X_{13}, X_{10}$ | | | 33.7 | 9.4 | 34.7 | 83.0 | 0.40 |
| W10 | $X_{12}, X_{13}, X_9$ | $X_8$ | | 39.9 | 15.8 | 40.3 | 79.5 | 0.30 |
| W11 | $X_{12}, X_{13}, X_9$ | | $X_8$ | 32.1 | 10.1 | 30.9 | 69.2 | 0.27 |
| W12 | $X_{12}, X_{13}$ | $X_8, X_{10}$ | | 39.4 | 15.9 | 39.1 | 70.5 | 0.23 |
| W13 | $X_{12}, X_{13}$ | | $X_8, X_{10}$ | 31.1 | 10.2 | 31.1 | 66.5 | 0.22 |
| W14 | $X_{12}$ | $X_8, X_{10}, X_{13}$ | | 39.4 | 15.4 | 38.8 | 54.5 | 0.12 |
| W15 | $X_{12}$ | | $X_8, X_{10}, X_{13}$ | 30.4 | 9.5 | 27.9 | 66.2 | 0.22 |

statistics for the fitted models in Tables 7 and 8. In Tables 7 and 8, models are labelled as M1–M15 for men's data and W1–W15 for women's data and indicate which predictors are included as linear terms and which are modelled by using monotone or *P*-spline assumptions.

From the numbers in Tables 7 and 8, we see that xGAIC selects similar models in men and women: model M14 for men and W14 for women. Models M14 and W14 are defined by using monotone assumptions for three auxiliaries and one linearity assumption for the fourth, as well as providing the smallest $\hat{\sigma}_u^2$-values in men and women.

In contrast, cGAIC selects model M13 in the case of men and model W9 in the case of women. The first is a model with two linear terms and two *P*-splines terms, whereas the latter is the Fay–Herriot model defined by using the four predictors.

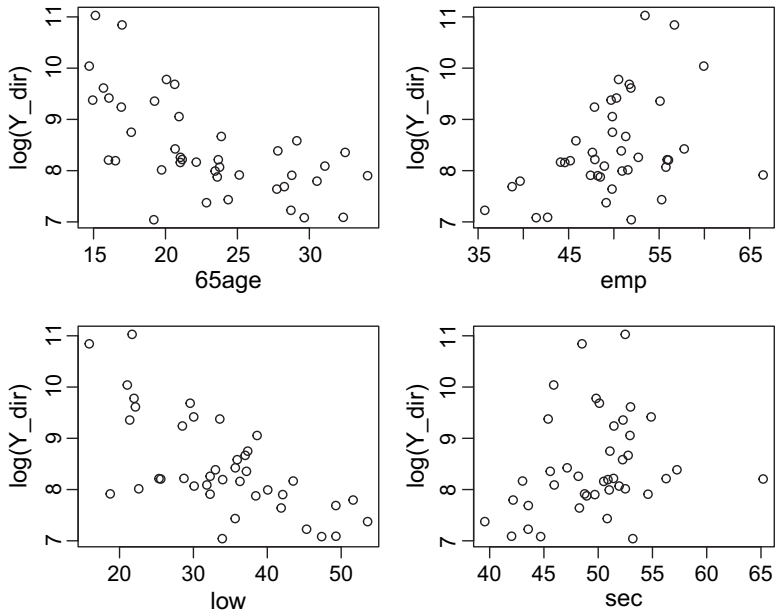


**Fig. 3.** Relationship between the auxiliary variables and the response variable ($\log(Y_{\text{dir}})$) in men

In this case, unlike the first application, the linearity assumptions do not seem to be sensible, at least for the four predictors (Fig. 3) which, jointly with the coincidence with the smallest $\hat{\sigma}_u^2$, evidence that xGAIC performs well, and better than cGAIC.

## 6. Conclusions

A mixed Akaike information measure xGAIC has been defined that is an important novelty to be incorporated in the debate on the use of marginal or conditional approaches. xGAIC is a compromise solution derived from a quasi-log-likelihood and an empirical estimator of GDF.

There is a broad range of other settings involving models with random effects, outside small area problems, where versions of the xGAIC-criteria, defined as suggested in this paper, using a mixed approach, could be investigated. We have shown that xGAIC is easily derived for complex models and has very good behaviour in SAE applications. These properties induce optimism about the suitability of the new criterion in other settings.

We have shown, using simulation studies and two real cases, that the mixed approach outperforms the conditional approach, which had been, until now, the most recommended approach in the SAE literature. This conclusion comes from simulated realistic scenarios in which a linear relationship of the predictors with the response cannot, or should not, be assumed, and from considering parametric and non-parametric model formulations.

The simulations have shown that xGAIC performs remarkably better than cGAIC or vAIC for selecting the functional form of the fixed part of the model, and also as a variable selection criterion. This assertion is supported by a rather smaller classification error rate, when the real model is not linear, but also by a smaller RRMSE of the random-effect variance parameter. When a linear model is simulated, the success rate is slightly higher by using cGAIC. Nevertheless, the model selected by xGAIC is also valid and provides inferences with the same efficiency, or with a very small loss, with respect to the model that is selected by cGAIC.

The conclusions from the analysis of the two real cases are more interesting. The socio-economic case is an example in which only one predictor is considered and the linearity assumption is correct. According to simulations, the differences between the GAIC-values from different candidate models are very small. Surprisingly, cGAIC selects a monotone model with a higher $\hat{\sigma}_u$ than xGAIC, which selects a linear model.

In the second case, several predictors, which can hardly be assumed to have a linear relationship with the response, are considered. The differences between the xGAIC and cGAIC model selection are bigger in this case. Note that the values of $\hat{\sigma}_u$ that are obtained from cGAIC-models are bigger than the values of $\hat{\sigma}_u$ from the xGAIC-models.

One should take into account the fact that the two real cases that were analysed correspond to small random-effect variability scenarios and, in both cases, the model-based estimators are close to the direct estimators.

We should point out, and show the simulations as support, that the consequences of using a conditional GAIC to select the model can be serious because standard approaches to make inferences on small areas are very dependent on the model assumptions and these assumptions are wrong when GAIC incorrectly selects a simpler model than the true model.

In fact, this question of dependence on the model assumptions has been treated lightly in the SAE literature. The common practice of comparing estimators by comparing the estimators of the mean-squared errors is not the most desirable practice, as the mean-squared error estimator may be model dependent. Occasionally, in practice, estimators are proposed that are not better than the direct estimators because an incorrect model is being used. More often, better solutions than those provided by using a given model, usually linear, could be provided by carrying out a model selection check and by trying different (including non-parametric) models.

In SAE applications, model selection means estimator selection. A clear advantage of using AIC-measures for this purpose, over the use of mean-squared error estimation, is that the selection is less dependent on the model assumed. Besides, the GDF-estimator is interpretable as an absolute inverse *distance* measure between the model-based estimator and the direct estimator. More explicitly, in a problem with $p$ predictors, $D$ areas ($D > p$) and the choice of a random effect, GDF would range, from an approximated maximum value $D$, which corresponds to a case where the mixed estimators are equal to the direct estimators, to an approximated minimum $p$ corresponding to a model with linear fixed effects, and no random effects, which provides synthetic estimators that are at a maximum *distance* from the direct estimator.

A step further on the subject of model selection in SAE applications would be to focus on the estimation of the area parameters instead of model selection and, in particular, to explore the behaviour of different AIC-measures to solve the problem of choosing between mixed and synthetic estimation approaches, at the same time as the selection of predictors and the functional form. This will be part of our future work.

## Acknowledgements

This work was supported by the Galician Institute of Statistics, by the Galician Public Health Authority ('Xunta de Galicia'), by Ministeria de Economía, Industria y Competitividad grants 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 European Regional Development Fund. Thanks go to Professor D. Morales for advice on some technical details.

## Appendix A: Maximum likelihood estimators under Fay–Herriot model

For the Fay–Herriot model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

with $\mathbf{u} \sim N(0, \boldsymbol{\Sigma}_u = \sigma_u^2 \mathbf{I}_D)$ and $\mathbf{e} \sim N(0, \boldsymbol{\Sigma}_e)$ independent distributions. The ML estimators are obtained as follows. The log-likelihood is

$$l(\boldsymbol{\beta}, \sigma_u^2; \mathbf{y}) = -\tfrac{1}{2} D \log(2\pi) - \tfrac{1}{2} \log|\mathbf{V}| - \tfrac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

The first-order derivatives of $l$ are

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta},$$

$$\frac{\partial l}{\partial \sigma_u^2} = -\frac{1}{2}\frac{\partial \log|\mathbf{V}|}{\partial \sigma_u^2} - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \frac{\partial \mathbf{V}^{-1}}{\partial \sigma_u^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Applying the differentiation formulae:

$$\frac{\partial V}{\partial \sigma_u^2} = \mathbf{Z}'\mathbf{Z},$$

$$\frac{\partial \log|V|}{\partial \sigma_u^2} = \operatorname{tr}\left( V^{-1} \frac{\partial V}{\partial \sigma_u^2} \right) = \operatorname{tr}(V^{-1}\mathbf{Z}\mathbf{Z}'),$$

$$\frac{\partial V^{-1}}{\partial \sigma_u^2} = -V^{-1}\frac{\partial V}{\partial \sigma_u^2} V^{-1} = -V^{-1}\mathbf{Z}\mathbf{Z}'V^{-1}.$$

We obtain that

$$\frac{\partial l}{\partial \sigma_u^2} = -\tfrac{1}{2}\operatorname{tr}(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') + \tfrac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

We define

$$\mathbf{T}_u = (\mathbf{Z}'\boldsymbol{\Sigma}_e^{-1}\mathbf{Z} + \sigma_u^{-2}\mathbf{I}_D)^{-1}.$$

Applying the inversion formula $(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$, we obtain

$$\mathbf{T}_u = \sigma_u^2 \mathbf{I}_D - \sigma_u^4 \mathbf{Z}' V^{-1} \mathbf{Z}.$$

We can write

$$\operatorname{tr}(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') = \operatorname{tr}(\mathbf{Z}\mathbf{V}^{-1}\mathbf{Z}') = \frac{1}{\sigma_u^4}\operatorname{tr}(\sigma_u^4 \mathbf{Z}' V^{-1}\mathbf{Z}).$$

Adding and subtracting $\sigma_u^2 \mathbf{I}_D$ in the above expression, we obtain

$$\operatorname{tr}(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') = \frac{1}{\sigma_u^4}\operatorname{tr}(\sigma_u^2 \mathbf{I}_D - T) = \frac{1}{\sigma_u^2}\left\{ D - \frac{1}{\sigma_u^2}\operatorname{tr}(T_u) \right\}.$$

Then, the likelihood equations are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}$$

and

$$\frac{1}{\hat{\sigma}_u^2}\left\{ D - \frac{1}{\hat{\sigma}_u^2}\operatorname{tr}(\hat{T}_u) \right\} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}^{-1}\mathbf{Z}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Then, multiplying and dividing by $\hat{\sigma}_u^4$ in the above expression and defining

$$\hat{\mathbf{u}} = \hat{\boldsymbol{\Sigma}}_u \mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

the second ML update equation of $\sigma_u^2$ is

$$\frac{1}{\hat{\sigma}_u^2}\left\{D-\frac{1}{\hat{\sigma}_u^2}\mathrm{tr}(\hat{T}_u)\right\}=\frac{1}{\hat{\sigma}_u^2}\hat{\mathbf{u}}'\hat{\mathbf{u}},$$

or

$$\hat{\sigma}_u^2=\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{D-(1/\hat{\sigma}_u^2)\mathrm{tr}(\hat{\mathbf{T}}_u)}$$

where $\hat{\mathbf{T}}_u$ is the empirical version of $\mathbf{T}_u$.

## References


Akaike, H. (1973) Information theory and the maximum likelihood principle. In *Proc. Int. Symp. Information Theory* (eds B. N. Petrov and F. Csàki), pp. 267–281. Budapest: Akademiai Kiado.

Fay, R. and Herriot, R. (1979) Estimates of income for small places: an application of James–Stein procedures to census data. *J. Am. Statist. Ass.*, **70**, 311–319.

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

Greven, S. and Kneib, T. (2010) On the behaviour of marginal and conditional Akaike information criteria in linear mixed models. *Biometrika*, **97**, 773–789.

Han, B. (2013) Conditional Akaike information criterion in the Fay–Herriot model. *Statist. Methodol.*, **11**, 53–67.

Hansen, N. and Sokol, A. (2014) Degrees of freedom for nonlinear least squares estimation. *Preprint arXiv:1402.2997*.

Hodges, J. and Sargent, D. (2001) Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, **88**, 367–379.

Jiang, J. (1996) REML estimation: asymptotic behavior and related topics. *Ann. Statist.*, **24**, 255–286.

Jiang, J., Nguyen, T. and Rao, J. S. (2010) Fence method for nonparametric small area estimation. *Surv. Methodol.*, **36**, 3–11.

Jiang, J., Nguyen, T. and Rao, J. (2015) The e-ms algorithm: model selection with incomplete data. *J. Am. Statist. Ass.*, **110**, 1136–1147.

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

Li, W., Land, T., Keithly, L. and Kelsey, J. (2009) Small-area estimation and prioritizing communities for tobacco control efforts in Massachusetts. *Am. J. Publ. Hlth*, **99**, 470–479.

Marhuenda, Y., Morales, D. and Pardo, M. (2014) Information criteria for Fay–Herriot model selection. *Computnl Statist. Data Anal.*, **70**, 268–280.

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

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

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008) Non-parametric small area estimation using penalized spline regression. *J. R. Statist. Soc.* B, **70**, 265–286.

Overholser, R. and Xu, R. (2014) Effective degrees of freedom and its application to conditional AIC for linear mixed-effects models with correlated error structures. *J. Multiv. Anal.*, **132**, 160–170.

Pfeffermann, D. (2013) New important developments in small area estimation. *Statist. Sci.*, **28**, 40–68.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2016) nlme: linear and nonlinear mixed effects models. *R Package Version 3.1-127*. (Available from `http://CRAN.R-project.org/package=nlme`.)

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

R Core Team (2015) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

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

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

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

Rueda, C., Menéndez, J. A. and Gómez, F. (2010) Small area estimators based on restricted mixed models.. *Test*, **19**, 558–579.

Salvati, N., Chandra, H., Ranalli, M. G. and Chambers, R. (2010) Small area estimation using a nonparametric model-based direct estimator. *Computnl Statist. Data Anal.*, **54**, 2159–2171.

Shen, X. and Huang, H.-C. (2006) Optimal model assessment, selection, and combination. *J. Am. Statist. Ass.*, **101**, 554–568.

Sperlich, S. and Lombardía, M. (2010) Local polynomial inference for small area statistics: estimation, validation and prediction. *J. Nonparam. Statist.*, **22**, 633–648.

Srebotnjak, T., Mokdad, A. and Murray, C. J. (2010) A novel framework for validating and applying standardized small area measurement strategies. *Hlth Metr.*, **29**, 8–26.
Tibshirani, R. J. and Taylor, J. (2012) Degrees of freedom in lasso problems. *Ann. Statist.*, **40**, 1198–1232.
Torabi, M. and Shokoohi, F. (2015) Non-parametric generalized linear mixed models in small area estimation. *Can. J. Statist.*, **43**, 82–96.
Ugarte, M., Goicoa, T., Militino, A. and Durbán, M. (2009) Spline smoothing in small area trend estimation and forecasting. *Computnl Statist. Data Anal.*, **53**, 3616–3629.
Vaida, F. and Blanchard, S. (2005) Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.
Wood, S. (2006) *Generalized Additive Models: an Introduction with R*. Boca Raton: Chapman and Hall–CRC.
Ye, J. (1998) On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Ass.*, **93**, 120–131.
You, C., Muller, S. and Ormerod, J. (2016) On generalized degrees of freedom with application in linear mixed models selection. *Statist. Comput.*, **26**, 199–210.
Yu, D. and Yau, K. (2012) Conditional Akaike information criterion for generalized linear mixed models. *Computnl Statist. Data Anal.*, **56**, 629–644.
Zhang, B., Shenb, X. and Mumford, S. (2012) Generalized degrees of freedom and adaptive model selection in linear mixed-effects models. *Computnl Statist. Data Anal.*, **56**, 574–586.