This version of the article has been accepted for publication, after peer review and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/s11749-020-00712-4

Selection model for domains across time. Application to Labour Force Survey by economic activities. *

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

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

Abstract

This paper introduces a small area estimation approach that borrows strength across domains (areas) and time and is efficiently used to obtain Labour Force Estimators by economic activity. Specifically, the data across time is used to select dfferent models for each domain; such selection is done with an Aggregated Mixed Generalized Akaike Information Criterion statistic which is obtained using data across all time points and then is splitted into individual component for each domain. The approach makes a selection from different estimators, including the direct estimator, synthetic and mixed estimators derived from different models using auxiliary information. Results from several simulation experiments, some with original designs, show the good performance of the approach against standard small area approaches. In addition, it is shown the important practical advantages in the real application.

Key words and phrases: Akaike Information Criterion, Bootstrap, Fay-Herriot model, Generalized Degree of Freedom, monotone model, small area estimation, spline regression.

1 Introduction

Government authorities dealing with the economy of a region need to know the number of employed people in each of the economic activities and their evolution over time. This is necessary to influence economic sectors in decline or to encourage potentially emerging sectors. It is also important to study the productivity of the different activities and thus promote employment policies in those economically more productive sectors. ln the Regional Accounts, labour is important for accounts compilation because employment (the number of employees or hours worked) or compensation of employees are often used as regional indicators for the allocation of regional gross added value to regions. From a statistical point of view, on the Regional Economic accounts of Galicia (Spain), the estimates of employment in each of the activities are done through a study of the different statistical sources, taking at the end the most coherent estimate

*Supported by the MINECO grants MTM2017-82724- R, MTM2015-71217-R, 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.

†email: maria.jose.lombardia@udc.es

in terms of evolution, profit ratios, etc. Eurostat (Eurostat (2013)) recommends that a social statistical database, in which surveys and administrative data are combined and reconciled, might provide the necessary consistency of labour data. Eurostat also says that administrative data is potentially a very important source when combined and reconciled with other sources, for both national and regional accounts. With this work, we go a step further on the Eurostat line to find the most appropriate model to determine the evolution of employed people by economic activity, combining administrative registers and survey data.

The main aim of this work is to estimate, in each quarter, the employed people by economic activity in Galicia in the period from the third quarter of 2009 to the first quarter of 2016, using data from the Labour Force Survey (LFS). We consider as domains the economic activities (Agriculture, Forestry, Manufacture of metals, Insurance, etc.), defined by the Statistical classification of economic activities in the European Community (NACE Rev.2) (Eurostat (2008)) at the division level. There is a total of 84 domains for each 27 quarters from 2009 to 2016.

The problem in the LFS case is that the sampling has been designed to estimate parameters associated with planned (big) domains and this implies very low sample sizes and inaccurate estimators for some small domains. The goal of the Small Area Estimation (SAE) techniques is to derive more accurate estimators for each domain (or area) in that scenario, by borrowing strength from other domains. Some of the most relevant references in SAE are the monographs 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).

When the sample information is available in more than one dimension, such as space or time, the small area estimators can also be derived by borrowing strength from one or more of these additional dimensions. Should data be available from several time periods, graphical displays of the domain estimators across time are very interesting outputs for practitioners. A smooth pattern is desirable in most applications, as domain estimators are expected to change slowly from one time point to the next. However, it may happen that the trend pattern is more abrupt than desired for direct estimators due to the sampling variability.

This is the LFS case, where very different temporal patterns, some of them quite abrupt, are exhibited depending on the economic activity. This is illustrated in Figure 1, where data from log observed rates (Y) and the log total number of people registered in the social security system (X), in consecutive quarters, is displayed for three economic activities: Forestry, Manufacture of metals and Employment activities. In addition, Figure 1 also shows an unequal behaviour of employment, depending on the economic activities, regarding the sample variability and the relationship with the auxiliary information, in addition to the temporal evolution. The simple strategy of using standard model-based estimators does not solve the problem; again, abrupt temporal patterns are exhibited, as can be seen in Section 4, where we study the real application in detail. Moreover, the option to consider time series models (models that borrow strength across time instead of across domains) may generate patterns which are too smooth and prevent major trend changes from being detected.

In this paper we propose, for these heterogeneity contexts, a simple and useful strategy that consists of selecting specific estimators for each domain, borrowing strength across time as well as across domains. This strategy lacks the drawback of time series models, taking into account the heterogeneous domains and providing less abrupt temporary patterns in many domains.

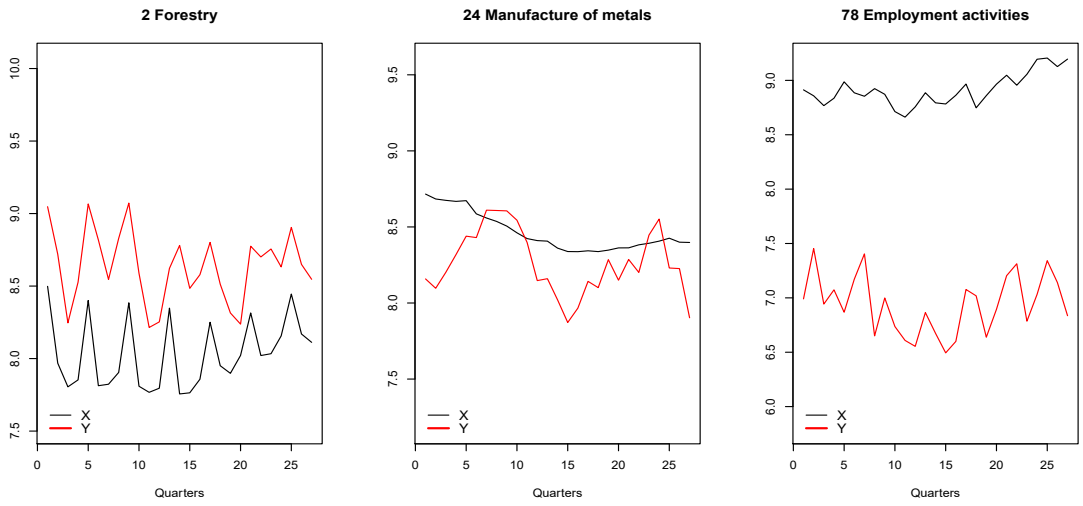

Figure 1: Patterns across time.

The use of AIC statistics is one of the most popular approaches to model selection (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 selected model within a set of candidates is the one with the lowest AIC. The penalty term plays an important role in statistical modeling. Different terms have been used for different models and authors: the number of parameters in the model, degrees of freedom, divergence, effective degrees of freedom, generalized degrees of freedom. Most of these terms are the same for simple models, such as the normal linear regression model, but not for complex models, such as the constrained, lasso or mixed models (Kato (2009), Rueda (2013), Tibshirani and Taylor (2012)). An important contribution to the subject is contained in Ye (1998), where the Generalized Degrees of Freedom (GDF) is defined as a measure of the sensitivity of each fitted value, $\hat{m}_{dt}$, to the perturbation in the corresponding observed value Y_{dt}. Since then, different versions of Generalized Akaike Information Criteria ($GAICs$), with either conditional or marginal log-likelihoods and different estimators of the GDF originally defined in Ye (1998), have been considered in the literature. For mixed models, some references of interest are Vaida and Blanchard (2005), Muller et al. (2013) and You et al. (2016), among others. In the particular problem of SAE, the problem has been only tentatively studied. We highlight the review of new important developments in Pfeffermann (2013) and the contributions of Han (2013), Marhuenda et al. (2014) and Lombardía et al. (2017).

In this work, we consider the Mixed Generalized Akaike Information Criterion ($xGAIC$) introduced by Lombardía et al. (2017) to select the model-based estimator for each particular domain.

Moreover, we take a step ahead on the subject, when data across a second dimension (time in our application) are also provided. We propose a specific domain measure by aggregating the individual components for the domain across time, using the property that AIC measures, including $xGAIC$, can be formulated as a sum of the individual components from each of the d domains, considering the individual log-likelihood component plus the $1/D$ fraction of the penalty term.

We show, using two simulation experiments, that the new approach for selecting a specific domain estimator outperforms the usual practice of considering the same model for deriving estimates for all the domains simultaneously. In addition, the approach is satisfactorily used to derive Labour Force estimates by economic activity and quarter in Galicia. In particular, temporal patterns neither too smooth nor too abrupt are exhibited, which provides a very interesting solution for decision makers. In the SAE literature there are other models that consider time and space as for example Rao and Yu (1994), which we initially considered but discarded for not improving the results, as discussed in Section 5.1.

The original design of the simulation experiments is also an interesting contribution of this paper. A generator model, which does not match any working model, is defined. It is shown that the generator model manages to replicate the real data in a faithful way.

This issue is particularly relevant in our problem because the sample data illustrate a complex generator process, as the data can only be replicated using different models across domains. To deal with the bias-variance tradeoff, statisticians often consider a working model simpler than the generator model. However, in simulations, when the questions are to select from a set of candidate models and validate the selection approach, the data must be generated with models that approximate to the real one as much as possible in order to achieve reliable comparisons.

We organize the remainder of the paper as follows. In Section 2, we describe the candidate models with and without random effects and the corresponding model-based estimators. In Section 3, we introduce a new $GAIC$ statistics to choose the appropriate model for heterogeneous data using information across time. Sections 4 and 5 describe numerical studies, in the former section, Labour Force estimates by economic activity and quarter in Galicia are obtained using the new approach; in the latter, a complete numerical study is designed, where different scenarios are considered, including one that imitates the real case. Section 6 is devoted to the conclusions. Finally, in the Appendix, we give the Labour Force estimates for each economic activity defined by NACE (division level).

2 Model-based estimators

To model the LFS data, and similar applications, we assume a response vector $\mathbf{Y}_t = (y_{1t}, \ldots, y_{Dt})'$, $d = 1, ...D,\ t = 1, ...T$; where t is the quarter, D is the number of domains of study, T is the number of time periods, and $\mathbf{Y}_t \sim N(\mathbf{m}_t, \mathbf{V}_t)$. We also use $\mathbf{Y}_d = (y_{d1}, \ldots, y_{dT})'$ for the response vector of domain d that is is the logarithm of the total of employed people by economic activity and time. Here, $\mathbf{m}_d = (m_{d1}, \ldots, m_{dT})'$ and $\mathbf{m}_t = (m_{1t}, \ldots, m_{Dt})'$. We consider a vector of covariates $X_{dt}, d = 1, ...D; t = 1, ...T$.

In this paper, we distinguish between generator and working models for $\mathbf{m}$. It is assumed that a real and unknown model exists, from which $\mathbf{m}$ is derived. Otherwise, working models are those from which the estimators for $\mathbf{m}$ are derived. Then, a working model, M, is defined as a mapping from $\mathbf{R}^D$ to $\mathbf{R}^D$, which produces a set of fitted values $\hat{\mathbf{m}}_t = (\hat{m}_{1t}, \ldots, \hat{m}_{Dt})'$ from $\mathbf{Y}_t$. To simplify the notation, the dependence of parameters estimators on M is assumed and will be implicit. The candidate set of working models is defined by $\aleph$. The models in $\aleph$ considered in this paper are additive models that differ in the functional form relating each covariate with the response, and in the inclusion or not of random effects. In particular, we consider a set with six members, as functional forms we take the linear, monotone and spline function, and may be synthetic (derived from a model without random effects) or mixed (derived from a model with

random effects).

Below is shown the candidate models in $\aleph$ and the corresponding estimators for m_{dt}. To simplify the notation the models are introduced for a single time period, however, the extension to the various time periods is simple as the models are fitted independently at each time point. Moreover, models use auxiliary information in terms of explanatory variables; however, as in the LFS case, only one explanatory variable is used. We present the models for this simple case as the extension to the case of more auxiliaries is straightforward.

As will be seen in Section 4, this family of models covers a wide enough range of functional relationships for the application at hand. However, the list can be expanded with other candidates, or even reduced, depending on the application and the researcher. In particular, when more explanatory variables are used, the number of members in $\aleph$ is expanded by combining the different types of models with a relevant combination of explanatory variables.

The models can be defined as follows:

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

where $\theta_d = f(x_d)$; u_d are independent and identically distributed as $N(0, \sigma_u^2)$ with σ_u^2 unknown and possibly zero, and $e_d \sim N(0, \sigma_d^2)$, with σ_d^2 assumed known.

The domain estimators are also defined as $\hat{m}_d = \hat{\theta}_d + \hat{u}_d$, where $\hat{\theta}_d$ is the estimator of the fixed effect θ and $\hat{u}_d$ the predictor of the random effect u_d.

Depending on the functional form $f()$ and whether σ_u^2 is zero or not, the result is different small area models from which the corresponding model-based estimators are derived.

Next, we define the estimators derived from the six models in $\aleph$, starting with the models for f as the linear function, which give the following estimators:

- Synthetic linear (SL): $\sigma_u = 0$,

$$\hat{m}_d = \widehat{\theta}_d = x_d\hat{\beta}, \quad d = 1, \ldots, D;$$

  $\hat{\boldsymbol{\beta}}$ is the standard linear regression vector of coefficient estimators and x_d is the explanatory variable in the domain d.

- Mixed linear (ML), : $\sigma_u \geq 0$,

$$\hat{m}_d = x_d\hat{\beta} + \hat{u}_d, \quad d = 1, \ldots, D.$$

$\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are derived using Maximum Likelihood, for details see the monograph Rao (2003) and Rao and Molina (2015). This is one of the most famous models in SAE, due to Fay and Herriot (1979). There is a lot of literature about it.

The next two estimators are derived from monotone models. Monotonicity is a simple and intuitive property stating that the greater (or smaller) an auxiliary information is, the greater the response must be. The incorporation of restrictions generates more efficient and more robust estimators and in many cases achieve interpretable solutions. Monotone estimators have proven efficient in a diversity of applications, including Small Areas problems, as it is shown in Rueda et al. (2010), Lombardía et al. (2017), Wagner et al. (2017) or Chetverikov et al. (2018) among others.

- Synthetic monotone (SM): $\sigma_u = 0$,

$$\widehat{m}_d = \widehat{\theta}_d = P(\mathbf{Y}|\mathbf{K}).$$

where $P(\mathbf{Y}|\mathbf{K})$ is the projection of $\mathbf{Y}$ onto $\mathbf{K}$ 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)).

- Mixed monotone (MM): $\sigma_u \geq 0$,

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

In the case where σ_u^2 is unknown, we propose an iterative procedure to obtain $\widehat{\boldsymbol{\theta}} = P(\mathbf{Y}|\mathbf{K})$ and $\widehat{\sigma}_u^2$ following the ideas of Rueda et al. (2010) and Lombardía et al. (2017).

Finally, we present the estimators for m_d using P-splines. In this case, the working model is the following

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

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 $\mathbf{X} = [\mathbf{1}, \mathbf{x}, \ldots, \mathbf{x}^{(l-1)})]$, with l 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. Noted that this strategy is justified only when the true underlying is not a spline, which would be the case in most applications, see more details in the monograph Jiang (2010). 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 the corresponding synthetic and mixed estimators:

- Synthetic Spline (SS): $\sigma_u = 0$,

$$\hat{m}_d = \mathbf{x}_d\hat{\boldsymbol{\beta}} + \mathbf{z}_d\hat{\mathbf{v}}, \quad d = 1, \ldots, D.$$

- Mixed Spline (MS): $\sigma_u \geq 0$,

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

where $\mathbf{x}_d$ and $\mathbf{z}_d$ are the rows of X and Z corresponding to the ith small area, respectively. In both cases, we fit a mixed linear model, with one or two random effects, respectively. More details of the estimation process can be seen in Lombardía et al. (2017).

As has been shown, we have a direct relationship beween the models and the proposed estimators. In the following sections we consider various time periods, the notation is adapted by including a time subscript as follows: $\hat{m}_{dt} = \hat{\theta}_{dt}$ for the synthetic estimators, and for the mixed estimators $\hat{m}_{dt} = \hat{\theta}_{dt} + \hat{u}_{dt}$, where $d = 1, \ldots, D$ and $t = 1, \ldots, T$.

3 GAIC statistics

Once the models and estimators that play a role in the approach have been introduced, we can go on to describe the methodology proposed to estimate m_{dt}, consisting of selecting for each d, the specific working model from $\aleph$, that provides the most accurate predictions. For the selection, we introduce a simple procedure in Section 3.1 that uses the value of a new Akaike Information Criterion (AIC) statistic, specific for each domain, and is derived from the information across time.

As we have commented in the introduction, the use of AIC type statistics has been one of the most popular approaches to model selection since Akaike (1973). Different versions have been defined depending on the penalty and loss function, but in all the cases, the selection mechanism is similar, the value of the AIC statistic is derived for a family of models and the model with the lowest is selected.

In this section, we introduce the $xGAIC$ statistic derived in Lombardía et al. (2017), which combines a quasi-log-likelihood with a bootstrap GDF-estimator. In this paper, the good performance of $xGAIC$ compared with other popular AIC measures is shown, particularly in SAE applications; moreover, it is also shown how $xGAIC$ is flexible for handling conditional or marginal log-likelihoods. This property is relevant for the selection problem presented in this paper, because the family of candidate models includes synthetic and mixed models and considering these candidates simultaneously is not a usual practice, although of great interest in SAE applications, where the objective is not the model but the domain estimators.

Besides, in Section 3.1, a new domain specific measure $xGAIC$, using data across time, is defined.

A general AIC statistic for model M is defined as $AIC(M) = -2log(l(M)) + 2P$, where $l(M)$ is the model likelihood and P is a penalty term. To calculate the likelihood, two popular approaches are used to measure the goodness of fit in mixed effects models: the marginal likelihood and the conditional likelihood. For the marginal focus, we consider $\mathbf{Y} \sim N(\boldsymbol{\theta}, \mathbf{V}_y)$, where $\mathbf{V}_y = Var(\mathbf{Y})$. Then, the marginal log-likelihood is calculated as

$$\log(l_m(M)) = -\frac{1}{2}D\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_y| - \frac{1}{2}(\mathbf{Y}-\boldsymbol{\theta})'\mathbf{V}_y^{-1}(\mathbf{Y}-\boldsymbol{\theta}).$$

On the other hand, 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 $\mathbf{V}_{y|u} = Var(\mathbf{Y}|\mathbf{u})$. Thus, the conditional log-likelihood is calculated as

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

Lombardía et al. (2017) introduce a quasi-likelihood which considers the focus in the random effect and the total variability as follows:

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

We now combine the previous loss functions with a penalty term. The GDF have been used by several authors in complex modeling procedures, where it is assumed that $E(Y) = m$ (Ye (1998), Vaida and Blanchard (2005), Greven and Kneib (2010), You et al. (2016), among others).

This is a measure of the sensitivity of each fitted value $\hat{m}_d$ to perturbation in the corresponding observed value Y_d, for $d = 1, \ldots, D$, without considering the time effect:

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

Different expectations have been defined in the literature to estimate GDF. From the conditional mean estimator and conditional expectation (Vaida and Blanchard (2005), Lombardía et al. (2017)), we get:

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

From the marginal mean estimator and marginal expectation, we get:

$$mGDF = \sum_{d=1}^{D} \frac{\partial E_{Y}(\hat{\theta}_d)}{\partial \theta_d}.$$

And from the conditional mean estimator and the marginal expectation (You et al. (2016) and Lombardía et al. (2017)), we get:

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

Finally, we select a model $M \in \aleph$ with the smallest value of $GAIC$:

$$GAIC(M) = -2\log(l(\hat{M})) + \widehat{GDF},$$

where $l(\hat{M})$ depends on the parameters estimated under the model M. Although GDF is a known quantity in simple models, it is unknown in complex modeling procedures. You et al. (2016) and Lombardía et al. (2017) propose estimating it by bootstrap, $\widehat{GDF}$.

In mixed models, an AIC based on the marginal likelihood and $mGDF$ is typically used, but some authors, such as Vaida and Blanchard (2005), consider the purely conditional measure, using l_c and $cGDF$. Others, such as You et al. (2016), combine conditional or marginal log-likelihood with the same expectation to estimate the GDF, depending on the objective. In this paper, we consider a mixed measure, $xGAIC$, introduced by Lombardía et al. (2017) and defined as follows:

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

All the $GAIC$ measures cited above were studied and compared in Lombardía et al. (2017) and the authors showed that $xGAIC$ has a rather smaller classification error rate than the others. They also proved its usefulness for the selection of variables, as they were better in the comparison. Moreover, $xGAIC$ can be used to choose between synthetic and mixed models, because both terms, the log-likelihood and the penalty term in $xGAIC$, are well defined for both types of model, giving different but comparable results. The other $GAIC$ measures are not so flexible, as they use the conditional or marginal measure.

3.1 GAIC using data across time

When data across time are available, several selection model approaches can be considered. First, we define the simple approach, which consists in selecting a unique model for each time point as follows and through all domains. For each t, select the model $M_{\bullet t} \in \aleph$ with the smallest value of $GAIC$. In particular, for the $xGAIC$:

$$xGAIC^t(M) = -2log(l_x^t(\hat{M})) + x\widehat{GDF}^t,$$

and

$$M_{\bullet t} = \arg\min_{M \in \aleph} xGAIC^t(M). \tag{1}$$

Remember that $\aleph$ is the candidate set of working models as in Section 2.

Alternatively, let us define the specific-domain $xGAIC$ measure across time as:

$$xGAIC_d(M) = \sum_{t=1}^{T} \left(-2log(l_x^t(\hat{M}, d)) + \frac{1}{D} x\widehat{GDF}^t \right),$$

where $log(l_x^t(\hat{M}, d)$ is the d-th component of the log-likelihood for time t. This measure can be interpreted as goodness of fit for model M for the domain d across time. In this case, we take

$$M_{d\bullet} = \arg\min_{M \in \aleph} xGAIC_d(M) \tag{2}$$

After an estimator is selected by $xGAIC^t(M)$ or $xGAIC_d(M)$, the MSE can be estimated using bootstrap, following the ideas of González-Manteiga et al. (2008), Hall and Maiti (2006) and Lombardía et al. (2017) among others. The good behaviour of these measures for selecting the best estimator will be shown in the next sections for the real case and simulations. We anticipate here that the model-based domain estimators, using $M_{d\bullet}$, outperforms, in terms of the MSE, the estimators based on $M_{\bullet t}$ and the estimators based on any of the models in $\aleph$. However, we must keep in mind that this model selector approach is not suitable when there is a strong spatial or temporal dependence, it is not reasonable to apply different models for each one of them.

4 Application to real data

Returning to the real problem at hand, our objective is to estimate the employed people by economic activity in Galicia. This is a region in the north-west of Spain with an important amount of people working in the retail trade, construction and food and beverage service activities, among others. Remember that in the Regional Accounts, the labour force is often used as a regional indicator for the allocation of regional gross value added, so it is important to have a good measure of employment by economic activity. We use data from the LFS of Galicia in the period from the third quarter of 2009 to the first quarter of 2016. We have D domains, where each one is an economic activity. P_{dt} is the population in each of the economic activities, i.e., the people employed in this activity d and time t and the unemployed who, in their last job, were employed in that activity. Our goal is to estimate the total number of employed people by economic activity:

$$Z_{dt} = \sum_{j \in P_{dt}} z_{jt},$$

where $z_{jt} = 1$ means the j-th person in the domain d and time t is employed, and $z_{jt} = 0$ in another case.

The LFS does not produce official estimates at the domain level, but the analogous direct estimates of the total Z_{dt}, the mean $\bar{Z}_{dt} = Z_{dt}/N_{dt}$ and the size N_{dt} are:

$$\hat{Z}_{dt}^{dir} = \sum_{j \in S_{dt}} w_{jt}\, z_{jt},\ \hat{\bar{Z}}_{dt}^{dir} = \hat{Z}_{dt}^{dir}/\hat{N}_{dt}^{dir},\ \hat{N}_{dt}^{dir} = \sum_{j \in S_{dt}} w_{jt}$$

where S_{dt} is the sample in domain d and time t and w_{jt} is the official calibrated sampling weight. Considering that the sample weights w_{jt} correspond to the inverse of the probability of selecting the individual j, π_j, then if $\pi_j \neq 0$, the variance estimator of the direct estimator would be:

$$\hat{V}_{\pi}(\hat{Z}_{dt}^{dir}) \approxeq \frac{1}{\hat{N}_{dt}^2} \sum_{j \in S_{dt}} w_{jt}(w_{jt} - 1) \left(z_{jt} - \hat{\bar{Z}}_{dt}^{dir}\right)^2$$

The last formula is obtained with the simplifications: $w_{jt} = \frac{1}{\pi_j}$, $\pi_{jj} = \pi_j$ and $\pi_{ij} = \pi_i \pi_j$ for $i \neq j$ in the second order inclusion probabilities. In this work, we use the variance of the logarithm of the direct estimator of the total; this can be approximated by a Taylor linearization with the following formula

$$\hat{V}_{\pi}(log(\hat{Z}_{dt}^{dir})) = \frac{1}{(\hat{Z}_{dt}^{dir})^2} \hat{V}_{\pi}(\hat{Z}_{dt}^{dir}).$$

LFS is designed to obtain precise estimates in the activity sector: Agriculture, Industry, Construction and Services. The problem is when the domains are below the planned level (for example: Forestry, Manufacture of metals, Insurance, ...). Then, the LFS has very low sample sizes in these domains and, therefore, very high sampling errors in the direct estimator of each domain. For the first quarter of 2016, the minimum sample size in the domain is 4, the first quartile is 22 and the median 44; so, for some domains with the direct estimator, a reliable estimate of our objective cannot be obtained. To obtain reliable estimates in these domains, it is necessary to propose models that use auxiliary information that allows us to provide more information to improve the estimation process.

In this study, we use domain level models, such as those defined in Section 2, and their corresponding model-based estimators. These models have the advantage that they only need aggregate auxiliary information, which can usually be found in administrative registers. Therefore, as explanatory variables for the estimation of the target variable, we have considered the total number of people registered in the Social Security System ($S3$) by economic activity. This information is available for each quarter and for all variables of interest, from the third quarter of 2009 to the first quarter of 2016. We take 77 domains, after discarding seven domains for lack of information. We construct the data set with the following information:

- Z_{dt}: denotes, for simplicity of notation, the direct estimator of the total number of employed people in each economic activity d and quarter t, obtained from the LFS.
- $\sigma^2_{LFS,dt}$: is the variance of the direct estimator of $log(Z)$ in each economic activity d and quarter t.
- NACE: is the variable that indicates the economic activity (agriculture, forestry, manufacture of metals, Insurance, ...).

- $S3_{dt}$: is the auxiliary information, the people registered in the social security system in each economic activity and quarter.

To better fit the normality error assumptions, the response variable for the model is the logarithm of employed people in each economic activity and quarter, $Y_{dt} = log(Z_{dt})$, and the explanatory variable is $log(S3_{dt})$. The correlation between Y_{dt} and $\log(S3_{dt})$ is 0.92. Figure 2 shows the relationship between both variables, which seems to be close to linearity.

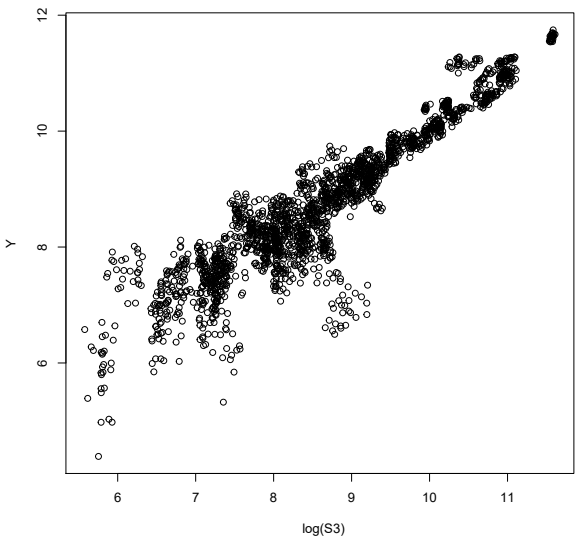

Figure 2: Scatter plot between the target variable Y_{dt} and the explanatory variable $\log(S3_{dt})$.

In Figure 3, we show the evolution along the 27 quarters of Y_{dt} and $\log(S3_{dt})$ in six domains. These figures show that there are different patterns, some of them quite abrupt. The behaviour of each economic activity is quite different, indicating that it may be convenient to consider different small area estimation approaches for different activities.

In order to see how the different estimation approaches fit the data, we consider the models presented in Section 2 for each of the quarters of our data ($t = 1, \ldots, 27$) and the corresponding estimators are derived. Those for the same six economic activities as in Figure 3 are plotted in Figure 4. This figure shows the similar behaviour of mixed estimators across domains; in fact, the lines that remain hidden correspond to mixed estimators that behave in a very similar way. Besides, the figure also shows how close the mixed estimators are to the direct estimators and that both exhibit an abrupt pattern in most cases.

On the other hand, the synthetic estimators are smoother, close to the explanatory variable and with a heterogeneous behaviour across activities.

It is relevant to remember at this point in the discussion that stability is a property highly valued by the Statistical Offices when publishing survey results. However, it is also worth remembering that overly smoothed patterns can prevent major trend changes being detected. Keeping these comments in mind, and in view of the graphical displays, a sensible approach could be to select different estimators for each domain.

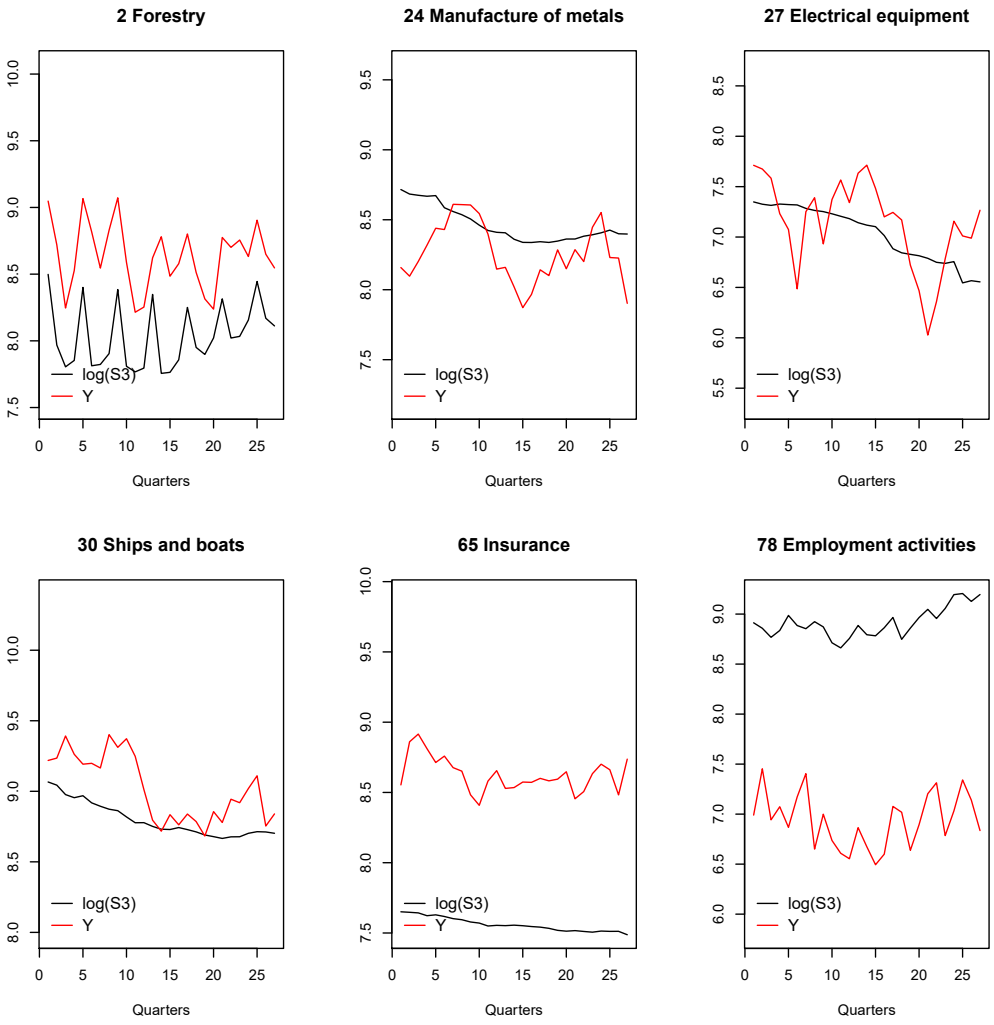

Figure 3: Evolution of Y_{dt} and $\log(S3_{dt})$ along the quarters in six domains.

In Table 1, we show the results for eight activities, as an illustration, but the complete results are shown in the Appendix for the 77 activities of the NACE and the last quarter. For each activity, we give the explanatory variable ($S3$), the value of the direct estimator (Z), of the usual estimator calculated in SAE (ML) and of the model-based estimator under the model (2) (for simplicity we denote the estimator as the model, $M_{d\bullet}$), and also the corresponding selected estimator by $xGAIC$ (labeled as "Est."). For NACE= 2 in Table 1 (Forestry activity), the selected estimator is the MM, which is a good choice because the distance between the direct estimator and the synthetic estimator is high in this domain (see Figure 4). The Figure 4 shows the synthetic estimator has a systematic and significative bias, pointing out the right choice for each domain. For NACE $= 24, 27$ and 30 (i.e. Manufacture of metals, Electrical equipment and Ships and boats), $M_{d\bullet}$ correspond to the synthetic estimators: SM, SS and SL, respectively. Again the choice is a good one, as the behaviour of the mixed and the direct estimators are very abrupt and, in addition, the synthetic is not far from the direct. In these domains, the smoother pattern exhibited by the chosen estimators facilitates the interpretation in the real world. The domains NACE $= 65$ and 78 have a similar behaviour to Forestry, i.e. important difference are shown between the direct and the explanatory values, so the choice of a mixed estimator, in this case MM, is a good one. Besides, it is also interesting to note, in the Forestry and Employment activities, that the estimators maintain seasonality.

We have also included in Table 1 the results for NACE $= 55$ and 85, from which we know what its relationship with the auxiliary information is like. It is well known that the explanatory variable underestimates employment for NACE $= 85$ (Education), but is a very good estimate for NACE $= 55$ (Accommodation). Then, we can see that the $xGAIC$ selects the most appropriate model to build the model-based estimator, a mixed estimator for the NACE $= 85$, close to the direct estimator, and a synthetic estimator for the NACE $= 55$, simply a function of the explanatory variable. These are only two examples of how the methodology works using the auxiliary information in a correct way.

Indeed, Figure 5 plots for the last quarter the direct estimates of Y_d against $M_{d\bullet}$, the model-based estimates under the model (2). We observe that there are some differences between both estimates, some quite important, especially in the activities with few employees. For example, NACE $= 24, 36, 38, 72, 80$, among others. The differences can be seen in detail in Tables 5 and 6 in the Appendix.

Moreover, in order to study the stability, we consider different number of time periods. Results using data from $T = 12, 20, 27$ periods of time has been compared. In these three scenarios, and for all the activities, either the same model is selected or a model that behaves in a similar way is selected. In fact, for a particular activity, the type of model selected turned out to be always synthetic or always mixed, independently of the number of time periods used.

Finally, in order to validate the goodness of the estimators selected, a simulation experiment, included in the next section, has been designed to imitate the real case. The results there show that $M_{d\bullet}$ is the best choice in terms of MSE for most of the domains as well as globally. In practice, the MSE can be estimated using bootstrap, following the ideas of González-Manteiga et al. (2008), Hall and Maiti (2006), Lombardía et al. (2017) among others. Once the model is selected, we apply the resampling method to calculate the MSE under this model. In addition, Lombardía et al. (2018) proved that $xGAIC$ select estimators with the lowest MSE. Another alternative to estimate the MSE is the unified Monte-Carlo jackknife method for small area

estimation introduced in Jiang et al. (2018), which will be the subject of future research.

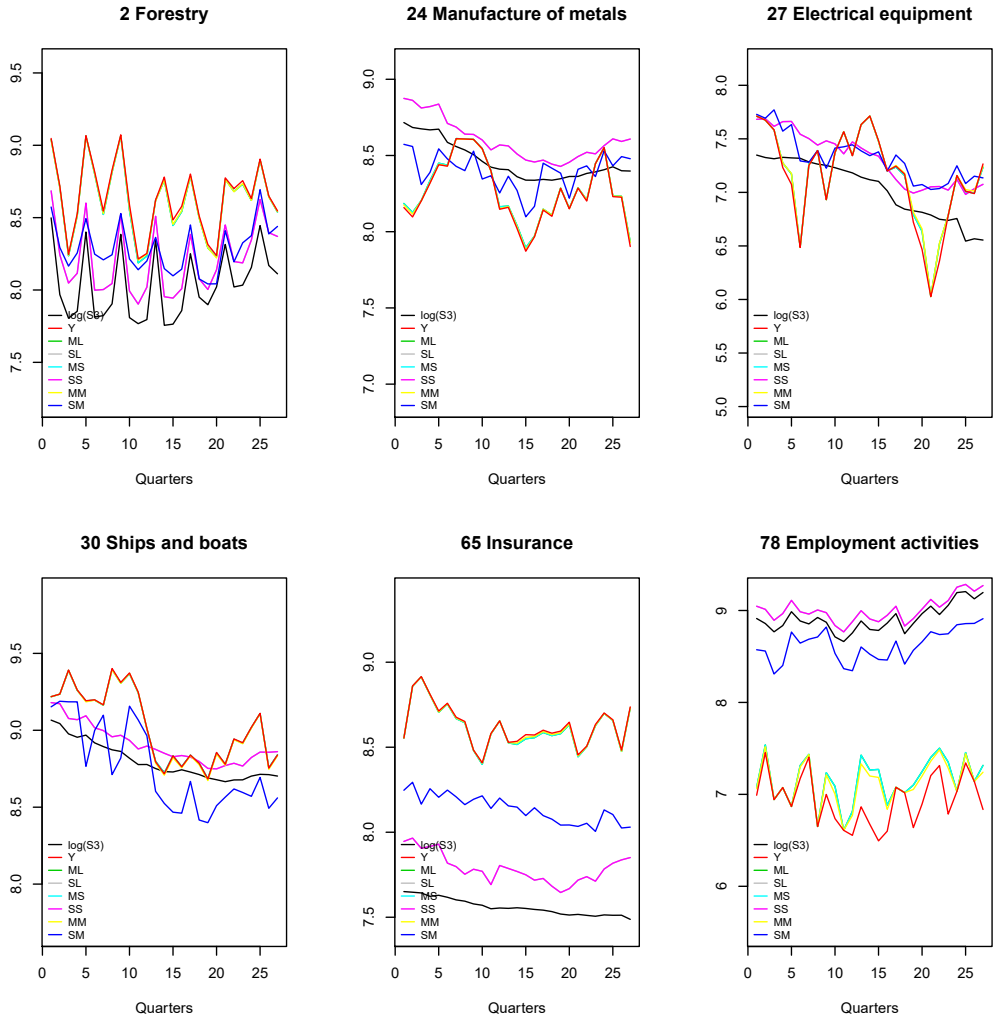

Figure 4: Evolution of Y_{dt}, $\log(S3_{dt})$ and the estimators of Y_{dt} along the quarters and in six domains.

5 Simulation studies

5.1 Simulation 1

The objective of this simulation is to study how the $xGAIC_d$, using data across time, selects the best estimator for each domain d and across time, designing a scenario very close to the real case. We use mixed and synthetic estimators obtained from linear, monotone and P-spline models.

NACE	Activity	$S3$	Z	ML	$M_{d\bullet}$	Est.
2	Forestry	3334	5150	5108	5351	MM
24	Manufacture of metals	4438	2705	2792	5023	SM
27	Electrical equipment	703	1430	1386	1235	SS
30	Ships and boats	6017	6902	6905	7363	SL
55	Accommodation	7458	12048	11988	8801	SL
65	Insurance	1786	6230	6187	6469	MM
78	Employment activities	9854	930	1503	1456	MM
85	Education	41982	74103	74070	77347	MM

Table 1: For eight selected activities, the people registered in the Social Security System ($S3$), the estimations of the employed people Z, ML and $M_{d\bullet}$, and the type of estimator selected by the $xGAIC$ (Est.) are shown.

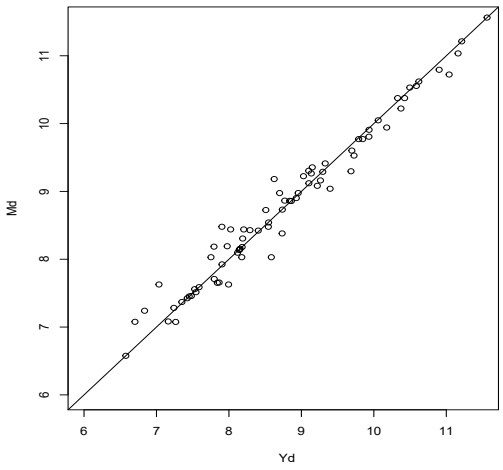

Figure 5: Scatter-plot of the direct vs the $M_{d\bullet}$ estimates in logarithm scale.

We consider generating the data the following model:

$$y_{dt} = m_{dt} + e_{dt} = \alpha_d + x_{dt}\beta_t + v_{dt} + e_{dt}, \quad d = 1, ..., D,\ t = 1, ..., T;$$

where $D = 77$ domains, $T = 27$ periods of time, x_{dt} the explanatory variable in the real example, $v_{dt} \sim N(0, \sigma^2_{v_d})$ and $e_{dt} \sim N(0, \sigma^2_{dt})$. We take $\sigma^2_{dt} = \sigma^2_{LFS,dt}$, the values in the real case, and we also consider, for comparative proposes, $\sigma^2_{dt} = \sigma^2_{LFS,dt} * 10$. For β_t, we take the estimated values for the linear model for each quarter in the real application. Finally, we use the following values for α_d and $\sigma^2_{v_d}$

$$\alpha_d = \frac{1}{T}\sum_t (y_{dt} - x_{dt}\beta_t) \quad d = 1, \ldots, D;$$

$$\sigma^2_{v_d} = \sigma^2_{r_d} - \frac{1}{T}\sum_t (\sigma^2_{dt}) \quad d = 1, \ldots, D$$

where $\sigma^2_{r_d}$ is the variance of the T residuals corresponding to the domain d and y_{dt} is the response

variable in the LFS data set.

We generate and analyze the data as follows:

- Repeat $I = 100$ times ($i = 1, \ldots, 100$)
  - Generate samples $\{(y_{dt}, x_{dt})\}$, $d = 1, \ldots, D$; $t = 1, \ldots, T$.
  - For each $d = 1, \ldots, D$ and $t = 1, \ldots, T$, fit the linear model, the monotone model and the P-spline model, and calculate for each: $\widehat{m}_{dt}$ and $xGAIC_{dt}$. We denote the estimators as in Section 2: SL (synthetic linear), ML (mixed linear), SM (synthetic monotone), MM (mixed monotone), SS (synthetic spline) and MS (mixed spline).
  - For each domain d and previous estimators, calculate the error:

$$E_d = (\widehat{\mathbf{m}}_d - \mathbf{m}_d)^{'}(\widehat{\mathbf{m}}_d - \mathbf{m}_d),$$

    with $\widehat{\mathbf{m}}_d = (\widehat{m}_{d1}, \ldots, \widehat{m}_{dT})$ the vector of model-based estimations of $\mathbf{m}_d = (m_{d1}, \ldots, m_{dT})$.
  - For each time t, record the estimator M^t, which is the model-based estimator under the model (1), see details in Section 3.1. For simplicity we denote the estimator the same as the model. We also calculate its error as before.
  - For each domain d, record the estimator M_d, which is the model-based estimator under the model (2). See details in Section 3.1. We also calculate its error as before.

- Derive global statistics:
  - Calculate the empirical MSE:

$$MSE_d = \frac{1}{I}\sum_{i=1}^{I} E_d^{(i)}, \quad d = 1, \ldots, D;$$

$$MSE = \frac{1}{D}\sum_{d=1}^{D} MSE_d.$$

First, using Figure 6 and Figure 7, we illustrate how the simulated data from the generating model used in the simulation, imitate the real data. In Figure 6, we present the scatter-plot for the response variable against the explicative variable corresponding to the real data in the left hand graph, and the same scatter-plot for the data generated in the first and second iteration, in the middle and right hand graphs, respectively.

On the other hand, in Figure 7, we show the evolution of the response variable along the 27 quarters for the activities selected in Section 4. In the first row, we present the results of the real data and the second and third rows present the results for the data generated in the first and second iteration. From these figures, we conclude that the generating approach can replicate the real data quite well. We note that several simpler models have been tested as generators, but they have failed to replicate the real data as well as the selected one does.

Next, we present the results of the simulation in Figure 8 and Table 2. Figure 8 plots the MSE_d for a list of estimators defined in this paper: $SL, ML, SM, MM, SS, MS, M_{\bullet t}, M_{d\bullet}$ and we add the direct estimator (Direct). The results are shown for the scenario where $\sigma^2_{dt} = \sigma^2_{LFS,dt}$, as similar conclusions are derived in the other case.

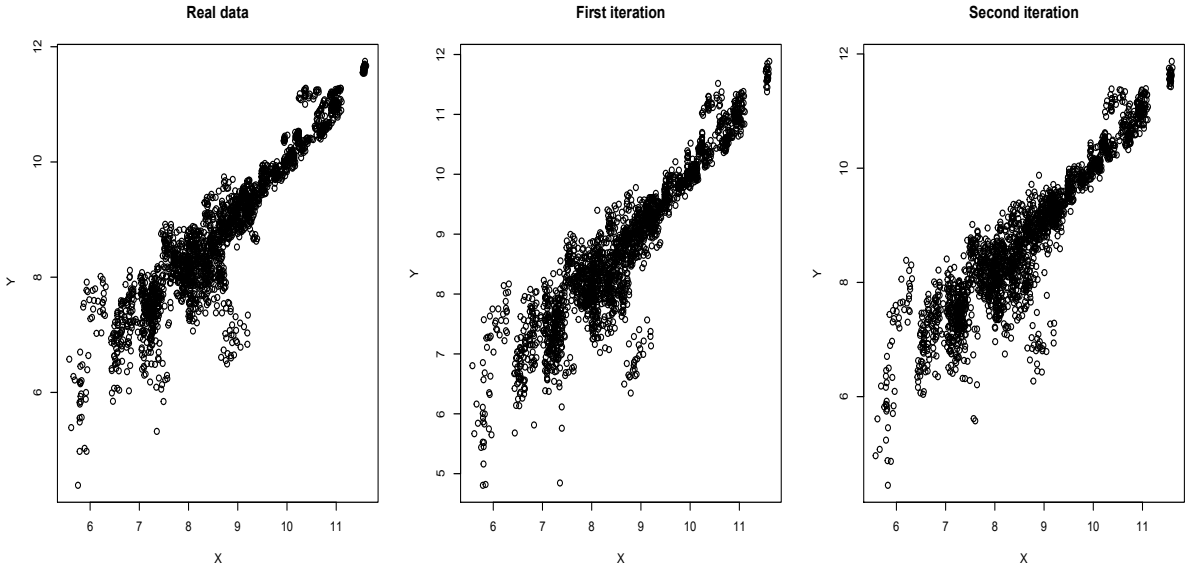

Figure 6: Scatter-plot of the response variable vs the explicative variable, for the real data and for the first and second iteration.

In Figure 8, we see that the variability of synthetic estimators is greater than that of both the mixed estimators and the Direct; also that the mixed estimators and the Direct have very similar MSE distributions. In fact, we have found that these estimators are very similar to each other. This last fact is shown in Table 2, where specific values for eight domains are presented (those analyzed in Section 4). The table is divided in to two parts, on the top, the MSE_d values for the different estimators are given. $M_{d\bullet}$ gives the lowest MSE_d values in domains NACE $= 2, 24, 27, 30, 55, 65, 78, 85$. Moreover, when all the domains are considered (the last column), the MSE of $M_{d\bullet}$ is reduced by a minimum of 20%.

On the bottom of the table, the results of the iterations are synthesized. For the estimators more often chosen, the table gives the percentage of times the estimator is chosen in the penultimate row and the same percentage, but attending only to the type (synthetic or mixed), in the last row. The numbers show that the same type, synthetic or mixed, is selected in most iterations and that only in one domain, NACE $= 2$, does the type change from one iteration to the next; however, even in this case, a mixed estimator is selected 82% of the times.

Comparing this table with Table 1, it is clear that the same estimators selected with the real data are also chosen with the generated data, except for domain 27, where the real data gives SS and the simulations give SL 76% of the times. In any case, the MSE_d of both estimators are quite similar (0.020 and 0.027) and the values for both estimators, across time, are quite close (see Figure 4).

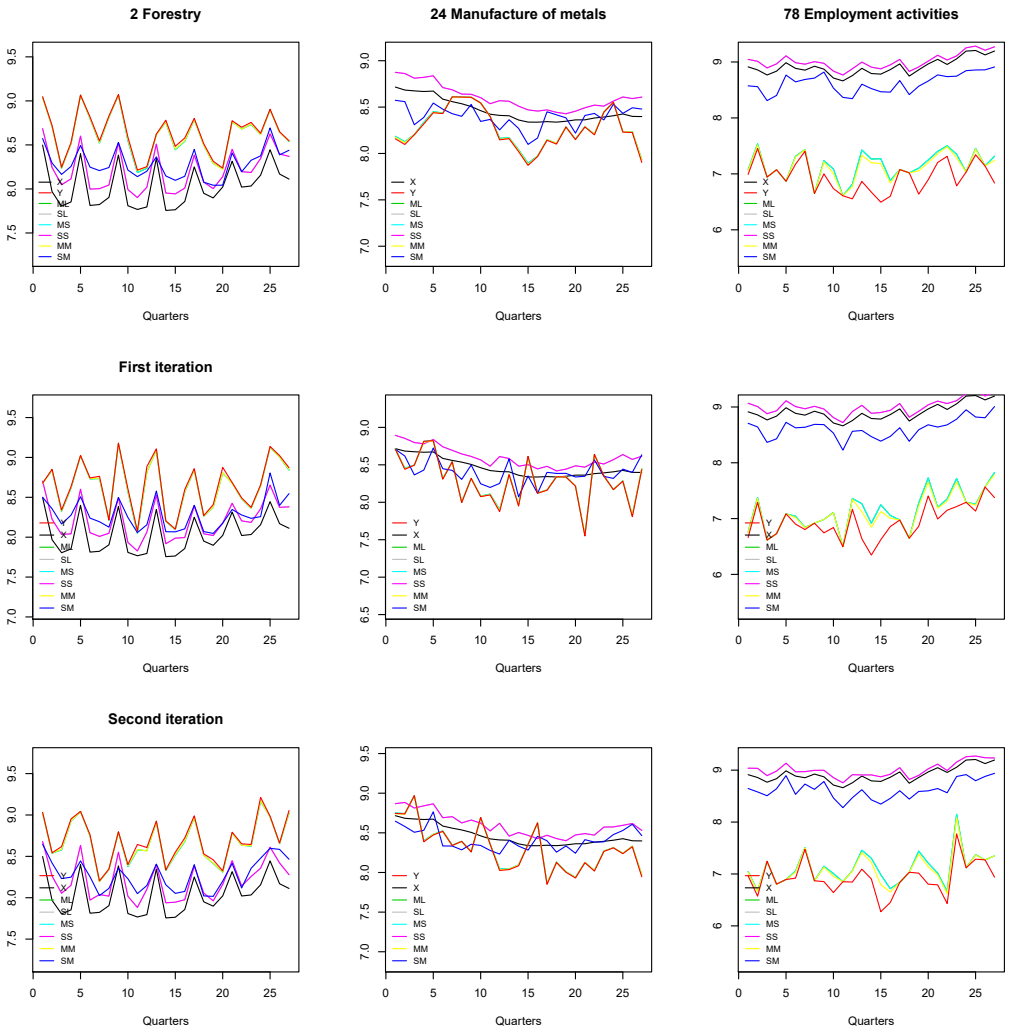

Figure 7: Values of Y across the 27 quarters. In the first row we have the real data, in the second and third rows we have the first and second iterations of Simulation 1, respectively.

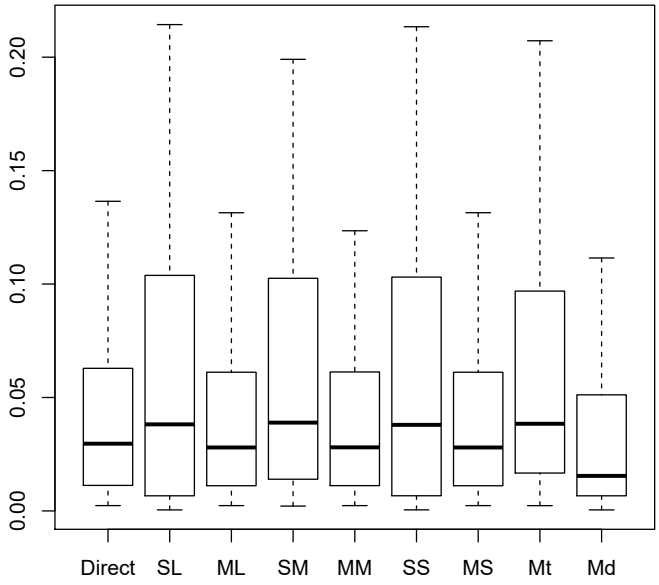

Figure 8: MSE_d for the estimators.

Domain	2	24	27	30	55	65	78	85	Total
				MSE					
Direct	0.041	0.053	0.137	0.028	0.019	0.011	0.063	0.009	0.053
SL	0.187	0.113	0.027	0.015	0.014	0.719	4.280	0.621	0.179
ML	0.038	0.051	0.120	0.027	0.018	0.011	0.132	0.009	0.048
SM	0.145	0.036	0.058	0.103	0.046	0.270	2.843	0.019	0.127
MM	0.038	0.051	0.123	0.027	0.018	0.011	0.112	0.009	0.050
SS	0.188	0.113	0.139	0.015	0.018	0.719	4.273	0.619	0.177
MS	0.038	0.051	0.020	0.027	0.014	0.011	0.135	0.009	0.047
$M_{\bullet t}$	0.038	0.051	0.122	0.027	0.018	0.011	0.120	0.009	0.049
$M_{d\bullet}$	0.063	0.036	0.046	0.015	0.014	0.011	0.112	0.009	0.040
				Selection					
$M_{d\bullet}$	MM	SM	SL	SL	SL	MM	MM	MM	
%	(79%)	(100%)	(76%)	(100%)	(99%)	(99%)	(99%)	(100%)	
Type	M	S	S	S	S	M	M	M	
%	(82%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	

Table 2: MSE_d ($d = 2, 24, 27, 30, 55, 65, 78, 85$) and MSE at the top. It shows the estimator selected by $xGAIC_d$ ($M_{d\bullet}$) and the type (synthetic or mixed) with the percentage of times selected, at the bottom.

5.2 Simulation 2

The objective of this simulation is to study how the $xGAIC_d$ selects the best estimator for each domain d using data across time in different scenarios. We consider the combination of the parameter values from real data with different values of random effects, with the aim of including the different cases that can be found in practice. In this case, the generated model is

$$y_{dt} = m_{dt} + e_{dt} = \alpha_d + f(x_{dt}) + v_{dt} + u_{dt} + e_{dt} \quad d = 1, ..., D;\ t = 1, ..., T;$$

where $D = 77$, $T = 27$ and the explanatory variable x_{dt} are those of the real example. For the random part of the model, we take $u_{dt} \sim N(0, \sigma_u^2)$ with $\sigma_u^2 = 0.2$, $v_{dt} \sim N(0, \sigma_{v_d}^2)$ with $\sigma_{v_d}^2$ the same as in Simulation 1, and $e_{dt} \sim N(0, \sigma_{dt}^2)$ with σ_{dt}^2 as follows. We consider different options:

- With regard to $f()$, three different functional forms are considered :
  - F1: $f(x_{dt}) = \beta_{0t} + \beta_t x_{dt}$
  - F2:
$$f(x_{dt}) = \begin{cases} \beta_{1t} & \text{if } x_{dt} \leq b_{1t}; \\ \beta_{2t} & \text{if } b_{1t} < x_{dt} \leq b_{2t}; \\ \beta_{3t} & \text{if } b_{2t} < x_{dt} \leq b_{3t}; \\ \beta_{4t} & \text{if } b_{3t} < x_{dt} \leq b_{4t}; \\ \beta_{5t} & \text{if } b_{4t} < x_{dt}. \end{cases}$$
  - F3: $f(x_{dt}) = \beta_{0t} + 5 + 10\beta_t \sin(\pi \frac{x_{dt} - \min x_{dt}}{\max x_{dt} - \min x_{dt}})$

  The parameters β_{0t} and β_t being the fitted values for the mixed linear model in the real application for each time. We take β_{it} $(i = 1, 2, 3, 4, 5)$ as the minimum value, the quantile 20, quantile 40, quantile 60 and quantile 80 of $(y_{1t}, ..., y_{Dt})$, which are the values of Y in the real case. Furthermore, we take b_{it} $(i = 1, 2, 3, 4)$ as the quantile 20, quantile 40, quantile 60 and quantile 80 of $(x_{1t}, ..., x_{Dt})$, which is calculated from the variable S in the real case.

- With regard to the number of domains where the random term $\sigma_u^2 \neq 0$, three options are considered:
  - U1 : 77 domains with $\sigma_u^2 \neq 0$;
  - U2 : 55 domains with $\sigma_u^2 \neq 0$;
  - U3 : 13 domains with $\sigma_u^2 \neq 0$.

- With regard to the number of domains *out of the model* (with $\alpha_d \neq 0$ or/and $\sigma_{v_d}^2 \neq 0$), four options are considered:
  - O1 : 0 domains with $\alpha_d \neq 0$ $\sigma_{v_d}^2 \neq 0$;
  - O2 : 5 domains with $\alpha_d \neq 0$ and $\sigma_{v_d}^2 \neq 0$;
  - O3 : 10 domains with $\alpha_d \neq 0$ and $\sigma_{v_d}^2 \neq 0$;
  - O4 : 20 domains with $\alpha_d \neq 0$ and $\sigma_{v_d}^2 \neq 0$.

- With regard to σ_{dt}^2, three options are considered:
  - V1: $\sigma_{LFS,dt}^2$;

– V2: $\sigma^2_{LFS,dt} * 10$;

– V3: σ^2_{dt} that decreases with x_{dt}.

A total of 108 scenarios are designed by combining each of the options above; 3 options for $f()$, 3 for σ^2_u, 4 for α_d and $\sigma^2_{v_d}$, and 3 for σ^2_{dt}. We repeat the procedure 100 times.

Tables 3 and 4 and Figures 9, 10 and 11 show the main results. Tables 3 and 4 give the average MSE values for different estimators defined in the paper (rows) and for the combination of scenarios (columns). $M_{d\bullet}$ gives the smallest MSE in all the columns (the MSE reduction being at least 18%) except for the case of scenarios O1, where SS gives a smaller MSE. Figure 9 compares SS with $M_{d\bullet}$ in the 108 simulated scenarios. Points above the diagonal represent scenarios where $M_{d\bullet}$ outperforms SS and points below are scenarios where the opposite happens. In a similar way, Figure 10 compares $M_{d\bullet}$ to $M_{\bullet t}$, the estimator selected using the quarter information, while Figure 11 compares $M_{d\bullet}$ to ML, the estimator commonly used in SAE. In both plots, almost all the points are above the diagonal, which implies that $M_{d\bullet}$ outperforms $M_{\bullet t}$ and ML in most scenarios, the differences being more important when the funtional form $f()$ moves away from lineality.

	Scenarios						
	F1	F2	F3	O1	O2	O3	O4
Direct	0.214	0.209	0.202	0.245	0.200	0.177	0.211
SL	0.106	0.358	4.985	1.845	1.785	1.818	1.816
ML	0.130	0.145	0.211	0.180	0.152	0.149	0.166
SM	0.106	0.115	3.237	1.145	1.134	1.168	1.164
MM	0.142	0.128	0.189	0.171	0.145	0.141	0.156
SS	0.105	0.289	0.100	0.067	0.197	0.229	0.167
MS	0.131	0.145	0.112	0.133	0.127	0.122	0.136
$M_{\bullet t}$	0.136	0.128	0.113	0.129	0.124	0.122	0.127
$M_{d\bullet}$	0.084	0.105	0.090	0.109	0.087	0.087	0.089

Table 3: MSE by rows, taking different functional forms on the left hand side of the table, and for the different values out of model on the right hand side.

	Scenarios					
	V1	V2	V3	U1	U2	U3
Direct	0.124	0.196	0.304	0.202	0.144	0.278
SL	1.732	1.731	1.985	1.826	1.798	1.825
ML	0.121	0.168	0.196	0.158	0.105	0.223
SM	1.083	1.101	1.274	1.159	1.133	1.166
MM	0.121	0.155	0.185	0.149	0.096	0.215
SS	0.163	0.168	0.164	0.169	0.154	0.171
MS	0.118	0.134	0.136	0.121	0.083	0.184
$M_{\bullet t}$	0.115	0.136	0.126	0.119	0.079	0.179
$M_{d\bullet}$	0.074	0.098	0.106	0.098	0.063	0.118

Table 4: MSE by rows, taking different values of the error variance on the left hand side of the table, and for the different scenarios of σ^2_u on the right hand side.

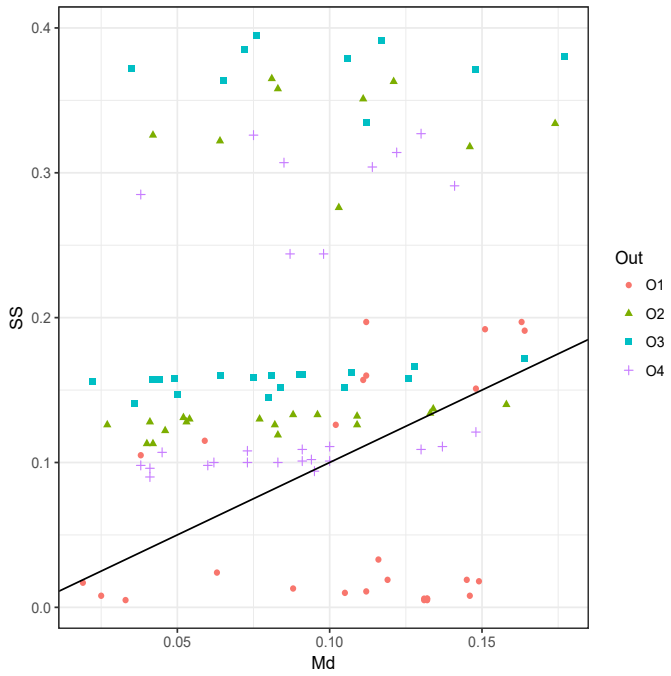

Figure 9: $MSE_d(M_{d\bullet})$ vs $MSE_d(SS)$, depending on the number of domains out of the model.

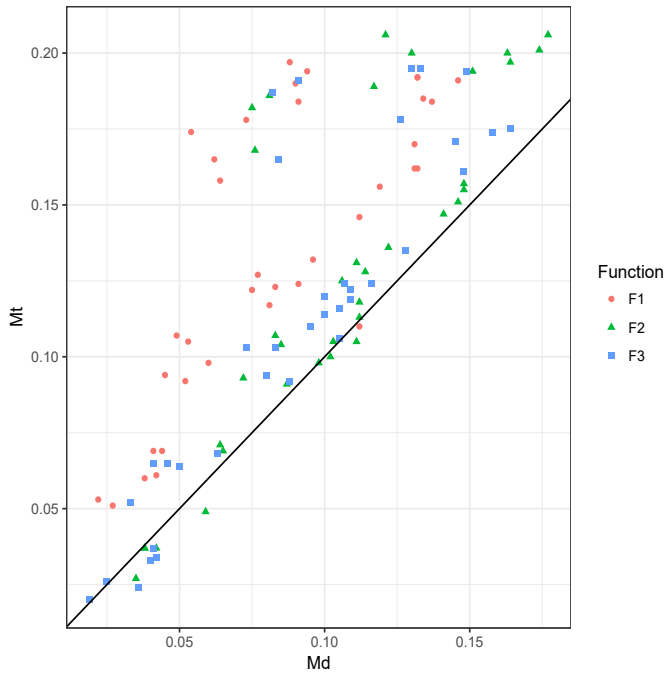

Figure 10: $MSE_d(M_{d\bullet})$ vs $MSE_d(M_{\bullet t})$, depending on the functional form.

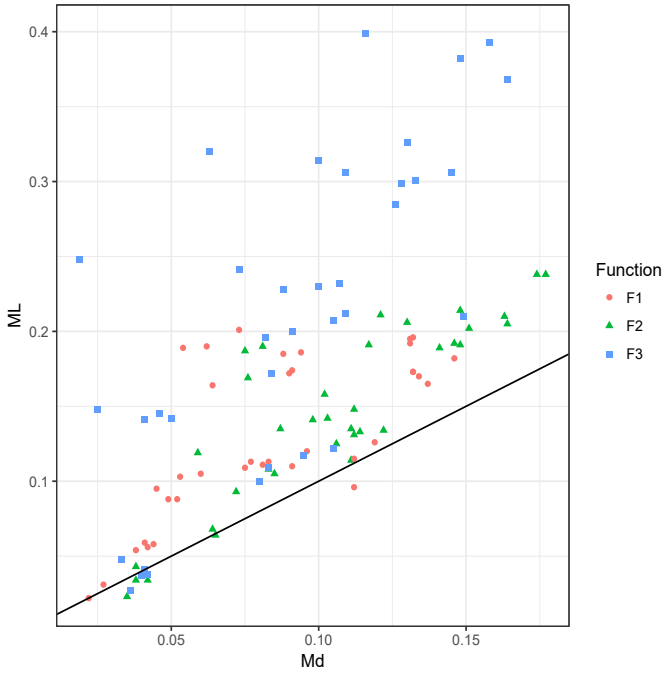

Figure 11: $MSE_d(M_{d\bullet})$ vs $MSE_d(ML)$, depending on the functional form.

6 Conclusions

The first contribution of this paper, and the most important from the practical point of view, is the derivation of accurate labour Force estimators across domains and time points. The estimators are model-based and have been derived using different models. A wide range of models, exhibiting a different functional relationship between response and explicative, and which may include random terms or not, are considered as a candidate for each domain. The different behaviour of the estimators in the LFS data justifies the consideration of this variety of models. As a final output, we have proposed for each economic activity the most appropriate model to determinate the evolution of the employed people and the models combining administrative registers and survey data, following the recommendations of Eurostat (Eurostat (2013)).

The question of selecting different models arises from the observation of different temporal patterns of observed rates across time for the domains, which gives an insight to the different generating models in the domains. In fact, we have been able to reproduce different patterns between domains across time in simulations by using a complex generator model. Furthermore, from these simulations, when using the same model to obtain the estimates for the different domains, the fact that the errors might be not negligible in some domains has been verified. The design of simulations using complex generator models is a second interesting contribution of this paper.

The third and most important contribution, from the theoretical point of view, is the design of an approach to determine which model-based estimator is suitable for each domain. The approach is based on the definition of the $xGAIC$ measure, specific for each domain, which is obtained using the bootstrap. The good performance of the procedures is evidenced by the

results of two simulation experiments. The new proposal of using different estimators across domains translates into a considerable decrease in the MSE, which is particularly relevant when the underlying model is not linear or the sampling variability is high.

The proposed approach can be considered as one that model the data over space and time jointly. Among the approaches in this line, considered in the literature, one of the most popular is the model of Rao and Yu (1994) which has been considered during the research for comparative purposes. However, this model has not been finally included, as we have checked that is not a good data generator and provide small area estimators that are quite similar to those given using any of the mixed models considered in the paper.

Finally, we would like to do some recommendations, first, when it is suspected that there are correlations between domains, the approach is not recommended as it is not reasonable to apply different models for each domain. Conglomerates of domains with different behaviour would then be taken. However, the approach is recommended when the correlations are between temporary moments. Second, regarding the number of periods recommended for this methodology, we note that it depends on the type of data. In this real case analysed here, the data are quarterly and our experience suggests using at least 12 quarters.

To the best of our knowledge, this is the first time this proposal of selecting different models for each domain has been considered in the literature of SAE.

7 Appendix

Tables 5 and 6 give the estimation for the employed people in the first quarter of 2016 (the last quarter of LFS data) in each economic activity (domain). So, for each domain, we give its NACE number, the explanatory variable ($S3$), the direct estimator (Z), the usual estimator calculated in SAE (ML), the estimator calculated under $M_{d\bullet}$ ($M_{d\bullet}$) and the corresponding selected estimator (labeled as "Est.").

References

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

Chetverikov, D., Santos, A., and Shaikh, A. M. (2018). The econometrics of shape restrictions. *Annual Review of Economics*, 10:31–63.

Eurostat (2008). Nace rev. 2. statistical classification of economic activities in the european community. *European Comission*.

Eurostat (2013). Manual on regional accounts methods. *European Comission*.

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.

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. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate fay-herriot model. *Computational Statistics and Data Analysis*, 52:5242–5252.

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

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

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

Jiang, J. (2010). *Large Sample Techniques for Statistics*. Springer, New York.

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

Jiang, J., Lahiri, P., and Nguyen, T. (2018). A unified monte-carlo jackknife for small area estimation after model selection. *Annals of Mathematical Sciences and Applications*, page in press.

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

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.

Lombardía, M., López-Vizcaíno, E., and Rueda, C. (2018). Selection of small area estimators. *Special edition in honor of J.N.K. RAO*, In press.

Marhuenda, Y., Morales, D., and Pardo, M. (2014). Information criteria for fay-herriot model selection. *Computational Statistics and Data Analysis*, 70:268–280.

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.

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.

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

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

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.

Wagner, J., Munnich, R., Hill, J., Stoffels, J., and Udelhoven, T. (2017). Nonparametric small area models using shape-constrained penalized b-splines. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180:1089–1109.

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.

Activity (NACE)	$S3$	Z	ML	$M_{d\bullet}$	Est.
1	38690	40973	40965	42778	SM
2	3334	5150	5108	5351	MM
3	18759	20584	20577	18952	SL
8	1939	3559	3526	3206	SM
10	28290	30588	30571	33486	SM
11	2670	2426	2564	3747	SS
13	1586	3577	3571	3728	MS
14	9316	8341	8360	10590	SL
16	8221	8984	8986	9544	SL
17	1256	1858	1859	2001	SS
18	3084	3603	3615	4224	SL
19	705	815	815	1237	SL
20	2690	2911	2942	3770	SS
21	898	2966	2966	2144	SM
22	3086	1723	1723	1799	MM
23	6904	5999	6002	8254	SL
24	4438	2705	2792	5023	SM
25	14673	16262	16260	15450	SL
26	709	1291	1282	1243	SS
27	703	1430	1386	1235	SS
28	4351	5137	5144	5023	SM
29	13453	18882	18860	18294	SM
30	6017	6902	6905	7363	SL
31	3353	7056	7025	7340	MM
32	1283	1975	1975	2062	MM
33	6320	7555	7551	7670	SL
35	2195	5353	5318	3206	SM
36	1432	1135	1135	2144	SM
38	4110	3656	3656	4829	SM
41	22079	26309	26274	21701	SL
42	4111	10884	10826	11298	MM
43	44847	36173	36178	39119	SL
45	21210	20631	20630	20989	SL
46	46179	39612	39610	40081	SL

Table 5: Estimated employed people.

Activity (NACE)	$S3$	Z	ML	$M_{d\bullet}$	Est.
47	105069	105142	105110	109774	SM
49	30931	32024	32014	28722	SL
50	747	1677	1621	1751	SM
51	264	718	718	750	SM
52	10260	8962	8993	11474	SL
53	3376	6212	6188	4554	SL
55	7458	12048	11988	8801	SL
56	61469	54262	54244	50842	SL
58	2399	3358	3357	3428	SS
59	1192	1892	1885	1916	SS
60	1414	2604	2586	2209	SS
61	3578	3985	3991	4779	SL
62	7857	10097	10097	9191	SL
64	9781	9293	9300	11027	SL
65	1786	6230	6187	6469	MM
66	6376	2710	2794	2886	MM
68	3624	3052	3111	4830	SL
69	13430	16751	16747	14354	SL
70	3550	4467	4468	4748	SL
71	10893	9417	9440	12061	SL
72	2769	2327	2327	3206	SM
73	3005	3435	3455	3571	MM
74	6915	6438	6472	7385	SM
75	1594	1557	1586	1657	MM
77	3233	1395	1480	1519	MM
78	9854	930	1503	1456	MM
79	1506	2434	2429	2327	SS
80	5113	4960	4976	6430	SL
81	25091	23398	23396	24135	SL
82	11796	5568	5762	10152	SM
84	52769	70450	70404	64716	SM
85	41982	74103	74070	77347	MM
86	56514	62285	62275	47410	SL
87	10165	16072	16051	11386	SL
88	11682	11257	11266	12782	SL
90	1977	3458	3431	3595	MM
91	351	1774	1662	1810	MM
92	1406	2540	2529	2198	SL
93	8642	10544	10531	9949	SL
94	6914	7755	7756	8265	SL
95	4319	3480	3480	3633	MM
96	17986	17876	17874	18300	SL
97	28108	33682	33651	33486	SM

Table 6: Estimated employed people (continuation).