ISSN: 1696-2281
SORT 33 (1) January - June (2009)

SORT

Statistics and Operations Research Transactions

Sponsoring institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Supporting institution

Spanish Region of the International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Published on paper
bearing the certificate
of sustainable forest
management

SORT

Volume 33 (1), January-June 2009

Formerly Qüestiió

Invited article (with discussion)

Articles

Erratum

Book review

Information for authors and subscribers

SORT 33 (1) January-June 2009, 3-30

Modelling consumer credit risk via survival analysis

Ricardo Cao, Juan M. Vilar and Andrés Devia

Universidade da Coruña[*]

Abstract

Credit risk models are used by financial companies to evaluate in advance the insolvency risk caused by credits that enter into default. Many models for credit risk have been developed over the past few decades. In this paper, we focus on those models that can be formulated in terms of the probability of default by using survival analysis techniques. With this objective three different mechanisms are proposed based on the key idea of writing the default probability in terms of the conditional distribution function of the time to default. The first method is based on a Cox's regression model, the second approach uses generalized linear models under censoring and the third one is based on nonparametric kernel estimation, using the product-limit conditional distribution function estimator by Beran. The resulting nonparametric estimator of the default probability is proved to be consistent and asymptotically normal. An empirical study, based on modified real data, illustrates the three methods.

MSC: 62P20, 91B30, 62G05, 62N01

Keywords: Probability of default, Basel II, nonparametric regression, conditional survival function, generalized product-limit estimator.

1 Introduction

Determining the probability of default, *PD*, in consumer credits, loans and credit cards is one of the main problems to be addressed by banks, savings banks, savings cooperatives and other credit companies. This is a first step needed to compute the capital in risk of insolvency, when their clients do not pay their credits, which is called *default*. The risk coming from this type of situation is called *credit risk*, which has been the object of research since the middle of last century. The importance of credit risk, as part of

[*] Departamento de Matemáticas. Facultad de Informática. Universidade da Coruña. Campus de Elviña, s/n. A Coruña 15071, Spain

Received: November 2008

financial risk analysis, comes from the New Basel Capital Accord (Basel II), published in 1999 and revised in 2004 by the Basel Committee for Banking Supervision (BCBS). This accord consists of three parts, called pillars. They constitute a universal theoretical framework for the procedures to be followed by credit companies in order to guarantee minimal capital requirements, called *statistical provisions for insolvency* (SPI).

Pillar I of the new accord establishes the parameters that play some role in the credit risk of a financial company. These are the probability of default, *PD*, the exposition after default, *EAD*, and the loss given default, *LGD*. The quantitative methods that financial entities can use are those used for computing credit risk parameters and, more specifically, for computing *PD*. These are the standard method and the internal ratings based method (IRB). Thus, credit companies can elaborate and use their own credit qualification models and, by means of them, conclude the Basel implementation process, with their own estimations of SPI.

There is an extensive literature on quantitative methods for credit risk, since the classical *Z*-score model introduced by Altman (1968). Nowadays there exist plenty of approaches and perspectives for modelling credit risk starting from *PD*. Most of them have provided better predictive powers and classification error rates than Altman's discriminant model, for credit solicitors (*application scoring*), as well as for those who are already clients of the bank (*behavioural scoring*). This is the case of logistic regression models, artificial neural networks (*ANN*), support vector machines (*SVM*), as well as hybrid models, as mixtures of parametric models and *SVM*. For the reader interested in a more extended discussion on the evolution of these techniques over the past 30 years we mention the work by Altman and Saunders (1998), Saunders (1999), Crouhy et al. (2000), Hand (2001), Hamerle et al. (2003), Hanson and Schuermann (2004), Wang et al. (2005), and Chen et al. (2006).

The main aim of this paper is to introduce an alternative approach for modelling credit risk. More specifically, we will focus on estimating *PD* for consumer credits and personal credits using survival analysis techniques.

The idea of using survival analysis techniques for constructing credit risk models is not new. It started with the paper by Narain (1992) and, later, was developed by Carling et al. (1998), Stepanova and Thomas (2002), Roszbach (2003), Glennon and Nigro (2005), Allen and Rose (2006), Baba and Goko (2006), Malik and Thomas (2006) and Beran and Djaïdja (2007). A common feature of all these papers is that they use parametric or semiparametric regression techniques for modelling the time to default (*duration models*), including exponential models, Weibull models and Cox's proportional hazards models, which are very common in this literature. The model established for the time to default is then used for modelling *PD* or constructing the scoring discriminant function.

In this paper we propose a basic idea to estimate *PD*, which is performed by three different methods. The first one is based on Cox's proportional hazards model, the second one uses generalized linear models, while the third one consists in using a random design nonparametric regression model. In all the cases, some random right

censoring mechanism appears in the model, so survival analysis techniques are natural tools to be used.

The conditional survival function used for modelling credit risk opens an interesting perspective to study default. Rather than looking at default or not, we look at the time to default, given credit information of clients (endogenous covariates) and considering the indicators for the economic cycle (exogenous covariates). Thus, the default risk is measured via the conditional distribution of the random variable time to default, T, given a vector of covariates, X. The variable T is not fully observable due to the censoring mechanism.

In practice, since the proportion of defaulted credits is small, the proportion of censored data is large, which may cause poor performance of the statistical methods. On the other hand, the sample size is typically very large. This alleviates somehow the problem of the large proportion of censoring.

In order to estimate empirically the conditional distribution function of the time to default, we use the generalized product-limit estimator by Beran (1981). This estimator has been extensively studied by Dabrowska (1987), Dabrowska (1989), González-Manteiga and Cadarso-Suárez (1994), Van Keilegom and Veraverbeke (1996), Iglesias-Pérez and González-Manteiga (1999), Li and Datta (2001), Van Keilegom et al. (2001) and Li and Van Keilegom (2002), among other authors.

The rest of the paper proceeds as follows. Section 2 presents some conditional functions, often used in survival analysis, and explains how they can be used for credit risk analysis. The estimation of the probability of default is considered in Section 3, under different models: Cox's proportional hazards model, generalized linear models and a nonparametric model. Special attention is given to the study of the theoretical properties of the nonparametric estimator for *PD*, denoted by $\widehat{PD}^{NPM}$. Its asymptotic bias and variance, as well as uniform consistency and asymptotic normality are stated in Section 4. An application to a real data set, with a brief discussion about the empirical results obtained, is presented in Section 5. Finally, Section 6 contains the proofs of the results included in Section 4.

2 Conditional survival analysis in credit risk

The use of survival analysis techniques to study credit risk, and more particularly to model *PD*, can be motivated via Figure 1. It presents three common situations that may occur in practice when a credit company observes the "lifetime" of a credit.

Let us consider the interval $[0, \tau]$ as the horizon of the study. Case (a) shows a credit with default before the endpoint of the time under study (τ). In this case, the lifetime of the credit, T, which is the time to default of the credit, is an observable variable. Cases (b) and (c) show two different situations. In both of them it is not possible to observe the time instant when a credit enters into default, which causes a lack of information coming

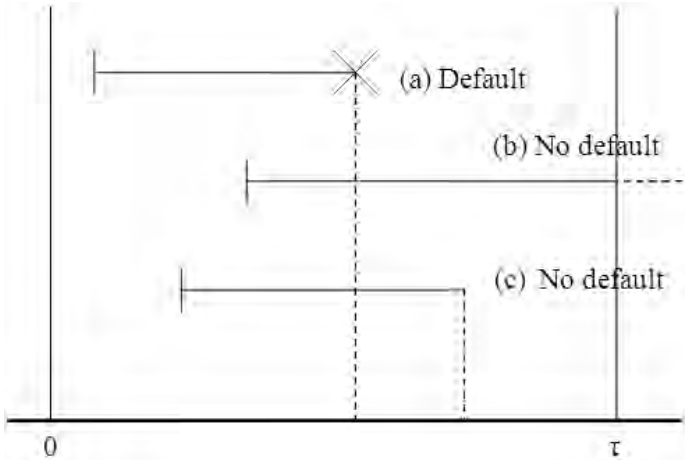

Figure 1: *Time to default in consumer credit risk.*

from right censoring. In case (b) it is only the time from the start of the credit to the end of the study, while (c) accounts for a situation where anticipated cancellation or the end of the credit occurs before default.

The available information to model the *PD* is a sample of n iid random variables $\{(Y_1,X_1,\delta_1),\ldots,(Y_n,X_n,\delta_n)\}$, of the random vector $\{Y,X,\delta\}$, where $Y=\min\{T,C\}$ is the observed maturity, T is the time to default, C is the time to the end of the study or anticipated cancellation of the credit, $\delta=I(T\leq C)$ is the indicator of noncensoring (default) and X is a vector of explanatory covariates. In this survival analysis setting we will assume that there exists an unknown relationship between T and X. We will also assume that the random variables T and C are conditionally independent given X.

In the previous setup it is possible to characterize completely the conditional distribution of the random variable T using some common relations in survival analysis. Thus the conditional survival function, $S(t|x)$, the conditional hazard rate, $\lambda(t|x)$, the conditional cumulative hazard function, $\Lambda(t|x)$, and the conditional cumulative distribution function, $F(t|x)$, are related as follows:

$$S(t|x)=P(T>t|X=x)=\int_t^{\infty} f(u|x)du$$

$$\lambda(t|x)=\lim_{\Delta t\to 0}\frac{P(t\leq T<t+\Delta t|T\geq t,X=x)}{\Delta t}=\frac{f(t|x)}{S(t|x)}$$

$$\Lambda(t|x)=\int_0^t \lambda(u|x)du=\int_0^t \frac{f(t|x)}{S(t|x)}du$$

$$S(t|x) = e^{-\Lambda(t|x)}$$

$$F(t|x) = 1 - S(t|x)$$

3 Probability of default in consumer portfolio

In the literature devoted to credit risk analysis there are not many publications on modelling the credit risk in consumer portfolios or personal credit portfolios. Most of the research deals with measuring credit risk by *PD* modelling in portfolios of small, medium and large companies, or even for financial companies. There exist, however, several exceptions. In the works by Carling et al. (1998), Stepanova and Thomas (2002) and Malik and Thomas (2006), the lifetime of a credit is modelled with a semiparametric regression model, more specifically with Cox's proportional hazards model.

In the following we present three different approaches to model the probability of default, *PD*, using conditional survival analysis. All the models are based on writing *PD* in terms of the conditional distribution function of the time to default. Thus *PD* can be estimated, using this formula, either by (i) Cox's proportional hazards model, where the estimator of the survival function is obtained by solving the partial likelihood equations in Cox's regression model, which gives $\widehat{PD}^{PHM}$, by (ii) a generalized linear model, with parameters estimated by the maximum likelihood method, which gives $\widehat{PD}^{GLM}$, or by (iii) using the nonparametric conditional distribution function estimator by Beran, which gives the nonparametric estimator of the default probability, $\widehat{PD}^{NPM}$.

3.1 Modelling the probability of default via the conditional distribution function

Following Basel II, credit scoring models are used to measure the probability of default in a time horizon $t+b$ from a maturity time, t. A typical value is $b = 12$ (in months). Thus, the following probability has to be computed:

$$\begin{aligned} PD(t|x) &= P(t \leq T < t+b | T \geq t, X = x) \\ &= \frac{P(T < t+b|X=x) - P(T \leq t|X=x)}{P(T \geq t|X=x)} \\ &= \frac{F(t+b|x) - F(t|x)}{1 - F(t|x)} = 1 - \frac{S(t+b|x)}{S(t|x)} \end{aligned} \tag{1}$$

where t is the observed maturity for the credit and x is the value of the covariate vector, X, for that credit.

3.2 Proportional hazards model

In this section, a semiparametric approach to perform the study of *PD* is given. Here we use Cox's proportional hazards approach to model the conditional survival function $S(t|x)$. The key in this method rests on the estimation of the cumulative conditional hazard function, $\Lambda(t|x)$, using maximum likelihood.

We follow the idea introduced by Narain (1992) for the estimation of $S(t|x)$, but we apply it in the definition of *PD*, as we have stated above in formula (1). The objective is to build a conditional model for the individual $PD(t|x)$, which is defined in terms of $\Lambda(t|x)$. In order to describe $\widehat{PD}^{PHM}$, we define the following expressions relative to Cox's regression theory.

The estimator of the conditional hazard rate function is defined as:

$$\hat{\lambda}(t|x) = \hat{\lambda}_0(t) \exp\left(x^{\mathrm{T}} \hat{\boldsymbol{\beta}}\right),$$

where $\hat{\lambda}_0(t)$ is an estimator of the hazard rate baseline function, $\lambda_0(t)$, and $\hat{\boldsymbol{\beta}}$ is an estimator of the parameter vector, $\boldsymbol{\beta}$.

Thus, under the assumption of a proportional hazards model, *PD* is estimated by:

$$\widehat{PD}^{PHM}(t|x) = \frac{\hat{F}_{\hat{\boldsymbol{\beta}}}(t+b|x) - \hat{F}_{\hat{\boldsymbol{\beta}}}(t|x)}{1 - \hat{F}_{\hat{\boldsymbol{\beta}}}(t|x)} = 1 - \frac{\hat{S}_{\hat{\boldsymbol{\beta}}}(t+b|x)}{\hat{S}_{\hat{\boldsymbol{\beta}}}(t|x)}, \tag{2}$$

where

$$1 - \hat{F}_{\hat{\boldsymbol{\beta}}}(t|x) = \hat{S}_{\hat{\boldsymbol{\beta}}}(t|x) = \exp\left(-\hat{\Lambda}(t|x)\right)$$

The estimation method for this model consists of two steps. In the first step the cumulative baseline hazard function, $\Lambda_0(t)$, is estimated by:

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{1\{Y_i \leq t, \delta_i = 1\}}{\sum_{j=1}^{n} 1\{Y_j \geq Y_i\}},$$

then the parameter $\boldsymbol{\beta}$ is estimated by

$$\hat{\boldsymbol{\beta}}^{PHM} = \arg\max_{\beta} L(\boldsymbol{\beta}),$$

where the partial likelihood function is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp\left(x_i^{\mathrm{T}} \boldsymbol{\beta}\right)}{\left(\sum_{j=1}^{n} 1_{\{Y_j > Y_i\}} \exp\left(x_j^{\mathrm{T}} \boldsymbol{\beta}\right)\right)}$$

Thus, the conditional cumulative hazard function estimator is given by

$$\hat{\Lambda}(t|x) = \int_0^t \hat{\lambda}(s|x)ds = \exp\left(x^\mathsf{T}\hat{\boldsymbol{\beta}}^{PHM}\right)\hat{\Lambda}_0(t).$$

The asymptotic properties of this estimator can be found, for instance, in the book by Fleming and Harrington (1991). As a consequence of these, similar properties can be obtained for the estimator of the default probability defined in (2).

Remark 1 *Narain (1992) and many other authors defined the probability of default as the complement of the conditional survival function evaluated at the forecast horizon,* $1-S(t+b|x)$*. According to this, the formulation by Narain does not take into account the fact that the credit should not be into default at maturity* t*.*

3.3 Generalized linear model

A generalized linear model can be assumed for the lifetime distribution:

$$P(T \leq t|X = x) = F_\theta(t|x) = g\left(\theta_0 + \theta_1 t + \theta^\mathsf{T} x\right),$$

where $\theta = (\theta_2, \theta_3, \ldots, \theta_{p+1})^\mathsf{T}$ is a p-dimensional vector and g is a known link function, like the logistic or the probit function. Thus, this model characterizes the conditional distribution of the lifetime of a credit, T, in terms of the unknown parameters. Once this parameters are estimated, an estimator of the conditional distribution function is obtained, $F_{\hat{\theta}}$, and, finally, an estimator of PD can be computed by plugging this estimator in equation (1), i.e.

$$\widehat{PD}^{GLM}(t|x) = \frac{F_{\hat{\theta}}(t+b|x) - F_{\hat{\theta}}(t|x)}{1 - F_{\hat{\theta}}(t|x)} = 1 - \frac{S_{\hat{\theta}}(t+b|x)}{S_{\hat{\theta}}(t|x)},$$

where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{GML}$ is the maximum likelihood estimator of the parameter vector.

Let us consider the one-dimensional covariate case. Then $\theta = \theta_2$ and the conditional distribution given by the model is $F(t|x) = g(\theta_0 + \theta_1 t + \theta_2 x)$, with density $f(t|x) = \theta_1 g'(\theta_0 + \theta_1 t + \theta_2 x)$. Since we are given a random right censored sample, the conditional likelihood function is a product of terms involving the conditional density, for the uncensored data, and the conditional survival function, for the censored data:

$$L(Y, X, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(Y_i|X_i)^{\delta_i}\left(1 - F(Y_i|X_i)\right)^{1-\delta_i},$$

where Y_i is the maturity of the i-th credit and δ_i is the indicator of default for the i-th credit. Thus, the log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \ln(L(Y,X,\boldsymbol{\theta})) = \sum_{i=1}^{n} [\delta_i \ln(f(Y_i|X_i)) + (1-\delta_i)\ln(1-F(Y_i|X_i))]$$
$$= \sum_{i=1}^{n} [\delta_i \ln(\theta_1 g'(\theta_0+\theta_1 Y_i+\theta_2 X_i)) + (1-\delta_i)\ln(1-g(\theta_0+\theta_1 Y_i+\theta_2 X_i))]$$
$$= \sum_{i=1}^{n} \delta_i [\ln(\theta_1) + \ln(g'(\theta_0+\theta_1 Y_i+\theta_2 X_i))] + \sum_{i=1}^{n} (1-\delta_i)\ln(1-g(\theta_0+\theta_1 Y_i+\theta_2 X_i))$$

Finally, the estimator is found as the maximizer of the log-likelihood function:

$$\hat{\boldsymbol{\theta}}^{GML} = \arg\max_{\theta} \ell(\boldsymbol{\theta}).$$

The works by Jorgensen (1983) and McCullagh and Nelder (1989) deal with generalized linear models in a regression context. These models can be adapted to the conditional distribution function setup.

3.4 Nonparametric conditional distribution estimator

First of all a nonparametric estimator of the conditional distribution function is obtained. This estimator, say $\hat{S}_h(t|x)$, is used to derive an estimator of $PD(t|x)$, say $\widehat{PD}^{NPM}(t|x)$, for the desired values of t and x.

Since we have a sample of right censored data for the lifetime distribution of a credit, we use the estimator proposed by Beran (1981) for the conditional survival function of T given $X = x$:

$$\hat{S}_h(t|x) = \prod_{i=1}^{n} \left(1 - \frac{1_{\{Y_i \le t, \delta_i = 1\}} B_{ni}(x)}{1 - \sum_{j=1}^{n} 1_{\{Y_j < Y_i\}} B_{nj}(x)}\right),$$

where Y_i is the observed lifetime of the i-th credit, δ_i is the indicator of observing default of the i-th credit (uncensoring) and X_i is the vector of explanatory covariates for the i-th credit. The terms $B_{ni}(x)$ are Nadaraya-Watson nonparametric weights:

$$B_{ni}(x) = \frac{K((x-X_i)/h)}{\sum_{j=1}^{n} K((x-X_j)/h)},\ 1 \le i \le n,$$

and $h \equiv h_n$ is the smoothing parameter that tends to zero as the sample size tends to infinity.

To estimate the probability of default at time t given a covariate vector x, we replace, in (1), the theoretical value of the conditional survival function by its estimator $\hat{S}_h$:

$$\widehat{PD}^{NPM}(t|x) = \frac{\hat{F}_h(t+b|x) - \hat{F}_h(t|x)}{1-\hat{F}_h(t|x)} = 1 - \frac{\hat{S}_h(t+b|x)}{\hat{S}_h(t|x)} \tag{3}$$

The asymptotic properties of this estimator will be studied in the next section.

4 Asymptotic results for the nonparametric approach

The asymptotic properties for the nonparametric estimator of the default probability, $\widehat{PD}^{NPM}$, have been obtained from the analogous properties for the conditional distribution function estimator under censoring, already obtained by Dabrowska (1989), Iglesias-Pérez and González-Manteiga (1999), Van Keilegom and Veraverbeke (1996) and Van Keilegom et al. (2001).

Using equation (3) the asymptotic bias, variance and mean squared error of the estimator $\widehat{PD}^{NPM}$ can be obtained via some expansions. Consistency and asymptotic normality can also be derived.

To simplify our notation, let us define $\varphi(t|x) = PD(t|x)$ and $\hat{\varphi}_n(t|x) = \widehat{PD}^{NPM}(t|x)$. Then, the nonparametric estimator of the default probability function is

$$\hat{\varphi}_n(t|x) = 1 - \frac{\hat{S}_h(t+b|x)}{\hat{S}_h(t|x)}. \tag{4}$$

Before stating the asymptotic results concerning $\hat{\varphi}_n$ we need to present some definitions and assumptions. Most of these assumptions were already required by Iglesias-Pérez and González-Manteiga (1999) and Dabrowska (1989) to obtain their results.

The function $G(t|x) = P(C \leq t|X = x)$ is the conditional distribution of the censoring random variable given the covariate X and $H(t|x) = P(Y \leq t|X = x)$ is the conditional distribution of the observed lifetime of the credit given the covariate X. The random lifetime, T, and the censoring time, C, are conditionally independent given the covariate X. As a consequence, $1 - H(t|x) = (1 - F(t|x))(1 - G(t|x))$. The conditional subdistribution function of the observed lifetime for default credits is denoted by $H_1(t|x) = P(Y \leq t, \delta = 1|X = x) = \int_0^t (1 - G(u|x))dF(u|x)$. Similarly, for nondefaulted credits, $H_0(t|x) = P(Y \leq t, \delta = 0|X = x) = \int_0^t (1 - F(u|x))dG(u|x)$. The distribution function and the density function of the covariate X are denoted by $M(x)$ and $m(x)$. The set $\Omega_X = \{x \in \mathbb{R}^+ : m(x) > 0\}$ will denote the support of m. The lower and upper endpoints of the support of any distribution function L will be denoted by $\underline{\tau}_L = \inf\{t : L(t) > 0\}$ and $\overline{\tau}_L = \sup\{t : L(t) < 1\}$.

The following assumptions are needed for the asymptotic results.

A.1 The kernel K is a symmetric density function with support $[-1, 1]$ and bounded variation.

A.2 Let us consider Ω_X, the support of the density m, and let $I = [x_1, x_2]$ be an interval contained in Ω_X, such that there exist $\alpha, \beta, \delta > 0$ with $\alpha\delta \leq \beta\delta < 1$,

$$\alpha \leq \inf\{m(x) : x \in I_\delta\} \leq \sup\{m(x) : x \in I_\delta\} \leq \beta,$$

where $I_\delta = [x_1 - \delta, x_2 + \delta]$. Then the functions $m'(x)$ and $m''(x)$ are continuous and bounded in the set I_δ.

A.3 There exist positive real numbers θ and τ_H^*, such that

$$0 < \theta \leq \inf_{0 \leq t \leq \tau_H^*} \{1 - H(t|x) : x \in I_\delta\}$$

A.4 The functions $H'(t|x) = \frac{\partial H(t|x)}{\partial x}$, $H''(t|x) = \frac{\partial^2 H(t|x)}{\partial x^2}$, $H_1'(t|x) = \frac{\partial H_1(t|x)}{\partial x}$ and $H_1''(t|x) = \frac{\partial^2 H_1(t|x)}{\partial x^2}$ exist, are continuous and bounded in $(t,x) \in [0, +\infty) \times I_\delta$.

A.5 The functions $\dot{H}(t|x) = \frac{\partial H(t|x)}{\partial t}$, $\ddot{H}(t|x) = \frac{\partial^2 H(t|x)}{\partial t^2}$, $\dot{H}_1(t|x) = \frac{\partial H_1(t|x)}{\partial t}$, $\ddot{H}_1(t|x) = \frac{\partial^2 H_1(t|x)}{\partial t^2}$ exist, are continuous and bounded in $(t,x) \in [0, \tau_H^*) \times I_\delta$.

A.6 The functions $\dot{H}'(t|x) = \frac{\partial^2 H(t|x)}{\partial x \partial t} = \frac{\partial^2 H(t|x)}{\partial t \partial x}$, $\dot{H}_1'(t|x) = \frac{\partial^2 H_1(t|x)}{\partial x \partial t} = \frac{\partial^2 H_1(t|x)}{\partial t \partial x}$ exist, are continuous and bounded in $(t,x) \in [0, \tau_H^*) \times I_\delta$.

A.7 The smoothing parameter h satisfies $h \to 0$, $(\ln n)^3/nh \to 0$ and $nh_n^5/\ln n = O(1)$, when $n \to \infty$.

The consistency and asymptotic normality of the nonparametric estimator $\widehat{\varphi}_n$ are stated in the next two theorems. The proofs of these results are given in Section 6.

Theorem 1 *Fix some t and x for which $0 < \varphi(t|x) < 1$. Under the assumptions A.1-A.7, $\hat{\varphi}_n(t|x)$ is a strongly consistent estimator of the default probability function, $\varphi(t|x)$. Moreover if $b < \tau_H^*$ and $\inf_{x \in I} S(\tau_H^*|x) > 0$, the consistency is uniform in $(t,x) \in [0, \tau_H^* - b] \times I$, i.e.*

$$\sup_{t \in [0, \tau_H^* - b]} \sup_{x \in I} |\hat{\varphi}_n(t|x) - \varphi(t|x)| \to 0 \ a.s.$$

Theorem 2 *Assume conditions A.1-A.7. Then the mean squared error of the nonparametric estimator for the default probability is*

$$MSE\left(\hat{\varphi}_n(t|x)\right) = b(t|x)^2 h^4 + \frac{1}{nh} v(t|x) + o\left(h^4 + \frac{1}{nh}\right), \tag{5}$$

where

$$b(t|x) = \frac{1}{2} c_K \left(1 - \varphi(t|x)\right) B_H(t, t+b|x), \tag{6}$$

$$v(t|x) = \frac{d_K D_H(t, t+b|x)}{m(x)} \left(1 - \varphi(t|x)\right)^2, \tag{7}$$

$c_K = \int K(u)^2 \, du$, $d_K = \int u^2 K(u) \, du$,

$$\begin{aligned} B_H(t, t+b|x) = & \int_t^{t+b} \left[\ddot{H}(s|x) + 2\frac{m'(x)}{m(x)} \dot{H}(s|x) \right] dH_1(s|x) \\ & + \left(1 + 2\frac{m'(x)}{m(x)}\right) \int_t^{t+b} \frac{d\dot{H}_1(s|x)}{1 - H(t|x)}, \end{aligned} \tag{8}$$

$$D_H(t|x) = \int_0^t \frac{dH_1(s|x)}{\left(1 - H(s|x)\right)^2}. \tag{9}$$

Furthermore, if $nh^5 \to c \in (0, \infty)$*, the limit distribution of* $\hat{\varphi}_n(t|x)$ *is given by*

$$\sqrt{nh}\left(\hat{\varphi}_n(t|x) - \varphi(t|x)\right) \xrightarrow{d} N\left(c^{1/2} b(t|x), v(t|x)\right).$$

Remark 2 *As a consequence, the bandwidth that minimizes the dominant terms of the MSE in (5) is*

$$h_0 = \left(\frac{v(t|x)}{4b(t|x)^2}\right)^{1/5} n^{-1/5}. \tag{10}$$

5 Application to real data

In this section we apply the estimation methods given in Section 3 to a real data set. Our goal is to show the results obtained from the application of the three models to the estimation of default probabilities in a sample of consumer loans. An empirical comparison between the models is given through the descriptive statistics as well as the estimated default rate curves. In all cases, the curves were constructed taking into account the recommendations from the Basel statements, i.e., *PD* estimates with maturity of one year forward.

The data consist of a sample of 25 000 consumer loans from a Spanish bank registered between July 2004 and November 2006. To preserve confidentiality, the data

were selected in order to provide a large distortion in the parameters describing the true solvency situation of the bank.

The sample represents two populations, non-defaulted loans and defaulted loans, where the observed cumulative default rate was 7.2%. The variables treated here are the following:

Y = maturity or loan lifetime. Here, maturity means time to default (T), when time is uncensored or time to withdrawal (C), in any other case. Time was measured in months.

X = scoring (credit ratio) observed for each debtor. Its range lies inside the interval $[0, 100]$. In this paper, X is an univariate endogenous measure of propensity to default. The closer to zero the better the credit behaviour.

δ = default indicator (uncensoring indicator). This variable takes value 1 if loan defaults or 0 if not.

Figure 2 shows that the scoring characteristics of debtors are clearly different in the two groups (defaulted and non-defaulted). The moment-based characteristics like the kurtosis (2.68 and 4.29) and the skewness (0.51 and 1.37) of these two subsamples are very different each other and they also reflect non-normal distributions. A large proportion (about 75%) of debtors belonging to the sample of non-defaulted loans show a credit scoring varying between 0.0 and 11.07. This whole range is below the first quartile (approximately 20.93) of the scoring in the group of defaulted loans.

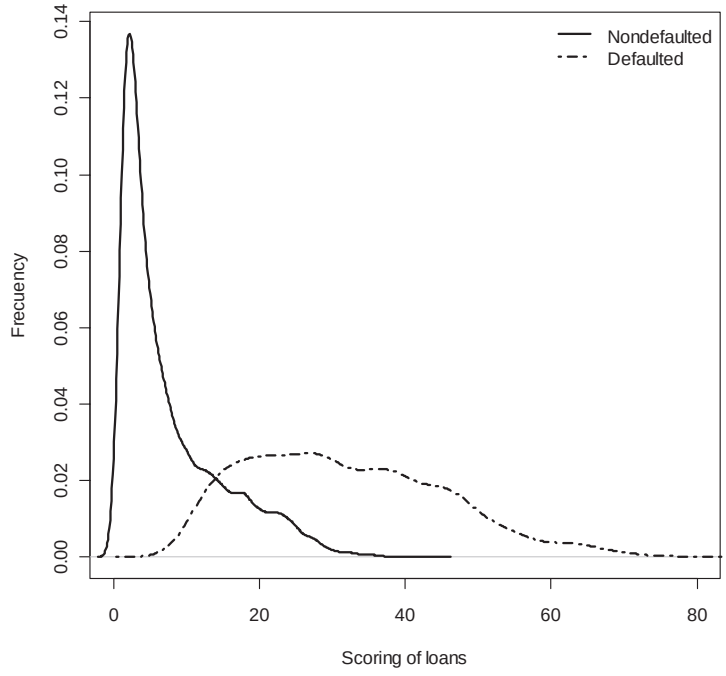

Figure 2: *Kernel density estimates for defaulted and non-defaulted loans.*

Table 1: *Descriptive statistics for maturity and covariate (X) in defaulted loans (DL), non-defaulted loans (NDL) and aggregated loans (AL).*

Sample		min.	1st. Q.	median	mean	3rd. Q.	max.
DL	maturity (T)	0.033	2.933	5.500	7.458	11.15	24.767
	X	8.398	20.295	30.066	31.817	41.167	77.819
NDL	maturity (C)	0.000	6.767	11.367	13.455	20.033	29.500
	X	0.150	2.412	4.857	7.688	11.070	43.920
AL	maturity (Y)	0.000	6.500	10.870	13.020	19.570	29.500
	X	0.150	2.540	5.440	9.425	13.405	77.819

The data show that the random variable X is a reasonable predictor to study loan default. This is also evident when observing the descriptive statistics for both groups of loans in Table 1.

Figure 3 shows curves for the empirical default rates obtained directly from the sample. These curves can be thought as the result of a naïve nonparametric estimator for the unconditional default rates curves. The study of this estimator is not the goal of this paper. Focusing the attention in the right panel in Figure 3, it is clear that the unconditional estimates of *PD* become constant when the loan maturity gets large. Naive approximations to *PD* do not behave well because of the lack of sensitivity to variations in the scoring characteristics of debtors. This result show that the unconditional approach may not be used to predict *PD* because the estimation accuracy on the right tail seems to be poor. This fact motivates the use of the conditional framework to obtain more realistic estimations for *PD*.

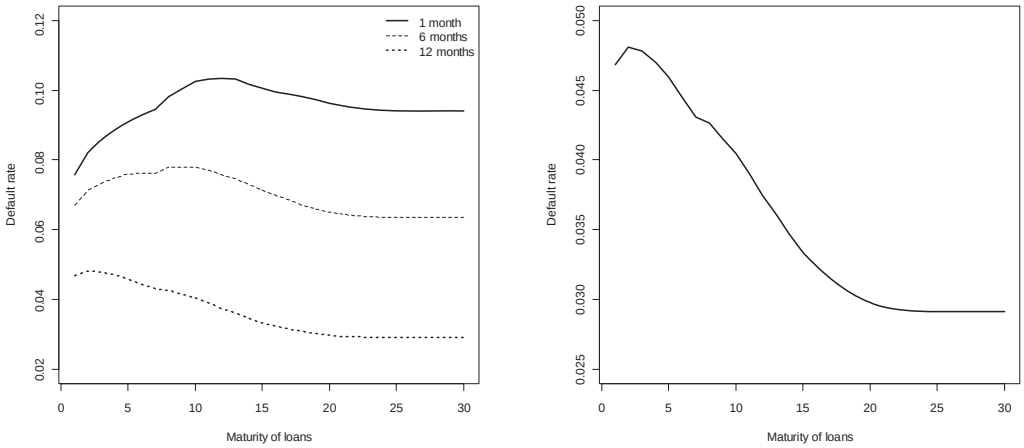

Figure 3: *Empirical default rates with different forecasting periods. Left panel shows default rates curves for 1, 6, and 12 months forward horizons, while the right panel shows the particular case of a default rate curve with 1 year forward horizon, which is a very common decision tool in credit risk analysis.*

5.1 Empirical results and discussion

The plots included in this section give a graphical description of the estimators proposed in this paper concerned with the conditional approach in consumer credit risk. All these results show that a reasonable improvement can be achieved when a survival analysis approach is used to model the credit quality in terms of "lifetime of loans".

5.1.1 Results for the proportional hazards model

Estimating the *PD* under the proportional hazards model presents clear differences with the results for the unconditional setting (Figure 3). It is easy to see that a clear disadvantage of using an unconditional approach is that the *PD* forecasts do not change with X. The conditional approach gives more realistic estimates using the scoring information, which is a reasonable covariate, as was established at the beginning of this section. The covariate X explains the propensity to defaults in loan obligations. Figure 4 shows that the *PD* estimates increase as the customer scoring increases.

A careful look at Figure 4 shows that the estimator of *PD* is zero when the time to default gets close to the maximum of the maturity observed in the sample (approximately 25). This effect on the *PD* curve is due to the heavy censoring and the lack of depth in the sample. As a consequence, the accuracy of the estimator at the right tail of *PD* is poor. Nevertheless, Cox's proportional hazards model gives more realistic default probabilities than the unconditional approximation (see Figure 3) when the bank previously knows the scoring characteristics of the portfolio customers.

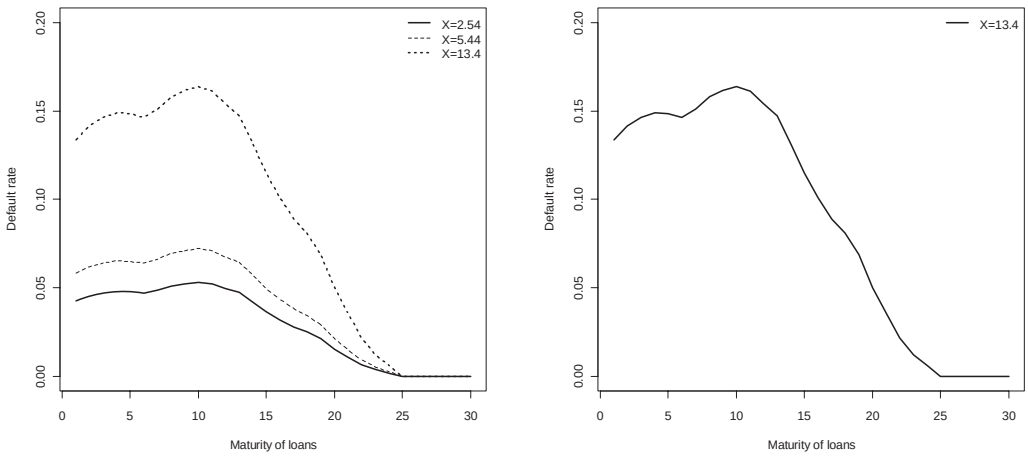

Figure 4: *$\widehat{PD}$ with maturity 1 year forward given $X = 2.54, 5.44, 13.4$ (left panel) and given the mean value of X (right panel).*

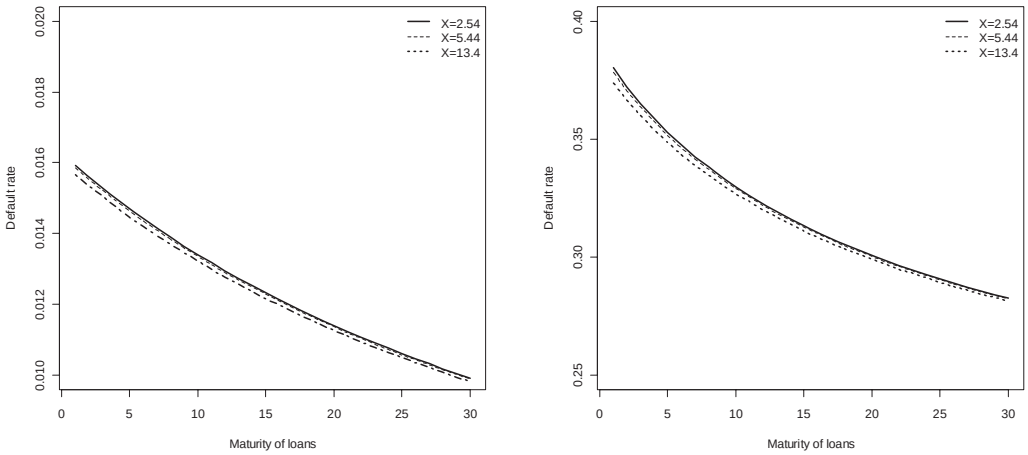

Figure 5: *$\widehat{PD}$ estimated with the Pareto (left panel) and Snedecor's F (right panel) link function.*

5.1.2 Results for the generalized linear model

Figure 5 show the results obtained for the *PD* estimated with the *GLM* model using two parametric links: Pareto and Snedecor's F. The range of the estimated *PD* lies within the interval $[0.0, 0.016]$ when the link function is Pareto and grows up to the interval $[0.0, 0.378]$ when the link function is $F_{10,50}$, as it can be seen in Table 2. The *PD* curves obtained with this model are exponentially decreasing, as expected, but in this case it seems that there is no a significantly contribution of the variable X in the accuracy of the estimated default probability curves. Furthermore, the estimated curves are all above the range of the observed default rate with maturity one year forward. The results achieved by using these two parametric links do not fit as well as expected to the data, when compared to the empirical default rate curves depicted in Figure 3. In spite of this, the *GLM* method may be useful to study the *PD* horizon in the long run.

Other link distributions belonging to the exponential family have been used to fit these data via *GLM*. The normal distribution, the Weibull distribution and the Cauchy distribution were used, among others. The results obtained were even worse than those presented in Figure 5 above.

5.1.3 Results for the nonparametric estimator

The results for the nonparametric method presented in (3) are collected in this subsection. In practice, we have used a k-nearest neighbour (KNN) type of bandwidth, which consists in fixing some positive integer k and define the parameter as follows:

$$h = h^{KNN}(x) = d(x, X_{[k]})$$

where $d(x, X_{[k]})$ is the k-th order statistic of the sample of distances from x to those X_i with $\delta_i = 1$. In other terms $h^{KNN}(x)$ is the k-th smallest distance from x to the uncensored observations of the X sample.

Figures 6-7 show the behaviour of the nonparametric estimator introduced in Section 3. In Figure 6 the value of the number of nearest neighbours has remained fixed ($k = 100$) and the estimator $\widehat{PD}(t|x)$ has been computed for three different values of X ($x = 2.54$, 5.44, 13.4). The reverse situation is showed in Figure 7, i.e., the curves $\widehat{PD}(t|x)$ were obtained for two fixed values of X ($x = 9.43$, 20) and varying the number of nearest neighbours ($k = 100$, 300, 500).

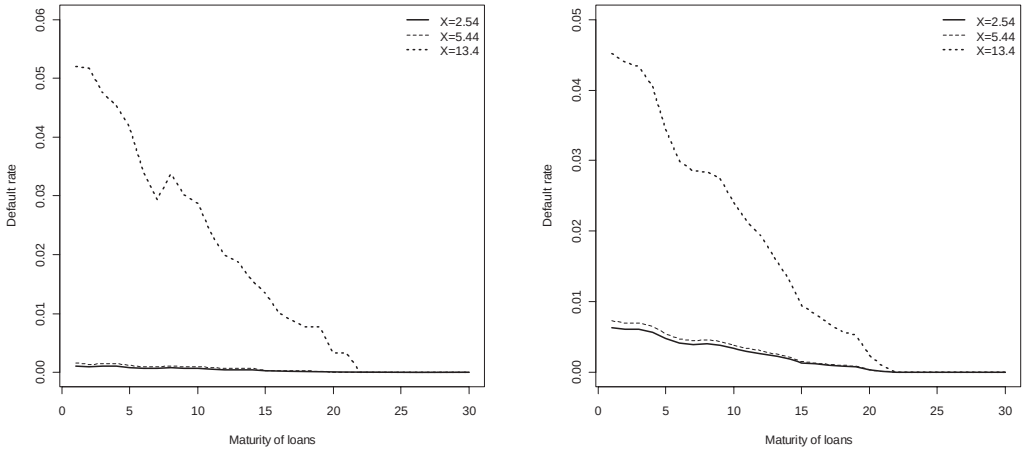

Figure 6: *$\widehat{PD}$ with fixed bandwidth parameter $k = 100$ (left panel) and $k = 400$ (right panel) given three scoring values.*

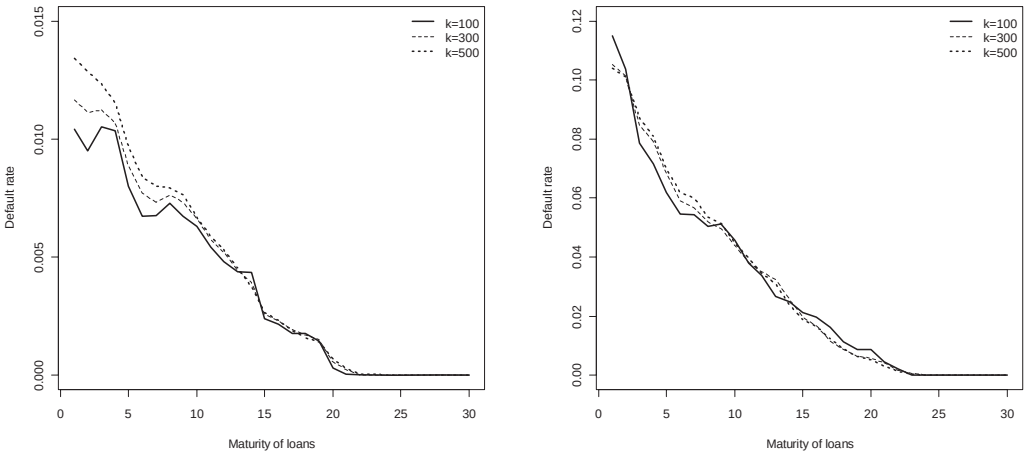

Figure 7: *$\widehat{PD}$ with three different bandwidth parameters, given $X = 9.43$ (left panel) and given $X = 20$ (right panel).*

The first two curves show much smaller values for PD when the values of X are close to or below the first quartile of the distribution. For $k = 100$ (see Figure 6, left panel) there is an apparent undersmoothing effect for the estimated default probability curve. The situation improves in the right panel of Figure 6. There, since $k = 400$, the $\widehat{PD}$ is much smoother. The estimates of the PD have a large sensitivity to small changes in the scoring variable. As a consequence the PD can be overestimated at the beginning of loan lifetime. A possible reason for this is the heavy censoring that usually affects consumer credit loans.

Table 2: *Descriptive statistics for the empirical default rates (EDR) and for the PD estimates obtained by Cox's proportional hazards model (PHM), the generalized linear model (GLM) and the nonparametric model (NPM).*

Model			min.	1st. Q.	median	mean	3rd. Q.	max.
		Maturity (months)						
		1	0.0758	0.0940	0.0947	0.0952	0.0994	0.1033
EDR		6	0.0636	0.0638	0.0692	0.0698	0.0755	0.0779
		12	0.0292	0.0292	0.0329	0.0359	0.0424	0.0481
		x						
		2.54	0.0000	0.0045	0.0343	0.0286	0.0138	0.0159
PHM		5.44	0.0000	0.0062	0.0467	0.0389	0.0138	0.0159
		13.4	0.0000	0.0147	0.1080	0.0891	0.0136	0.0157
	Link	x						
		2.54	0.0099	0.0110	0.0122	0.0125	0.0139	0.0159
	Pareto	5.44	0.0010	0.0109	0.0122	0.0124	0.0138	0.0159
GLM		13.4	0.0098	0.0108	0.0121	0.0123	0.0136	0.0157
		2.54	0.2826	0.2950	0.3120	0.3183	0.3370	0.4468
	$F_{10,50}$	5.44	0.2823	0.2946	0.3114	0.3176	0.3361	0.3784
		13.4	0.2815	0.2935	0.3098	0.3156	0.3336	0.3738
	k	x						
	100	2.54	0.0000	0.0000	0.0002	0.0004	0.0007	0.0012
	100	5.44	0.0000	0.0000	0.0003	0.0005	0.0010	0.0015
	100	13.4	0.0000	0.0000	0.0118	0.0175	0.0300	0.0520
	400	2.54	0.0000	0.0000	0.0012	0.0021	0.0039	0.0064
	400	5.44	0.0000	0.0001	0.0014	0.0024	0.0045	0.0073
NPM	400	13.4	0.0000	0.0001	0.0089	0.0152	0.0282	0.0452
	100	9.43	0.0000	0.0000	0.0023	0.0037	0.0067	0.0105
	300	9.43	0.0000	0.0000	0.0024	0.0040	0.0073	0.0117
	500	9.43	0.0000	0.0001	0.0025	0.0042	0.0079	0.0134
	100	20	0.0000	0.0005	0.0205	0.0301	0.0509	0.1149
	300	20	0.0000	0.0009	0.0183	0.0302	0.0514	0.1054
	500	20	0.0000	0.0006	0.0177	0.0306	0.0531	0.1040

Figure 7 includes the default probability conditional to just a single value of X, using three different levels of smoothness. Visual inspection of Figure 7 shows that, for a fixed bandwidth, the larger the scoring, the smoother the estimated *PD* curve. It is also clear that the variability of the *PD* reduces when the scoring gets large.

5.1.4 Comparison

A summary with a descriptive comparison of the three models is given in Table 2. Fixed values for the covariate X (first, second and third quartiles) were used for the conditional distributions. Of course, the empirical default rate does not depend on the value of X.

Although no goodness-of-fit tests have been applied for the proposed models, the results of the estimation can be checked by simple inspection of Figures 4–7 and the descriptive statistics collected in Table 2. The results for each model can be compared with those of the aggregated default rates in the whole portfolio. Such values should be considered as a reference value for the three models.

6 Proofs

Proof of Theorem 1

Recall equations (1) and (4). Let us write

$$\varphi(t|x) = 1 - \frac{P}{Q},$$
$$\hat{\varphi}_n(t|x) = 1 - \frac{\hat{P}}{\hat{Q}},$$

with $P = S(t+b|x)$, $Q = S(t|x)$, $\hat{P} = \widehat{S}_h(t+b|x)$ and $\hat{Q} = \widehat{S}_h(t|x)$. Using Theorem 2 in Iglesias-Pérez and González-Manteiga (1999) we have

$$\left(\hat{P}, \hat{Q}\right) \longrightarrow (P, Q) \text{ a.s.}$$

Since the function $g(x,y) = \dfrac{x}{y}$ is continuous in (P,Q), then we obtain

$$\hat{\varphi}_n(t|x) \longrightarrow \varphi(t|x) \text{ a.s.}$$

and the first part of the proof is concluded.

For the second part of the proof we use Corollary 2.1 in Dabrowska (1989) to obtain

$$\sup_{t\in\left[0,\tau_H^*-b\right]}\sup_{x\in I}\left|\hat{S}_h(t+b|x)-S(t+b|x)\right| \to 0 \text{ a.s.} \tag{11}$$

$$\sup_{t\in\left[0,\tau_H^*-b\right]}\sup_{x\in I}\left|\hat{S}_h(t|x)-S(t|x)\right| \to 0 \text{ a.s.} \tag{12}$$

We now use the following identity:

$$\frac{1}{z}=1-(z-1)+\cdots+(-1)^p(z-1)^p+(-1)^{p+1}\frac{(z-1)^{p+1}}{z}, \tag{13}$$

that is valid for any $p\in\mathbb{N}$. Applying (13) with $p=1$ and $\frac{1}{z}=\frac{Q}{\hat{Q}}$ we obtain:

$$\begin{aligned}1-\hat{\varphi}_n(t|x) &= \frac{\hat{P}}{\hat{Q}}=\frac{\hat{P}}{Q}\frac{Q}{\hat{Q}}\\ &= \frac{\hat{P}}{Q}\left[1-\left(\frac{\hat{Q}}{Q}-1\right)+\frac{Q}{\hat{Q}}\left(\frac{\hat{Q}}{Q}-1\right)^2\right]\\ &= \frac{\hat{P}}{Q}-\frac{\hat{P}\left(\hat{Q}-Q\right)}{Q^2}+\frac{\hat{P}}{\hat{Q}}\frac{\left(\hat{Q}-Q\right)^2}{Q^2},\end{aligned}$$

thus

$$\left|(1-\hat{\varphi}_n(t|x))-(1-\varphi(t|x))\right|\le A_1+A_2+A_3 \tag{14}$$

where

$$\begin{aligned}A_1 &= \frac{\left|\hat{P}-P\right|}{Q},\\ A_2 &= \frac{\hat{P}\left|\hat{Q}-Q\right|}{Q^2},\\ A_3 &= \frac{\hat{P}}{\hat{Q}}\frac{\left(\hat{Q}-Q\right)^2}{Q^2}.\end{aligned}$$

On the other hand if $x\in I$ and $t\le\tau_H^*-b$,

$$A_1 \le \frac{\sup_{t\in\left[0,\tau_H^*-b\right]}\sup_{x\in I}\left|\hat{S}_h(t+b|x)-S(t+b|x)\right|}{\inf_{x\in I}S(\tau_H^*|x)}, \tag{15}$$

$$A_2 \leq \frac{\sup\limits_{t\in\left[0,\tau_H^*-b\right]}\sup\limits_{x\in I}\left|\hat{S}_h(t|x)-S(t|x)\right|}{\inf\limits_{x\in I}S(\tau_H^*|x)^2}, \tag{16}$$

$$A_3 \leq \frac{\sup\limits_{t\in\left[0,\tau_H^*-b\right]}\sup\limits_{x\in I}\left|\hat{S}_h(t|x)-S(t|x)\right|^2}{\inf\limits_{x\in I}S(\tau_H^*|x)^2}. \tag{17}$$

Finally using (11) and (12) in (15), (16) and (17), equation (14) gives

$$\sup_{t\in\left[0,\tau_H^*-b\right]}\sup_{x\in I}|\hat{\varphi}_n(t|x)-\varphi(t|x)| \to 0 \text{ a.s.}$$

and the proof is concluded.

Proof of Theorem 2

To study the bias, we use (13) for $p=1$ and $\frac{1}{z}=\frac{E\left(\hat{Q}\right)}{\hat{Q}}$ to obtain:

$$\begin{aligned}
1-\hat{\varphi}_n(t|x) &= \frac{\hat{P}}{\hat{Q}}=\frac{\hat{P}}{E\left(\hat{Q}\right)}\frac{E\left(\hat{Q}\right)}{\hat{Q}} \\
&= \frac{\hat{P}}{E\left(\hat{Q}\right)}\left[1-\left(\frac{\hat{Q}}{E(\hat{Q})}-1\right)+\frac{E\left(\hat{Q}\right)}{\hat{Q}}\left(\frac{\hat{Q}}{E\left(\hat{Q}\right)}-1\right)^2\right] \\
&= \frac{\hat{P}}{E\left(\hat{Q}\right)}-\frac{\hat{P}\left(\hat{Q}-E\left(\hat{Q}\right)\right)}{\left(E\left(\hat{Q}\right)\right)^2}+\frac{\hat{P}}{\hat{Q}}\frac{\left(\hat{Q}-E\left(\hat{Q}\right)\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}.
\end{aligned} \tag{18}$$

As a consequence

$$E\left(1-\hat{\varphi}_n(t|x)\right)=A_1+A_2+A_3, \tag{19}$$

with

$$A_1 = \frac{E\left(\hat{P}\right)}{E\left(\hat{Q}\right)}, \tag{20}$$

$$A_2 = -\frac{Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^2}, \tag{21}$$

$$A_3 = \frac{E\left[\frac{\hat{P}}{\hat{Q}}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^2\right]}{\left(E\left(\hat{Q}\right)\right)^2}. \tag{22}$$

Theorem 2 and Corollary 3 in Iglesias-Pérez and González-Manteiga (1999) give

$$E\left(\hat{P}\right) = P\left(1-\frac{1}{2}c_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)\right), \tag{23}$$

$$E\left(\hat{Q}\right) = Q\left(1-\frac{1}{2}c_K A_H\left(t|x\right)h^2+o\left(h^2\right)\right), \tag{24}$$

where

$$\begin{aligned} A_H\left(t|x\right) = & \int_0^t \left[\ddot{H}(s|x)+2\frac{m'(x)}{m(x)}\dot{H}(s|x)\right] dH_1(s|x) \\ & + \left(1+2\frac{m'(x)}{m(x)}\right)\int_0^t \frac{d\dot{H}_1(s|x)}{1-H(t|x)}. \end{aligned} \tag{25}$$

Recall expressions (8) and (25). Then equations (23) and (24) can be used to find asymptotic expressions for (20) and (21):

$$\begin{aligned} A_1 = & \frac{P\left(1-\frac{1}{2}c_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)\right)}{Q\left(1-\frac{1}{2}c_K A_H\left(t|x\right)h^2+o\left(h^2\right)\right)} \\ = & \left(1-\varphi(t|x)\right)\frac{1-\frac{1}{2}c_K A_H\left(t+b|x\right)h^2+o\left(h^2\right)}{1-\frac{1}{2}c_K A_H\left(t|x\right)h^2+o\left(h^2\right)} \\ = & \left(1-\varphi(t|x)\right)\left[1-\frac{1}{2}c_K\left(A_H\left(t+b|x\right)-A_H\left(t|x\right)\right)h^2\right]+o\left(h^2\right) \\ = & \left(1-\varphi(t|x)\right)-\frac{1}{2}c_K B_H\left(t,t+b|x\right)\left(1-\varphi(t|x)\right)h^2+o\left(h^2\right), \end{aligned} \tag{26}$$

$$A_2 = -\frac{Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^2} = O\left(\frac{1}{nh}\right). \tag{27}$$

Finally, since $1-\hat{\varphi}_n(t|x) = \dfrac{\hat{P}}{\hat{Q}} \in [0,1]$, the term (22) can be easily bounded:

$$0 \le A_3 \le \frac{Var\left[\hat{Q}\right]}{\left(E\left(\hat{Q}\right)\right)^2} = O\left(\frac{1}{nh}\right). \tag{28}$$

Using (26), (27), (28) and (6) in (19) we get

$$E\left(\hat{\varphi}_n(t|x)\right)-\varphi(t|x) = b(t|x)h^2+o\left(h^2\right). \tag{29}$$

To deal with the variance we use (13) for $p=3$ and $\frac{1}{z}=\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}$ to obtain:

$$\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}=1+\sum_{i=1}^{3}(-1)^i\left(\frac{\hat{Q}^2-E\left(\hat{Q}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}\right)^i+\left(\frac{\hat{Q}^2-E\left(\hat{Q}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}\right)^4\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}. \tag{30}$$

On the other hand

$$\hat{Q}^2-\left(E\left(\hat{Q}\right)\right)^2=\left[\hat{Q}-E\left(\hat{Q}\right)\right]^2+2E\left(\hat{Q}\right)\left[\hat{Q}-E\left(\hat{Q}\right)\right]$$

gives

$$\begin{aligned}\left(\frac{\hat{Q}^2-\left(E\left(\hat{Q}\right)\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}\right)^i &= \sum_{j=0}^{i}\binom{i}{j}\left[\frac{\left(\hat{Q}-E\left(\hat{Q}\right)\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}\right]^j\left[\frac{2E\left(\hat{Q}\right)\left[\hat{Q}-E\left(\hat{Q}\right)\right]}{\left(E\left(\hat{Q}\right)\right)^2}\right]^{i-j}\\ &= \sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+i}}{\left(E\left(\hat{Q}\right)\right)^{j+i}}\end{aligned} \tag{31}$$

Substituting (30) in (31) we obtain:

$$\begin{aligned}\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2} &= 1+\sum_{i=1}^{3}(-1)^i\sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+i}}{\left(E\left(\hat{Q}\right)\right)^{j+i}}\\ &\quad+\sum_{j=0}^{4}\binom{4}{j}\frac{2^{4-j}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+4}}{\left(E\left(\hat{Q}\right)\right)^{j+4}}\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}.\end{aligned} \tag{32}$$

Equation (32) is useful to obtain an expansion for the second moment:

$$\begin{aligned}E\left[\left(1-\hat{\varphi}_n(t|x)\right)^2\right] &= E\left(\frac{\hat{P}^2}{\hat{Q}^2}\right)=E\left(\frac{\hat{P}^2}{\left(E\left(\hat{Q}\right)\right)^2}\frac{\left(E\left(\hat{Q}\right)\right)^2}{\hat{Q}^2}\right)\\ &= \frac{E\left[\left(\hat{P}-E\left(\hat{P}\right)\right)^2\right]}{\left(E\left(\hat{Q}\right)\right)^2}+\frac{E\left(\hat{P}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2}\\ &\quad+\sum_{i=1}^{3}(-1)^i\sum_{j=0}^{i}\binom{i}{j}\frac{2^{i-j}E\left[\hat{P}^2\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+i}\right]}{\left(E\left(\hat{Q}\right)\right)^{j+i+2}}\\ &\quad+\sum_{j=0}^{4}\binom{4}{j}\frac{2^{4-j}E\left[\frac{\hat{P}^2}{\hat{Q}^2}\left(\hat{Q}-E\left(\hat{Q}\right)\right)^{j+4}\right]}{\left(E\left(\hat{Q}\right)\right)^{j+4}}.\end{aligned} \tag{33}$$

Defining, for $i, j = 0, 1, \ldots$, the notation

$$A_{ij} = E\left[\left(\hat{P} - E\left(\hat{P}\right)\right)^i \left(\hat{Q} - E\left(\hat{Q}\right)\right)^j\right], \tag{34}$$

$$B_{ij} = E\left[\hat{P}^i \left(\hat{Q} - E\left(\hat{Q}\right)\right)^j\right], \tag{35}$$

$$C_i = \left(E\left(\hat{Q}\right)\right)^i, \tag{36}$$

$$D_{ij} = E\left[\left(1 - \hat{\varphi}_n(t|x)\right)^i \left(\hat{Q} - E\left(\hat{Q}\right)\right)^j\right] \tag{37}$$

and using

$$A_{2j} = B_{2j} - 2B_{10}A_{1j} + B_{10}^2 A_{0j},$$

expression (33) can be rewritten as

$$\begin{aligned} E\left[\left(1 - \hat{\varphi}_n(t|x)\right)^2\right] &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} + \sum_{i=1}^{3} (-1)^i \sum_{j=0}^{i} \binom{i}{j} 2^{i-j} \frac{B_{2\ i+j}}{C_{j+i+2}} \\ &\quad + \sum_{j=0}^{4} \binom{4}{j} 2^{4-j} \frac{D_{2\ j+4}}{C_{j+4}} \\ &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} \\ &\quad + \sum_{i=1}^{3} (-1)^i \sum_{j=0}^{i} \binom{i}{j} 2^{i-j} \frac{A_{2\ i+j} + 2B_{10}A_{1\ i+j} - B_{10}^2 A_{0\ i+j}}{C_{j+i+2}} \\ &\quad + \sum_{j=0}^{4} \binom{4}{j} 2^{4-j} \frac{D_{2\ j+4}}{C_{j+4}} \end{aligned} \tag{38}$$

It is easy, but long and tedious, to show that

$$E\left[\left(\hat{P} - E\left(\hat{P}\right)\right)^i\right] = o\left(\frac{1}{nh}\right), \text{ for } i \geq 3,$$

$$E\left[\left(\hat{Q} - E\left(\hat{Q}\right)\right)^i\right] = o\left(\frac{1}{nh}\right), \text{ for } i \geq 3.$$

Now recalling (34), (35), (36) and (37), and using Cauchy-Schwartz inequality and straight forward bounds, it can be proven that

$$A_{01} = A_{10} = 0, \tag{39}$$

$$A_{ij} = o\left(\frac{1}{nh}\right), \text{ whenever } i + j \geq 3, \tag{40}$$

$$B_{ij} = o\left(\frac{1}{nh}\right), \text{ for } j \geq 3, \tag{41}$$

$$D_{ij} = o\left(\frac{1}{nh}\right), \text{ for } j \geq 3. \tag{42}$$

Using (39), (40), (41) and (42) in (38), we conclude:

$$\begin{aligned} E\left[\left(1-\hat{\varphi}_n(t|x)\right)^2\right] &= \frac{A_{20}}{C_2} + \frac{B_{10}^2}{C_2} - \frac{4B_{10}A_{11}}{C_3} - \frac{3B_{10}^2A_{02}}{C_4} + o\left(\frac{1}{nh}\right) \\ &= \frac{Var\left(\hat{P}\right)}{\left(E\left(\hat{Q}\right)\right)^2} + \frac{E\left(\hat{P}\right)^2}{\left(E\left(\hat{Q}\right)\right)^2} - \frac{4E\left(\hat{P}\right)Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^3} \\ &\quad + \frac{3E\left(\hat{P}\right)^2Var\left(\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^4} + o\left(\frac{1}{nh}\right) \end{aligned} \tag{43}$$

On the other hand, plugging (18) in the term A_3 of expression (19), using (39), (40), (41) and (42) and some simple algebra gives:

$$\begin{aligned} E\left(1-\hat{\varphi}_n(t|x)\right) &= \frac{B_{10}}{C_1} - \frac{A_{11}}{C_2} + \frac{A_{12}+B_{10}A_{02}}{C_3} - \frac{A_{13}+B_{10}A_{03}}{C_4} + \frac{D_{14}}{C_4} \\ &= \frac{E\left(\hat{P}\right)}{E\left(\hat{Q}\right)} - \frac{Cov\left(\hat{P},\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^2} \\ &\quad + \frac{E\left(\hat{P}\right)Var\left(\hat{Q}\right)}{\left(E\left(\hat{Q}\right)\right)^3} + o\left(\frac{1}{nh}\right) \end{aligned} \tag{44}$$

The residual term $R'_n(y|x)$ in Theorem 2 of Iglesias-Pérez and González-Manteiga (1999) was proved to be uniformly negligible almost surely. A uniform rate for the moments of $R'_n(y|x)$ can be also obtained similarly. As a consequence of this, Theorem 2 and Corollary 4 in Iglesias-Pérez and González-Manteiga (1999) are applicable to obtain asymptotic expressions for the covariance structure of the process $\hat{S}_h(\cdot|x)$. This can be used to find and asymptotic expression for variances of $\hat{P}$ and $\hat{Q}$:

$$Var\left(\hat{P}\right) = \frac{1}{nh}v_1(t+b|x) + o\left(\frac{1}{nh}\right), \tag{45}$$

$$Var\left(\hat{Q}\right) = \frac{1}{nh}v_1(t|x) + o\left(\frac{1}{nh}\right), \tag{46}$$

$$Cov\left(\hat{P},\hat{Q}\right) = \frac{1}{nh}v_2(t,t+b|x) + o\left(\frac{1}{nh}\right), \tag{47}$$

where

$$v_1(t|x) = \frac{(1-F(t|x))^2}{m(x)} d_K C_H(t|x), \tag{48}$$

$$v_2(t,s|x) = \frac{(1-F(t|x))(1-F(s|x))}{m(x)} d_K C_H(t\wedge s|x), \tag{49}$$

$$C_H(t|x) = \int_0^t \frac{dH_1(s|x)}{(1-H(s|x))^2}. \tag{50}$$

Now using the orders found in (45), (46) and (47) in expressions (43) and (44) gives:

$$Var(\hat{\varphi}_n(t|x)) = Var(1-\hat{\varphi}_n(t|x)) = \frac{Var(\hat{P})}{(E(\hat{Q}))^2} - \frac{2E(\hat{P})Cov(\hat{P},\hat{Q})}{(E(\hat{Q}))^3} + \frac{(E(\hat{P}))^2 Var(\hat{Q})}{(E(\hat{Q}))^4} + o\left(\frac{1}{nh}\right).$$

Finally, the asymptotic expressions (23), (24), (45), (46) and (47) and the definitions (48), (49), (50), (9) and (7) can be used to conclude:

$$\begin{aligned} Var(\hat{\varphi}_n(t|x)) &= \frac{1}{nh}\frac{v_1(t+b|x)}{(S(t|x))^2} - \frac{2}{nh}\frac{v_2(t,t+b|x)S(t+b|x)}{(S(t|x))^3} + \\ &\quad \frac{1}{nh}\frac{v_1(t|x)(S(t+b|x))^2}{(S(t|x))^4} + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}\frac{d_K C_H(t|x)}{m(x)}\frac{(S(t+b|x))^2 - 2(S(t+b|x))^2 + (S(t+b|x))^2}{(S(t|x))^2} \\ &\quad + \frac{1}{nh}\frac{d_K[C_H(t+b|x) - C_H(t|x)]}{m(x)}\left(\frac{S(t+b|x)}{S(t|x)}\right)^2 + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}\frac{d_K D_H(t,t+b|x)}{m(x)}(1-\varphi(t|x))^2 + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}v(t|x) + o\left(\frac{1}{nh}\right). \end{aligned} \tag{51}$$

Finally collecting expressions (29) and (51) we conclude (5). The formula for the asymptotic optimal bandwidth, (10), can be easily derived from (5).

To prove the last part of Theorem 2, we use Corollaries 3 and 4 in Iglesias-Pérez and González-Manteiga (1999) to show that

$$\sqrt{nh}\left[(\hat{P},\hat{Q})^t - (P,Q)^t\right] \xrightarrow{\mathrm{d}} N_2(\mathbf{b},\mathbf{V}),$$

where

$$\mathbf{b} = (b_1, b_2)^t = -c^{1/2}\frac{1}{2}c_K \left(A_H(t+b|x)P, A_H(t|x)Q\right)^t,$$
$$\mathbf{V} = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} = \begin{pmatrix} v_1(t+b|x) & v_2(t,t+b|x) \\ v_2(t,t+b|x) & v_1(t|x) \end{pmatrix}.$$

Now applying the continuous function $g(u,v) = \frac{u}{v}$ to the sequence of the bivariate random variable above and using the delta method, simple but long and tedious algebra gives

$$\sqrt{nh}\left(\frac{\hat{P}}{\hat{Q}} - \frac{P}{Q}\right) \xrightarrow{\mathrm{d}} N\left(\mu, \sigma^2\right), \tag{52}$$

with

$$\begin{aligned}
\mu &= \left(\frac{\partial g(u,v)}{\partial u}, \frac{\partial g(u,v)}{\partial v}\right)\Bigg|_{(u,v)=(P,Q)} \mathbf{b} \\
&= \frac{1}{Q}b_1 - \frac{P}{Q^2}b_2 = -c^{1/2}\frac{1}{2}c_K\frac{P}{Q}\left(A_H(t+b|x) - A_H(t|x)\right) \\
&= -c^{1/2}b(t|x), \\
\sigma^2 &= \left(\frac{\partial g(u,v)}{\partial u}, \frac{\partial g(u,v)}{\partial v}\right)\Bigg|_{(u,v)=(P,Q)} \mathbf{V} \left(\frac{\partial g(u,v)}{\partial u}, \frac{\partial g(u,v)}{\partial v}\right)^t\Bigg|_{(u,v)=(P,Q)} \\
&= \frac{1}{Q^2}v_1(t+b|x) - \frac{2P}{Q^3}v_2(t,t+b|x) + \frac{P^2}{Q^4}v_1(t|x) \\
&= v(t|x).
\end{aligned}$$

This concludes the proof by substituting $\frac{\hat{P}}{\hat{Q}} = 1 - \hat{\varphi}_n(t|x)$ and $\frac{P}{Q} = 1 - \varphi(t|x)$ in (52).

Acknowledgements

This research was partially supported by Grants 07SIN01205PR from Xunta de Galicia (Spain) and MTM2009-00166 from Ministerio de Ciencia e Innovación (Spain).

7 References

Allen, L. N. and Rose, L. C. (2006). Financial survival analysis of defaulted debtors, *Journal of Operational Research Society*, 57, 630-636.

Altman, E. I. (1968). Finantial ratios, discriminant analysis, and the prediction of corporate bankruptcy, *Journal of Finance*, 23, 589-611.

Altman, E. I. and Saunders, A. (1998). Credit risk measurement: developments over the last 20 years, *Journal of Banking and Finance*, 21, 1721-1742.

Baba, N. and Goko, H. (2006). Survival analysis of hedge funds, Bank of Japan, Working Papers Series No. 06-E-05.

Basel Comitee on Banking Supervison (1999). International convergence of capital measurement and capital standards, Bank for International Settlements.

Basel Comitee on Banking Supervision (2004). International convergence of capital measurement and capital standards: a revised framework, Bank for International Settlements.

Beran, R. (1981). Nonparametric regression with randomly censored survival data, Unpublished technical report, University of California, Berkeley.

Beran, J. and Djaïdja, A. K. (2007). Credit risk modeling based on survival analysis with inmunes, *Statistical Methodology*, 4, 251-276.

Carling, K., Jacobson, T. and Roszbach, K. (1998). Duration of consumer loans and bank lending policy: dormancy versus default risk, Working Paper Series in Economics and Finance No. 280, Stockholm School of Economics.

Chen, S., Härdle, W. K. and Moro, R. A. (2006). Estimation of default probabilities with support vector machines, Center of Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin, Germany, SFB 649 discussion paper No. 2006-077.

Crouhy, M., Galai, D. and Mark, R. (2000). A comparative analysis of current credit risk models, *Journal of Banking and Finance*, 24, 59-117.

Dabrowska, D. (1987). Non-parametric regression with censored survival time data, *Scandinavian Journal of Statistics*, 14, 181-197.

Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate, *Annals of Statistics*, 17, 1157-1167.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.

Glennon, D. and Nigro, P. (2005). Measuring the default risk of small business loans: a survival analysis approach, *Journal of Money, Credit, and Banking*, 37, 923-947.

González-Manteiga, W. and Cadarso-Suárez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications, *Journal of Nonparametric Statistics*, 4, 65-78.

Hamerle, A., Liebig, T. and Rösch, D. (2003). Credit risk factor modeling and the Basel II IRB Approach, Deutsche Bundesbank Discussion Paper Series 2, Banking and Financial Supervision, document No. 02/2003.

Hand, D. J. (2001). Modelling consumer credit risk, *IMA Journal of Management Mathematics*, 12, 139-155.

Hanson, S. and Schuermann, T. (2004). Estimating probabilities of default, Federal Reserve Bank of New York, Staff Report No. 190.

Iglesias Pérez, M. C. and González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications, *Journal of Nonparametric Statistics*, 10, 213-244.

Jorgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models, *Biometrika*, 70, 19-28.

Li, G. and Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data, *Annals of the Institute of Statistical Mathematics*, 53, 708-729.

Li, G. and Van Keilegom, I. (2002). Likelihood ratio confidence bands in non-parametric regression with censored data, *Scandinavian Journal of Statistics*, 29, 547-562.

Malik, M. and Thomas L. (2006). Modelling credit risk of portfolio of consumer loans, University of Southampton, School of Management Working Paper Series No. CORMSIS-07-12.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* , 2nd. Ed., Chapman and Hall, London.

Narain, B. (1992). Survival analysis and the credit granting decision. In: Thomas L., Crook, J. N. and Edelman, D. B. (eds.). *Credit Scoring and Credit Control*. OUP: Oxford, 109-121.

Roszbach, K. (2003). Bank lending policy, credit scoring and the survival of loans, Sverriges Riksbank Working Paper Series No. 154.

Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data, *Operations Research*, 50, 277-289.

Saunders, A. (1999). *Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms*, John Wiley & Sons, New York.

Van Keilegom, I. and Veraverbeke, N. (1996). Uniform strong results for the conditional Kaplan-Meier estimators and its quantiles, *Communications in Statistics: Theory and Methods*, 25, 2251-2265.

Van Keilegom, I., Akritas, M. and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study, *Computational Statistics & Data Analysis*, 35, 487-500.

Wang, Y., Wang, S. and Lai, K. K. (2005). A fuzzy support vector machine to evaluate credit risk, *IEEE Transactions on Fuzzy Systems*, 13, 820-831.

Discussion of "Modelling consumer credit risk via survival analysis" by Ricardo Cao, Juan M. Vilar and Andrés Devia

Noël Veraverbeke
Centrum Voor Statistiek
Universiteit Hasselt, Belgium

The present paper deals with the estimation of the probability of default (PD) which is a very important parameter in many models for consumer credit risk in the literature.

If T denotes the time to default of a client, then it is immediately clear that in many cases T will not be observed, due to the ending of the observation period or the occurrence of some other event that happens earlier in time. This perfectly fits into the classical model of right random censoring in survival analysis. Here the observations are $Y = \min(T,C)$ and $\delta = I(T \leq C)$ where T is the time to default and C is the censoring time.

Classical survival analysis tools like Kaplan-Meier estimation and Cox estimation allow to obtain estimates for the distribution function of T. Moreover it is also possible to incorporate a vector X of explanatory variables and to estimate the conditional distribution function of T, given that $X = x$.

Since the probability of default just the conditional residual life distribution function (see Veraverbeke (2008)), it can be expressed as a simple function of the conditional distribution function and different estimation methods of the latter lead to different estimators for the PD.

Three methods are explored in this paper. The first is based on Cox's proportional hazards regression model, the second on a generalized linear model and the third on Beran's (1981) nonparametric product limit estimator for the conditional distribution function. For the third method, some new asymptotic properties are derived for the conditional residual life distribution function estimator. The illustration with real data clearly shows that the covariate information is essential and that methods 1 and 3 give a good fit.

I want to congratulate the authors for their contribution to this field of modelling credit risk using regression techniques from survival analysis. The results are very promising and I hope to see further work in that direction. My comments/questions below are meant to stimulate this.

1) It would be interesting to explore the use of time-dependent covariates. In particular, how could this be done for the nonparametric method?
2) The theoretical results and also the real data application are shown for one single covariate. Is the extension to more than one covariate straightforward?

3) An assumption throughout is the conditional independence of T and C, given X. But there are more and more examples in survival analysis where this is questionable. See, for example, Zheng and Klein (1995), Braekers and Veraverbeke (2005). How realistic is the independence assumption in credit risk modelling and how could this assumption possibly be relaxed?
4) Is it possible to generalize the asymptotic normality result in Theorem 2 in order to obtain practical confidence bands for the default rate curves?
5) The third method relies on a good choice for the bandwidth. Is there a suggestion for an optimal choice?

It was a pleasure for me to be invited as a discussant for this interesting paper.

References

Braekers, R. and Veraverbeke, N. (2005). Copula-graphic estimator for the conditional survival function under dependent censoring. *Canadian Journal of Statistics*, 33, 429-447.

Veraverbeke, N. (2008). Conditional residual life under random censorship. In: B. C. Arnold, U. Gather, S. M. Bendre (eds). *Advances in Statistics: Felicitation Volume in Honour of B. K. Kale*. MacMillan India, New Dehli, pp.174-185.

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82, 127-138.

Jean-Philippe Boucher
Département de Mathématiques
Université du Québec à Montréal, Québec, Canada

The paper deals the default intensity of consumer to determine the probability of defaults. Because the authors use the fact that a large proportion of consumers does not have default, they use censored models in the estimation.

I would like to point out to authors some important details. In consumer credit data, the amount of information from real data is very small. Indeed, depending on the definition of what is a ***default***, we can suppose that a default can arise continuously. However the default will only be observable on a small period of time since consumer only pays his debt at the beginning of each month. Consequently, we must deal with even less information than what is assumed in the paper. For the Cox proportional hazard model, Malik and Thomas (2006) worked with a modified likelihood function when dealing with this situation.

This problem shares similarities with insurance data. Indeed, with aggregate insurance data, it is impossible to know at what time insureds had their accident (see for example Boucher and Denuit (2007)). A major difference between credit and claim count analysis is the fact that a default of credit happens only once, while it is possible to see more than one claim in a single insurance period. However, even with this difference, for a parametric approach such as the GLM model proposed by the authors, it is possible to construct credit risk models based on models of Boucher and Denuit (2007).

Conceptually, let τ be the waiting time between the beginning of the loan and the default. Let $I(t)$ be the indicator of a default during the interval $[0,t]$. Hence,

$$\begin{aligned} P(I(t)=0) &= P(\tau > t) \\ P(I(t)=1) &= 1-P(\tau > t) \end{aligned} \tag{1}$$

For a loan of one year, we only have up to 12 partial informations on the credit default. Consequently, we then observed intervals $[0, \frac{1}{12}]$, $]\frac{1}{12}, \frac{2}{12}]$, $]\frac{2}{12}, \frac{3}{12}]$, $\ldots$, $]\frac{11}{12}, \frac{12}{12}]$.

In count data, duration dependence occurs when the outcome of an experiment depends on the time that has elapsed since the last success (Winkelmann (2003)). Then, the occurrence of an event modifies the expected waiting time to the next occurrence of the event. For credit risk, a positive (negative) duration dependence would mean that the

probability of default decreases (increases) over time. Consequently, the true probability depends on which interval the default happens and can be expressed:

$$P(I(1)=y) = \begin{cases} P(\tau \leq \frac{1}{12}) & \text{for a default } y \text{ occuring in } [0, \frac{1}{12}] \\ P(\frac{1}{12} < \tau \leq \frac{2}{12}) & \text{for a default } y \text{ occuring in }]\frac{1}{12}, \frac{2}{12}] \\ \ldots & \\ P(\frac{11}{12} < \tau \leq \frac{12}{12}) & \text{for a default } y \text{ occuring in }]\frac{11}{12}, \frac{12}{12}] \\ 1 - P(\tau > \frac{12}{12}) & \text{for a non-registered default} \end{cases}, \quad (2)$$

By comparison with this last equation, for an individual i, the contribution of conditional likelihood function of the authors, involving the conditional density, was written as $f(\tau)^{\delta}(1-F(\tau))^{1-\delta}$, where $\delta = 1$ if an individual did a default. Less information is available with (2) since differences in cumulative distributions is used rather than density functions.

Except when working with the Exponential distribution that is known to be memoryless, it is not possible to simply express all the probabilities of a default as:

$$P(I(t)=0) = P(\tau > \frac{1}{12}),$$

for all possible values of t because it is only valid for the first interval. Indeed, for illustration, let us assume that τ is Gamma distributed, with density

$$f(\tau;\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\lambda\tau}, \quad (3)$$

where $\Gamma(\cdot)$ is the gamma function, $\alpha > 0$ and $\lambda > 0$. Note that the Gamma hazard function is increasing for $\alpha > 1$ and is decreasing for $\alpha < 1$ (and shows constant hazard of $\alpha = 1$). Thus, for $\alpha \leq 1$, the model exhibits positive duration, while $\alpha \geq 1$ implies negative duration (and does not show duration dependence for $\alpha = 1$, from which we find the Exponential distribution). It can be interesting to see how well real data can estimate these parameters. Indeed, with the use of (2), the model can be expressed by using this useful notation:

$$P(t_a < \tau \leq t_b) = \frac{1}{\Gamma(\alpha)}\int_{t_a}^{t_b} \lambda^{\alpha}\nu^{\alpha-1}e^{-\lambda\nu}d\nu$$

where the integral is known as an incomplete gamma function. This probability can be evaluated using integrations approximations or asymptotic expansions (Abramowitz and Stegun (1968)).

It would be interesting to apply the unusual non-parametric approach of the authors using equation (0.2).

References

Abramowitz, M. and Stegun, I. A. (1968). *Handbook of Mathematical Functions*. National Bureau of Standards, Applied Mathematics Series Nr. 55, Washington, D.C.

Boucher, J.-P. and Denuit, M. (2007). Duration Dependence Models for Claim Counts. *Deutsche Gesellschaft fur Versicherungsmathematik (German Actuarial Bulletin)*, 28, 29-45.

Malik, M. and Thomas, L. (2006). Modeling Credit Risk of Portfolio of Consumer Loans. *University of Southampton, School of Management, Working Paper, Series No. CORMSIS-07-12*.

Winkelmann, R. (2003). *Econometric Analysis of Count Data*. Springer-Verlag, Berlin, 4th ed.

Jan Beran
Department of Mathematics and Statistics
University of Konstanz, Germany

Since Basel II, the modeling of credit risks has become an important practical issue that has to be addressed in a mathematically tractable manner while taking into account particular characteristics of the market and available data. One general approach discussed in the literature is modelling probability of default (PD) by applying survival analysis. The idea is quite natural, since in the financial context the time until default can be interpreted directly as a survival time and such data are readily available. As in usual survival analysis, the observed times until default are partially censored. In spite of the obvious analogy to the biological context, survival analysis may not be very well known among practitioners in finance. The paper by Cao, Vilar and Devia is therefore a welcome contribution.

The authors essentially discuss three methods of estimation:

1. Cox's proportional hazard model;
2. generalized linear models; and
3. nonparametric conditional distribution estimation.

For the third method, the asymptotic mean squared error and a formula for the asymptotically optimal bandwidth h_o are given. While 1 and 2 and their properites are well known, the asymptotic result for the third method appears to be new. From the practical point of view, the question is which of the three methods perform best when applied to real data, and also whether there may be any alternative methods that even outperform any of these. Before answering this question, one needs to define a criterion for judging the performance. In the paper here, empirical and estimated PDs are compared. Thus, the criterion is simply to what extent a model fits the data. More interesting would be to use predictive out of sample criteria and also financial risk measures. Furthermore, the fitted PDs reported in table 2 are of varying quality. One may therefore ask whether any of the models considered here reflect the underlying mechanism with sufficient accuracy. In particular, the perfomance of standard models in survival analysis depends on the amount of censoring. Typically for credit default data, a large majority (often more than 95%) of the observations are censored. In such situations, maximum likelihood estimates based on unimodal distributions tend to be highly biased. For this reason, Beran and Djaïdja (2007) adopted an idea originally introduced by Maller and Zhou (1996) in a medical context. Observations are assumed to come from a mixture distributions consisting of a usually large proportion p of "immunes" and a smaller

proportion $1-p$ of clients who may default. Thus, the time until a randomly chosen client number i defaults can be written as

$$Y_i = \varsigma_i \cdot \infty + (1-\varsigma_i)W_i$$

where $P(\varsigma_i = 1) = 1 - P(\varsigma_i = 0) = p$ and W_i is a continuous distribution $F_W(.;\lambda)$ with density $f_W(.;\lambda)$ $(\lambda \in \Lambda \subseteq \mathbb{R}^k)$ on $\mathbb{R}_+$. Conditionally on the censoring constants c_i, the maximum likelihood estimate of $\theta = (p,\lambda)$ is obtained from observed survival times $x_i = y_i \wedge c_i$ by maximizing

$$L(\theta) = n_1 \log(1-p) + \sum_{i \in I} \log f_W(y_i;\lambda) + \sum_{i \in I^c} \log\left[1-(1-p)F_W(c_i;\lambda)\right]$$

where $I = \{i : y_i \leq c_i\}$ and $n_1 = |I|$. In practice estimates of PDs and prediction of defaults turned out to be much more accurate in the case of retail clients where defaults are (or used to be) very rare. It may therefore be worth the effort to see whether the same applies to the consumer loans considered in this discussion paper.

References

Beran, J. and Djaïdja (2007). Credit risk modeling based on survival analysis with immunes. *Statistical Methodology*, 4, 251-276.

Maller, R. A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

Rejoinder

First of all we would like to thank the discussants for their kind words and suggestions concerning this paper. According to the topics mentioned in the comments we will organize this rejoinder in four sections. These sections deal with other censoring models, predictive criteria, bandwidth selection and extensions to other settings.

1 Other censoring models

As mentioned by Prof. Beran, some other alternative models are available for heavy censoring situations like in credit risk. In this rejoinder we will adopt the approach by Maller and Zhou (1996) and Beran and Djaïdja (2007) for the generalized linear model presented in the paper. Using the notation of Subsection 3.3, we have considered the model

$$F(t|x) = (1-p)\, g\,(\theta_0 + \theta_1 t + \theta_2 x)\,, \tag{1}$$

where $p \in (0,1)$ is the proportion of credits that are immune to default and F is any of the two parametric distributions considered in Subsection 5.1.2 of the paper. Using equation (1), the log-likelihood function in Subsection 3.3 results in

$$\begin{aligned}\ell\,(\theta_0,\theta_1,\theta_2,p) \;=\; & [\ln(1-p)+\ln\theta_1]\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\delta_i \ln g'\,(\theta_0+\theta_1 Y_i+\theta_2 X_i) \\ & + \sum_{i=1}^{n}(1-\delta_i)\ln\left[1-(1-p)\, g\,(\theta_0+\theta_1 Y_i+\theta_2 X_i)\right]\end{aligned}$$

For dealing with the high complexity of the model and the minimization of the log-likelihood equations, we have used a differential evolution based program called *DEoptim*, implemented in R. See, for instance, Price et al. (2005) for details about this numerical optimization approach.

Figure 1 shows the estimated *PD* using these heavy censoring models when conditioning to $X = 5.44$, the median value of the covariate. The $\widehat{PD}$ curves for the *GLM* and the modified *GLM* (*MGLM*) are shown in a range of maturity times given by the depth of the sample. The *GLM* curves in the left panel are those presented in Figure 5 of the paper. Using the same link functions, the heavy censoring models with

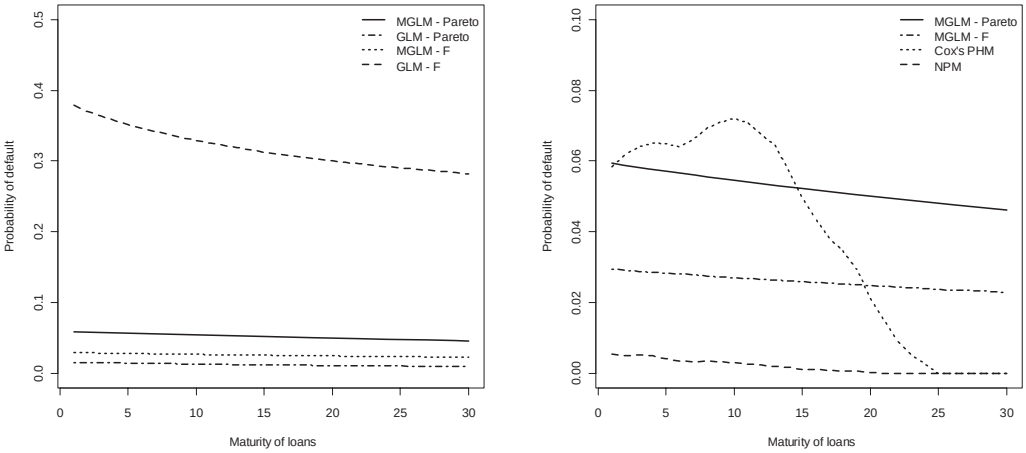

Figure 1: *Left panel: $\widehat{PD}$ curves for the GLM and the MGLM. Right panel: $\widehat{PD}$ curves for the MGLM, Cox's proportional hazard model and the nonparametric approach. All these $\widehat{PD}$ curves were obtained conditioning on $X = 5.44$.*

single parameter Pareto and Snedecor's F distributions are also plotted in the left panel. The estimated parameters were $\alpha = 0.6$ and $p = 0.01$ for the Pareto distribution and 8.553 and 0.525 for the degrees of freedom of Snedecor's F distribution with $p = 0.01$. The right panel plots a graphical comparison of the *MGLM*, Cox's proportional hazard model and the nonparametric approach.

The results obtained with the *GLM* approach are not good in general, and the modified version proposed in equation (1) did not produced a significant improvement in the estimated *PD* for our data set. The $\widehat{PD}$ curve computed with the F link fits better than that with the Pareto link function for a range of covariate values. Thus, in the following we will only present results concerning the *MGLM* approach with the Snedecor's F link function. The estimated default probabilities with both links were extremely large for those values of X smaller than 1, or extremely small for values of X larger than the third quartile (28.2703).

An alternative way to deal with heavy censoring, not considered here, is to use the transfer tail models introduced by Van Keilegom and Akritas (1999) and Van Keilegom, Akritas and Veraverbeke (2001). This consists in using nonparametric regression residuals to transfer tail information from regions of light censoring to regions of heavy censoring in conditional distribution function estimation.

The possible discrete nature of the defaults, mentioned by Prof. Boucher, gives rise to an interval censored model for the time to default (see his equation (2)). This censoring model is very useful when the defaults are reported in multiples of a given time unit (e.g., a month). This is not the case for our data set with 1800 defaults corresponding to 576 different values. The highest frequency of these values is only 15 and the average frequency of these 576 different values is only 3.156.

Our data set has been facilitated by a financial company. This company records the contract date and sends a payment order on a fixed date in the second month following the contract formalization date. This fixed date may change from month to month. When a client does not make one of these payments and this situation is maintained for more than 90 days, the 91st day after the due payment date is considered as the default time. However there are even a few exceptions in which default may be considered even before than four months from the contract date. For all these reasons it is virtually impossible, at least for this data set, that default times occur in multiples of one month.

Nevertheless, there may exist practical situations where defaults exhibit a discrete nature. In these cases the nonparametric estimator given by Beran (1981) can be extended to interval-censored response lifetimes. The idea is to adapt the estimator proposed by Turnbull (1976) to the conditional setting, in a general framework of censoring and truncation (which includes interval censoring). This adaptation could be very similar to the one used in Beran (1981) to extend the Kaplan-Meier estimator to a conditional setup.

As Professor Veraverbeke points out, one could consider more general censoring models that allow for some sort of conditional dependence between the censoring time, C, and the life time, T, of a credit. The hypothesis of conditional independence is very common in survival analysis and it is also very convenient in credit risk applications.

In principle, when the censoring times come from time from contract formalization to end of the study, the conditional independence assumption seems a natural one. However, this is not the only source of censored data. For instance credit cancellation, which also causes censoring, may be correlated to possible time to default. Unfortunately it is often very difficult to test such an assumption from real data. This is because most of the times there is no available information about jointly observed values of (C,T). As Professor Veraverbeke mentions, copula models are useful tools for constructing more flexible models that allow for conditional dependence. An interesting future study would be to extend the results on nonparametric estimation of default probability to copula models as those proposed in Braekers and Veraverbeke (2005).

2 Predictive criteria

As Professor Beran explains in his report interesting model adequacy tests for a financial firms are based on predictive criteria. The estimated probability of default can be used to classify a credit in default or nondefault. Using the three methods proposed in the paper and fixing a maturity time of $t = 5$ months and a forecast time horizon of $b = 12$ months, the estimated PD has been computed for every single credit of a real loan portfolio.

Starting from the sample of credits alive at time t, the two subsamples of defaulted and non-defaulted credits at time $t + b$ have been considered. In order to study the discrimination power of the three models, we have considered the pertaining estimated

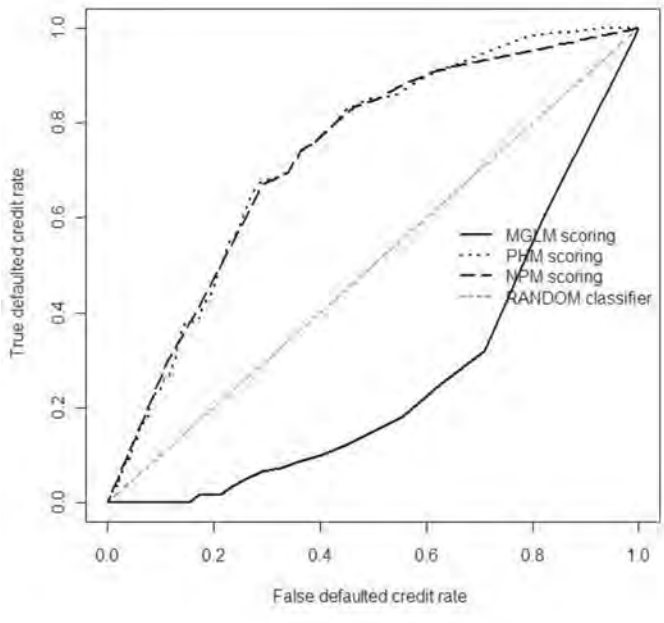

Figure 2: *ROC curves for the three PD approaches: MGLM, Cox's PHM, and NPM.*

PD and computed the *ROC* curves. This tool has been used in financial setups by Thomas (2000), Stein (2005), Blöchlinger and Leippold (2006) and Engelmann (2006), among others. The area under the *ROC* curve (*AUC*), which is a measure of the the discrimination power of the methods, has also been computed.

The study was performed by just dividing our data set of size 25000 in a training sample of size 20000 and a test sample of size 5000. The choice of these two samples was made at random. The test sample was split up into defaulted and non-defaulted credits. The *PD* estimates, obtained for the three approaches using the training sample, were applied to the test sample and the out-of-sample *ROC* curves are plotted in Figure 2. The areas under these curves and their confidence intervals are collected in Table 1.

Figure 2 shows a surprisingly poor discrimination power of the *MGLM* model. This is also reflected by the *AUC* values in Table 1. An open question is how important is the choice of the link function in order to produce much better results. The performance of Cox's proportional hazard model and the nonparametric approach is very comparable. Their discrimination power (measured via the *AUC*) is about 74%.

A first conclusion is that the modification of the original *GLM* approach was not able to produce the expected improvement in the original *GLM* setting, but it may be interesting to study the problem of choice of the link function in a deeper way. On the other hand *PD* estimates obtained by Cox's proportional hazards model and the nonparametric approach provide quite powerful discrimination between default and non-default credits.

Table 1: *Area under the ROC curves for the three approaches computed by using the validation sample.*

Model	*AUC*	95% asymptotic confidence interval	
MGLM	0.265	0.234	0.297
Cox's *PHM*	0.735	0.703	0.766
NPM	0.738	0.706	0.770

3 Bandwidth selection

As Professor Veraverbeke points out, the nonparametric approach relies on a good choice for the bandwidth. Direct plug-in methods for the selection of the smoothing parameter require the estimation of plenty of population functions involved in equation (10): $H_1(t|x)$, $H(t|x)$, $\dot{H}(t|x)$, $\ddot{H}(t|x)$, $m(x)$, $m'(x)$ and $\varphi(t|x)$. This turns out to be a tedious procedure. Furthermore, since the method is based on an asymptotic expression, it may not produce accurate results for samples with a moderate number of uncensored data. See, for instance Cao, Janssen and Veraverbeke (2001) for similar ideas in a different context.

A good alternative for bandwidth selection in this context is the bootstrap method. This method can be used to find a bootstrap analogue of the mean squared error of $\varphi(t|x) = PD(t|x)$ (see, for instance, Cao (1993) for the use of the bootstrap for estimation of the mean integrated squared error in a different context). This method would require the use of two pilot bandwidths, g_1 and g_2, for estimating $F(t|x)$ and $G(t|x)$ and a pilot bandwidth, g_3, for the density m. The method proceeds as follows:

1. Compute, $\hat{F}_{g_1}(t|x)$, Beran's estimator of $F(t|x)$ and $\hat{G}_{g_2}(t|x)$, Beran's estimator of $G(t|x)$.
2. Estimate $m(x)$ by $\hat{m}_{g_3}(x)$.
3. Draw a sample $(X_1^*, X_2^*, \ldots, X_n^*)$ from $\hat{m}_{g_3}(x)$.
4. For every $i = 1, 2, \ldots, n$, draw T_i^* from $\hat{F}_{g_1}(t|x)$ and C_i^* from $\hat{G}_{g_2}(t|x)$.
5. Compute, for every $i = 1, 2, \ldots, n$, $Y_i^* = \min\{T_i^*, C_i^*\}$ and $\delta_i^* = \mathbf{1}_{\{T_i^* \leq C_i^*\}}$.
6. Use the sample $\{(Y_1^*, \delta_1^*, X_1^*), (Y_2^*, \delta_2^*, X_2^*), \ldots, (Y_n^*, \delta_n^*, X_n^*)\}$ to compute $\hat{\varphi}_h^*(t|x)$, the bootstrap analogue of $\hat{\varphi}_h(t|x)$.
7. Approximate the mean squared error of $\hat{\varphi}_h(t|x)$ by its bootstrap version:

$$MSE_{t,x}^*(h) = E^*\left[\left(\hat{\varphi}_h^*(t|x) - \hat{\varphi}_{g_1}(t|x)\right)^2\right].$$

8. This bootstrap MSE can be approximated by drawing a large number, B, of bootstrap replications following steps 4-6 and computing

$$\frac{1}{B}\sum_{j=1}^{B}\left(\hat{\varphi}_h^{*j}(t|x)-\hat{\varphi}_{g_1}(t|x)\right)^2.$$

9. Finally the bootstrap bandwidth, $h^*_{MSE,t,x}$, is the minimizer of $MSE^*_{t,x}(h)$ in h.

Since this resampling plan may be very time consuming, a possible way to make this approach feasible for very large sample sizes (like $n = 25000$) is the following. Fix some smaller subsample size (for instance $m = 2500$), i.e., $n = \lambda m$, with λ typically large (in this example $\lambda = 10$). Use the bootstrap resampling plan to get a bootstrap bandwidth, $h^*_{MSE,m,t,x}$, for sample size m. Based on the asymptotic formula (10), in the paper, obtain $h^*_{MSE,n,t,x} = \lambda^{-1/5} h^*_{MSE,m,t,x}$.

Consistency and practical behaviour of this bootstrap method is left for future work.

4 Extensions to other settings

Professor Veraverbeke raises the question of extension of the nonparametric default probability estimator to the multiple covariate case. We believe that this extension is rather straightforward, as it is for the conditional distribution estimator. From the theoretical viewpoint, it is expected that the convergence rate gets worse when the dimension of the covariate vector increases. In fact, it is very likely that the *PD* nonparametric estimator is worthless for covariates of dimension larger to 3 or 4, except for huge sample sizes (curse of dimensionality). A possible way to overcome this problem is to use the dimension reduction ideas proposed by Hall and Yao (2005) to produce a semiparametric estimator of the default probability that is free of the curse of dimensionality. At the same time, the projection of the covariate vector obtained by such a procedure would probably be interpretable as a kind of overall scoring that accounts for propensity of credits to default.

The time-dependent covariate case mentioned by Professor Veraverbeke can be treated using ideas of McKeague and Utikal (1990), who extended Beran's estimator to time-dependent covariates. Last, but not least, although convergence of the default probability process could be studied and used to derive asymptotic theory for confidence bands, in our opinion this is out of the scope of the present paper. On the other hand we believe that, for practical reasons, financial companies are more interested (for prediction) in the estimation of the default probability at a given maturity and with fixed values of the covariates, than in a confidence band.

We would like to finish this rejoinder by thanking, once again, the discussants for their suggestions and comments. We are also grateful to the Editor in Chief of SORT, Montserrat Guillén, for her kind invitation to write this paper and for her efficiency along the editing process. Her support has helped us a lot to improve the quality of this paper.

References

Beran, J. and Djaïdja, A. K. (2007). Credit risk modeling based on survival analysis with inmunes, *Statistical Methodology*, 4, 251-276.

Beran, R. (1981). Nonparametric regression with randomly censored survival data, Unpublished technical report, University of California, Berkeley.

Blöchlinger, A. and Leippold, M. (2006). Economic benefit of powerful credit scoring, *Journal of Banking and Finance*, 30, 851-873.

Braekers, R. and Veraverbeke, N. (2005). Copula-graphic estimator for the conditional survival function under dependent censoring, *Canadian Journal of Statistics*, 33, 429-447.

Cao, R. (1993). Bootstrapping the mean integrated squared error, *Journal of Multivariate Analysis*, 45, 137-160.

Cao, R., Janssen, P. and Veraverbeke, N. (2001). Relative density estimation and local bandwidth selection with censored data, *Computational Statistics & Data Analysis*, 36, 497-510.

Engelmann, B. (2006). Measures of a rating's discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. *The Basel Risk Parameters: Estimation, Validation, and Stress Testing.* Springer, New York.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction, *The Annals of Statistics*, 33, 1404-1421.

Maller, R.A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*, Wiley, New York.

McKeague, I. W. and Utikal, K. (1990). Inference for a nonlinear counting process regression model, *The Annals of Statistics*, 18, 1172-1187.

Price, K., Storn, R. and Lampinen, J. (2005). *Differential Evolution – a Practical Approach to Global Optimization*. Springer, New York.

Stein, R. (2005). The relationship between default prediction and lending profits: integrating the ROC analysis and loan pricing, *Journal of Banking and Finance*, 29, 1213-1236.

Thomas, L. (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16, 149-172.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society, Series B*, 38, 290-295.

Van Keilegom, I. and Akritas, M. G. (1999). Transfer of tail information in censored regression models, *The Annals of Statistics*, 27, 1745-1784.

Van Keilegom, I., Akritas, M. G. and Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study, *Computational Statistics & Data Analysis*, 35, 487-500.

SORT 33 (1) January-June 2009, 49-70

Estimating unemployment in very small areas

Ugarte[1*], M.D., Goicoa[1], T., Militino[1], A.F., and Sagaseta-López, M.[2]

Abstract

In the last few years, European countries have shown a deep interest in applying small area techniques to produce reliable estimates at county level. However, the specificity of every European country and the heterogeneity of the available auxiliary information, make the use of a common methodology a very difficult task. In this study, the performance of several design-based, model-assisted, and model-based estimators using different auxiliary information for estimating unemployment at small area level is analyzed. The results are illustrated with data from Navarre, an autonomous region located at the north of Spain and divided into seven small areas. After discussing pros and cons of the different alternatives, a composite estimator is chosen, because of its good trade-off between bias and variance. Several methods for estimating the prediction error of the proposed estimator are also provided.

MSC: 62D05, 62J12, 62F40

Keywords: Finite population, Prediction theory, Labour Force Survey

1 Introduction

The Spanish Labour Force Survey (SLFS), called "Encuesta de Población Activa" in Spanish or, in short "EPA", is a quarterly survey of households living at private addresses in Spain. Its purpose is to provide information on the Spanish labour market that can then be used to develop, manage, evaluate, and report on labour market policies. It is conducted by the Spanish Statistical Institute (INE). Yet there are multiple aims achieved with this survey, of which the estimation of unemployment is one of the most relevant. The survey follows a stratified two-stage cluster design and, for each province, a separate

[1] Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra. Campus de Arrosadía, 31006 Pamplona

[2] Instituto de Estadística de Navarra. Avda. Carlos III 36, 2º Dcha, 31003 Pamplona

[*] E-mail: lola@unavarra.es

Received: November 2007
Accepted: July 2008

sample is drawn. In the first stage sampling, 3,588 primary sampling units (PSUs) called "Secciones Censales" in Spanish, are selected with probabilities proportional to size according to the number of households. In the second stage sampling, the secondary sampling units (SSUs) are households and a simple random sampling is applied to draw 18 SSUs from each PSU selected. This sampling design generates self-weighted samples at stratum level and then, every household has the same probability of being drawn. In every Spanish province there is a fixed number of PSUs that have to be selected. Navarre is a small autonomous region (with a single province) located in northern Spain. It has an area of 10,000 km^2 and only 600,000 inhabitants, irregularly distributed in seven small areas. In Navarre, 91 PSUs are selected in the first stage, and, for each PSU, 18 households are drawn obtaining a total of 1,638 households. In this study, interest is focussed on evaluating the performance of design-based, model-assisted, and model-based estimators for estimating unemployment in Navarre using different auxiliary information. We are dealing with a scenario where the number of small areas is very limited (seven small areas), the incidence of the study variable is scarce, and the sample size cannot be modified because it is already determined. To evaluate the alternative estimators, a Monte Carlo study has been conducted drawing 500 samples from the 2001 Spanish Census following the same sampling design of the SLFS. Similar simulation studies over the scenario of the Labour Force Census of companies affiliated to the Social Security system in Catalonia have been recently published by Costa *et al.* (2003). These authors conclude that a composite estimator with estimated weights based on the assumption of heterogeneity of biases and variances across small areas is superior to its competitors: direct and indirect estimators. The sample size effect on these composite estimators has also been studied by Costa *et al.* (2004). Morales *et al.* (2007) have analyzed the performance of model-assisted and model-based estimators on the Spanish Labour Force data from the Canary Islands. In their paper, model-based estimators are more competitive than model-assisted estimators, but they do not consider composite estimators. The impact of supplementary samples sizes on the Spanish Labour Force has also been analyzed by Costa *et al.* (2006), who evaluate the performance of some composite estimators when the sample size of the small areas is boosted.

The rest of this paper is organized as follows. Section 2 describes several design-based, model-assisted, and model-based estimators that might be used for estimating unemployment in very small areas. Section 3 describes a Monte Carlo study, on the scenario of the employment survey in Navarre, and gives the accuracy measures for the estimators presented in Section 2. Section 4 provides three different procedures to calculate the mean squared error (MSE) of the estimator chosen as optimal in Section 3. The paper ends up with some conclusions.

Figure 1: *Small Areas of Navarre*

2 Alternative estimators of unemployment

The variable of interest is the total number of unemployed in the seven small areas (called "comarcas", in Spanish) of Navarre (see Figure 1). Some of these small areas, mainly those located at the north of the province, are scarcely populated. In this section alternative estimators of unemployment are briefly described. They will be compared in the next section through a Monte Carlo study.

2.1 Design-based estimators

In the design-based theory, the variable of interest is a fixed quantity and the probability distribution is induced by the sampling design. It is a distribution-free method mainly focused on obtaining estimates for domains with large samples. A direct estimator only uses observations coming from the domain of interest, whereas an indirect estimator takes information outside of the domain. In the design-based theory, unbiasedness and design-consistency are desirable properties pursued by the majority of estimators. An estimator $\hat{Y}$ of Y is design-unbiased if $E[\hat{Y}] = Y$ and it is design-consistent if it is unbiased and its variance tends to zero as the sample size increases (Rao, 2003). The use of auxiliary information is a common tool for improving the precision of design-based estimators. Here, it consists of age-sex groups (E), with six categories which are a combination of age groups $(16-24, 25-54, >55)$ and sex; Stratum (S), that represents

municipality sizes and takes nine possible values in Spain, although in Navarre only six of those nine possible strata are available: (1) capital of the province, (5) between 20000 and 49999 inhabitants, (6) between 10000 and 19999 inhabitants, (7) between 5000 and 9999 inhabitants, (8) between 2000 and 4999 inhabitants and (9) with less than 2000 inhabitants; educational level (N) has two categories: (1) for illiterate, primary, and secondary school, and (2) for technical workers and professionals; employment status (P) in the Navarre Employment Register (SNE) with two categories: (1) occupied or inactive, and (2) unemployed; claimant of employment (C) taking the value 1 if he/she is registered in the SNE and 0, otherwise.

In this paper, the following design-based estimators are considered:

(a) Two direct estimators: the so-called direct, and the post-stratified estimator
(b) Five indirect estimators: one synthetic, and four composite estimators

Direct estimators use only data in the domain of interest. Although they are design-unbiased, the variability is usually big enough to be considered appropriate in small-area estimation.

The **direct estimator** of the total unemployment Y in the dth small area takes the form

$$\hat{Y}_d^{direct} = \hat{\bar{Y}}_d^{direct} N_d = \frac{\sum_{j=1}^{n_d} w_j y_j}{\sum_{j=1}^{n_d} w_j} N_d, \quad d = 1, \ldots, D,$$

where in area d, N_d is the population aged 16 or more, n_d is the number of people aged 16 or more in the sample, and the sampling weight w_j is the inverse of the inclusion probability for person j. The sampling weights are given by $w_j = N_h/n_h$ for $j = 1, \ldots, n_d$. This means that every person belonging to the same stratum h ($h = 1, \ldots, H$) has the same weight. Detailed expressions on how to obtain these weights are given by Morales *et al.* (2007). The variable y_j takes the value 1 if person j is unemployed and 0 otherwise. The total number of small areas is denoted by D.

The **post-stratified estimator** of the total unemployment Y in the dth small area is given by

$$\hat{Y}_d^{post} = \sum_{g=1}^{G} \hat{\bar{Y}}_{dg} N_{dg} = \sum_{g=1}^{G} \frac{\sum_{j=1}^{n_{dg}} w_j y_j}{\sum_{j=1}^{n_{dg}} w_j} N_{dg}, \quad d = 1, \ldots, D, \tag{1}$$

where G is defined by the categories of the different auxiliary variables. For instance, G has six categories when the variable age-sex group (E) is considered. The number of sampled people in the dth region belonging to the gth group is n_{dg} while N_{dg} is the corresponding population value. The post-stratified estimator is a direct estimator that only uses information from the domain of interest, yet it may also be considered as a model-assisted estimator as it can be derived from a linear model where the explanatory variable is a group indicator variable.

The synthetic estimator (González, 1973) is an indirect estimator used in small areas under the assumption that the small areas have the same characteristics as the large area, with regard to the variable of interest. When this does not happen, synthetic estimators are usually biased. The **synthetic estimator** used here takes the form

$$\hat{Y}_d^{synt} = \sum_{g=1}^{G} \hat{\bar{Y}}_g N_{dg} = \sum_{g=1}^{G} \frac{\sum_{j=1}^{n_g} w_j y_j}{\sum_{j=1}^{n_g} w_j} N_{dg}, \quad d = 1, \ldots, D, \tag{2}$$

where n_g is the number of sampled people in the whole province belonging to the gth group.

A natural way to balance the potential bias of a synthetic estimator and the instability of a direct estimator is to take a weighted average of the two estimators, what is called a composite estimator. The name of composite estimator has a more general meaning, corresponding to any kind of linear combination of estimators. In our case, **sample size dependent composite estimators** (Drew *et al.*, 1982) are considered. They are defined as a linear combination of a post-stratified and a synthetic estimator. Namely

$$\hat{Y}_d^{comp} = \lambda_d \hat{Y}_d^{post} + (1 - \lambda_d)\hat{Y}_d^{synt}$$

where

$$\lambda_d = \begin{cases} 1 & \text{if } \hat{N}_d \geq \alpha N_d \\ \dfrac{\hat{N}_d}{\alpha N_d} & \text{otherwise.} \end{cases}$$

$0 \leq \lambda_d \leq 1$, $\hat{N}_d = \sum_{j=1}^{d} w_j$ is a direct expansion estimator of N_d that increases with the domain sample size, and α is chosen to control the contribution of the synthetic estimator, and can take the following values: $\alpha = 2/3, 1, 1.5, 2$. Note that there are four possible composite estimators, one for each value of α. Hereafter in the paper, they will be denoted by composite 1 ($\alpha = 2/3$), 2 ($\alpha = 1$), 3 ($\alpha = 1.5$), and 4 ($\alpha = 2$), respectively. When the sample size increases λ_d is close to 1, and $\hat{Y}_d^{comp}$ is similar to the post-stratified estimator $\hat{Y}_d^{post}$, otherwise more weight is given to the synthetic estimator.

2.2 Model-assisted estimators

These estimators take account of the auxiliary information through the use of regression models as a means to obtain design-consistent estimators (Särndal *et al.*, 1992). They are more efficient than design-based estimators as auxiliary information is explicitly used at the estimation stage. Therefore, an important reduction of bias is attained. The most well-known model-assisted estimators are the generalized regression estimators (GREG), and, here, GREG estimators assisted in three different models are considered:

a linear model (see (3)), a logit model (see (4)), and a logit mixed model (see (5)) . Let us consider the following linear model

$$y_{jd} = \mathbf{x}_{jd}^{\mathsf{T}}\boldsymbol{\beta} + \epsilon_{jd}, \quad j = 1,\ldots,n_d, \quad d = 1,\ldots,D, \tag{3}$$

where for every small area d, y_{jd} takes the value 1 if the jth person is unemployed, and 0 otherwise, $\mathbf{x}_{jd} = (x_{jd,1}, x_{jd,2}, \ldots, x_{jd,p})^{\mathsf{T}}$ is the vector of the p auxiliary variables, and $\epsilon_{jd} \sim N(0, \sigma^2/w_{jd})$.

The GREG estimator of the total number of unemployed in the dth area, assisted in model (3), is given by

$$\hat{Y}_d^{LinearGREG} = N_d \left(\hat{\bar{Y}}_d^{direct} + (\bar{\mathbf{X}}_d - \hat{\bar{\mathbf{X}}}_d^{direct}) \right)^{\mathsf{T}} \hat{\boldsymbol{\beta}},$$

where $\bar{\mathbf{X}}_d = (\bar{X}_{d1}, \bar{X}_{d2}, \ldots, \bar{X}_{dp})^{\mathsf{T}}$ is the vector of the p auxiliary population means. The parameter vector $\boldsymbol{\beta}$ is estimated by generalized least squares with all the province observations.

Assuming that $y_{jd} \sim Bernoulli(P_{jd})$, where P_{jd} is the probability that the jth person in the dth area is unemployed, it seems more appropriate to be assisted in a logit model where

$$\text{logit}(P_{jd}) = \log\left(\frac{P_{jd}}{1-P_{jd}}\right) = \mathbf{x}_{jd}^{\mathsf{T}}\boldsymbol{\beta}. \tag{4}$$

The GREG estimator of the total number of unemployed in the dth area, assisted in model (4), is now given by

$$\hat{Y}_d^{LogitGREG} = \sum_{j=1}^{N_d} \frac{e^{\mathbf{x}_{jd}^{\mathsf{T}}\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_{jd}^{\mathsf{T}}\hat{\boldsymbol{\beta}}}} + \frac{N_d}{\hat{N}_d} \sum_{j=1}^{n_d} w_{jd} \left(y_{jd} - \frac{e^{\mathbf{x}_{jd}^{\mathsf{T}}\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_{jd}^{\mathsf{T}}\hat{\boldsymbol{\beta}}}} \right).$$

Usually, $\boldsymbol{\beta}$ is estimated by maximum likelihood (ML) using Fisher or Newton-Raphson algorithms (see for example, McCullagh and Nelder, 1989).

Estimators may be also assisted in mixed models where area random effects are considered. Assuming that $y_{jd}|u_d \sim Bernoulli(P_{jd})$ where u_d is the area random effect with $u_d \sim N(0, \sigma_u^2)$, a logit mixed model takes the form

$$\text{logit}(P_{jd}) = \log\left(\frac{P_{jd}}{1-P_{jd}}\right) = \mathbf{x}_{jd}^{\mathsf{T}}\boldsymbol{\beta} + u_d. \tag{5}$$

The model fitting was carried out maximizing an approximation of the likelihood integrated over the random effects using adaptive Gaussian quadrature.

The GREG estimator assisted in model (5) of the total number of unemployed in the dth area takes the form

$$\hat{Y}_d^{LogitMixedGREG} = \sum_{i=1}^{N_d} \frac{e^{\mathbf{x}_{jd}^T\hat{\boldsymbol{\beta}}+\hat{u}_d}}{1+e^{\mathbf{x}_{jd}^T\hat{\boldsymbol{\beta}}+\hat{u}_d}} + \frac{N_d}{\hat{N}_d}\sum_{j=1}^{n_d} w_{jd}\left(y_{jd} - \frac{e^{\mathbf{x}_{jd}^T\hat{\boldsymbol{\beta}}+\hat{u}_d}}{1+e^{\mathbf{x}_{jd}^T\hat{\boldsymbol{\beta}}+\hat{u}_d}}\right).$$

The choice between fixed or random effects when using models is not a trivial task neither from a theoretical nor from a practical point of view. In principle, if there are a sensible number of small areas and one expects a different behaviour from them, it makes sense to introduce random effects to avoid model overparameterization. An interesting discussion about the use of fixed or random effects models in a real context is given by Militino *et al.*(2007a).

Table 1: *Mean of the absolute relative bias (MARB), and mean of the relative root mean square error (MRMSE) for the post-stratified, synthetic, and direct estimators evaluated in the eight groups of auxiliary variables.*

		Accuracy Measures of Design-Based Estimators					
		Males		Females		Total	
		MARB	MRMSE	MARB	MRMSE	MARB	MRMSE
Poststratified	E	1.119	47.695	2.022	38.778	1.377	30.964
	EC	6.770	44.306	5.709	36.311	6.152	29.737
	EN	1.521	48.343	3.095	38.445	2.277	30.966
	EP	9.292	44.320	6.178	36.718	7.491	29.855
	ES	2.665	47.862	3.350	39.790	3.039	31.444
	ESC	17.380	45.021	11.367	37.627	13.919	31.656
	ESN	4.562	48.221	5.634	39.322	5.150	31.330
	ESP	17.620	45.703	12.526	38.048	14.688	32.136
Synthetic	E	17.246	22.248	13.585	17.956	13.451	16.811
	EC	12.114	17.867	12.426	16.482	10.746	14.265
	EN	17.869	22.778	13.064	17.679	13.589	16.949
	EP	13.645	18.900	11.526	15.646	10.765	14.171
	ES	6.151	22.312	8.017	18.962	6.022	15.451
	ESC	7.941	22.063	8.471	18.679	6.741	15.283
	ESN	5.419	22.184	8.158	19.150	5.973	15.504
	ESP	7.986	21.791	7.476	18.251	7.534	15.313
Direct		1.043	47.307	1.770	38.789	1.232	30.828

2.3 Model-based estimators

Model-based estimators are essentially based and derived from models. To estimate in a particular area, the models "borrow strength" from other related areas, improving the quality and efficiency of the estimation procedure. In this regard, many classical inferential tools are available in small-area estimation (SAE). Frequently, the goal in SAE is to obtain the best linear unbiased prediction (BLUP) estimators. The BLUP estimators minimize the mean squared error (MSE) among the class of linear unbiased estimators. These estimators usually depend on the covariance matrix of the random effects that can be estimated by several methods as maximum likelihood, restricted maximum likelihood or the method of fitting of constants. When we estimate the variance components and plug these values in the BLUP estimator, the resulting estimator is called empirical BLUP or EBLUP (see for instance Rao, 2003, pp. 95).

Both, model-based and model-assisted estimators use models. However, model-assisted estimators are built to produce design-consistent estimators because they are derived under the design-based theory, while model-based estimators are developed under the prediction theory (see, for instance, Valliant *et al.*, 2000). This means that the sampling scheme is usually ignored in the model-based perspective. There are some recent attempts in the literature to introduce sampling weights in model-based estimators (see, for example, You and Rao, 2002; Militino *et al.*, 2006, Militino *et al.*, 2007b). In summary, model-assisted and model-based procedures produce competitor estimators that are used by statistical offices, and both have mixed reviews, as one may find fans and detractors of both procedures in the literature. An important key-point is the different way of calculating the mean squared prediction error.

The model-based theory, called prediction theory, considers $y_1, \ldots, y_N$ as realizations of the random variables $Y_1, \ldots, Y_N$. Splitting the population of area d in sample (s_d) and non-sample units (r_d), the total of Y in area d, called T_d, can be expressed as

$$T_d = \sum_{j \in s_d} y_{jd} + \sum_{j \in r_d} y_{jd}$$

The task of estimating T_d becomes one of predicting the value of $\sum_{j \in r_d} y_{jd}$ for the non-observed variable $\sum_{j \in r_d} Y_{jd}$, and therefore the estimator is written as the sum of the sample and predicted observations

$$\hat{T}_d = \sum_{j \in s_d} y_{jd} + \sum_{j \in r_d} \hat{Y}_{jd}.$$

If the sampling fractions are negligible, the above predictor can be written as

$$\hat{T}_d = \sum_{j=1}^{N_d} \hat{Y}_{jd}.$$

To obtain estimators under the prediction theory, different models may be used. In this paper, linear models, logit models, and some mixed models have been considered. When the linear model (3) is assumed, and the sampling fractions are negligible, the predictor of the total number of unemployed in area d is given by

$$\hat{T}_d^{linear} = \sum_{j=1}^{N_d} \hat{Y}_{jd} = \mathbf{X}_d \hat{\boldsymbol{\beta}} \tag{6}$$

where $\mathbf{X}_d = (X_{d1}, X_{d2}, \ldots, X_{dp})^{\mathsf{T}}$ is the total population vector of the p covariates in area d. Alternative estimators can be obtained depending on the use of sampling weights to estimate $\boldsymbol{\beta}$, and the inclusion of the areas as random or fixed effects in the model. In this section, fixed effects are also considered because the reduced number of small areas in Navarre produces a lack of significance of the variance component of the random effects in some models. The following alternative estimators based on linear models are considered in the Monte Carlo study

(a) A synthetic estimator, called Linear Synthetic, assuming that in model (3), $\epsilon_{jd} \sim N(0, \sigma^2)$.

(b) A synthetic estimator, called Linear Synthetic W, based on a weighted linear model where $\epsilon_{jd} \sim N(0, \sigma^2/w_{jd})$.

(c) An estimator, called Linear F, based on a linear model with an area fixed effect, and $\epsilon_{jd} \sim N(0, \sigma^2)$.

(d) An estimator, called Linear WF, based on a weighted linear model with an area fixed effect similar to model (c), but also assuming that $\epsilon_{jd} \sim N(0, \sigma^2/w_{jd})$.

When the logit model (4) is assumed, the estimator of the total number of unemployed in the dth area is given by

$$\hat{T}_d^{logit} = \sum_{j=1}^{N_d} \frac{\mathrm{e}^{\mathbf{x}_{jd}^{\mathsf{T}} \hat{\boldsymbol{\beta}}}}{1 + \mathrm{e}^{\mathbf{x}_{jd}^{\mathsf{T}} \hat{\boldsymbol{\beta}}}}. \tag{7}$$

The following alternative models are also considered:

(e) A synthetic estimator, called Logit Synthetic, based on a logit model that does not incorporate sampling weights in the estimation process.

(f) A synthetic estimator, called Logit Synthetic W, based on a weighted logit model similar to (e) but including weights w_{jd} when estimating $\boldsymbol{\beta}$.

(g) An estimator, called Logit F, based on a logit model with an area fixed effect that estimates $\boldsymbol{\beta}$ without sampling weights.

Table 2: *Mean of absolute relative bias (MARB), and mean of the relative root mean square error (MRMSE) for the composite design-based estimators evaluated in the 8 groups of auxiliary variables and the direct estimator for comparison purposes.*

		Accuracy Measures of the Design-Based Estimators					
		Males		Females		Total	
		MARB	MRMSE	MARB	MRMSE	MARB	MRMSE
Composite 1	E	1.258	47.030	1.776	38.104	1.099	30.424
($\alpha = 2/3$)	EC	6.221	43.683	5.386	35.673	5.762	29.199
	EN	1.024	47.743	2.751	37.880	1.921	30.477
	EP	8.762	43.634	5.814	36.106	7.107	29.310
	ES	2.624	47.299	3.148	39.141	2.853	30.944
	ESC	16.956	44.488	11.072	37.024	13.606	31.165
	ESN	4.250	47.727	5.323	38.779	4.900	30.894
	ESP	17.242	45.121	12.175	37.461	14.384	31.646
Composite 2	E	1.447	43.943	1.233	35.799	1.047	28.641
($\alpha = 1$)	EC	5.174	40.808	4.442	33.337	4.830	27.327
	EN	1.119	44.657	2.170	35.619	1.300	28.721
	EP	7.558	40.686	4.836	33.801	6.156	27.484
	ES	2.676	44.604	2.577	37.063	2.470	29.371
	ESC	15.836	42.124	9.987	34.887	12.583	29.467
	ESN	3.891	45.158	4.517	36.799	4.306	29.384
	ESP	16.355	42.687	11.127	35.349	13.473	29.983
Composite 3	E	5.668	33.199	4.089	27.074	4.116	21.754
($\alpha = 1.5$)	EC	3.099	30.150	4.088	24.984	3.277	20.125
	EN	5.470	33.718	3.558	26.895	3.644	21.762
	EP	3.579	29.847	2.925	25.081	2.892	20.035
	ES	3.397	35.194	3.105	29.120	2.858	23.042
	ESC	11.620	33.272	6.644	27.448	8.901	23.004
	ESN	3.717	35.656	3.524	29.021	3.654	23.111
	ESP	12.716	33.759	7.731	27.705	9.990	23.398
Composite 4	E	8.562	27.675	6.429	22.551	6.450	18.567
($\alpha = 2$)	EC	4.447	24.432	5.398	20.637	4.636	16.604
	EN	8.570	28.084	5.933	22.301	6.121	18.534
	EP	4.565	24.127	4.812	20.425	4.301	16.425
	ES	4.086	30.124	4.333	24.850	3.586	19.795
	ESC	9.062	28.586	5.686	23.541	6.757	19.704
	ESN	4.101	30.460	4.200	24.800	3.860	19.872
	ESP	10.487	28.967	5.810	23.636	7.954	20.004
Direct		1.043	47.307	1.770	38.789	1.232	30.828

(h) An estimator, called Logit WF, based on a weighted logit model with an area fixed effect but including the sampling weights w_{jd} in the estimation process of $\boldsymbol{\beta}$.

(i) An estimator, called EB Mixed Logit, based on a weighted logit model with an area random effect.

The EB Mixed Logit estimator of the total number of unemployed in the dth area is given by

$$\hat{T}_d^{logitMixed} = \sum_{j=1}^{N_d} \frac{e^{\mathbf{x}_{jd}^T \hat{\boldsymbol{\beta}} + \hat{u}_d}}{1 + e^{\mathbf{x}_{jd}^T \hat{\boldsymbol{\beta}} + \hat{u}_d}}.$$

To produce real small-area estimates, official statistical agencies must adjust the small-area estimates to make them coherent with more accurate values at some level

Table 3: *Mean of the absolute relative bias (MARB), and mean of the relative root mean square error (MRMSE) for model-assisted estimators evaluated in the 8 groups of auxiliary variables and the direct estimator for comparison purposes.*

		Accuracy Measures of the Model-Assisted Estimators					
		Males		Females		Total	
		MARB	MRMSE	MARB	MRMSE	MARB	MRMSE
Linear GREG	E	1.011	46.713	1.733	38.126	1.149	30.236
	EC	0.445	43.395	1.724	36.271	0.865	28.998
	EN	0.960	46.686	1.710	38.151	1.129	30.265
	EP	0.870	42.793	1.347	35.975	0.929	28.608
	ES	0.993	46.192	0.983	37.543	0.582	29.893
	ESD	2.010	42.514	1.346	35.480	0.761	28.387
	ESN	1.050	46.162	0.989	37.565	0.564	29.909
	ESP	1.040	42.094	1.150	35.319	1.319	28.209
Logit GREG	E	1.011	46.713	1.733	38.126	1.149	30.236
	EC	0.442	43.616	1.772	36.356	0.948	29.169
	EN	1.006	46.657	1.714	38.138	1.130	30.261
	EP	0.749	42.962	1.345	36.076	0.983	28.819
	ES	0.940	46.531	1.708	37.946	1.034	30.087
	ESD	1.206	43.253	2.007	35.920	1.499	28.859
	ESN	0.963	46.478	1.702	37.963	1.045	30.103
	ESP	3.459	42.909	2.443	35.847	2.604	28.754
Mixed Logit GREG	E	1.005	46.732	1.706	38.158	1.116	30.244
Direct		1.043	47.307	1.770	38.789	1.232	30.828

Table 4: *Mean of the absolute relative bias (MARB), and mean of the relative root mean square error (MRMSE) for the linear model-based estimators evaluated in the eight groups of auxiliary variables. The direct estimator is also included for comparison purposes.*

		Accuracy Measures of the Model-Based Estimators					
		Males		Females		Total	
		MARB	MRMSE	MARB	MRMSE	MARB	MRMSE
Linear Synthetic	E	18.752	23.589	14.035	18.603	14.604	17.884
	EC	12.897	18.428	12.606	16.765	11.382	14.771
	EN	19.429	24.346	13.234	17.931	14.349	17.736
	EP	14.384	19.521	11.862	15.981	11.286	14.709
	ES	6.182	21.700	8.735	18.506	6.218	15.096
	ESC	7.726	19.985	9.338	17.788	7.367	14.315
	ESN	5.539	21.494	8.615	18.489	6.250	15.117
	ESP	7.961	20.056	7.876	17.090	7.479	14.441
Linear Synthetic W	E	17.246	22.248	13.585	17.956	13.451	16.811
	EC	12.202	17.904	12.458	16.441	10.784	14.308
	EN	17.829	22.841	12.893	17.394	13.206	16.702
	EP	13.466	18.755	11.558	15.648	10.704	14.164
	ES	6.155	21.787	8.653	18.554	6.176	15.146
	ESC	7.708	20.096	9.248	17.831	7.303	14.368
	ESN	5.510	21.580	8.529	18.534	6.215	15.169
	ESP	7.937	20.210	7.794	17.159	7.465	14.517
Linear F	E	3.613	43.794	3.815	33.671	2.215	27.742
	EC	6.817	38.015	5.339	30.544	4.062	24.984
	EN	3.736	43.743	3.857	33.716	2.256	27.757
	EP	5.580	38.522	4.792	30.988	3.552	25.310
	ES	3.645	43.192	4.267	33.555	2.346	27.410
	ESD	7.555	37.607	6.019	30.428	4.462	24.806
	ESN	3.757	43.099	4.292	33.602	2.382	27.424
	ESP	6.034	38.043	5.334	30.944	3.827	25.098
Linear WF	E	2.654	43.702	3.298	33.663	1.801	27.683
	EC	6.041	37.972	4.981	30.546	3.652	24.941
	EN	2.730	43.638	3.362	33.700	1.827	27.693
	EP	4.673	38.475	4.312	30.999	3.018	25.286
	ES	3.390	43.279	4.031	33.645	2.169	27.481
	ESC	7.218	37.702	5.797	30.534	4.206	24.874
	ESN	3.500	43.184	4.062	33.687	2.208	27.493
	ESP	5.686	38.153	5.078	31.029	3.536	25.178
Direct		1.043	47.307	1.770	38.789	1.232	30.828

Table 5: *Mean of the absolute relative bias (MARB), and mean of the relative root mean square error (MRMSE) for the logit model-based estimators evaluated in the 8 groups of auxiliary variables. The direct estimator is also included for comparison purposes.*

		Accuracy Measures of the Model-Based Estimators					
		Males		Females		Total	
		MARB	MRMSE	MARB	MRMSE	MARB	MRMSE
Logit Synthetic	E	18.752	23.589	14.035	18.603	14.604	17.884
	EC	13.797	19.060	12.448	16.948	12.382	15.339
	EN	19.181	24.073	13.336	18.038	14.422	17.772
	EP	15.536	20.364	12.043	16.316	12.322	15.382
	ES	6.046	22.259	7.971	18.838	5.973	15.340
	ESC	7.340	21.463	8.351	18.315	6.555	14.909
	ESN	5.526	22.067	8.110	18.861	5.997	15.363
	ESP	7.823	21.481	7.319	17.834	7.297	15.068
Logit Synthetic W	E	17.247	22.248	13.585	17.957	13.451	16.811
	EC	12.885	18.471	12.254	16.535	11.570	14.743
	EN	17.598	22.595	12.966	17.493	13.279	16.736
	EP	14.402	19.518	11.712	15.899	11.437	14.697
	ES	6.056	22.253	7.980	18.839	5.982	15.342
	ESC	7.350	21.460	8.351	18.313	6.555	14.909
	ESN	5.527	22.061	8.129	18.864	6.011	15.369
	ESP	7.833	21.485	7.328	17.839	7.299	15.069
Logit F	E	1.410	46.985	1.755	38.242	1.174	30.481
	EC	1.385	44.263	2.003	36.749	1.120	29.434
	EN	1.411	46.948	1.711	38.282	1.152	30.516
	EP	1.917	44.010	1.664	36.414	1.182	29.161
	ES	0.815	46.880	1.717	38.351	1.036	30.377
	ESC	1.168	44.305	1.789	36.843	1.029	29.413
	ESN	0.815	46.820	1.690	38.389	1.022	30.404
	ESP	1.350	44.034	1.477	36.565	0.981	29.153
Logit WF	E	1.059	46.879	1.842	38.271	1.179	30.410
	EC	1.130	44.315	1.829	36.783	0.922	29.432
	EN	1.042	46.835	1.816	38.310	1.162	30.445
	EP	1.636	43.978	1.567	36.472	1.046	29.150
	ES	0.942	47.038	1.738	38.504	1.077	30.469
	ESC	1.893	44.177	2.131	36.824	1.488	29.364
	ESN	0.954	46.979	1.721	38.538	1.066	30.497
	ESP	4.222	43.623	2.583	36.462	2.644	29.034
EB mixed logit	E	28.909	33.919	17.840	23.217	16.279	21.109
Direct		1.043	47.307	1.770	38.789	1.232	30.828

of aggregation. This adjustment is necessary because when aggregating small-area estimates within the same region (province), the sum of these small-area estimates do not generally coincide with the estimate obtained using an appropriate estimator for the larger region. This adjustment process is called benchmarking and in this study it should be done to the provincial estimate. Ugarte *et al.* (2008) show how to introduce constraints into a linear mixed model to produce final benchmarked estimates in small areas.

3 Monte Carlo simulation study

In this section the performance of the estimators described in Section 2 to estimate unemployment in the small areas of Navarre is evaluated. We have drawn $K = 500$ samples from the 2001 Census following the same sampling design as the SLFS. To assess the estimators bias the mean absolute relative bias (*MARB*) over the D small areas is computed. Namely

$$MARB(\hat{Y}) = \frac{1}{D}\sum_{d=1}^{D} |RB_d(\hat{Y})| \quad \text{where} \quad RB_d(\hat{Y}) = \frac{1}{K}\sum_{k=1}^{K} \frac{\hat{Y}_d(k) - Y_d}{Y_d} 100.$$

To evaluate the estimators prediction errors the mean of the square root of the relative mean squared error (*MRMSE*) over the D small areas is considered

$$MRMSE(\hat{Y}) = \frac{1}{D}\sum_{d=1}^{D} RMSE_d(\hat{Y}) \quad \text{where} \quad RMSE_d(\hat{Y}) = \left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\hat{Y}_d(k) - Y_d}{Y_d}\right)^2\right)^{\frac{1}{2}} 100.$$

The estimator giving a better balance between bias and prediction error will be selected as the best proposal for estimating unemployment in the small areas of Navarre. The expertise of the local statistical office has also been considered and this aspect will be discussed at the end of this section.

Now, let us recall that there are a total of seven small areas in Navarre for which we have evaluated the seven design-based estimators (including the direct), three GREG estimators, and nine model-based estimators. A total of eight combinations of auxiliary variables have been used: (E) age-sex, (EC) age-sex-claimant, (EN) age-sex-educational level, (EP) age-sex-employment status in the SNE, (ES) age-sex-stratum, (ESC), age-sex-stratum-claimant status in the SNE, (ESN) age-sex-stratum-educational level, and (ESP) age-sex-stratum-employment status in the SNE.

Tables 1 and 2 display the MARB and the MRMSE of the design-based estimators for the eight combinations of auxiliary variables. In general, the post-stratified estimator exhibits neither large biases nor low predictions errors as it is expected because it is a

direct estimator. However, both post-stratified estimators ESC and ESP present biases around 17% (for males), and around 14% (for females). The synthetic estimators E, EC, EN, and EP present large biases (for both males and females), indicating that the assumption of a similar behaviour of the small areas with respect to the large region does not hold. We also observe that the bias is small when the stratum (S) auxiliary variable is considered. The behaviour of composite 1, composite 2, and composite 3 estimators (see Table 2) is similar to the post-stratified estimator, and they have, in general, small biases, except ESC and ESP that present larger biases. This is not surprising because, as α is not very large, these composite estimators give more weight to the post-stratified component. This also means that the error is high. Composite 4 estimator is slightly more biased than the rest of composite estimators, but its error is reduced up to a half. Then, it might be a good option to achieve a compromise between bias and error.

Table 3 shows the MARB and the MRMSE for the model-assisted estimators (GREG). All of them exhibit practically the same results. They have negligible bias, but the error is not reduced with regard to the direct estimator. Note that the linear GREG estimator is a modified direct estimator, and then it is approximately unbiased. However, it does not increase the effective sample size (see Rao, 2003, chapter 2, p. 20), and then, the error is not decreased. The mixed logit GREG suffers from the same deficiencies as both the linear and the logit GREG.

The results for the model-based estimators are displayed in Tables 4 and 5. The model-based synthetic estimators (linear and logit) are very similar with regard to bias and error. When the auxiliary variable stratum (S) is not considered, the bias is large, but it is notably reduced when stratum is introduced in the models. The error has been decreased, and it is about one half the error of the direct estimator. Those model-based estimators making use of a fixed area effect are practically unbiased, but the error is unacceptably high (as large as the error of the direct estimator).

For an optimal choice between the estimators, we first select the estimator with a smaller MARB – excluding, of course, the classical direct estimator that has an enormous MSE in small areas – and later we select those with a smaller MRMSE. It does not seem reasonable to use just the MRMSE as the single option to choose a sensible estimator, because the statistical office will be reluctant to use an estimator with a large bias. They would agree to accept some bias to reduce variability when estimating in small areas, but not a big amount of bias, because traditionally they are used to work with unbiased estimators. The choice among estimators is then not easy, however, looking at Tables 1, 2, 3 and 4 – the synthetic estimators with auxiliary variables (EC, EN, ES, ESC, ESN), the composite 4 estimators with auxiliary variables (EC and EP), and the linear synthetic estimators with auxiliary variables (ES and ESN) outperform the rest. Figure 2 shows the accuracy measures for these estimators. Unfortunately, the performance of the model-based estimators is not as promising as it should be. In particular, those model-based estimators introducing a fixed area effect exhibit errors as large as the direct estimator. The bias is also too big when introducing a random area effect (as in the EB mixed logit). This error is reduced if the area effect is removed,

but in this situation, the bias is large if the auxiliary variable stratum is not included in the model. In this case, the performance of the model-based estimators is pretty similar to the design-based synthetic estimator. A different behaviour is observed for the post-stratified and the composite 1, 2, and 3 estimators. In general, they exhibit low bias (except in two cases with the auxiliary variable stratum), but they present large errors. The best balance between bias and error is achieved with composite 4 estimator. More precisely, the composite 4 EP. The bias is around 5% and the error is one half the error of the direct estimator. Therefore, we consider it a reasonable estimator for estimating unemployment in the small areas of Navarre. In addition, this estimator is very appealing to the local statistical office as it combines two sources of measuring unemployment: namely, data from the SLFS and data from the Navarre Employment Register.

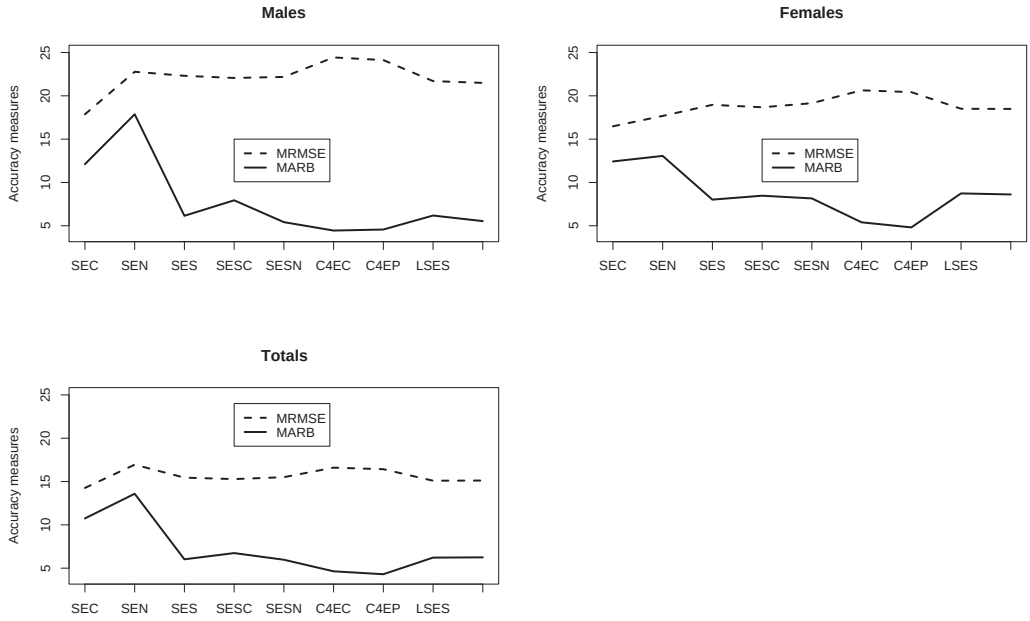

Figure 2: *Mean of the absolute relative bias (MARB) and mean of the relative square error (MRMSE) of the synthetic estimator with auxiliary variables (EC, EN, ES, ESC, ESN), Composite 4 estimator with auxiliary variables (EC and EP) and Logit Synthetic estimator with auxiliary variables (ES, ESN) for males, females, and totals over the seven small areas of Navarre*

4 Estimators of the mean squared error

The main interest of this article is to choose an appropriate estimator to estimate unemployment in very small areas in the province of Navarre, where sometimes theoretical assumptions are not well fulfilled, and practical performance needs to be explored. In the last section composite 4 EP has been chosen as the appropriate

estimator, and so this section is devoted to the estimation of its mean squared error. Here, three alternative MSE estimators are derived using three well-known procedures in the literature: the variance linearization method, and two resampling methods, the jackknife and the bootstrap.

4.1 The variance linearization method

The variance linearization method, or delta method, consists of applying a Taylor series expansion to a function of the estimators of the total, and calculating the variance of this function through the variance of its derivatives with regard to these totals (Woodruff, 1971). Let us define the following indicator variables $I_k(h,i,j) = 1$ if person j ($j = 1,\ldots,m_{hi}$) of household i, ($i = 1,\ldots,n_h$) and stratum h, ($h = 1,\ldots,H$) is in group k, and 0 otherwise, $z_{hij} = y_{jd} I_k(h,i,j)$ and $v_{hij} = w_{jd} I_k(h,i,j)$.

Post-stratified and synthetic estimators of the mean of Y_d can be written as

$$\widehat{\bar{Y}}_d^k = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij} z_{hij} \right) / v_{...} \text{ , where } \qquad v_{...} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij}, \tag{8}$$

because when k is the group g in the domain d, $\bar{Y}_d^k$ is the post-stratified estimator of the mean (see expression (1)), and when $k = g$, $\bar{Y}_d^k$ is the synthetic estimator of the mean (see expression (2)). For both of them the linearized estimator of the variance is the following

$$\begin{aligned} &\widehat{\text{Var}}_L(\widehat{\bar{Y}}_d^k) = \sum_{h=1}^{H} \widehat{\text{Var}}_h(\widehat{\bar{Y}}_d^k), \quad \text{where} \quad \widehat{\text{Var}}_h(\widehat{\bar{Y}}_d^k) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (U_{hi.} - \bar{U}_{h..})^2, \\ &U_{hi.} = \frac{1}{v_{...}} \sum_{j=1}^{m_{hi}} v_{hij} \left(z_{hij} - \widehat{\bar{Y}}_d^k \right), \quad \text{and} \quad \bar{U}_{h..} = \frac{1}{n_h} \sum_{i=1}^{n_h} U_{hi.}. \end{aligned} \tag{9}$$

The estimator of the total of Y in area d, is given by

$$\widehat{Y}_d^k = \sum_{g=1}^{G} \widehat{\bar{Y}}_d^k N_{dg}. \tag{10}$$

Then, the variance linearized estimator is calculated as a weighted sum such that

$$\widehat{\text{Var}}_L(\hat{Y}_d^k) = \sum_{g=1}^{G} \widehat{\text{Var}}_L(\widehat{\bar{Y}}_d^k) N_{dg}^2. \tag{11}$$

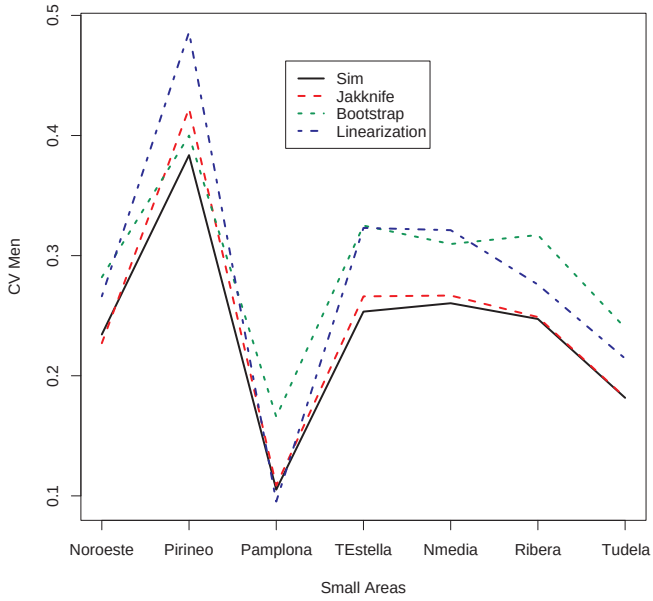

Figure 3: *Coefficient of variation of composite 4 EP (males)*

To calculate the bias of the synthetic estimator given by $\text{Bias}(\hat{Y}_d^{sint}) = -\sum_{j=1}^{N_d} \epsilon_{jd}$, the following estimator is provided (see Ghosh and Särndal, 2001)

$$\widehat{\text{Bias}}(\hat{Y}_d^{sint}) = -N_d \frac{1}{n_d} \sum_{j=1}^{n_d} \hat{\epsilon}_{jd}, \qquad \text{where} \qquad \hat{\epsilon}_{jd} = y_{jd} - \hat{\bar{Y}}_g. \tag{12}$$

Therefore

$$\widehat{MSE}_L(\hat{Y}_d^{sint}) = \widehat{\text{Var}}_L(\hat{Y}_d^{sint}) + \widehat{Bias}^2(\hat{Y}_d^{sint}),$$

and finally

$$\widehat{MSE}_L(\hat{Y}_d^{comp}) \doteq \lambda_d^2 \widehat{MSE}_L\left(\hat{Y}_d^{post}\right) + (1-\lambda_d)^2 \widehat{MSE}_L\left(\hat{Y}_d^{sint}\right). \tag{13}$$

Note that the $\widehat{MSE}_L\left(\hat{Y}_d^{post}\right)$ is calculated using expression (11) for k defined as a combination of groups of g and d, and the covariance between the post-stratified and the synthetic estimator has been considered negligible.

4.2 The jackknife estimator

The jackknife method was introduced by Quenouille (1949, 1956) as a method to reduce the bias, and later Tukey (1958) proposed its use for estimating variances and confidence

intervals. In the jackknife method, we take as many sub-samples as clusters (census sections) are in the sample, because sub-samples are obtained leaving one cluster out every time from the original sample. Let $\hat{Y}^k_{d(hi)}$ be the estimator $\hat{Y}^k_d$ obtained by dropping a cluster i from the hth stratum. Then, original weights w_{jd} must be substituted by $w_{jd(hi)}$, where

$$w_{jd(hi)} = \begin{cases} w_{jd} & \text{if } j \text{ is not in the } h \text{ stratum} \\ 0 & \text{if } j \text{ is in cluster } i \text{ of the } h \text{ stratum} \\ \frac{n_h}{n_h-1} w_{jd} & \text{if unit } j \text{ is in the } h \text{ stratum but not in cluster } i \end{cases} \tag{14}$$

The jackknife estimator of the MSE of $\hat{Y}^k_d$ can be obtained as

$$\widehat{MSE}_{JK}(\hat{Y}^k_d) \doteq \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} [\hat{Y}^k_{d(hi)} - \hat{Y}^k_{d(h.)}]^2, \tag{15}$$

where $\hat{Y}^k_{d(h.)} = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{Y}^k_{d(hi)}$ and superscript k indicates the composite estimator. Note that Expression (15) is approximate as we are ignoring the possible bias of the composite estimator.

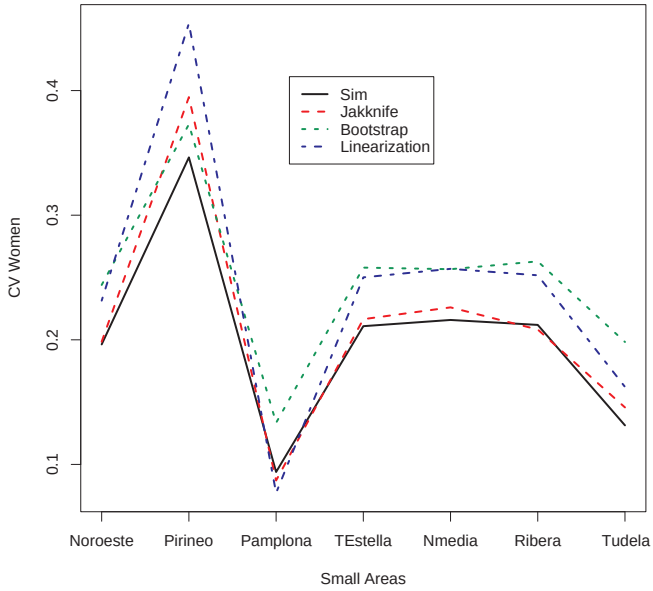

Figure 4: *Coefficient of variation of composite 4 EP (females)*

4.3 The bootstrap estimator

In the bootstrap, the sub-samples are obtained by random sampling with replacement. Similarly to the jackknife, for every sub-sample new weights are defined in each step. The rescaled bootstrap estimator in a stratified random sampling has been provided by Rao and Wu (1988). It assumes the following steps

1. Given the h stratum we have a sample of n_h clusters. From the sample of the h stratum, a sub-sample of $n_h - 1$ clusters is drawn by random sampling with replacement.
2. For every sub-sample r ($r = 1, 2, \ldots, R$) a new weight is defined

$$w_{jd}(r) = w_{jd} \frac{n_h}{n_h - 1} m_i(r), \tag{16}$$

   where $m_i(r)$ is the number of times that cluster i is chosen in the sub-sample.
3. Repeat steps (1) and (2) R times.

To derive the bootstrap estimator of the MSE of $\hat{Y}_d^k$ we calculate

$$\widehat{MSE}_B(\hat{Y}_d^k) = \frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{Y}_{rd}^{*k} - \hat{Y}_d^k\right)^2, \tag{17}$$

where $\hat{Y}_{rd}^{*k}$ is similar to (10) but using the new weight $w_{jd}(r)$ when estimating the mean (8) and k indicates the composite estimator.

In the Monte Carlo study the composite 4 EP estimator has been computed with 500 samples. For the bootstrap mean squared error estimator, R has taken the values $200, 500, 1000$, and 4000. Small values of R lead to differences in the estimator performance, but values of R equal to 1000 and higher provide similar results. Both Figures 3 and 4 show the coefficients of variation for males and females respectively. The coefficient of variation obtained from the census data is depicted using a continuous line. All of the methods proposed here tend to overestimate the MSE, particularly when the sample size is small. However, the best behaviour corresponds to the jackknife estimator, because the corresponding coefficients of variation are very close to the real ones.

Conclusions

Small area estimation is becoming a challenge in European statistical offices because of the increasing demand of precise estimates at county or regional level. Unfortunately,

procedures used in other regions and/or countries seem not to be directly applicable everywhere, because they are based on a large number of small areas, and the availability of the auxiliary information is not the same for every country. In some regions such as Navarre, the task of estimating unemployment in very small areas is not easy, not only because of the reduced number of small areas and the great heterogeneity between them, but also because unemployment has a low incidence in the population. Both reasons can worsen the performance of model-assisted and model-based estimators, which were promising in other scenarios.

In this work a composite 4 EP estimator has shown to be a reasonable alternative for estimating unemployment in Navarre. This estimator comes from a linear combination of a direct estimator (a poststratified estimator) and an indirect estimator (a synthetic estimator). The accuracy measures evaluated through a Monte-Carlo study have shown its good trade-off in terms of bias and MSE. This estimator is easy to calculate and interpret, and the MSE can be derived using jackknife. The composite 4 estimator uses as auxiliary information the age-sex (E) groups and the employment register in Navarre (P). Although it is known that this later register might overestimate unemployment, the combination of the two sources of data to estimate unemployment (namely, the SLFS and the Navarre employment register) is very appealing to the local statistical office.

Acknowledgments

We thank a referee for the helpful comments that have contributed to improving an earlier version of this paper. This research has been supported by the Spanish Ministry of Science and Education (project MTM2005-00511) and the Spanish Ministry of Science and Innovation (MTM2008-03085).

References

Costa, A., Satorra, A. and Ventura, E. (2002). Estimadores compuestos en estadística regional: Aplicación para la tasa de variación de la ocupación en la industria, *Qüestió*, 26, 213-243.

Costa, A., Satorra, A. and Ventura, E. (2003). An empirical evaluation of small area estimators. *SORT (Statistics and Operations Research Transactions)*, 27, 113-135.

Costa, A., Satorra, A. and Ventura, E. (2004). Using composite estimators to improve both domain and total area estimation. *SORT (Statistics and Operations Research Transactions)*, 28, 69-86.

Costa, A., Satorra, A. and Ventura, E. (2006). Improving small area estimation by combining surveys: new perspectives in regional statistics. *SORT (Statistics and Operations Research Transactions)*, 28, 69-86.

Drew, D, Singh, M. P. and Choudhry, G. H. (1982). Evaluation of small area estimation techniques for the Canadian labor Force Survey. *Survey Methodology*, 8, 17-47.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

Ghosh, N. and Särndal, C. E. (2001). *Lecture Notes on Estimation for Population Domains and Small Areas*. Statistics Finland, vol. 48.

González, M. E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section* 33-36. American Statistical Association. Washington, D.C.

McCullagh P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall.

Militino A. F.,Ugarte, M. D., Goicoa, T. and González-Audícana, M. (2006). Using small area models to estimate the total area occupied by olive trees. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 450-461.

Militino A. F., Ugarte, M. D. and Goicoa, T. (2007a). A BLUP synthetic versus an EBLUP estimator: an empirical study of a small area estimation problem. *Journal of Applied Statistics*, 34, 153-165.

Militino A. F., Ugarte, M. D. and Goicoa, T. (2007b). Combining sampling and model weights in agricultural small area estimation. *Environmetrics*, 18, 87-99.

Morales, D., Esteban, M. D., Sánchez, A. Santamaría, L., Marhuenda, Y., Pérez, A., Saralegui, J. and Herrador, M. (2007). Estimación en áreas pequeñas con datos de la Encuesta de Población Activa en Canarias. *Estadística Española*, 49, 301-332.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society Series B*, 11, 18-84.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Särndal, C. E., Swensson B. and Wretman, J. H. (1992). *Model Assisted Survey Sampling.* Springer.

Tukey, J. (1958). Bias and confidence in not quite large samples (abstract). *Annals of Mathematical Statistics*, 29, 614.

Ugarte, M. D., Militino, A. F. and Goicoa, T. (2008). Benchmarked estimates in small areas using linear mixed models with restrictions, *Test*, in press. DOI 10.1007/s11749-008-0094-x

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference: a Prediction Approach.* Wiley Series in Survey Methodology.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414

You, Y. and Rao J. N. K. (2002). A pseudo-empirical best linear prediction approach to small-area estimation using survey weights. *Canadian Journal of Statistics*, 30, 431-439.

SORT 33 (1) January-June 2009, 71-84

A general procedure of estimating the population mean in the presence of non-response under double sampling using auxiliary information

Housila P. Singh and Sunil Kumar

Abstract

In the present study, we propose a general class of estimators for population mean of the study variable in the presence of non-response using auxiliary information under double sampling. The expression of mean squared error (MSE) of the proposed class of estimators is derived under double (two-stage) sampling. Some estimators are also derived from the proposed class by allocating the suitable values of constants used. Comparisons of the proposed strategy with the usual unbiased estimator and other estimators are carried out. The results obtained are illustrated numerically using an empirical sample considered in the literature.

MSC: 62D05

Keywords: Double sampling, Mean squared error, Non-response, Study variable, Auxiliary variable.

1 Introduction

In practice almost all surveys suffer from non-response. The problem of non-response often happens due to the refusal of the subject, absenteeism and sometimes due to the lack of information. The pioneering work of Hansen and Hurwitz (1946), assumed that a sub-sample of initial non-respondents is recontacted with a more expensive method, suggesting the first attempt by mail questionnaire and the second attempt by a personal interview. In estimating population parameters such as the mean, total or ratio, sample survey experts sometimes use auxiliary information to improve

School of Studies in Statistics, Vikram University, Ujjain-456010, M. P., India
Received: January 2009
Accepted: April 2009

precision of the estimates. Sodipo and Obisesan (2007) have considered the problem of estimating the population mean in the presence of non-response, in sample survey with full response of an auxiliary character x. Other authors such as Cochran (1977), Rao (1986, 1987), Khare and Srivastava (1993, 1995, 1997), Okafor and Lee (2000) and Tabasum and Khan (2004, 2006) and Singh and Kumar (2008a,b) have studied the problem of non-response under double (two-stage) sampling.

From a finite population $U = (U_1, U_2, ..., U_N)$, a large first phase sample of size n' is selected by simple random sampling without replacement (SRSWOR). A smaller second phase sample of size n is selected from n' by SRSWOR. Non-response occurs on the second phase sample of size n in which n_1 units respond and n_2 units do not. From the n_2 non-respondents, by SRSWOR a sample of $r = n_2/k$; $k > 1$ units is selected where k is the inverse sampling rate at the second phase sample of size n. All the r units respond this time round. The auxiliary information can be used at the estimation stage to compensate for units selected for the sample that fail to provide adequate responses and for population units missing from the sampling frame. In a household survey, for example, the household size can be used as an auxiliary variable for the estimation of, say, family expenditure. Information can be obtained completely on the family size during the household listing while there may be non-response on the household expenditure.

An unbiased estimator for the population mean $\bar{Y}$ of the study variable y, proposed by Hansen and Hurwitz (1946), is defined by

$$\bar{y}^* = w_1\bar{y}_1 + w_2\bar{y}_{2r},$$

where $w_1 = n_1/n$ and $w_2 = n_2/n$. The variance of $\bar{y}^*$ is given by

$$Var(\bar{y}^*) = \left(\frac{1-f}{n}\right) S_y^2 + \frac{W_2(k-1)}{n} S_{y(2)}^2,$$

where $f = n/N$, $S_y^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2 \Big/ (N-1)$, $S_{y(2)}^2 = \sum_{i=1}^{N_2} (y_i - \bar{Y}_2)^2 \Big/ (N_2 - 1)$, $\bar{Y} = \sum_{i=1}^{N} y_i/N$, $\bar{Y}_2 = \sum_{i=1}^{N_2} y_i/N_2$, $W_2 = N_2/N$, $\bar{y}_{2r} = \sum_{i=1}^{r} y_i/r$, N_1 and $N_2\,(= N - N_1)$ are the sizes of the responding and non-responding units from the finite population N.

It is well known that in estimating the population mean, sample survey experts sometimes use auxiliary information to improve the precision of the estimates. Let x denote an auxiliary variable with population mean $\bar{X} = \sum_{i=1}^{N} x_i/N$. Let $\bar{X}_1 = \sum_{i=1}^{N_1} y_i/N_1$ and $\bar{X}_2 = \sum_{i=1}^{N_2} x_i/N_2$ denote the population means of the response and non-response groups

(or strata). Let $\bar{x} = \sum_{i=1}^{n} x_i/n$ denote the mean of all the n units. Let $\bar{x}_1 = \sum_{i=1}^{n_1} x_i/n_1$ and $\bar{x}_2 = \sum_{i=1}^{n_2} x_i/n_2$ denote the means of the n_1 responding units and n_2 non-responding units. Further, let $\bar{x}_{2r} = \sum_{i=1}^{r} x_i/r$ denote the mean of the $r\,(=n_2/k), k>1$ sub-sampled units. With this background we define an unbiased estimator of population mean $\bar{X}$ as

$$\bar{x}^* = w_1\bar{x}_1 + w_2\bar{x}_{2r}.$$

The variance of $\bar{x}^*$ is given by

$$Var\,(\bar{x}^*) = \left(\frac{1-f}{n}\right) S_x^2 + \frac{W_2\,(k-1)}{n} S_{x(2)}^2,$$

where $S_x^2 = \sum_{i=1}^{N} (x_i - \bar{X})^2 \big/ (N-1)$, $S_{x(2)}^2 = \sum_{i=1}^{N_2} (x_i - \bar{X}_2)^2 \big/ (N_2 - 1)$.

In the present study, we propose a general class of estimators for the population mean $\bar{Y}$ of the study variable y in the presence of non-response under double sampling using auxiliary information. The expressions of bias and variance have been obtained to the first degree of approximation of the proposed class of estimators, which will enable us to obtain these expressions for any member of this family. Some estimators are shown to be particular members of this family. Comparison of the proposed strategy with the usual unbiased estimator and other estimators are carried out. An empirical study is presented to expound the performance of the proposed class of estimators.

2 The proposed family of estimators

We define a class of estimators for the population mean $\bar{Y}$ of the study variable y as

$$T_{DS} = \bar{y}^* \left(\frac{a\bar{x}^* + b}{a\bar{x}' + b}\right)^{\alpha} \left(\frac{a\bar{x} + b}{a\bar{x}' + b}\right)^{\beta}, \tag{1}$$

where $\bar{x}'$ denote the sample mean of x based on first phase sample of size n', $a\,(\neq 0)\,, b$ are either real numbers or functions of known parameters such as standard deviation (σ), Coefficient of variation (C_x), Correlation coefficient (ρ) etc. of the auxiliary variable x, and α, β are suitable chosen constants.

To obtain the bias and variance of the class of estimators T_{DS}, we write

$$\bar{y}^* = \bar{Y}(1+\varepsilon_0), \quad \bar{x}^* = \bar{X}(1+\varepsilon_1), \quad \bar{x}' = \bar{X}(1+\varepsilon_1'), \quad \bar{x} = \bar{X}(1+\varepsilon_2)$$

such that

$$E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_1') = E(\varepsilon_2) = 0$$

and

$$E(\varepsilon_0^2) = \lambda S_y^2 + \lambda^* S_{y(2)}^2, \quad E(\varepsilon_1^2) = \lambda S_x^2 + \lambda^* S_{x(2)}^2, \quad E\left(\varepsilon_1^{\prime 2}\right) = \lambda' S_x^2, \quad E(\varepsilon_2^2) = \lambda S_x^2,$$
$$E(\varepsilon_0\varepsilon_1) = \lambda \rho_{yx} S_y S_x + \lambda^* \rho_{yx(2)} S_{y(2)} S_{x(2)}, \quad E(\varepsilon_2\varepsilon_1') = \lambda' S_x^2, \quad E(\varepsilon_1\varepsilon_1') = \lambda' S_x^2,$$
$$E(\varepsilon_1\varepsilon_2) = \lambda S_x^2, \quad E(\varepsilon_0\varepsilon_2) = \lambda \rho_{yx} S_y S_x, \quad E(\varepsilon_0\varepsilon_1') = \lambda' \rho_{yx} S_y S_x,$$

where ρ_{yx} and $\rho_{yx(2)}$ are respectively the correlation coefficient of response and non-response group between study variable y and auxiliary variable x,

$$\lambda = \left(\frac{1-f}{n}\right), \quad \lambda' = \left(\frac{1-f'}{n'}\right), \quad \lambda^* = \frac{W_2(k-1)}{n} \quad \text{and} \quad f' = n'/N.$$

Now expressing T_{DS} in terms of $\varepsilon' s$ we have

$$T_{DS} = \bar{Y}(1+\varepsilon_0)(1+\phi\varepsilon_1)^{\alpha}(1+\phi\varepsilon_1')^{-\alpha}(1+\phi\varepsilon_2)^{\beta}(1+\phi\varepsilon_1')^{-\beta}, \tag{2}$$

where

$$\phi = \left(\frac{a\bar{X}}{aX+b}\right).$$

We assume that $|\phi\varepsilon_1| < 1$, $|\phi\varepsilon_1'| < 1$ and $|\phi\varepsilon_2| < 1$ so that the right hand side of (2) is expandable. Now, expanding the right hand side of (2) to the first degree of approximation, we have

$$\begin{aligned}(T_{DS} - \bar{Y}) = \bar{Y}\{&\varepsilon_0 + \beta\phi(\varepsilon_2 + \varepsilon_0\varepsilon_2 - \varepsilon_0\varepsilon_1') + \alpha\phi(\varepsilon_1 + \varepsilon_0\varepsilon_1 - \varepsilon_0\varepsilon_1') - (\alpha+\beta)\phi\varepsilon_1' \\ &+ \alpha\beta\phi^2\left(\varepsilon_1^{\prime 2} - \varepsilon_1'\varepsilon_2 - \varepsilon_1\varepsilon_1' + \varepsilon_1\varepsilon_2\right) - \phi^2\left(\beta^2\varepsilon_2\varepsilon_1' + \alpha^2\varepsilon_1\varepsilon_1'\right) \\ &+ \frac{\beta(\beta+1)}{2}\phi^2\left(\varepsilon_1^{\prime 2} + \varepsilon_2^2\right) + \frac{\alpha(\alpha+1)}{2}\phi^2\left(\varepsilon_1^{\prime 2} + \varepsilon_1^2\right)\Big\}.\end{aligned} \tag{3}$$

Taking expectations of both sides of (3), we get that the bias of T_{DS} to the first degree of approximation is given by

$$B(T_{DS}) = \bar{Y}\left[(\lambda-\lambda')\phi\left\{\alpha\left(K_{yx}+\frac{\alpha-1}{2}\phi\right)+\beta\left(K_{yx}+\alpha\phi+\frac{\beta-1}{2}\phi\right)\right\}C_x^2 + \lambda^*\alpha\phi\left(K_{yx(2)}+\frac{\alpha-1}{2}\phi\right)C_{x(2)}^2\right], \tag{4}$$

where

$$C_x^2=\frac{S_x^2}{X^2},\quad C_{x(2)}^2=\frac{S_{x(2)}^2}{X^2},\quad K_{yx}=\frac{\beta_{yx}}{R},\quad K_{yx(2)}=\frac{\beta_{yx(2)}}{R},\quad R=\frac{\bar{Y}}{X},$$

$$\beta_{yx}=\frac{S_{yx}}{S_x^2},\quad \beta_{yx(2)}=\frac{S_{yx(2)}}{S_{x(2)}^2},\quad S_{yx}=\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})/(N-1),$$

$$S_{yx(2)}=\sum_{i=1}^{N_2}(x_i-\bar{X}_2)(y_i-\bar{Y}_2)/(N_2-1).$$

Squaring both sides of (3) and neglecting terms of $\varepsilon's$ involving power greater than two, we have

$$\begin{aligned}(T_{DS}-\bar{Y})^2 &= \bar{Y}^2\{\varepsilon_0+\alpha\phi\varepsilon_1+\beta\phi\varepsilon_2-(\alpha+\beta)\phi\varepsilon_1'\}^2\\ &= \bar{Y}^2\left\{\varepsilon_0^2+\left(\alpha^2\varepsilon_1^2+\beta^2\varepsilon_2^2+2\alpha\beta\varepsilon_1\varepsilon_2\right)\phi^2+(\alpha+\beta)^2\phi^2\varepsilon_1'^2\right.\\ &+ \left.2\phi(\alpha\varepsilon_0\varepsilon_1+\beta\varepsilon_0\varepsilon_2)-2(\alpha+\beta)\phi\varepsilon_1'\varepsilon_0-2(\alpha+\beta)\phi^2(\alpha\varepsilon_1\varepsilon_1'+\beta\varepsilon_2\varepsilon_1')\right\}.\end{aligned} \tag{5}$$

Taking expectations of both sides of (5), we get the variance of T_{DS} to the first degree of approximation, we get

$$Var(T_{DS}) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2+(\alpha+\beta)\phi\left((\alpha+\beta)\phi+2K_{yx}\right)C_x^2\right\} + \lambda^*\left\{C_{y(2)}^2+\alpha\phi\left(\alpha\phi+2K_{yx(2)}\right)C_{x(2)}^2\right\}+\lambda'C_y^2\right], \tag{6}$$

The variance of T_{DS} is minimized for

$$\begin{aligned}\alpha &= -\frac{K_{yx(2)}}{\phi}\\ \beta &= \left(\frac{1}{\phi}\right)\left(K_{yx(2)}-K_{yx}\right)=-\left(\frac{1}{\phi}\right)\left(K_{yx}-K_{yx(2)}\right),\end{aligned} \tag{7}$$

Substituting (7) in (1), we get the asymptotically optimum estimator (AOE) as

$$T_{DS(opt)}=\bar{y}^*\left\{\frac{(a\bar{x}'+b)^2}{(a\bar{x}^*+b)(a\bar{x}+b)}\right\}^{K_{yx(2)}/\phi}\left(\frac{a\bar{x}'+b}{a\bar{x}+b}\right)^{K_{yx}/\phi}. \tag{8}$$

The minimum variance of T_{DS} is given by

$$\begin{aligned}\min Var\left(T_{DS}\right) &= \bar{Y}^2\left[\left(\lambda-\lambda'\right)\left(1-\rho_{yx}^2\right)C_y^2+\lambda^*\left(1-\rho_{yx(2)}^2\right)C_{y(2)}^2+\lambda' C_y^2\right] \\ &= Var\left(T_{DS(opt)}\right). \end{aligned} \tag{9}$$

2.1 Some members of the proposed class of estimators T_{DS}

The following are the estimators of the population mean which can be obtained by suitable choices of constants α, β, a and b.

Estimator	α	β	a	b
$T_{DS}^{(0)}=\bar{y}^*$ Usual unbiased estimator	0	0	a	b
$T_{DS}^{(1)}=\bar{y}^*\left(\frac{\bar{x}'}{\bar{x}^*}\right)$ Khare and Srivastava (1993), Tabasum and Khan's (2004) ratio estimator	-1	0	1	0
$T_{DS}^{(2)}=\bar{y}^*\left(\frac{\bar{x}'}{\bar{x}}\right)$ Khare and Srivastava (1993), Tabasum and Khan's (2006) ratio estimator	0	-1	1	0
$T_{DS}^{(3)}=\bar{y}^*\left(\frac{\bar{x}'}{\bar{x}^*}\right)\left(\frac{\bar{x}'}{\bar{x}}\right)$ Singh and Kumar's (2008a) ratio estimator	-1	-1	1	0
$T_{DS}^{(4)}=\bar{y}^*\left(\frac{\bar{x}^*}{\bar{x}'}\right)\left(\frac{\bar{x}}{\bar{x}'}\right)$ Singh and Kumar's (2008a) product estimator	1	1	1	0
$T_{DS}^{(5)}=\bar{y}^*\left(\frac{\bar{x}'+C_x}{\bar{x}^*+C_x}\right)$	-1	0	1	C_x
$T_{DS}^{(6)}=\bar{y}^*\left(\frac{\bar{x}'+C_x}{\bar{x}+C_x}\right)$	0	-1	1	C_x
$T_{DS}^{(7)}=\bar{y}^*\left(\frac{\bar{x}'+C_x}{\bar{x}^*+C_x}\right)\left(\frac{\bar{x}'+C_x}{\bar{x}+C_x}\right)$	-1	-1	1	C_x
$T_{DS}^{(8)}=\bar{y}^*\left(\frac{\bar{x}'+\rho}{\bar{x}^*+\rho}\right)$	-1	0	1	ρ
$T_{DS}^{(9)}=\bar{y}^*\left(\frac{\bar{x}'+\rho}{\bar{x}+\rho}\right)$	0	-1	1	ρ
$T_{DS}^{(10)}=\bar{y}^*\left(\frac{\bar{x}'+\rho}{\bar{x}^*+\rho}\right)\left(\frac{\bar{x}'+\rho}{\bar{x}+\rho}\right)$	-1	-1	1	ρ

where C_x is the coefficient of variation of the auxiliary variable x and ρ the correlation coefficient between the study variable y and the auxiliary variable x.

Many more estimators can also be generated from the proposed estimator in (1) just by putting different values of α, β, a and b. The expressions of bias and variance of the above estimators can be obtained by mere substituting the values of α, β, a and b in (4) and (6), respectively. Up to the first degree of approximation, the bias and variance expressions of various estimators are

$$B\left(T_{DS}^{(0)}\right) = \bar{Y}, \tag{10}$$

$$B\left(T_{DS}^{(1)}\right) = \bar{Y}\left\{(\lambda-\lambda')(1-K_{yx})C_x^2 + \lambda^*\left(1-K_{yx(2)}\right)C_{x(2)}^2\right\}, \tag{11}$$

$$B\left(T_{DS}^{(2)}\right) = \bar{Y}(\lambda-\lambda')(1-K_{yx})C_x^2, \tag{12}$$

$$B\left(T_{DS}^{(3)}\right) = \bar{Y}\left\{(\lambda-\lambda')(3-2K_{yx})C_x^2 + \lambda^*\left(1-K_{yx(2)}\right)C_{x(2)}^2\right\}, \tag{13}$$

$$B\left(T_{DS}^{(4)}\right) = \bar{Y}\left\{(\lambda-\lambda')(1+2K_{yx})C_x^2 + \lambda^*K_{yx(2)}C_{x(2)}^2\right\}, \tag{14}$$

$$B\left(T_{DS}^{(5)}\right) = \bar{Y}\left\{(\lambda-\lambda')\phi'(\phi'-K_{yx})C_x^2 + \lambda^*\phi'\left(\phi'-K_{yx(2)}\right)C_{x(2)}^2\right\}, \tag{15}$$

$$B\left(T_{DS}^{(6)}\right) = \bar{Y}(\lambda-\lambda')\phi'(\phi'-K_{yx})C_x^2, \tag{16}$$

$$B\left(T_{DS}^{(7)}\right) = \bar{Y}\left\{(\lambda-\lambda')\phi'(3\phi'-2K_{yx})C_x^2 + \lambda^*\phi'\left(\phi'-K_{yx(2)}\right)C_{x(2)}^2\right\}, \tag{17}$$

$$B\left(T_{DS}^{(8)}\right) = \bar{Y}\left\{(\lambda-\lambda')\phi^*(\phi^*-K_{yx})C_x^2 + \lambda^*\phi^*\left(\phi^*-K_{yx(2)}\right)C_{x(2)}^2\right\}, \tag{18}$$

$$B\left(T_{DS}^{(9)}\right) = \bar{Y}(\lambda-\lambda')\phi^*(\phi^*-K_{yx})C_x^2, \tag{19}$$

$$B\left(T_{DS}^{(10)}\right) = \bar{Y}\left\{(\lambda-\lambda')\phi^*(3\phi^*-2K_{yx})C_x^2 + \lambda^*\phi^*\left(\phi^*-K_{yx(2)}\right)C_{x(2)}^2\right\}, \tag{20}$$

$$Var\left(T_{DS}^{(0)}\right) = \bar{Y}^2\left\{\lambda C_y^2 + \lambda^* C_{y(2)}^2\right\}, \tag{21}$$

$$Var\left(T_{DS}^{(1)}\right) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2 + (1-2K_{yx})C_x^2\right\} + \lambda^*\left\{C_{y(2)}^2 + \left(1-2K_{yx(2)}\right)C_{x(2)}^2\right\} + \lambda' C_y^2\right], \tag{22}$$

$$Var\left(T_{DS}^{(2)}\right) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2 + (1-2K_{yx})C_x^2\right\} + \lambda^* C_{y(2)}^2 + \lambda' C_y^2\right], \tag{23}$$

$$Var\left(T_{DS}^{(3)}\right) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2 + 4(1-K_{yx})C_x^2\right\} + \lambda^*\left\{C_{y(2)}^2 + \left(1-2K_{yx(2)}\right)C_{x(2)}^2\right\} + \lambda' C_y^2\right], \tag{24}$$

$$Var\left(T_{DS}^{(4)}\right) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2 + 4(1+K_{yx})C_x^2\right\} + \lambda^*\left\{C_{y(2)}^2 + \left(1+2K_{yx(2)}\right)C_{x(2)}^2\right\} + \lambda' C_y^2\right], \tag{25}$$

$$Var\left(T_{DS}^{(5)}\right) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2 + \phi'(\phi'-2K_{yx})C_x^2\right\} + \lambda^*\left\{C_{y(2)}^2 + \phi'\left(\phi'-2K_{yx(2)}\right)C_{x(2)}^2\right\} + \lambda' C_y^2\right], \tag{26}$$

$$Var\left(T_{Ds}^{(6)}\right) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2 + \phi'(\phi'-2K_{yx})C_x^2\right\} + \lambda^* C_{y(2)}^2 + \lambda' C_y^2\right], \tag{27}$$

$$Var\left(T_{DS}^{(7)}\right) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2 + 4\phi'(\phi'-K_{yx})C_x^2\right\} + \lambda^*\left\{C_{y(2)}^2 + \phi'\left(\phi'-2K_{yx(2)}\right)C_{x(2)}^2\right\} + \lambda' C_y^2\right], \tag{28}$$

$$Var\left(T_{DS}^{(8)}\right) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2 + \phi^*(\phi^*-2K_{yx})C_x^2\right\} + \lambda^*\left\{C_{y(2)}^2 + \phi^*\left(\phi^*-2K_{yx(2)}\right)C_{x(2)}^2\right\} + \lambda' C_y^2\right], \tag{29}$$

$$Var\left(T_{DS}^{(9)}\right) = \bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2 + \phi^*(\phi^*-2K_{yx})C_x^2\right\} + \lambda^* C_{y(2)}^2 + \lambda' C_y^2\right], \tag{30}$$

$$Var\left(T_{DS}^{(10)}\right)=\bar{Y}^2\left[(\lambda-\lambda')\left\{C_y^2+4\phi^*\left(\phi^*-K_{yx}\right)C_x^2\right\}+\lambda^*\left\{C_{y(2)}^2+\phi^*\left(\phi^*-2K_{yx(2)}\right)C_{x(2)}^2\right\}+\right.$$
$$\left.+\lambda' C_y^2\right], \tag{31}$$

where $\phi'=\left(\frac{\bar{X}}{\bar{X}+C_x}\right)$ and $\phi^*=\left(\frac{\bar{X}}{\bar{X}+\rho}\right)$.

2.2 Efficiency comparison

The proposed class of estimators T_{DS} is more efficient than

(i) usual unbiased estimator $T_{DS}^{(0)}=\bar{y}^*$if

$$\left.\begin{array}{c} 0<\alpha<\min.\left\{-\left(\frac{2K_{yx}}{\phi}+\beta\right);-\left(\frac{2K_{yx(2)}}{\phi}\right)\right\} \\ or\ \max.\left\{-\left(\frac{2K_{yx}}{\phi}+\beta\right);-\left(\frac{2K_{yx(2)}}{\phi}\right)\right\}<\alpha<0 \end{array}\right\}, \tag{32}$$

(ii) usual ratio estimator $T_{DS}^{(1)}$ if

$$\left.\begin{array}{c} 0<\alpha<\min.\left\{-\left(\frac{2K_{yx}-1}{\phi}+\beta\right);-\left(\frac{2K_{yx(2)}-1}{\phi}\right)\right\} \\ or\ \max.\left\{-\left(\frac{2K_{yx}-1}{\phi}+\beta\right);-\left(\frac{2K_{yx(2)}-1}{\phi}\right)\right\}<\alpha<0 \end{array}\right\}, \tag{33}$$

(iii) the ratio estimator $T_{DS}^{(2)}$ if

$$\left.\begin{array}{c} 0<\alpha<\min.\left\{-\left(\frac{2\left(K_{yx}-1\right)}{\phi}+\beta\right);-\left(\frac{2K_{yx(2)}}{\phi}\right)\right\} \\ or\ \ \max.\left\{-\left(\frac{2\left(K_{yx}-1\right)}{\phi}+\beta\right);-\left(\frac{2K_{yx(2)}}{\phi}\right)\right\}<\alpha<0 \end{array}\right\}, \tag{34}$$

(iv) the ratio estimator $T_{DS}^{(3)}$ if

$$\left.\begin{array}{c} 0<\alpha<\min.\left[-\left\{\frac{2\left(K_{yx}-1\right)}{\phi}+\beta\right\};-\left(\frac{2K_{yx(2)}-1}{\phi}\right)\right] \\ or\ \ \max.\left[-\left\{\frac{2\left(K_{yx}-1\right)}{\phi}+\beta\right\};-\left(\frac{2K_{yx(2)}-1}{\phi}\right)\right]<\alpha<0 \end{array}\right\}, \tag{35}$$

(v) the product estimator $T_{DS}^{(4)}$ if

$$\left.\begin{array}{c} 0<\alpha<\min.\left[-\left\{\dfrac{2\left(K_{yx}+1\right)}{\phi}+\beta\right\};-\left(\dfrac{2K_{yx(2)}+1}{\phi}\right)\right] \\ \text{or}\quad \max.\left[-\left\{\dfrac{2\left(K_{yx}+1\right)}{\phi}+\beta\right\};-\left(\dfrac{2K_{yx(2)}+1}{\phi}\right)\right]<\alpha<0 \end{array}\right\}, \tag{36}$$

(vi) the estimator $T_{DS}^{(5)}$ if

$$\left.\begin{array}{c} 0<\alpha<\min.\left[-\left\{\left(\dfrac{2K_{yx}-\phi'}{\phi}\right)+\beta\right\};-\left(\dfrac{2K_{yx(2)}-\phi'}{\phi}\right)\right] \\ \text{or}\quad \max.\left[-\left\{\left(\dfrac{2K_{yx}-\phi'}{\phi}\right)+\beta\right\};-\left(\dfrac{2K_{yx(2)}-\phi'}{\phi}\right)\right]<\alpha<0 \end{array}\right\}, \tag{37}$$

(vii) the estimator $T_{DS}^{(6)}$ if

$$\left.\begin{array}{c} 0<\alpha<\min.\left\{-\left(\dfrac{2K_{yx}-\phi'}{\phi}+\beta\right);-\left(\dfrac{2K_{yx(2)}}{\phi}\right)\right\} \\ \text{or}\quad \max.\left\{-\left(\dfrac{2K_{yx}-\phi'}{\phi}+\beta\right);-\left(\dfrac{2K_{yx(2)}}{\phi}\right)\right\}<\alpha<0 \end{array}\right\}, \tag{38}$$

(viii) the estimator $T_{DS}^{(7)}$ if

$$\left.\begin{array}{c} 0<\alpha<\min.\left[-\left\{\dfrac{2\left(K_{yx}-\phi'\right)}{\phi}+\beta\right\};\left(\dfrac{2K_{yx(2)}-\phi'}{\phi}\right)\right] \\ \text{or}\quad \max.\left[-\left\{\dfrac{2\left(K_{yx}-\phi'\right)}{\phi}+\beta\right\};\left(\dfrac{2K_{yx(2)}-\phi'}{\phi}\right)\right]<\alpha<0 \end{array}\right\}, \tag{39}$$

(ix) the estimator $T_{DS}^{(8)}$ if

$$\left.\begin{array}{c} 0<\alpha<\min.\left\{-\left(\dfrac{2K_{yx}-\phi^*}{\phi}+\beta\right);-\left(\dfrac{2K_{yx(2)}-\phi^*}{\phi}\right)\right\} \\ \text{or}\quad \max.\left\{-\left(\dfrac{2K_{yx}-\phi^*}{\phi}+\beta\right);-\left(\dfrac{2K_{yx(2)}-\phi^*}{\phi}\right)\right\}<\alpha<0 \end{array}\right\}, \tag{40}$$

(x) the estimator $T_{DS}^{(9)}$ if

$$\left.\begin{aligned} 0<\alpha<\min.\left\{-\left(\frac{2K_{yx}-\phi^{*}}{\phi}+\beta\right);-\left(\frac{2K_{yx(2)}}{\phi}\right)\right\} \\ or\ \ \min.\left\{-\left(\frac{2K_{yx}-\phi^{*}}{\phi}+\beta\right);-\left(\frac{2K_{yx(2)}}{\phi}\right)\right\}<\alpha<0 \end{aligned}\right\}, \tag{41}$$

(xi) the estimator $T_{DS}^{(10)}$ if

$$\left.\begin{aligned} 0<\alpha<\min.\left[-\left\{\frac{2\left(K_{yx}-\phi^{*}\right)}{\phi}+\beta\right\};-\left(\frac{2K_{yx(2)}-\phi^{*}}{\phi}\right)\right] \\ or\ \ \max.\left[-\left\{\frac{2\left(K_{yx}-\phi^{*}\right)}{\phi}+\beta\right\};-\left(\frac{2K_{yx(2)}-\phi^{*}}{\phi}\right)\right]<\alpha<0 \end{aligned}\right\}. \tag{42}$$

3 Empirical study

To illustrate the properties of the proposed estimators of the population mean $\bar{Y}$, we consider a real data set considered before by Srivastava (1993). The description of the sample is given below.

The sample of 100 consecutive trips (after omitting 20 outlier values) measured by two fuel meters for a small family car in normal usage given by Lewis et al (1991) has been taken into consideration. The measurement of turbine meter (in ml) is considered as main character *y* and the measurement of displacement meter (in cm^3) is considered as auxiliary character *x*. We treat the last 25% values as non-response values. The values of the parameters are as follows:

$$\bar{Y}=3500.12,\quad \bar{X}=260.84,\quad S_y=2079.30,\quad S_x=156.40,\quad \bar{Y}_2=3401.08,$$
$$\bar{X}_2=259.96,\quad S_{y(2)}=1726.02,\quad S_{x(2)}=134.36,\quad \rho=0.985,\quad \rho_2=0.995.$$

Here, we have computed (i) the absolute relative bias of different suggested estimators of $\bar{Y}$ using the formula:

$$ARB\left(.\right)=\left|\frac{Bias\left(.\right)}{\bar{Y}}\right|;$$

(ii) the percent relative efficiencies (PRE) of different suggested estimators with respect to the usual unbiased estimator $\bar{y}^{*}$, for different values of *k*.

Table 1: *Absolute relative bias (ARB) of different proposed estimators.*

Estimators	$N=100,\quad n'=50,\quad n=30$			
	$(1/k)$			
	$(1/5)$	$(1/4)$	$(1/3)$	$(1/2)$
$\bar{y}^*=T_{DS}^{(0)}$	0.0000	0.0000	0.0000	0.0000
$T_{DS}^{(1)}$	0.00053	0.00043	0.00032	0.00022
$T_{DS}^{(2)}$	0.00012	0.00012	0.00012	0.00012
$T_{DS}^{(3)}$	0.00543	0.00533	0.00522	0.00512
$T_{DS}^{(4)}$	0.02251	0.02041	0.01831	0.01621
$T_{DS}^{(5)}$	0.00050	0.00040	0.00030	0.00020
$T_{DS}^{(6)}$	0.00010	0.00010	0.00010	0.00010
$T_{DS}^{(7)}$	0.00536	0.00527	0.00517	0.00507
$T_{DS}^{(8)}$	0.00048	0.00038	0.00029	0.00019
$T_{DS}^{(9)}$	0.00010	0.00010	0.00010	0.00010
$T_{DS}^{(10)}$	0.00532	0.00523	0.00513	0.00503

Table 2: *Percent relative efficiency (PRE) of the estimators with respect to* $\bar{y}^*$.

Estimators	$N=100,\quad n'=50,\quad n=30$			
	$(1/k)$			
	$(1/5)$	$(1/4)$	$(1/3)$	$(1/2)$
$\bar{y}^*=T_{DS}^{(0)}$	100.00	100.00	100.00	100.00
$T_{DS}^{(1)}$	433.12	381.95	330.09	277.54
$T_{DS}^{(2)}$	138.74	146.79	159.06	180.07
$T_{DS}^{(3)}$	185.74	163.18	140.48	117.65
$T_{DS}^{(4)}$	35.15	30.83	26.49	22.17
$T_{DS}^{(5)}$	309.02	284.54	257.42	227.21
$T_{DS}^{(6)}$	132.15	138.46	147.83	163.25
$T_{DS}^{(7)}$	322.99	298.09	270.38	239.33
$T_{DS}^{(8)}$	433.56	382.27	330.30	277.65
$T_{DS}^{(9)}$	138.75	146.80	159.08	180.09
$T_{DS}^{(10)}$	187.38	164.60	141.70	118.66
$T_{DS(opt)}$	435.74	383.77	331.23	278.12

It is observed from Table 1 that the absolute relative bias (ARB) of the estimators $T_{DS}^{(j)};\ j=1,3,4,5,7,8,10$ decreases with the increase of $(1/k)$ while it remains

constant for $T_{DS}^{(2)}$, $T_{DS}^{(6)}$ and $T_{DS}^{(9)}$. The ARB of the estimator $T_{DS}^{(4)}$ is larger than all other estimators. It may be due to positive correlation. The estimator $T_{DS}^{(4)}$ is a product type estimator which is appropriate in the situations where the correlation between y and x is negative. It is further observed from Table 1 that

$$ARB\left(T_{DS}^{(0)}\right) < ARB\left(T_{DS}^{(9)}\right) = ARB\left(T_{DS}^{(6)}\right) < ARB\left(T_{DS}^{(2)}\right) < ARB\left(T_{DS}^{(8)}\right) < ARB$$
$$\left(T_{DS}^{(5)}\right) < ARB\left(T_{DS}^{(1)}\right) < ARB\left(T_{DS}^{(10)}\right) < ARB\left(T_{DS}^{(7)}\right) < ARB\left(T_{DS}^{(3)}\right) < ARB\left(T_{DS}^{(4)}\right)$$

which clearly indicates that the estimator $T_{DS}^{(8)}$ (based on the knowledge of correlation coefficient) has least magnitude of relative bias followed by $T_{DS}^{(1)}$.

From Table 2, we see that the proposed general class of estimators is more desirable over all the considered estimators under optimum condition. It is observed from Table 2 that the percent relative efficiency of the estimators $T_{DS}^{(1)}$, $T_{DS}^{(3)}$, $T_{DS}^{(4)}$, $T_{DS}^{(5)}$, $T_{DS}^{(7)}$, $T_{DS}^{(8)}$ and $T_{DS(opt)}$ decreases as $(1/k)$ increases, but for the estimators $T_{DS}^{(2)}$, $T_{DS}^{(6)}$ and $T_{DS}^{(9)}$, it increases with the increase in the value of $(1/k)$.

From Tables 1 and 2, it is further observed that the estimator $T_{DS}^{(8)}$ (based on known correlation coefficient) seems to be more appropriate estimator in comparison to others as it has appreciable efficiency (close to the efficiency of the optimum estimator $T_{DS(opt)}$) as well as negligible magnitude of relative bias. However, the estimators $T_{DS}^{(5)}$ and $T_{DS}^{(7)}$ (based on known coefficient of variation) are also appropriate choices among the estimators as they have considerable gain in efficiency as well as lower relative bias.

Finally, we conclude that the proposed estimator $T_{DS}^{(5)}$, $T_{DS}^{(7)}$ and $T_{DS}^{(8)}$ are better alternatives of the optimum estimator $T_{DS(opt)}$.

Acknowledgement

Authors are thankful to the referee for his valuable constructive suggestions regarding improvement of the paper.

References

Cochran, W. G. (1977). *Sampling Techniques*, 3rd *ed.*, John Wiley and Sons, New York.

Hansen, M. H. and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *J. Amer. Statist. Assoc.*, 41, 517-529.

Khare, B. B. and Srivastava, S. (1993). Estimation of population mean using auxiliary character in presence of non-response. *Nat. Acad. Sc. Letters, India*, 16, 111-114.

Khare, B. B. and Srivastava, S. (1995). Study of conventional and alternative two-phase sampling ratio, product and regression estimators in presence of non-response. *Proc. Nat. Acad. Sci. India,* 65(A), 195-203.

Khare, B. B. and Srivastava, S. (1997). Transformed ratio type estimators for the population mean in the presence of non-response. *Comm. Statist.-Theory Methods*, 26, 1779-1791.

Lewis, P. A., Jones, P. W., Polak, J. W. and Tillotson, H. T. (1991). The problem of conversion in method comparison studies. *Applied Statistics*, 40, 105-112.

Okafor, F. C. and Lee, H. (2000). Double sampling for ratio and regression estimation with sub sampling the non-respondent. *Survey Methodology*, 26, 183-188.

Rao, P. S. R. S. (1986). Ratio Estimation with sub sampling the non-respondents. *Survey Methodology*, 12, 217-230.

Rao, P. S. R. S. (1987). Ratio and regression estimates with sub-sampling the non-respondents. *Paper presented at a special contributed session of the International Statistical Association Meetings*, September, 2-16, Tokyo, Japan.

Singh, H. P. and Kumar, S. (2008a). Estimation of mean in presence of non-response using two phase sampling scheme. *Statistical Papers*, DOI10.1007/s00362-008-0140-5.

Singh, H. P. and Kumar, S. (2008b). A regression approach to the estimation of finite population mean in presence of non-response. *Aust. N. Z. J. Stat.*, 50, 395-408.

Sodipo, A. A. and Obisesan, K. O. (2007). Estimation of the population mean using difference cum ratio estimator with full response on the auxiliary character. *Res. J. Applied Sci*., 2, 769-772.

Srivastava, S. (1993). Some problems on the estimation of population mean using auxiliary character in presence of non-response in sample surveys. *Thesis submitted to Banaras Hindu University, Varanasi, India*.

Tabasum, R. and Khan, I. A. (2004). Double sampling for ratio estimation with non-response. *J. Ind. Soc. Agril. Statist.*, 58, 300-306.

Tabasum, R. and Khan, I. A. (2006). Double sampling ratio estimator for the population mean in presence of non-response. *Assam Statist. Review*, 20, 73-83.

SORT 33 (1) January-June 2009, 85-104

On the performance of small-area estimators: fixed vs. random area parameters*

Alex Costa[1], Albert Satorra and Eva Ventura[2]

Abstract

Most methods for small-area estimation are based on composite estimators derived from design- or model-based methods. A composite estimator is a linear combination of a direct and an indirect estimator with weights that usually depend on unknown parameters which need to be estimated. Although model-based small-area estimators are usually based on random-effects models, the assumption of fixed effects is at face value more appropriate. Model-based estimators are justified by the assumption of random area effects; in practice, however, areas can not be substituted for one another in a random manner (we say, they are not interchangeable). In the present paper we empirically assess the quality of several small-area estimators in the setting in which the area effects are treated as fixed. We consider two settings: one that draws samples from a theoretical population, and another that draws samples from an empirical population of a labour force register maintained by the National Institute of Social Security (NISS) of Catalonia. We distinguish two types of composite estimators: ***a)*** those that use weights that involve area specific estimates of bias and variance; and, ***b)*** those that use weights that involve a common variance and a common squared bias estimate for all the areas. We assess their precision and discuss alternatives to optimizing composite estimation in applications.

MSC: 62G10,62J02

Keywords: Small area estimation, composite estimator, Monte Carlo study, random effect model, BLUP, empirical BLUP

1 Introduction

Sample surveys are often used to estimate quantities related not only to the total population but also to a variety of small-area domains. Small-area estimation is concer-

* This research was supported by IDESCAT (Statistical Institute of Catalonia). The views expressed in this paper are those of the authors and do not necessarily represent the policies of IDESCAT. Detailed and very helpful comments by Nicholas T. Longford on a previous version of this paper are acknowledged. Constructive and detail reviewers' comments are also appreciated. Partial support by the Spanish Ministry of Science and Technology grant SEJ2006-13537 to A. Satorra and E. Ventura is also acknowledged.

[1] Idescat (Statistical Institute of Catalonia).

[2] Department of Economics and Business, Universitat Pompeu Fabra and Barcelona GSE.

Received: February 2008

Accepted: December 2008

ned with estimating population quantities associated with a partition of the domain (population) into subdomains (small areas or districts) $j = 1, \ldots, J$. Nowadays there is a large body of methodology for small-area estimation; see, e.g., Platek, Rao, Särndal and Singh (1987), Isaki (1990), Ghosh and Rao (1994), Singh, Gambino and Mantel (1994), and Rao (2003).

Large-scale (national) surveys are usually designed to yield estimates of a small number of key national population quantities (means, proportions and the like) that have sufficient precision, without having to adopt any assumptions other than the sampling design. Insisting on a large sample for each district is not realistic, especially when there are many districts, and several of them form a very small fraction of the population.

When estimating a domain quantity, we refer to a *direct estimator* if it is based only on the domain-specific sample. A domain (area) is regarded as small if the direct estimate for the area does not have adequate precision. For a small area one could use *indirect estimators* that borrow strength from values of the variable of interest from related areas and/or time periods. An implicit or explicit model is used to link the different areas and/or time periods, often through the use of auxiliary information such as a census count or some administrative records. An initial classification of small-area estimation divides the methods into *design-based* and *model-based*.

Design-based methods are based solely on the sampling design and do not make use of distributional (model) assumptions about the observed variables. Sampling variation, that is, variation across hypothetical replications of drawing a sample, arises only due to the variation of the specific units that are selected into the sample, and not due to variation of the population characteristics of interest (such as the small-area means) which are considered fixed because they are constant across replications. In contrast, model-based methods assume stochastic models governing the population values that are the target of the estimation process. Models are used to mediate the process of borrowing strength across the districts (small areas). That is, inference about a district that is represented in the sample by very few observations is supported by the information in the other districts' subsamples. This is most effective when the districts are very similar. Similarity can be enhanced by adjustment for other variables, opening up the potential of regression models.

Borrowing strength, as defined originally by Efron and Morris (1973), is based on the assumption of random effects. In the simplest setting with no covariates, the deviations of the district-level means θ_j from their national mean θ are assumed to be a random sample from a centred distribution with a finite variance, such as $\mathcal{N}(0, \sigma_u^2)$.

Model-based methods for small-area estimation associate the districts with random effects. In applications, however, the districts have their names (labels), and the target quantities θ_j could in principle be established by enumeration. In an hypothetical replication of the survey, the same districts, with the same subpopulations and the same values of θ_j would be involved. Therefore, it is natural to associate the districts with fixed effects. Longford (2007) argues that the assumption of fixed or random effect has a profound effect on standard errors of model-based small-area estimators. In the present paper we consider both design- and model-based estimators, and assess their accuracy

in the case of the fixed-effect assumption. Accuracy refers not to average MSE across areas, but to MSE for the particular (fixed) areas. This departs from previous studies in which accuracy was assessed by averaging MSE across areas (see, e.g., Costa, Satorra and Ventura (2003), and Santamaría, Morales and Molina (2004)).

In the model-based approach, the best linear unbiased predictor (BLUP) of the parameter of interest (the small-area parameter), is a linear combination of a direct and a synthetic estimator with weights that depend on two parameters that are usually unknown: the within- and between-area variances (possibly after controlling for other variables, regressors). Since both parameters are unknown quantities, these two variances have to be estimated, giving rise to the empirical BLUP (EBLUP). This estimation can distort the optimality of the EBLUP. In sections 3 and 4 we assess the consequences on accuracy of the substitution of model parameters by estimated values.

The purpose of the paper is to compare the performance of design- vs. model-based small-area estimators, with a focus on a specific (fixed) set of small areas. Monte Carlo methods are used for this investigation.

Two population frames will be considered in the Monte Carlo study: ***a)*** a theoretical population with varying distribution and sample size; ***b)*** an empirical population of labour statistics from the affiliation of firms in the NISS (National Institute of Social Security) registers. The choice of the NISS is motivated by current work at IDESCAT (Statistics Bureau of Catalonia).

The plan of the paper is as follows. Section 2 develops the notation and general context of small-area estimation, focusing on the distinction between design-based and model-based methods. Sections 3 and 4 describe the Monte Carlo studies using the theoretical and the empirical population, respectively. Section 5 concludes with a discussion of the results and the avenues for further research.

2 Approaches for small-area estimation

We consider a population stratified into J (small-area) domains (strata), $j = 1, 2, \ldots, J$, and we seek to estimate the stratum parameters θ_j as well as an overall population parameter θ. A direct estimator of θ_j uses sample data only from area j. An indirect or synthetic estimator of θ_j uses data also from outside area j. We suppose that there is a direct estimator $\hat{\theta}_{dj}$ of θ_j and that it is unbiased (but may have large variance), and a synthetic estimator $\hat{\theta}_{sj}$ that has small variance but may be biased for θ_j.

Two perspectives motivate the different small-area estimators. The first assumes that the θ_j are fixed values and that there is sampling variation only within each stratum. In the second, in addition to the random variation within strata, there is also random variation of the θ_j, that are supposed to be realizations from a specific sampling distribution. We now describe these two approaches, design-based (fixed θ_j) and model-based (random θ_j), respectively.

2.1 Fixed-area perspective

Following Rao (2003, Section 4.3), a natural way to balance the potential bias of a synthetic estimator $\hat{\theta}_{sj}$ of θ_j against the instability of a direct estimator $\hat{\theta}_{dj}$ of the same parameter is to take the composite estimator (weighted average)

$$\hat{\theta}_{cj}(\pi_j) = (1-\pi_j)\,\hat{\theta}_{dj} + \pi_j\,\hat{\theta}_{sj}, \tag{1}$$

a function of the weight $0 \leq \pi_j \leq 1$. This estimator has a mean square error (MSE) given by (Rao, 2003, formula (4.3.2)):[1]

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}_{cj},\theta_j) \;=\; & (1-\pi_j)^2\mathrm{MSE}(\hat{\theta}_{dj},\theta_j) + \pi_j^2\,\mathrm{MSE}(\hat{\theta}_{sj},\theta_j) \\ & +2\pi_j(1-\pi_j)\mathrm{E}\left\{(\hat{\theta}_{dj}-\theta_j)(\hat{\theta}_{sj}-\theta_j)\right\} \end{aligned} \tag{2}$$

where $\mathrm{MSE}(\hat{\delta},\delta)$ denotes the MSE of an estimator $\hat{\delta}$ with respect to the target δ. The expectation in the last term of (2) is taken with respect to the design-based sampling variation. In most applications, $\hat{\theta}_{dj}$ and $\hat{\theta}_{sj}$ are uncorrelated, so this last term vanishes. This is assumed throughout. Denote $\tilde{\theta}_{cj} = \hat{\theta}_{cj}(\tilde{\pi}_j)$.

The weight that minimizes the MSE of $\tilde{\theta}_{cj}$ is approximately (see Rao (2003, formula (4.3.3))

$$\tilde{\pi}_j \;= \frac{\mathrm{MSE}(\hat{\theta}_{dj},\theta_j)}{\mathrm{MSE}(\hat{\theta}_{dj},\theta_j)+\mathrm{MSE}(\hat{\theta}_{sj},\theta_j)} \tag{3}$$

in which case the (minimum) MSE is

$$\mathrm{MSE}(\tilde{\theta}_{cj},\theta_j) = \tilde{\pi}_j^2\,\mathrm{MSE}(\hat{\theta}_{sj},\theta_j); \tag{4}$$

and

$$\mathrm{MSE}(\tilde{\theta}_{cj},\theta_j) = (1-\tilde{\pi}_j)^2\mathrm{MSE}(\hat{\theta}_{dj},\theta_j); \tag{5}$$

so, the (optimal) composite estimator $\tilde{\theta}_{cj}$ is superior to both the synthetic estimator $\hat{\theta}_{sj}$, since $\tilde{\pi}_j < 1$, and the direct estimator $\hat{\theta}_{dj}$, since $\tilde{\pi}_j > 0$ and $\tilde{\theta}_{cj}(0) = \hat{\theta}_{dj}$. If there was covariation among the synthetic and the direct estimator, $\mathrm{cov}(\hat{\theta}_{dj},\hat{\theta}_{sj})$ would be subtracted once in the numerator and twice in the denominator.

The expression (2) (with the covariance term ignored) will be used in sections 3 and 4 to compute the exact MSE of various composite estimators arising in a Monte Carlo study. The exact values of the MSE can be computed since in Monte Carlo studies we

[1] As in Rao (2003, Section 4.3), and throughout this section, "Var", "MSE" and "E" should carry a subscript p that refers to variation with respect to the sampling design; this subscript has been suppressed for notational convenience.

know the population values of the parameters. In applications, the MSE will have to be estimated and several estimates are available. Longford (2007) discusses issues arising in the estimation of the MSE in the case of fixed area effects.

When the direct and synthetic estimators are unbiased for θ_j and θ respectively, and the variance of $\hat{\theta}_{sj}$ is small relative to the variance of the direct estimator, we have

$$\tilde{\pi}_j = \frac{\text{var}(\hat{\theta}_{dj})}{\text{var}(\hat{\theta}_{dj}) + (\theta - \theta_j)^2} \tag{6}$$

If $\hat{\theta}_{dj}$ is the sample mean, then $\text{var}(\hat{\theta}_{dj}) = \sigma^2_{j\epsilon}/n_j$, where $\sigma^2_{j\epsilon}$ is a within-domain variance and n_j is the sample size of the jth domain.[2] Then (6) becomes

$$\tilde{\pi}_j = \frac{\sigma^2_{j\epsilon}/n_j}{\sigma^2_{j\epsilon}/n_j + (\theta - \theta_j)^2} \tag{7}$$

For a synthetic estimator (unbiased for θ) whose variance is small compared with the variance of the direct estimator (unbiased for θ_j), we have

$$\text{E}(\hat{\theta}_{sj} - \hat{\theta}_{dj})^2 \approx (\theta - \theta_j)^2 + \text{var}(\hat{\theta}_{dj}) \tag{8}$$

and $\tilde{\pi}_j \approx \text{var}(\hat{\theta}_{dj})/(\hat{\theta}_{sj} - \hat{\theta}_{dj})^2$, suggesting the weight

$$\hat{\pi}^{\dagger}_j = \frac{\widehat{\text{var}}(\hat{\theta}_{dj})}{(\hat{\theta}_{sj} - \hat{\theta}_{dj})^2},$$

where $\widehat{\text{var}}(\hat{\theta}_{dj})$ is an unbiased estimator of $\text{var}(\hat{\theta}_{dj})$. This estimator is very unstable and it could even fall outside the interval $[0,1]$. In the Monte Carlo study in sections 3 and 4 we use instead the weight

$$\hat{\pi}^{*}_j = \frac{\widehat{\text{var}}(\hat{\theta}_{dj})}{(\hat{\theta}_{sj} - \hat{\theta}_{dj})^2 + \widehat{\text{var}}(\hat{\theta}_{dj})}, \tag{9}$$

which satisfies the condition $0 \leq \hat{\pi}^{*}_j \leq 1$.

The composite small-area estimators can be based on the assumption of homogeneity of the within-area variances, in which case they will use a common estimate of this variance, such as the estimator of (14) defined in the section below, or they may contemplate heteroscedasticity, in which case they may use an area specific estimate such as (16) (see section below).

[2] We could contemplate homoscedasticity and replace $\sigma^2_{j\epsilon}$ by σ^2_{ϵ}.

The optimal composite estimator that uses the weight in (7) is not feasible in practice because the bias term $(\theta_j - \theta)^2$ and the variance $\sigma^2_{j\epsilon}$ are unknown quantities that need to be estimated. We shall see that several alternatives to the estimation of the within-area variance do not induce much difference among estimators; in contrast, alternatives to the estimation of the squared area-bias term will lead to fundamental differences among estimators.

2.2 Random-area perspective

Alternative small-area estimators are based on models. Suppose

$$y_{ji} = X_{ji}\beta + Z_{ji}\gamma_j + \epsilon_{ji} \tag{10}$$

where $i = 1,2,\ldots,n_j$ and $j = 1,2,\ldots J$, i and j denoting primary and secondary level units, observations and areas, respectively. X_{ji} and Z_{ji} are vectors of attributes of observation i of area j, β is a vector of regression coefficients and γ_j is a vector of random area effects, independent of ϵ_{ji}, and usually both normally distributed with respective variances σ^2_u and σ^2_ϵ (variance matrix Σ_u, instead of σ^2_u, when Z_{ji} is a vector).

Since the issues we want to investigate arise already on the simplest of the random-area models, the one that has no covariates, our research will be made more transparent by using the simplest version of the model in (10), with Z_{ji} set to the indicator of area j and $X_{ji} = 1$ is empty (there are no covariates); that is,

$$y_{ji} = \mu + u_j + \epsilon_{ji} \tag{11}$$

Variables u_j and ϵ_{ij} are centred random variables with respective variances σ^2_u (variance "between") and σ^2_ϵ (variance "within").

Let $y_{j\cdot} = n_j^{-1} \sum_i y_{ij}$ and $y_{\cdot\cdot} = n^{-1} \sum_i \sum_j y_{ij}$ be the respective direct and synthetic estimators of $\theta_j = \mu + u_j$, where $n = \sum_j n_j$ is the overall sample size. Since $\operatorname{var}(y_{j\cdot}) = \sigma^2_u + (\sigma^2_\epsilon / n_j)$ and $\operatorname{cov}(y_{j\cdot}, u_j) = \sigma^2_u$, the best unbiased linear predictor (BLUP[3]) of θ_j given $y_{j\cdot}$ is (see, e.g., Neudecker and Satorra, 2003)

$$\operatorname{BLUP}(\theta_j \mid y_{j\cdot}) = \mu + \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_\epsilon / n_j}(y_{j\cdot} - \mu) = (1 - \omega_j) y_{j\cdot} + \omega_j \mu \tag{12}$$

where

$$\omega_j = 1 - \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_\epsilon / n_j} = \frac{\sigma^2_\epsilon / n_j}{\sigma^2_u + \sigma^2_\epsilon / n_j} = \frac{1}{1 + n_j \gamma}$$

[3] A common notation is also BLP, but since BLP is unbiased in the predictive sense, i.e. $\mathrm{E}\{\operatorname{BLP}(\theta_j) - \theta_j\} = 0$, the terminology of 'best linear unbiased predictor' (BLUP) will be used.

and $\gamma = \sigma_u^2/\sigma_\epsilon^2$. We used $E(y_{j.}) = \mu$. The empirical BLUP (EBLUP) is

$$\hat{\theta}_{cj}(\hat{\omega}) = \text{EBLUP}(\theta_j \mid y_{j.}) = (1-\hat{\omega}_j)y_{j.} + \hat{\omega}_j y_{..}$$

where $y_{..}$ (the overall mean) is used as an estimator of μ and

$$\hat{\omega}_j = \frac{\hat{\sigma}_\epsilon^2/n_j}{\hat{\sigma}_u^2 + \hat{\sigma}_\epsilon^2/n_j} = \frac{1}{1+n_j\hat{\gamma}} \tag{13}$$

as the estimator of ω_j, with $\hat{\gamma} = \hat{\sigma}_u^2/\hat{\sigma}_\epsilon^2$. Here

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n-J} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - y_{.j})^2 \tag{14}$$

and

$$\hat{\sigma}_u^2 = \frac{1}{J-1} \sum_{j=1}^{J} (y_{.j} - y_{..})^2 \tag{15}$$

are moment-matching estimators of the variances. The estimator $\hat{\sigma}_\epsilon^2$ could be alternatively written as a weighted mean, i.e.

$$\hat{\sigma}_\epsilon^2 = \sum_{j=1}^{J} ((n_j - 1)/(N-J))\, \hat{\sigma}_{\epsilon j}^2,$$

of the within-area variance estimates

$$\hat{\sigma}_{j\epsilon}^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - y_{.j})^2. \tag{16}$$

As an alternative to these estimators we could use maximum likelihood (ML) estimation of the mixed regression model. For the unbalanced case, this provides alternative EBLUP estimators. These will be evaluated in the Monte Carlo study of sections 3 and 4.

One could also question the quality of these EBLUP estimators when the model (11) deviates from the standard assumptions, such as normality of the within- and between-area distributions, or both, or when there is variation among the within-area variances while equality is assumed.

Note that all composite estimators we consider are “borrowing strength” estimators, with the distinction that the ones based on the “random effect” perspective use “average” type of estimates for the bias, while the ones based on the “fixed effect” assumption use area specific estimate of the bias. In the Monte Carlo study below, we will see that there is a sharp difference on performance for these two type of estimators.

The Monte Carlo study of Section 3 contemplates normal and highly skewed distributions, both for the first- and second-level distributions. Non-normality of the distribution within each area, and heteroscedasticity of the within-area variances, is present in the Monte Carlo study of Section 4 involving an empirical population.

3 Monte Carlo study: theoretical population

In the simplest set-up, data is generated from a two-level model in which the domain parameters θ_j are realizations of $\theta_j \sim \mathrm{N}(\mu = \theta, \sigma_u^2 = 3)$ and the observations y_{ji} (subject i in area j) are realizations of $y_{ji} \sim \mathrm{N}(\mu = \theta_j, \sigma_\epsilon^2 = 6)$. The number of small areas is 40. In some simulations the within-area sample sizes are equal to $n_j = 10$, while in other simulations n_j ranges from 6 to 40.

Next we list the estimators considered in the Monte Carlo study. The direct and synthetic estimators are respectively the sample mean of area j and the overall sample mean $\hat{\theta}$. The composite estimators can be classified according to whether or not the weights are known (*theoretical*) or estimated (*empirical*), and according to whether the estimator of the squared bias term $(\theta_j - \theta)^2$ is *area specific* (weights will be denoted by π_j) or *averaged* across the areas (weights denoted as ω_j). Except for the direct estimator, denoted by D, all estimators considered are composite estimators whose weights are specified as follows:

DESIGN-BASED ESTIMATORS

Theoretical composite: TC1

$$\tilde{\pi}_j = \frac{\sigma_{j\epsilon}^2/n_j}{(\theta_j - \theta)^2 + \sigma_{j\epsilon}^2/n_j}$$

Empirical composite: CA

$$\hat{\pi}_j^* = \frac{\hat{\sigma}_{j\epsilon}^2/n_j}{(\hat{\theta}_j - \hat{\theta})^2 + \hat{\sigma}_{j\epsilon}^2/n_j}$$

Note that TC1 and CA use area-specific values for the within-area variance $\sigma_{j\epsilon}^2$ (they allow for heteroscedasticity of this variance across areas).

MODEL-BASED ESTIMATORS

Theoretical composite: TC2

$$\tilde{\omega}_j = \frac{\sigma_\epsilon^2/n_j}{\sigma_u^2 + \sigma_\epsilon^2/n_j},$$

with population (true) values for the variances within σ_ϵ^2 and between σ_u^2.

Empirical composites: C

C is the composite estimator with $\hat{\omega}_j$ defined in (13), (14) and (15).

ML estimator: CML

Uses the estimator (12) with the population values μ, σ^2_ϵ and σ^2_u substituted by estimates obtained by fitting the model in (10) by ML.

3.1 Monte Carlo study: θ_j random

We first generate the area-level quantities θ_j as random draws from an assumed distribution $\mathrm{N}(\theta, \sigma^2_u)$, independently across replications. The areas cannot be distinguished by any features (they are exchangeable), and so their MSEs are the same for all the areas, for each estimator. As should be expected, the results summarized in Figure 1 indicate that the MSEs for the different methods are highly correlated. Within a method, the empirical MSE's are not constant because the number of replications is finite (it is 3000).

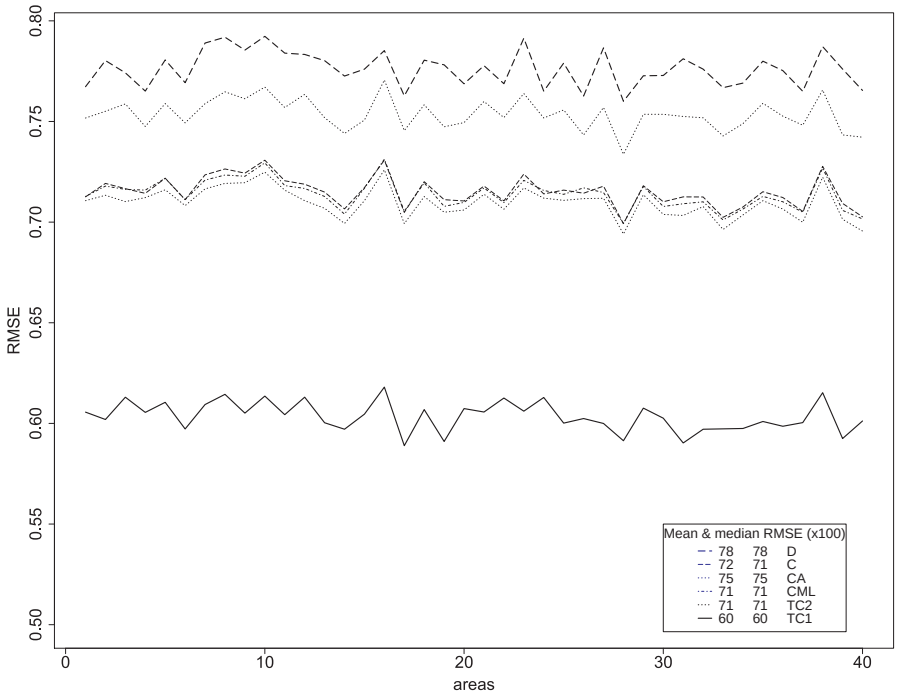

Figure 1: *Root MSE (RMSE) for areas $j = 1, \ldots, 40$ when each mean θ_j is random across replications. Within- and between-area distributions are normal (number of replications is* 3000*, sample size in each area is* 10*). The mean and the median of the RMSE across areas are shown in the legend.*

The theoretical design-based estimator TC1 is far more efficient than the others since it uses more information about the true values of the within-area variance and area-bias in each replication. The within-area variance is constant across replications, but this is not the case for the area-bias. The theoretical estimator TC2 that uses variance and bias parameters common across the areas performs similarly as the C and CML estimators (the last two estimators are equivalent, given that the n_j are equal across areas), which are the next in performance. The feasible design-based estimator CA performs poorly. Finally, the direct estimator has the poorest performance.

The gain of TC1 over TC2 can be explained by the fact that TC1 uses information about the squared area-bias $(\theta_j - \theta)^2$, which varies across replications, while TC2 uses only information about the true value of its expectation, the between-area variance (model parameter σ_u^2). Replacement of the parameters by their estimates in the model-based methods does not reduce this efficiency substantially; indeed, the RMSEs of TC2, C and CML are nearly indistinguishable in Figure 1. In contrast, the design-based estimator CA, which is based on substituting an estimate for the true value of the area-bias, incurs a severe loss of efficiency when compared with the theoretical estimator TC1.

3.2 Monte Carlo study: θ_j fixed

Now we assume that the θ_j are fixed across replications, in accordance with the empirical set-up in which the *eccentricity* of an area, i.e. the deviation of the area from the overall mean, is an (unknown) but fixed quantity that remains constant across replications.

Figure 2 reports the empirical root-MSE (RMSE) across replications for each area and for the different estimates. TC1 and TC2 are not feasible in practice since they use true values of population parameters that are not available in a typical application. However, the performance of TC1 and TC2 will shed light on the nature of the accuracy of the alternative estimators.

We see that the theoretical composite estimator TC1 that uses area-specific bias performs better than the theoretical composite TC2 that uses a single parameter (the variance between) to account for the squared bias averaged across the areas. In fact, TC2 performs poorly in all the areas, and as the worst estimator for areas with an extreme value of eccentricity (on the far right-hand side of the x-axis), that is, the areas for which the mean deviates highly from the overall area mean. To understand the variation of RMSE across the areas, these have been ordered according to their absolute deviation $| \theta_j - \theta |$, so that the extreme areas are located at the right-hand side. The lengths of the bars at the bottom of the graph are proportional to these deviations. We see that the largest difference between TC1 and the other statistics arises when $| \theta_j - \theta |$ is small; on the other hand, TC1 and TC2 nearly coincide when $| \theta_j - \theta |$ is approximately equal to the between-area variance. The empirical model-based composite estimators also perform poorly in all the areas. These results can be summarized as follows:

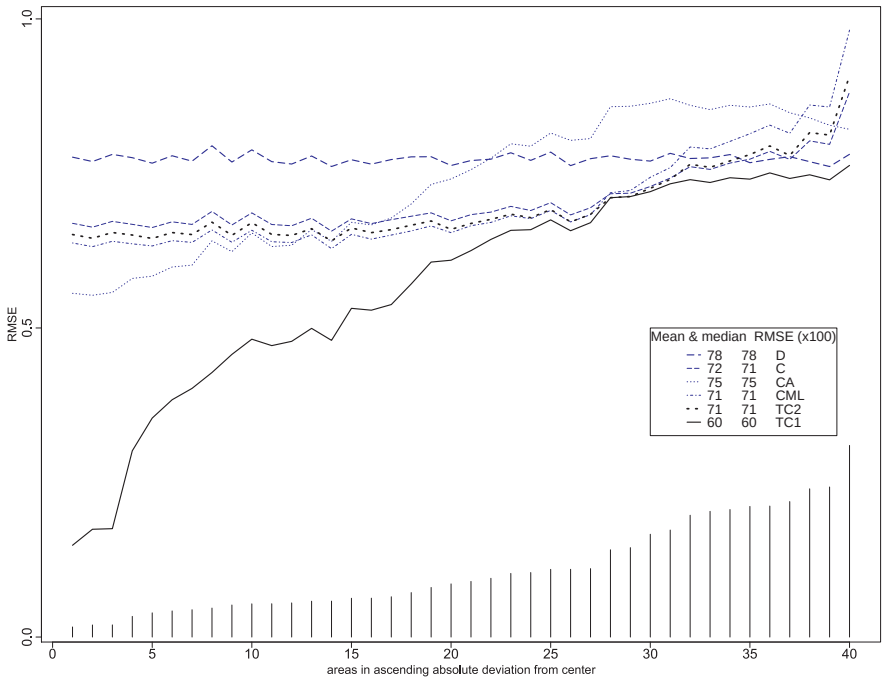

Figure 2: *Root-MSE (RMSE) of each area when the θ_j are fixed across replications. The within-area distribution is normal, with homocedastic within-area variances. The area sample size is constant and equal to* 10*. The number of replications is* 3000*. The legend shows the mean and the median of the RMSE across areas.*

- TC1 is the most efficient estimator for all the areas. This is a theoretical estimator, not feasible in applications. It provides a benchmark against which other estimators can be compared or related.
- CML and C are inefficient for all he areas and specially for those with the largest deviations from the centre (large eccentricity).
- For the model-based estimators (C, CML and TC2), using estimated or true values of the parameters makes very little difference. This is not the case for the design-based estimators; just compare TC1 with CA.
- CML performs poorly for all the areas and specially for those that deviate substantively from the centre. The accuracy of CA increases for small or extreme values of eccentricity.

The above difference among estimators can not be appreciated when observing RMSE averaged across areas.

We will see that these results hold in a variety of circumstances, when we vary the sample size, with large or small number of areas, and also with deviation from normality, both in the within-area distributions and in the distribution that generates the fixed realized values of the area means.

3.3 Non-normality and unequal sample sizes n_j

Now we consider the case where the θ_j have been sampled from an asymmetric distribution (and they stay fixed across replications), and the within-area distribution is non-normal (in fact, it is a chi-square distribution with 1 degree of freedom). The sample size ranges from 6 to 45, the number of areas is 40. The number of replications is 3000. The results are shown in Figure 3. For clarity of the graph, since across areas the maximum difference of the RMSE for C and CML is .03, only the RMSE for C is shown.

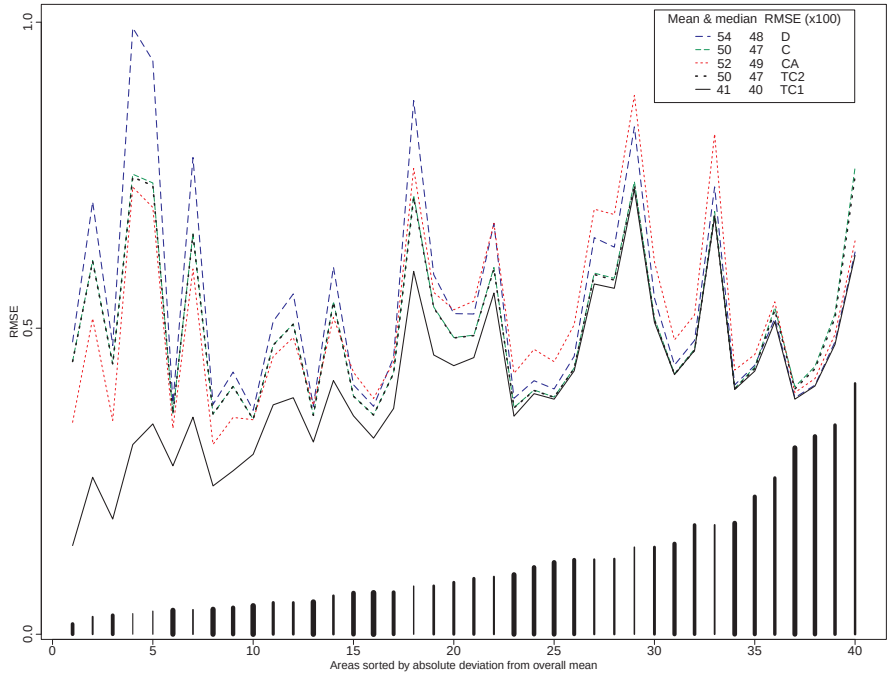

Figure 3: *Root-MSE (RMSE) of each area when the θ_j are fixed across replications. The within-area distribution is non-normal, with homocedastic within-area variances, and area sample size ranging from* 6 *to* 45*. Sample size variation is indicated by the thickness of the bars in the x-axis (thicker bar indicating larger sample size). The number of replications is* 3000*. The legend shows the mean and median of the RMSE across areas.*

Figures (2) and (3) show a similar pattern regarding the relative position of the estimators, though peaks are present in Figure (3) due to the variation of sample size across areas (sample size variation is proportional to the thickness of the bars in the x-axis). Note that the peaks correspond to areas with a relatively small sample size. From Figure (3) we conclude

1. The RMSE tends to increase with the eccentricity of the area.
2. TC1 is superior to all the estimators.

3. The feasible estimators C and CML are inefficient for all the areas, especially on those that deviate highly from the overall mean (high eccentricity), and so is the theoretical estimator TC2.
4. CA does not do as badly as C and CML for those areas with low values on eccentricity.
5. As expected, the RMSE tends to decrease with the sample size.

We also computed a version of the empirical composite C that estimates the variance-within σ_ϵ^2 as an (unweighted) mean of the within-area estimates $\hat{\sigma}_{\epsilon_j}^2$ of (16), but the difference in terms of MSE with the standard version of *C* was negligible.

We found that the true values of MSE computed according to formula (2) are indistinguishable from the (estimated) ones computed with 3000 replications and presented in Figure 3. Figure 4 displays the same graph with true RMSE for the three estimators D, TC1 and TC2. In both figures we see the superiority of the design-based estimators (TC1) over the model-based ones (TC2), not only for some areas that deviate highly from the overall mean, but also for those areas that exhibit a small value of eccentricity (the areas on the left of the x-axis).

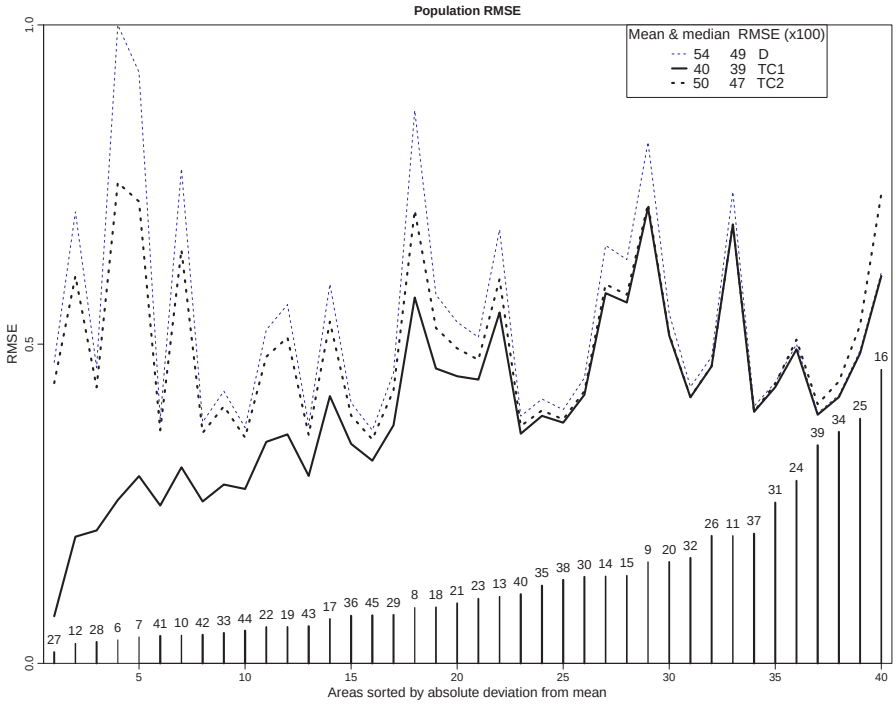

Figure 4: *Theoretical values of the RMSE for D, TC1 and TC2 and for each area, for θ_j fixed across replications. Area sample size ranging from* 6 *to* 45*. The legend shows the mean and the median of the RMSE across areas.*

Table 1: *Population characteristics*

Counties ('Comarques')	Code	Size N_j	Mean θ_j	Squared bias $(\theta_j - \theta)^2$	Variance $\sigma^2_{j\epsilon}$
Alt Camp	AC	1282	8.73	0.09	3250.37
Alt Empordà	AE	4712	5.28	14.11	294.27
Alt Penedès	AP	3052	8.91	0.02	1686.24
Alt Urgell	AU	745	4.71	18.70	158.25
Alta Ribagorça	AR	140	4.59	19.73	205.38
Anoia	AN	3264	7.86	1.37	801.64
Bages	BA	5698	8.24	0.63	1356.90
Baix Camp	BC	5530	6.47	6.59	6479.54
Baix Ebre	BB	2237	6.31	7.41	534.40
Baix Empordà	BE	4634	5.44	12.92	425.17
Baix Llobregat	BL	20541	9.73	0.48	1642.46
Baix Penedès	CP	2197	5.26	14.23	171.82
Barcelonès	BN	88331	10.63	2.55	10314.88
Berguedà	BG	1397	5.44	12.90	196.15
Cerdanya	CR	788	3.71	28.34	71.93
Conca de Barberà	CB	611	8.29	0.56	1388.95
Garraf	GR	3466	6.28	7.62	685.91
Garrigues	GS	516	5.24	14.42	96.89
Garrotxa	GX	1909	7.51	2.33	419.72
Gironès	GI	6369	9.82	0.62	2037.47
Maresme	MA	11718	6.46	6.64	605.07
Montsià	MO	1918	5.61	11.73	246.00
Noguera	NG	1128	5.12	15.30	93.29
Osona	OS	5494	7.09	3.77	774.65
Pallars Jussà	PJ	410	4.37	21.76	130.37
Pallars Sobirà	PS	272	4.06	24.76	55.46
Pla d'Urgell	PU	1106	6.59	5.95	271.85
Pla de l'Estany	PE	1160	6.07	8.79	143.37
Priorat	PR	254	4.11	24.26	180.17
Ribera d'Ebre	RE	620	5.71	11.07	418.72
Ripollès	RI	959	7.87	1.35	875.92
Segarra	SG	594	10.87	3.35	8171.41
Segrià	SR	7096	7.74	1.69	714.23
Selva	SV	4586	7.11	3.70	610.20
Solsonès	SO	508	5.58	11.93	157.58
Tarragonès	TG	7440	9.42	0.15	1675.66
Terra Alta	TA	297	4.25	22.87	40.28
Urgell	UG	1178	6.28	7.59	312.25
Val d'Aran	VA	503	5.28	14.08	270.11
Vallès Occidental	VC	26683	10.34	1.71	3026.89
Vallès Oriental	VR	11795	8.45	0.34	832.68

† The average number of affiliates in Catalonia (overall mean θ) is 9.04.

Figures 3 and 4 show the same ranking among the estimators according to their root-MSE. The same conclusions 1 to 4 that were drawn from Figure 3 apply also to Figure 4.

4 Simulation study on a real population

In this section we study the behaviour of several estimators through a Monte Carlo simulation in which we replicate samples from the Labour Force Census of Enterprises affiliated with the Social Security system in Catalonia. This census contains information on the number of employees who are registered in the Social Security system for each enterprise. The data is available on a quarterly basis from year 1992. We consider only the population in the first quarter of year 2000. The census contains 243,184 observations for Catalonia in year 2000, divided into 12 groups according to the economic sector to which each firm belongs, and into 41 counties (the 'comarques').

We ignore the sector-based classification and focus solely on the division by counties. Table 1 shows the number of enterprises (population size) and the mean and variance of the variable of interest (number of registered employees) in each county. The distribution of the enterprises across Catalonia is very uneven, as they are concentrated mainly in densely populated areas. In our set-up, the small areas are held fixed across resampling over the 1000 replications. In each replication, we extract a proportional stratified sample by county. We used sample sizes representing 10%, 5%, 2% and 1% of the population, which gives sample sizes close to those used by IDESCAT in several surveys. Table 2 summarizes the characteristics of these samples. Sample sizes for each county can be easily deduced from Table 1, applying the corresponding sample size percentage reported in each simulation.

Table 2: *Sample sizes of the Monte Carlo study (empirical population)*

Overall sample		Sample size in county			
% of pop.	Sample size	Mean	Median	Min.	Max.
1	2431	59.3	19	1	883
2	4863	118.6	38	3	1767
5	12159	296.6	95	7	4417
10	24316	593.1	191	14	8833

This population recreates conditions of non-normality, uneven sample sizes, and heterogeneity of within-area variances that are likely to appear in applications. So a Monte Carlo evaluation based on this population will assess the performance of competing estimators in a realistic setting.

We evaluate the performance of the theoretical estimators TC1 and TC2, and the empirical estimators C, CML and CA described in Section 2. For each estimator,

we computed the empirical relative root-MSE (RRMSE) across replications for each county. Using absolute (instead of relative) root-MSE gave the same pattern of performance as when using RRMSE. Even though the theoretical estimators TC1 and TC2 are unfeasible in practice (since the true values of the variances are unknown in a given application), we wanted the Monte Carlo study to illustrate the effect on accuracy when moving from BLUP to EBLUP. The graphs show clearly a shift on accuracy between the theoretical estimators TC1 and TC2 (BLUE estimators, from the fixed and random perspective, respectively) and the other empirical (EBLUP) estimators, with the theoretical estimators having, of course, lower MSE. In this Monte Carlo study, the theoretical estimators should be viewed as providing benchmarks for the accuracy of the EBLUP estimators. While TC1 is the best estimator of them all, its empirical counterpart CA performs worse in average. Also, the empirical C performs worse on average than its theoretical counterpart TC2. The discrepancy between the empirical estimators and their theoretical counterparts is larger when the areas are not too extreme. The CML estimator was computed using `proc xtmixed` of the software package `Stata 10.0`, employing the option `emonly`.[4] For each estimator, we computed the empirical relative root-MSE (RRMSE) across replications for each county. For the 10% and 5% sample sizes, the direct and the composite estimators have similar RRMSE values. For those sampling schemes, D has the smallest RRMSE among the feasible estimators and is more efficient than the theoretical model-based TC2 estimator. We therefore focus on the description of the 2% and 1% sampling designs.

Figure 5 plots the variation of the RRMSE for the estimators and areas for the 2% sampling design. For clarity, the RRMSE of CML is omitted as it is nearly indistinguishable from C. The same pattern of variation is observed for the 1% sample. Areas have been ordered with respect to their eccentricity, i.e. deviation of the area mean from the overall mean (the heights of the bars are proportional to the eccentricity of the area). We see that the RRMSE tends to increase as the areas become more extreme in terms of eccentricity. The area sample size is proportional to the thickness of the bar. We observe that RRMSE tends to decrease as the sample size increases.

The direct estimator D performs poorly. The design-based estimator CA has a very good performance across all the areas, being close to its theoretical counterpart, TC1, which is the most efficient. The model-based estimators TC2 and C (and CML) do better than the direct estimator D but worse than the theoretical design-based estimator TC1. The poor performance of D and other feasible model-based estimators for some counties, specially for the county SG ('Segarra") stands out. Segarra has both a huge value of the within-county variance and a very small sample size (see columns 3 and 5 of Table 1). In applications of small-area estimation, it should be of high concern that our area has such extreme features. If we knew the true values of the squared area-bias and

[4] Using `proc xtmixed` with default options produced a large percentage of replications with non-convergent solutions; however, the problem of non-convergence disappeared when we used the option `emonly` (expectation maximization algorithm when the gradient based routines did not converge). That is, a proper CML estimate was obtained in each replication.

the within-area variance (as in TC1 and TC2) then MSE would be reduced dramatically for SG.

The high fluctuation of the performance of the direct estimator is due to the variation of the sample sizes and within-area variances across areas. For extreme areas, the CA estimator performs similarly as the design-based estimator TC1. Estimation of the population parameters has a profound effect on the accuracy for areas that are not extreme. This is the case, for example, for Alt Camp (AC).

For completeness, Figure 6 shows the results for the 10% sampling design. We observe the same pattern of performance across areas as in Figure 5. The distance between the model-based and the design-based estimators is more obvious for areas with greater eccentricity. This graph shows that the direct estimator nearly matches the efficiency of TC1, a clear indication that for such a large sample size, small-area estimation is redundant.

Figures 5 and 6 show that the model-based estimators do not perform too badly when the small areas do not show a high value of eccentricity. But for areas that are very extreme these estimators do worse than the the direct estimator. Figure 5 shows that CA can be a good alternative in practice, because TC1 is unknown in real cases. Surprisingly, not even averaging across areas, the estimator C (or CML) can compete with the composite alternative CA.

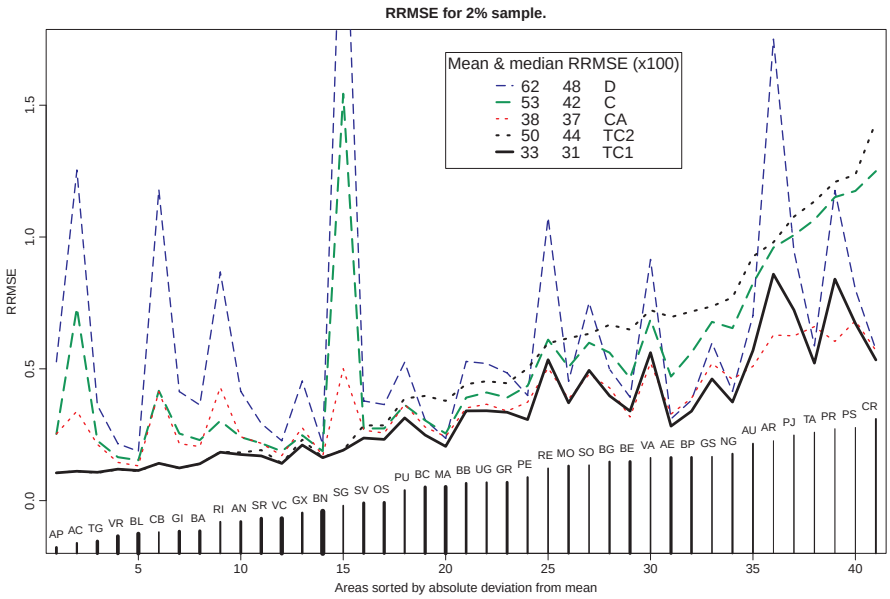

Figure 5: *For the Monte Carlo analysis with an empirical population the graph shows the RRMSE or each county, for the 2% sample. Heights of the bars are proportional to the deviations or the area-level means from the overall mean and their thicknesses are proportional to sample sizes. The legend shows the mean and the median of the RRMSE across areas.*

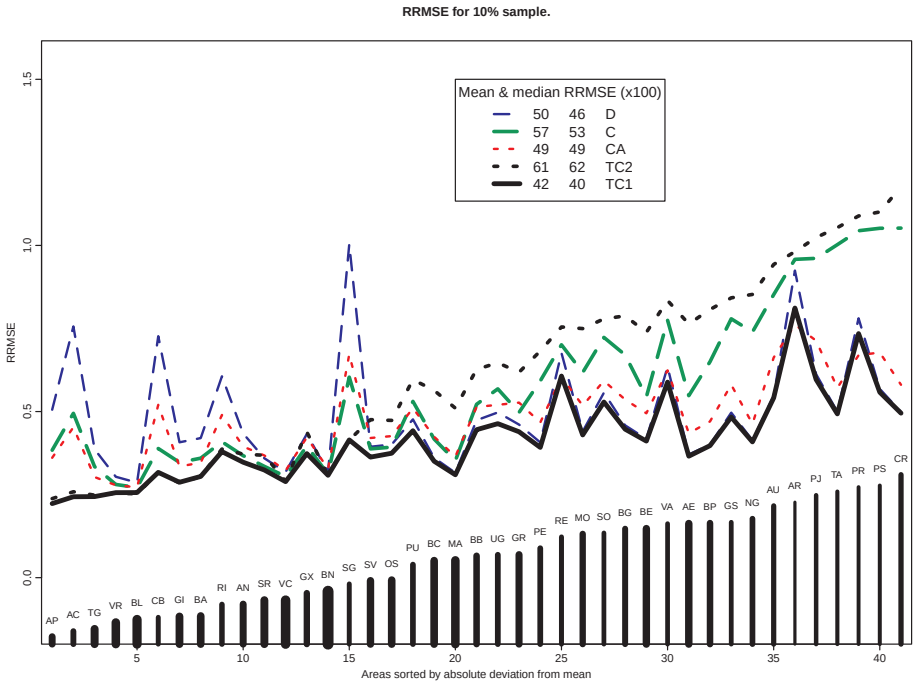

Figure 6: *For the Monte Carlo analysis with an empirical population the graph shows the RRMSE for each county, for the 10% sample. Heights of the bars are proportional to the deviations or the area-level means from the overall mean and their thicknesses are proportional to sample sizes. The legend shows the mean and the median of the RRMSE across areas.*

This Monte Carlo study with a real population provides a context in which we can recognize different scenarios encountered in an application.

1. For an area with a large sample size, all the small-area estimators are close to each other. This is the case of Barcelonès (BN).
2. In areas with a small sample size and a extreme within-area variance, not necessarily extreme in eccentricity, the empirical small-area estimators may perform very poorly. This is the case of Segarra (SG). For such areas, the incorporation of information on the magnitude of the between- and within-area variances may produce dramatic gains on RRMSE.
3. In an area with a small value of eccentricity and small sample size, model-based estimators are less efficient than the design-based estimators. This can be seen in Alt Camp (AC).
4. In an area with a high value of eccentricity and small sample size, design- and model-based estimators gives high gains over the direct estimator. This can be seen in Alta Ribagorça (AR).

5 Conclusions and agenda

We have seen that in the case of fixed areas, averaging MSE across areas does not provide the complete picture of the performance of alternative small-area estimators. Such averaging of MSEs can be used only to evaluate accuracy in the context of random-area parameters.

We conclude that a composite estimator that uses a common bias estimator for all the areas performs poorly on areas that are extreme. The same is true for the theoretical composite estimator TC2. The problem carries over to the mixed-effects regression, even when the model is not misspecified. Therefore, estimation of the squared bias term for each area becomes crucial.

We conjecture that issues regarding the estimation of the variances within the areas will be less critical; however, the exercise on a real population shows also the importance of recognizing non-normality and variation across areas of the within-area variance.

These findings indicate that the key to improve small-area estimation is to acknowledge the fixed-effect nature of the data and to improve estimation of the squared area-bias. Differences (heteroscedasticity) of the within-area variances seem to be also critical. Several alternatives arise:

1. Using auxiliary information (such as a census or a previous survey) to estimate the squared bias. Then the same simple and convenient composite estimators could be used.
2. Improving the alternative composite estimator by defining different groups of areas that share a common between-area variance.
3. Estimating the squared bias using small-area methods. This approach has already been used in Longford (2007) for estimating MSEs of model-based estimators.

Further work assessing these alternatives is needed. As a final remark, we should note that we have confined discussion to the most simple model set-up where covariates are not present in the model; additional work assessing the validity of our findings when the model is expanded is worth pursuing.

References

Costa, A., Satorra, A. and Ventura, E. (2003). An empirical evaluation of small area estimators. *SORT (Statistics and Operations Research Transactions)*, 27, 113-135.

Efron, B., and Morris C. E. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach, *Journal of the American Statistical Association*, 68, 117-130.

Ghosh M. and Rao, J. N. K. (1994). Small area estimation: an appraisal, *Statistical Science*, 9, 55-93.

Isaki, C. T. (1990). Small-area estimation of economic statistics, *Journal of Business & Economic Statistics*, 8, 435-41.

Longford, N. T. (2007). On standard errors of model-based small-area estimators, *Survey Methodology*, 33, 69-79.

Neudecker, H. and Satorra, A. (2003). On best affine prediction, *Statistical Papers*, 44, 257-266.

Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (Eds.) (1987). *Small Area Statistics: An International Symposium*, John Wiley and Sons: New York

Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley and Sons: New York

Santamaría, L., Morales, D. and Molina, I. (2004). A comparative study of small area estimators, *SORT*, 28, 215-230.

Singh, M. P., Gambino, J. and Mantel H. J. (1994). Issues and strategies for small area data, *Survey Methodology*, 20, Statistics Canada, 3-22.

Erratum

Erratum

In Ocaña et al. *SORT*, 32, 151–176, at the beginning of subsection 2.5 the second and third indexes in expressions should be interchanged. The correct text is:

2.5 The carry-over controversy

As has been mentioned, under model (1), $\bar{D}$ is an unbiased estimator of the true formulation effect ϕ only in absence of carry-over effect.

The analysis of the carry-over effect is straightforward. In order to estimate it, we first form the sums inside each individual, $Y_{i\cdot k} = Y_{i1k} + Y_{i2k}$. Simple computations from model (1) lead to the following expressions:

$$\begin{array}{l} var\left(Y_{i\cdot k}\right) = 4\sigma_S^2 + \sigma_R^2 + \sigma_T^2 = \sigma_+^2 \\ E\left(Y_{i\cdot 2}\right) - E\left(Y_{i\cdot 1}\right) = \kappa. \end{array} \tag{24}$$

Then, the difference:

$$\hat{\kappa} = \bar{Y}_{\cdot\cdot 2} - \bar{Y}_{\cdot\cdot 1} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{i\cdot 2} - \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{i\cdot 1} \tag{25}$$

is an unbiased estimator of the carry-over effect with standard error estimated by:

$$\widehat{se}_{\hat{\kappa}} = \sqrt{\frac{\sum_{i=1}^{n_1}\left(Y_{i\cdot 1} - \bar{Y}_{\cdot\cdot 1}\right)^2 + \sum_{i=1}^{n_2}\left(Y_{i\cdot 2} - \bar{Y}_{\cdot\cdot 2}\right)^2}{n_1 n_2 \left(N-2\right)/N}}. \tag{26}$$

Book review

Probability and statistics with R

María Dolores Ugarte, Ana F. Militino and Alan T Arnholt

Chapman & Hall / CRC

ISBN: 978-1-58488-891-9

On the one hand, this book offers an extensive and well structured introduction to statistics. On the other hand, in my opinion, the best aspect of this book is that it opts for R as the statistical package for the entire learning process and for performing the statistical analyses. The book is intended for a wide audience, but I think that teachers and students will especially find it a valuable resource for introductory courses on statistics. The chapters cover the content of a classic one-year course (or two semesters courses) on descriptive statistics, probability, basic inference and regression methods, giving special attention to practical issues through interesting examples that use mostly real data sets, most of them provided by the authors. Illustrated examples as well as proposed exercises are also useful pedagogical tools and reflect the authors' experience in university teaching during the last years. This book is also the evolution of a previous book in Spanish , but the current book is much more complete both from the theoretical and practical point of view so that the previous one has to be considered as the initial seed.

Firstly, a brief introduction to S is provided, mainly regarding data objects and their manipulation, but also some concepts about programming and producing graphs are introduced. Here, one important point that needs to be clarified is the difference between the S-PLUS and R languages. Even though the title of the book only refers to R, the authors use S to refer to both R and S-PLUS, which can be a bit confusing, specially for readers that have recently started to work with R, and who may not have ever heard about S or S-PLUS. The point is that most code in S is also valid in R and vice versa since S was the initial language but R is the open source version. However, nowadays R has evolved by itself and has become the "lingua franca" of computational statistics, as the authors remark and S-PLUS is a commercial evolution of S. After this preliminary introduction to R, chapters 2 to 5 offer a well structured introduction to descriptive statistics and basic probability, including univariate and multivariate probability distributions. Chapter 6 covers issues regarding sampling and sampling distributions and the next three chapters (7 to 9) provide methods for point estimation and classic inference, including confidence intervals and hypothesis testing. Non-parametric methods for performing statistical inference are presented in chapter 10 and

ANOVA models and other methods of experimental design are covered in chapter 11. Finally, chapter 12 provides an extensive introduction to linear models, with special attention to the mathematical aspects.

The authors of the book have also developed a R library, named PASWR, that contains the datasets and some functions used through the book, which makes the content really interactive and provides facilities for using this book for teaching. Scripts from the commands used in the chapter can be found at the following web page: http://www1.appstate.edu/∼arnholta/PASWR/. A generous number of exercises are proposed at the end of each chapter, most of which have to be solved using R. A manual with the complete solutions to the problems has also been edited, which is available only for teachers.

In conclusion, I strongly recommend this book, especially for students and teachers of statistics and also many other fields, since they will find a large amount of material for learning statistics and R. From my own experience of teaching statistics, I can state that R is one of the best options for supporting practical sessions with students. Although R can be difficult to handle at the initial learning stage, after a reasonable period it can become one of the most effective platforms for performing statistical analyses. In addition to this, I think that this book can become a reference book for learning statistics with R, jointly with the books of Venables and Ripley and Dalgaard. The authors, in the preface of the book, finally thank "the geniuses of this age who first conceived of the idea of an excellent open source software for statistics and those who reared the idea to adulthood, our gratitude is immeasurable. May the lighthouse of your brilliance guide travelers on the ocean of statistics for decades for come. Thank you, R Core Team ". In this sense, I think that the authors of this book have to be thanked for their important contribution that will help statistical practitioners, students, teachers and scientists to definitely adopt R as the gold-standard package for statistics.

References

Ugarte, M. D. and Militino, A. F. (2001). *Estadística Aplicada con S-PLUS. Pamplona*, Universidad Pública de Navarra.

Venables, W.N. and Ripley. B. D. (2002). *Modern Applied Statistics with S.* New York, Springer.

Dalgaard, P. (2008). *Introductory Statistics with R.* New York, Springer.

Joan Valls Marsal
joan.valls@iconcologia.net
Catalan Cancer Registry, Catalan Institute of Oncology, Barcelona
Department of Mathematics, Autonomous University of Barcelona, Barcelona

Information for authors and subscribers

Information for authors and subscribers

Submitting articles to SORT

Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.cat) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a LaTeX 2_ε .

In any case, upon request the journal secretary will provide authors with LaTeX 2_ε templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

Publishing rights and authors' opinions

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following examples:

Citations
Mahalanobis (1936), Rao (1982b)

Journal articles
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9, 73-84.

Books
Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium*. New York: John Wiley and Sons.

Parts of books
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Web files or "pages"
Nielsen, S. F. (2001). *Proper and improper multiple imputation*
http://www.stat.ku.dk/~feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel.: +34-93 412 09 24 or +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name _____

_____

Organisation _____

_____

Street Address _____

_____

Zip/Postal code _____ City _____

State/Country _____ Tel. _____

Fax _____ NIF/VAT Registration Number _____

E-mail _____

Date _____

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)*** for the year 2009 (volume 33)

Annual subscription rates:

– Spain: €22 (4% VAT included)
– Other countries: €25 (4% VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4% VAT included)
– Other countries: €17/issue (4% VAT included)

Method of payment:

☐ Bank transfer to account number 2013-0100-53-0200698577

☐ Automatic bank withdrawal from the following account number

☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Check made payable to the Institut d´Estadística de Catalunya

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Bank copy

Authorisation for automatic bank withdrawal in payment for

SORT *(Statistics and Operations Research Transactions)*

The undersigned _____

authorises Bank/Financial institution _____

located at (Street Address) _____

Zip/postal code _____ City _____

Country _____

to draft the subscription to **SORT *(Statistics and Operations Research Transactions)*** from my account

number ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Date _____

Signature

SORT *(Statistics and Operations Research Transactions)*
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45