

SEMIPARAMETRIC EMPIRICAL BEST PREDICTION FOR SMALL AREA ESTIMATION OF UNEMPLOYMENT INDICATORS

BY MARIA FRANCESCA MARINO^{*,1,4}, MARIA GIOVANNA RANALLI^{†,1},
NICOLA SALVATI^{‡,1,3} AND MARCO ALFÒ^{§,2}

University of Florence^{*}, *University of Perugia*[†], *University of Pisa*[‡] and
“Sapienza” University of Rome[§]

The Italian National Institute for Statistics regularly provides estimates of unemployment indicators using data from the labor force survey. However, direct estimates of unemployment incidence cannot be released for local labor market areas. These are unplanned domains defined as clusters of municipalities; many are out-of-sample areas, and the majority is characterized by a small sample size which renders direct estimates inadequate. The empirical best predictor represents an appropriate, model-based alternative. However, for non-Gaussian responses its computation and the computation of the analytic approximation to its mean squared error require the solution of (possibly) multiple integrals that, generally, have not a closed form. To solve the issue, Monte Carlo methods and parametric bootstrap are common choices, even though the computational burden is a nontrivial task. In this paper, we propose a semiparametric empirical best predictor for a (possibly) nonlinear mixed effect model by leaving the distribution of the area-specific random effects unspecified and estimating it from the observed data. This approach is known to lead to a discrete mixing distribution which helps avoid unverifiable parametric assumptions and heavy integral approximations. We also derive a second-order, bias-corrected analytic approximation to the corresponding mean squared error. Finite sample properties of the proposed approach are tested via a large scale simulation study. Furthermore, the proposal is applied to unit-level data from the 2012 Italian Labor Force Survey to estimate unemployment incidence for 611 local labor market areas using auxiliary information from administrative registers and the 2011 Census.

1. Introduction. The Italian National Institute for Statistics (ISTAT) regularly provides estimates of unemployment indicators based on data obtained

Received December 2017; revised August 2018.

¹Supported by the project PRIN–SURWEY (grant 2012F42NS8, Italy).

²Supported by the project “Mixture and latent variable models for causal inference and analysis of socioeconomic data” (grant RBFR12SHVV, FIRB–Futuro in Ricerca).

³Supported by the projects: InGRID 2 (grant agreement No 730998, EU) and “From survey-based to register-based statistics: a paradigm shift using latent variable models” (grant PRA2018-9-Progetto di Ricerca di Ateneo).

⁴Supported by the project “Dipartimenti Eccellenti 2018–2022” (Ministerial grant, Italy).

Key words and phrases. Binary data, Exponential Family, finite mixtures, general parameters, mixed logistic model, unit-level model.

through the Italian Labor Force Survey (ILFS). The ILFS allows to obtain quarterly estimates of the main aggregates regarding the labor market which are important both at the local and the central government levels for the development of labor market policies. These estimates are planned to be reliable at a given, chosen a priori, geographical level and may not be suitable to all needs. For example, direct estimates of unemployment indicators cannot be disseminated for local labor market areas (LLMAs). These are 611 unplanned domains obtained as clusters of municipalities and defined at the Census on the basis of daily working commuting flows. In this context, direct survey estimates of unemployment incidence cannot be computed and/or published for most LLMAs. This is due to the presence of out-of-sample areas and too many LLMAs having a small sample size which leads to estimates with an unacceptable large coefficient of variation. For these reasons, ISTAT has implemented the use of indirect, model-based small area estimators to produce official yearly estimates of unemployment incidence for Italian LLMAs [D'Alò, Falorsi and Solari (2017), D'Alò et al. (2012)].

Small area estimation (SAE) has received considerable attention in the past decades in terms of theoretical developments and applications to official statistics. An updated appraisal of available approaches for SAE is given in Rao and Molina (2015). In this context, generalized linear mixed models [GLMMs, Laird and Ware (1982)] represent a typical tool of analysis. Area-specific random effects are used to account for sources of unobserved heterogeneity that are not captured by the covariates and describe correlation between units within the same small area. For Gaussian data, Battese, Harter and Fuller (1988) introduced and Prasad and Rao (1990) developed an empirical best linear unbiased predictor (EBLUP) to estimate small area characteristics. Tailored to the purpose of the ILFS, D'Alò, Falorsi and Solari (2017) developed unit-level linear mixed models with area- and time-specific random effects, which, based on data from different survey cycles, implement estimation using aggregate data to manage a large number of records. In fact, the ILFS is a continuous survey that yearly collects information on almost 250,000 households in 1,400 municipalities for a total of 600,000 individuals. However, many survey variables, such as the unemployment status, are categorical in nature and, therefore, SAE methods based on linear mixed models may not be fully appropriate.

Jiang and Lahiri (2001) developed an empirical best prediction (EBP) method for the area-specific random effects under a mixed logistic model providing a second-order, bias-corrected estimator for the corresponding mean squared error (MSE). Jiang (2003) extended this approach to deal with GLMMs for general responses in the Exponential Family. Several functions of area-specific model effects are also investigated by the author. More recently, Boubeta, Lombardía and Morales (2016, 2017) derived the EBP and the corresponding (second-order) approximation to the MSE under an area-level mixed Poisson model for small area counts, while Hobza and Morales (2016) specifically focused on the development of an EBP for small area proportions under the unit-level mixed logistic model

according to Jiang (2003). They also investigated the empirical behavior of the proposal through a large scale simulation study. An extension of this latter approach to deal with longitudinal responses was also recently proposed by Hobza, Morales and Santamaría (2018).

In all of these approaches, the area-specific random effects are assumed to be independent and identically distributed (i.i.d.) draws from a Gaussian distribution. One of the drawbacks associated with this assumption entails the computational burden required to derive parameter estimates, compute the EBP and, in particular, provide the corresponding measure of reliability. For non-Gaussian responses, we need to deal with (possibly) multiple integrals that do not admit a closed form expression and, therefore, need to be approximated. Numerical approaches based, for example, on (adaptive) Gaussian quadrature or Laplace approximations [see, e.g., Pinheiro and Bates (1995)] or using Monte Carlo approximations [see, e.g., McCulloch (1997)] are frequently used for this purpose. To avoid computational issues, ad hoc alternatives, mainly based on plug-in predictors and Taylor linearizations, were proposed and are currently largely applied [González-Manteiga et al. (2007), Molina, Saei and Lombardía (2007), Saei and Chambers (2003), López-Vizcaíno, Lombardía and Morales (2013)].

In this paper, we describe a further alternative and develop a semiparametric EBP (sp-EBP) for the small area parameters of interest and a second-order, bias-corrected approximation to the corresponding MSE. In particular, we propose to leave the distribution of the area-specific random effects (the mixing distribution) unspecified and estimate it from the observed data via a nonparametric maximum likelihood approach [NPML, Laird (1978), Simar (1976), Lindsay (1983a, 1983b)]. This estimate is known to be a discrete distribution defined over a finite number of locations leading to a (semiparametric) finite mixture model with a conditional kernel in the Exponential Family. The proposed approach offers a number of advantages. First, it allows us to avoid unverifiable assumptions on the random effect distribution; second, since mixture parameters are directly estimated from the data and are completely free to vary over the corresponding support, extreme and/or asymmetric departures from the homogeneous model can be easily accommodated. Last and more important, the discrete nature of the mixing distribution allows us to avoid integral approximations and considerably reduces the computational effort. The gain with respect to the parametric alternatives is particularly evident when analyzing non-Gaussian responses.

We present the proposed approach for a general small area parameter, starting from a general response with density in the Exponential Family. Later, we focus on the relevant case of binary data. We compare our proposal to the EBP [Jiang (2003)] and to the plug-in estimator [e.g., González-Manteiga et al. (2007), Saei and Chambers (2003)] in terms of prediction accuracy and computational burden in a large scale simulation study. Then, we prove the benefits from using the proposed sp-EBP approach on data from the ILFS to estimate unemployment incidence for the 611 LLMA's using auxiliary information from administrative registers and the

2011 Census. We compare the proposed approach with direct estimates and with the two aforementioned approaches based on parametric mixed logistic models.

The paper is organized as follows. Section 2 presents the Italian Labor Force Survey, the estimation problem, and the auxiliary information available. Section 3 introduces the notation and a brief review of the EBP and its MSE approximation. In Section 4, we describe the proposed approach; Section 4.1 entails maximum likelihood estimation, while the proposed sp-EBP and its MSE approximation are detailed in Section 4.2. Section 5 focuses on the case of binary responses. Section 6 reports the results of the simulation study, while Section 7 entails the application of the proposed approach to the ILFS data. Last, Section 8 summarizes our findings and provides guidelines for future research.

2. The ILFS data. The ILFS is the most important statistical source of information on the Italian labor market. The target population includes the members of all Italian households who regularly live within the national borders, have Italian or foreign citizenship and are regularly enrolled in the municipal lists. Households registered as resident in Italy who habitually live abroad and permanent members of collective facilities (hospices, children's homes, religious institutions, barracks, etc.) are excluded. A two-stage municipality-household sampling design is used to collect data. Primary sampling units are stratified by province (LAU1) and population size. Secondary sampling units are selected with equal probabilities. All individuals with usual residence in the dwelling are interviewed.

The ILFS provides quarterly estimates of the main aggregates for the labor market, such as employment status, type of work and work experience by gender, age and region (NUTS2). Here, we focus on data from the first quarter of 2012 which consist of measurements taken on 93,217 units aged 15–65 and distributed in 453 LLMA. LLMA refer to 611 unplanned domains obtained as clusters of municipalities in which the bulk of the labor force lives and works, and where establishments can find the largest amount of the labor force necessary to occupy the offered jobs. They respond to the need for meaningfully comparable subregional labor market areas for the reporting and the analysis of statistics. LLMA are defined on a functional basis, the key criterion being the proportion of commuters who cross the LLMA boundary on their way to work. In 2011 with the last Census, LLMA were redefined by the analysis of daily working commuting flows using a new allocation process, an evolution of the previous algorithm. Nearly half of the LLMA stands in the size class from 10,000 up to 50,000 inhabitants, whereas the highest proportion of the population (35%) lives in LLMA with a dimension between 100,000 and 500,000 inhabitants. In 332 LLMA (over 70% of the national population), more than three quarters of the labor force lives and works in the same LLMA, that is, selfcontainment is more than 75%.

Figure 1(a) shows the distribution of the LLMA by sample size. This plot does not include the 158 areas with zero sample size. Among the observed LLMA, the

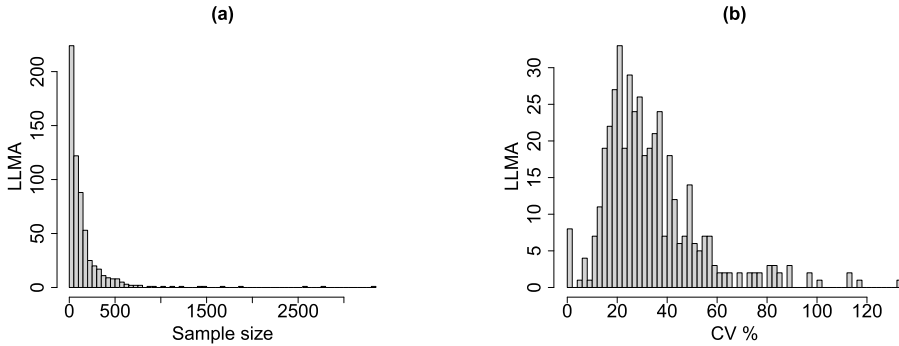


FIG. 1. Distribution of LLMA by sample size (a) and percent coefficient of variation of direct estimates of unemployment incidence (b). First quarter, 2012.

sample size ranges between 13 (Acqui Terme, Piedmont Region) and 3,301 (Milan, Lombardy Region). The mean value is equal to 205.8, while quartiles are 61 (25%), 122 (50%) and 223 (75%), respectively. That is, several LLMA are characterized by a very small sample size that hinders reliability of direct estimates. Figure 1(b) reports the distribution of the (percent) coefficient of variation (CV) for the direct estimates of unemployment incidence. The vast majority of estimates have a CV that is larger than 33% which is usually considered as a threshold for reliability.

Our main interest is on the *employment status* variable which can take one out of three different categories: employed (53.6%), unemployed (6.6%) and inactive (39.8%). Together with information on employment status for sampled individuals, the following explanatory variables are also available: *gender-age*, a categorical variable with six categories corresponding to female or male (F/M) and three age groups (15–24, 25–34 and 35–65); *educational level*, a categorical variable with four categories corresponding to no education or primary school diploma, secondary school diploma, high school diploma and university degree or beyond; and *U-count*, a discrete variable measuring the number of unemployed in a given gender-age group for each LLMA according to the 2011 Census.

To have a first insight on the data, we report in Tables 1–2 the sample distribution of the *employment status* by *gender-age* and *educational level*, respectively. From these tables, we may observe that unemployment incidence is generally higher for people aged 25–34 in the sample regardless of gender; for the other age groups, unemployment is less frequent among females. By looking at the last column of the table, we notice that females are more frequently *inactive* when compared to males regardless of the age group. This is likely due to their engagement in housekeeping and explains why unemployment incidence is lower in this group. Similarly, by looking at Table 2 we may observe that the percentage of unemployment is relatively higher for individuals with higher education. Also in this case, by looking at the last column it is evident that such a finding

TABLE 1

Sample percentage distribution with standard errors (S.E.) of unemployed status by gender-age

	Unemployed	S.E.	Employed	S.E.	Inactive	S.E.
M: 15–24	11.0	0.4	21.5	0.5	67.6	0.6
M: 25–34	11.7	0.4	71.6	0.5	16.7	0.4
M: 35–65	5.3	0.1	70.8	0.3	23.8	0.2
F: 15–24	8.8	0.3	14.1	0.4	77.1	0.5
F: 25–34	11.7	0.4	53.9	0.6	34.5	0.5
F: 35–65	4.2	0.1	48.3	0.3	47.5	0.3

is mainly related to job hunting. For instance, 70% of individuals with a primary school diploma or less are out of the job market as they are not actively looking for a job, and this can be explained by a relatively older age.

As highlighted before, the prediction of unemployment incidence cannot be based on direct survey estimation as direct estimates cannot be computed and/or published for most LLMAAs. For these reasons, unit-level SAE methods may provide a viable tool to obtain such estimates. In the following section, we introduce the EBP approach by Jiang (2003) for the estimation of small area parameters together with the approach to approximate the corresponding MSE. As stated in Section 1, one of the main drawbacks of such a method is the computational complexity we have to face with non-Gaussian data and a large number of observations/small areas as for the ILFS data. In Section 4, we develop a computationally efficient alternative based on a semiparametric approach.

3. The empirical best prediction. Let U denote a finite population of size N , which can be partitioned into m nonoverlapping small areas/domains, with U_i denoting the i th small area with size N_i , $i = 1, \dots, m$. For a given small area i , data consist of N_i measurements of a response variable Y_{ij} and a p -dimensional vector of covariates $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})'$, with $j = 1, \dots, N_i$. Also, let $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_m$ be i.i.d., q -dimensional, vectors of area-specific random effects ($q \leq p$) with density

TABLE 2

Sample percentage distribution with standard errors (S.E.) of unemployed status by educational level

	Unemployed	S.E.	Employed	S.E.	Inactive	S.E.
Primary school or less	4.6	0.2	24.9	0.4	70.5	0.5
Middle school	6.9	0.1	44.5	0.3	48.6	0.3
High school	7.2	0.1	62.8	0.3	30.0	0.2
University degree or beyond	5.6	0.2	75.7	0.4	18.8	0.4

$f_\alpha(\cdot)$, $E_\alpha(\alpha_i) = 0$ and $E_\alpha(\alpha_i \alpha_i') = \Sigma$, for all $i = 1, \dots, m$. Last, let \mathbf{w}_{ij} denote a q -dimensional subset of \mathbf{x}_{ij} associated to α_i .

We assume that a sample of size n is drawn from the above population and denote by s_i the set containing the n_i population indexes of sample units belonging to small area i with $n = \sum_{i=1}^m n_i$. On the other hand, the set $r_i \subseteq U_i$ contains the $N_i - n_i$ indexes for nonsampled units in small area i . For ease of notation, we assume that all areas are sampled, even though the presence of out of sample areas can be easily accommodated. We further assume that values of Y_{ij} are known only for the sample ($i = 1, \dots, m, j \in s_i$), while the values of \mathbf{x}_{ij} and \mathbf{w}_{ij} are known for all units in the population ($i = 1, \dots, m, j = 1, \dots, N_i$). This assumption can be quite restrictive in some real-world applications, since it implies the availability of individual population information. However, when the auxiliary variables are categorical and/or take a finite number of values, the assumption can be relaxed, as we will discuss in more details in the application to the ILFS data. Last, we assume that sampling is noninformative for the small area distribution of $Y_{ij} | \mathbf{x}_{ij}$ allowing us to use population level models with sample data.

3.1. *The model.* According to a local independence assumption, we assume that, conditional on the area-specific random effects α_i , responses Y_{ij} from the same small area i are independent with density in the Exponential Family

$$f_{y|\alpha}(y_{ij} | \alpha_i; \mathbf{x}_{ij}) = \exp\left\{ \frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi) \right\},$$

for $i = 1, \dots, m$ and $j = 1, \dots, N_i$. In the previous expression, ϕ is a dispersion parameter, $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, and θ_{ij} is the canonical parameter for the chosen member of the family. Let β denote a p -dimensional vector of fixed regression coefficients, and let us assume that θ_{ij} is modeled via the following regression model:

$$\theta_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{w}'_{ij}\alpha_i.$$

The joint distribution of $\mathbf{y}_i = (y_{i1}, \dots, y_{iN_i})'$ for the i th small area, conditional on the vector of area-specific random effects α_i , is obtained by exploiting conditional independence, while the marginal distribution of the area-specific sequence \mathbf{y}_i is obtained by integrating out α_i :

$$f_y(\mathbf{y}_i; \mathbf{X}_i) = \int_{\mathbb{R}^q} f_{y|\alpha}(\mathbf{y}_i | \alpha_i; \mathbf{X}_i) f_\alpha(\alpha_i) d\alpha_i,$$

with \mathbf{X}_i denoting the matrix of covariates associated to units in the i th area. Typically, a parametric specification for $f_\alpha(\alpha_i)$ is adopted with a common choice being the zero mean, multivariate Gaussian distribution. It is worth noticing that an implicit exogeneity assumption of observed covariates \mathbf{x}_{ij} is taken, that is, $f_\alpha(\alpha_i | \mathbf{X}_i) = f_\alpha(\alpha_i)$ or $E(\alpha_i | \mathbf{X}_i) = E(\alpha_i) = \mathbf{0}$. When this assumption is not fulfilled, the auxiliary regression approach by Mundlak (1978) can be adopted.

This slightly modifies the linear predictor above and produces area-specific random effects that are (linearly) free of \mathbf{X}_i ; see Neuhaus and McCulloch (2006). In the following, we assume that, if needed, such an approach is applied and that $f_\alpha(\boldsymbol{\alpha}_i | \mathbf{X}_i) = f_\alpha(\boldsymbol{\alpha}_i)$.

3.2. *EBP and MSE approximation.* We are interested in using sample data on responses Y_{ij} ($i = 1, \dots, m, j \in s_i$) and population data on covariates \mathbf{x}_{ij} ($i = 1, \dots, m, j = 1, \dots, N_i$) to predict a (possibly) nonlinear function of fixed and random effects, say $\zeta(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$, with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_m)$. According to Jiang (2003), the best predictor (BP) of ζ in terms of minimum MSE is given by

$$(1) \quad \tilde{\zeta}^{\text{BP}} = E_{\alpha|y}[\zeta(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) | \mathbf{y}] = \int_{\mathbb{R}^v} \zeta(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) f_{\alpha|y}(\boldsymbol{\alpha} | \mathbf{y}) d\boldsymbol{\alpha},$$

where

$$f_{\alpha|y}(\boldsymbol{\alpha} | \mathbf{y}) = \frac{\prod_{i=1}^m f_{y|\alpha}(\mathbf{y}_i | \boldsymbol{\alpha}_i; \mathbf{X}_i) f_\alpha(\boldsymbol{\alpha}_i)}{\prod_{i=1}^m \int_{\mathbb{R}^q} f_{y|\alpha}(\mathbf{y}_i | \boldsymbol{\alpha}_i; \mathbf{X}_i) f_\alpha(\boldsymbol{\alpha}_i) d\boldsymbol{\alpha}_i},$$

$f_{y|\alpha}(\mathbf{y}_i | \boldsymbol{\alpha}_i; \mathbf{X}_i) = \prod_{j \in s_i} f_{y_{ij}|\alpha}(\mathbf{y}_{ij} | \boldsymbol{\alpha}_i; \mathbf{x}_{ij})$, $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_m)'$ and $v = m \times q$. Since model parameters $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \phi, \boldsymbol{\Sigma})$ are unknown, they need to be estimated. Estimation can be accomplished by maximizing the observed data likelihood function:

$$(2) \quad L(\boldsymbol{\Phi}) = \prod_{i=1}^m f_y(\mathbf{y}_i; \mathbf{X}_i) = \prod_{i=1}^m \int_{\mathbb{R}^q} f_{y|\alpha}(\mathbf{y}_i | \boldsymbol{\alpha}_i; \mathbf{X}_i) f_\alpha(\boldsymbol{\alpha}_i) d\boldsymbol{\alpha}_i,$$

where, as before, $f_{y|\alpha}(\mathbf{y}_i | \boldsymbol{\alpha}_i; \mathbf{X}_i)$ refers to sample data only. To maximize equation (2), we need to evaluate an integral defined over the support of the area-specific random effects, and this can be directly done only in a few cases, for instance, when $f_{y|\alpha}(\cdot | \cdot)$ and $f_\alpha(\cdot)$ are conjugate. In all other cases numerical approximations (e.g., Gaussian quadrature techniques) or simulation based methods (e.g., Monte Carlo integration) need to be used, often leading to a nontrivial computational complexity. To overcome the issue, Jiang (1998) suggested to derive estimates by exploiting the method of moments. A penalized quasi-likelihood (PQL) approach [e.g., Breslow and Clayton (1993)] represents a further alternative which is less computationally demanding, even though it may provide inconsistent model parameter estimates [see, e.g., Rodriguez and Goldman (1995)].

Once parameters are estimated, we may compute the EBP of ζ , that is, $\hat{\zeta}^{\text{EBP}} = \tilde{\zeta}^{\text{BP}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Sigma}})$. To evaluate the quality of such predictions, the second-order MSE estimator detailed by Jiang (2003) can be considered. Under mild regularity conditions, the following decomposition holds:

$$(3) \quad \begin{aligned} \text{MSE}(\hat{\zeta}^{\text{EBP}}) &= E[(\hat{\zeta}^{\text{EBP}} - \zeta)^2] \\ &= E[(\tilde{\zeta}^{\text{BP}} - \zeta)^2] + E[(\hat{\zeta}^{\text{EBP}} - \tilde{\zeta}^{\text{BP}})^2] \\ &= d(\boldsymbol{\Phi}) + \frac{1}{m} e(\boldsymbol{\Phi}) + o_p(1/m), \end{aligned}$$

where

$$(4) \quad E[(\hat{\zeta}^{\text{EBP}} - \tilde{\zeta}^{\text{BP}})^2] = \frac{1}{m} e(\Phi) = \frac{1}{m} E_y \left[\left(\frac{\partial \tilde{\zeta}^{\text{BP}}}{\partial \Phi} \right)' m V(\hat{\Phi}) \left(\frac{\partial \tilde{\zeta}^{\text{BP}}}{\partial \Phi} \right) \right]$$

and

$$(5) \quad \begin{aligned} & E[(\tilde{\zeta}^{\text{BP}} - \zeta)^2] \\ &= d(\Phi) = E_\alpha[(\zeta)^2] - E_y[(\tilde{\zeta}^{\text{BP}})^2] \\ &= \int_{\mathbb{R}^v} \zeta(\beta, \alpha, \Sigma)^2 f_\alpha(\alpha) d\alpha - E_y \left[\left(\int_{\mathbb{R}^v} \zeta(\beta, \alpha, \Sigma) f_{\alpha|y}(\alpha | y) d\alpha \right)^2 \right]. \end{aligned}$$

In the last expression, $f_\alpha(\alpha) = \prod_{i=1}^m f_\alpha(\alpha_i)$ denotes the joint density of the random effects α_i , with $i = 1, \dots, m$. An estimator of $\text{MSE}(\hat{\zeta}^{\text{EBP}})$ can be obtained by replacing Φ in equation (3) by a consistent estimator, that is,

$$\widehat{\text{MSE}}(\hat{\zeta}^{\text{EBP}}) = d(\hat{\Phi}) + \frac{1}{m} e(\hat{\Phi}).$$

However, as outlined by Jiang (2003), while we get an error of order $o_p(m^{-1})$ when we replace Φ by $\hat{\Phi}$ into $e(\Phi)$, a bias correction is needed to obtain an unbiased estimator for $d(\Phi)$. We discuss this issue in more detail in the following.

As it is clear by looking at the above equations, computing the MSE requires the solution of (multiple) integrals that may not admit a closed form expression and this makes the computation extremely time-consuming. Bootstrap may represent an alternative to Monte Carlo approximations or Gaussian quadrature techniques, particularly when dealing with a limited number of small areas. However, when m is large, as in the case of the ILFS data, neither the analytic MSE approximation nor the bootstrap represent viable strategies due to computational issues. González-Manteiga et al. (2007) proposed a nonoptimal Prasad–Rao-type MSE estimator derived from a Taylor series approximation. This estimator fails when sample sizes are too small, while its behavior is proved to be reliable in the case of large sample sizes.

4. The semiparametric empirical best prediction. As highlighted before, deriving the EBP of small area parameters and the corresponding MSE approximation as detailed by Jiang (2003) is a nontrivial task. In this section, we develop a computationally convenient alternative that allows us to avoid unverifiable parametric assumption on the random effect distribution. In Section 4.1, we present the proposed approach to derive model parameter estimates within a maximum likelihood framework. In Section 4.2, we detail the proposed semiparametric empirical best predictor (sp-EBP) and the corresponding second-order, bias-corrected MSE estimator.

4.1. *Model parameter estimation.* When dealing with non-Gaussian responses and GLMMs with Gaussian random effects, maximum likelihood (ML) estimators, although optimal, can be time consuming as we need to approximate (possibly multidimensional) integrals that do not admit a closed form expression. An alternative may be based on leaving the distribution of α_i completely unspecified and follow the approach detailed by Aitkin (1996, 1999). The area-specific random effects are treated as nuisance parameters, and a NPML estimate of their distribution is derived. Different contributions to the theory of NPML can be found in the literature [Böhning (1982), Laird (1978), Lindsay (1983a), Simar (1976), Lindsay (1983b)]. Results by Lindsay (1983a, 1983b) show that, as long as the (log-) likelihood function is bounded, it is maximized by a discrete distribution defined on, at most, as many support points as the number of distinct area profiles in the sample. In particular the mixing distribution estimate is a discrete distribution, which puts masses $\pi_g > 0$ on locations $\xi_g = (\xi_{g1}, \dots, \xi_{gq})'$, $g = 1, \dots, G$, where the constraint $\sum_{g=1}^G \pi_g = 1$ holds. In a regression context the number of locations G is bounded from above by the number of different profiles (y_i, \mathbf{X}_i) in the sample. That is, in the presence of categorical covariates the number of locations does not necessarily grow with m .

Let Φ denote the global vector of model parameters, $\Phi = (\beta, \phi, \xi'_1, \dots, \xi'_G, \pi_1, \dots, \pi_{G-1})'$; the observed data likelihood is approximated by

$$(6) \quad L(\Phi) = \prod_{i=1}^m \int_{\mathbb{R}^q} f_{y|\alpha}(y_i | \alpha_i; \mathbf{X}_i) f_{\alpha}(\alpha_i) d\alpha_i \simeq \prod_{i=1}^m \sum_{g=1}^G f_{y|\alpha}(y_i | \xi_g; \mathbf{X}_i) \pi_g,$$

where

$$f_{y|\alpha}(y_i | \xi_g; \mathbf{X}_i) = \prod_{j \in s_i} f_{y|\alpha}(y_{ij} | \alpha_i = \xi_g; \mathbf{x}_{ij})$$

denotes the product of densities in the Exponential Family with canonical parameter θ_{ijg} defined by the (mixed) model

$$\theta_{ijg} = \mathbf{x}'_{ij} \beta + \mathbf{w}'_{ij} \xi_g.$$

As it is clear, expression (6) resembles the likelihood of a finite mixture of distributions, with weights $\pi_g = \Pr(\alpha_i = \xi_g)$. That is, $\alpha_i \sim \sum_{g=1}^G \pi_g \delta(\xi_g)$, where $\delta(a)$ is a one-point distribution putting a unit mass at a . It is worth noticing that, while the discrete nature of the estimate for $f_{\alpha}(\cdot)$ may seem unappealing, most approximation techniques applied when a parametric specification is considered are exactly of the type in equation (6). The only substantial difference is that locations ξ_g and masses π_g in the present proposal are estimated to best fit observed data.

To maximize the likelihood in (6), the EM algorithm [Dempster, Laird and Rubin (1977)] can be employed. A drawback of such an algorithm is that it does not directly provide estimates for the covariance matrix of model parameters. A frequent solution to this issue is based on the use of the Oakes' formula [Oakes

(1999)]; computational details on estimation of model parameters and standard errors are provided in Sections 1 and 2 of the online Supplementary Material [Marino, Ranalli, Salvati and Alfò (2019)], where the EM algorithm is described. A crucial point in the proposed approach is the choice of the number of mixture components G . A simple and frequently used solution is as follows: parameter estimates are computed for varying values of G and the model with the best fit, typically measured by penalized likelihood criteria (such as AIC or BIC), is retained. Typically, the optimal G increases either (i) when the variability of the random effect distribution increases or (ii) when the number of small areas increases as, in this case, this may lead to a higher number of distinct area profiles in the sample. As long as convergence is entailed, the order for the mixing distribution estimate is $O_p(m^{-1/4})$, as compared to $O_p(m^{-1/2})$ for ML parameter estimates in regular models [see Chen (1995)]. However, according to Lindsay and Lesperance (1995), some smooth functionals, such as the empirical Bayes estimates, can be estimated at the usual $O_p(m^{-1/2})$ rate. Furthermore, as shown by Redner and Walker (1984), when the order of the mixture is finite and known (that is, when $\alpha_i \sim \sum_{g=1}^G \pi_g \delta(\xi_g)$ is the true mixing, with G known), the usual ML asymptotics apply.

4.2. *Semiparametric EBP and MSE approximation.* Let us now turn to the main problem of interest, where we have a finite population of size N which can be partitioned into m nonoverlapping domains or small areas. Furthermore, let $\xi = (\xi'_1, \dots, \xi'_G)'$ and $\pi = (\pi_1, \dots, \pi_G)'$ denote the vectors of locations and masses of the finite mixture respectively. We aim at predicting a (possibly) non-linear function of fixed and random effects, $\zeta(\beta, \xi, \pi)$, by exploiting sample data on responses Y_{ij} and populations data on covariates x_{ij} . Under the proposed approach, the semiparametric best predictor of ζ is defined according to the following expression:

$$(7) \quad \tilde{\zeta}^{\text{sp-BP}} = E_{\alpha|y}[\zeta(\beta, \xi, \pi) | y] = \sum_{g_1 \dots g_m} \zeta(\beta, \xi_{g_1, \dots, g_m}, \pi) \prod_{i=1}^m \tau_{ig_i},$$

where $\sum_{g_1 \dots g_m}$ is a shorthand for $\sum_{g_1=1}^G \dots \sum_{g_m=1}^G$, $\xi_{g_1, \dots, g_m} = (\xi'_{g_1}, \dots, \xi'_{g_m})'$ and τ_{ig_i} denotes the posterior probability for the i th small area to belong to the g_i th component of the finite mixture. In particular, denoting by z_{ig_i} , $i = 1, \dots, m$, $g_i = 1, \dots, G$, the component membership indicator for the i th small area, τ_{ig_i} is defined by

$$(8) \quad \tau_{ig_i} = \Pr(z_{ig_i} = 1 | y_i) = \frac{\pi_{g_i} f_{y|\alpha}(y_i | \xi_{g_i}; \mathbf{X}_i)}{\sum_{l=1}^G \pi_l f_{y|\alpha}(y_i | \xi_l; \mathbf{X}_i)},$$

where, as before, $f_{y|\alpha}(y_i | \xi_{g_i}; \mathbf{X}_i)$ refers to sample data only. As it is clear, expression (7) denotes the expected value of $\zeta(\beta, \xi, \pi)$, with respect to the posterior distribution of the random effects α . Since this is a discrete distribution, the integral approximation which is required in equation (1) directly translates into simpler summations.

An estimate of $\tilde{\zeta}^{\text{sp-BP}}$ can be obtained by replacing model parameters β, ξ and π by consistent estimates. Here, we consider the estimates derived by the EM algorithm described in Section 1 of the online Supplementary Material. In the following, we will refer to such a quantity as the semiparametric empirical best predictor (sp-EBP) of ζ , denoted by $\hat{\zeta}^{\text{sp-EBP}} = \tilde{\zeta}^{\text{sp-BP}}(\hat{\beta}, \hat{\xi}, \hat{\pi})$.

To evaluate the quality of predictions, we develop an analytic approximation to the MSE of $\hat{\zeta}^{\text{sp-EBP}}$ based on the approach detailed by Jiang (2003) but considering a maximum likelihood estimator. Starting from equation (3), the MSE of the sp-EBP is given by

$$(9) \quad \text{MSE}(\hat{\zeta}^{\text{sp-EBP}}) = \frac{1}{m} e^{\text{sp}}(\Phi) + d^{\text{sp}}(\Phi) + o_p(1/m),$$

where the former term, $e^{\text{sp}}(\Phi)$, is defined according to expression (4) and can be derived by computing model derivatives with respect to β, α and π , together with the covariance matrix of model parameter estimates, $V(\hat{\Phi})$. See Section 2 in the online Supplementary Material for computational details. On the other hand, $d^{\text{sp}}(\Phi)$ can be derived as follows:

$$\begin{aligned} d^{\text{sp}}(\Phi) &= E_{\alpha}[(\zeta)^2] - E_y[(\tilde{\zeta}^{\text{sp-BP}})^2] \\ &= \sum_{g_1 \cdots g_m} \zeta(\beta, \xi_{g_1, \dots, g_m}, \pi)^2 \prod_{i=1}^m \pi_{g_i} \\ &\quad - E_y \left[\left(\sum_{g_1 \cdots g_m} \zeta(\beta, \xi_{g_1, \dots, g_m}, \pi) \prod_{i=1}^m \tau_{i g_i} \right)^2 \right]. \end{aligned}$$

The computational burden to obtain the above quantities is substantially lower than that required for the approach by Jiang (2003). Intractable integrals appearing in equations (4) and (5) all translate into simple summations which can be solved analytically.

An estimator of $\text{MSE}(\hat{\zeta}^{\text{sp-EBP}})$ is obtained by replacing Φ in (9) by a consistent estimator such as that obtained by maximizing the observed data likelihood in equation (6). That is,

$$(10) \quad \widehat{\text{MSE}}(\hat{\zeta}^{\text{sp-EBP}}) = \frac{1}{m} e^{\text{sp}}(\hat{\Phi}) + d^{\text{sp}}(\hat{\Phi}).$$

However, as we remarked before, this approach does not directly lead to an unbiased estimator of $\text{MSE}(\hat{\zeta}^{\text{sp-EBP}})$. When replacing $\hat{\Phi}$ in $d^{\text{sp}}(\Phi)$, we get an error of order $O_p(m^{-1/2})$ and a bias correction term needs to be considered. Jiang (2003) provided an explicit expression for such a term when model parameters are estimated by the method of moments. Clearly, under the current approach, these results do not directly hold but, rather, need to be adapted.

Let Φ_0 denote the “true” vector of model parameters, and let us consider a second-order Taylor expansion of $d^{SP}(\Phi)$ around Φ_0 evaluated at $\hat{\Phi}$:

$$\begin{aligned}
 d^{SP}(\hat{\Phi}) &= d^{SP}(\Phi_0) + \left(\frac{\partial d^{SP}}{\partial \Phi} \right)' \Big|_{\Phi_0} (\hat{\Phi} - \Phi_0) \\
 &+ \frac{1}{2} (\hat{\Phi} - \Phi_0)' \left(\frac{\partial^2 d^{SP}}{\partial \Phi \partial \Phi'} \right) \Big|_{\Phi_0} (\hat{\Phi} - \Phi_0) + o_p(m^{-1}),
 \end{aligned}
 \tag{11}$$

where d^{SP} is a shorthand for $d^{SP}(\Phi)$. From expression (11), it is easy to see that

$$E[d^{SP}(\hat{\Phi})] = d^{SP}(\Phi_0) + \frac{1}{m} b^{SP}(\hat{\Phi}) + o_p(m^{-1}),$$

where $b^{SP}(\hat{\Phi})$ denotes a bias correction defined as

$$\begin{aligned}
 b^{SP}(\hat{\Phi}) &= \left(\frac{\partial d^{SP}}{\partial \Phi} \right)' \Big|_{\Phi_0} m E(\hat{\Phi} - \Phi_0) \\
 &+ \frac{m}{2} E \left[(\hat{\Phi} - \Phi_0)' \left(\frac{\partial^2 d^{SP}}{\partial \Phi \partial \Phi'} \right) \Big|_{\Phi_0} (\hat{\Phi} - \Phi_0) \right] \\
 &= b_1^{SP}(\hat{\Phi}) + b_2^{SP}(\hat{\Phi}).
 \end{aligned}
 \tag{12}$$

As shown in Section 3 of the online Supplementary Material, denoting by K the total number of free model parameters in the model, the former term on the right-hand side of equation (12) is given by

$$b_1^{SP}(\hat{\Phi}) = - \left(\frac{\partial d^{SP}}{\partial \Phi} \right)' \Big|_{\Phi_0} \frac{m}{2} I_e(\Phi_0)^{-1} \text{tr} \left[I_e(\Phi_0)^{-1} \left(\frac{\partial I_e^k(\Phi)}{\partial \Phi'} \right) \Big|_{\Phi_0} \right]_{k=1}^K.$$

Here, $I_e^k(\Phi)$ denotes the k th row of the expected information matrix $I_e(\Phi)$ and $k = 1, \dots, K$, is used to indicate that trace values need to be stacked underneath each other to derive $b_1^{SP}(\hat{\Phi})$. On the other hand, it can be shown that the term $b_2^{SP}(\hat{\Phi})$ in equation (12) can be computed as

$$b_2^{SP}(\hat{\Phi}) = \frac{m}{2} \text{tr} \left\{ \left(\frac{\partial^2 d^{SP}}{\partial \Phi \partial \Phi'} \right) \Big|_{\Phi_0} V(\hat{\Phi}) \right\},$$

where $V(\hat{\Phi})$ denotes the covariance matrix of parameter estimates, as detailed in Section 2 of the online Supplementary Material.

A second-order, bias-corrected estimator of $\text{MSE}(\hat{\xi}^{SP-EBP})$ is then given by

$$\widehat{\text{MSE}}^*(\hat{\xi}^{SP-EBP}) = d^{SP}(\hat{\Phi}) + \frac{1}{m} [e^{SP}(\hat{\Phi}) - b^{SP}(\hat{\Phi})].
 \tag{13}$$

We report the computational details required to derive $b_1^{SP}(\hat{\Phi})$ and $b_2^{SP}(\hat{\Phi})$ in Section 3 of the online Supplementary Material.

5. A special case: Binary data. In this section we focus on the relevant case of binary responses modeled via a mixed logistic model with random intercepts. Let Y_{ij} denote the binary response associated to unit j in the i th small area ($i = 1, \dots, m, j = 1, \dots, N_i$), and let α_i denote an area-specific random effect. Again, let \mathbf{x}_{ij} denote a p -dimensional vector of covariates, and \mathbf{X}_i the matrix of covariates for the i th small area. We assume that, conditional on α_i , responses for units in the i th small area are independent Bernoulli random variables with success probability p_{ij} described by the following mixed logistic model:

$$(14) \quad \theta_{ij} = \log \frac{p_{ij}}{1 - p_{ij}} = \alpha_i + \mathbf{x}'_{ij}\boldsymbol{\beta}.$$

In the equation above $\boldsymbol{\beta}$ is a p -dimensional vector of fixed model parameters that describes the effect of observed covariates on the logit transform of p_{ij} . We consider the practical problem of predicting small area proportions

$$\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij},$$

using the GLMM in equation (14). To this end, we use the EBP for the quantity

$$(15) \quad p_i = \frac{1}{N_i} \sum_{j=1}^{N_i} p_{ij}.$$

In fact, since N_i is usually very large in most applications, the EBP for p_i can also be used to predict the indicator \bar{Y}_i .

Let us assume that responses Y_{ij} are observed for sampled units only ($i = 1, \dots, m, j \in s_i$), while covariates \mathbf{x}_{ij} are available at the population level ($i = 1, \dots, m, j = 1, \dots, N_i$). Following the approach detailed in the previous sections, we leave the distribution of the area-specific random effects in equation (14) unspecified and approximate it via a discrete distribution that puts masses $\pi_g > 0$ on locations ξ_1, \dots, ξ_G , with $\sum_{g=1}^G \pi_g = 1$. By adopting a canonical link function, the logistic transform of the success probability for a generic area i in the g th component of the finite mixture is given by

$$\theta_{ijg} = \log \frac{p_{ijg}}{1 - p_{ijg}} = \xi_g + \mathbf{x}'_{ij}\boldsymbol{\beta}.$$

Using the standard notation for the Exponential Family, the joint conditional density for the observed responses in the i th small area and the g th component is

$$f_{ig} = f_{y|\alpha}(\mathbf{y}_i | \xi_g; \mathbf{X}_i) = \exp \left\{ \sum_{j \in s_i} [y_{ij}\theta_{ijg} - \log(1 + e^{\theta_{ijg}})] \right\}.$$

Turning back to the problem of estimating p_i in equation (15), the corresponding sp-BP is given by

$$\begin{aligned} \tilde{p}_i^{\text{sp-BP}} &= \sum_{g=1}^G p_{ig} \frac{\exp[\sum_{j \in s_i} y_{ij} \theta_{ijg} - \sum_{j \in s_i} \log(1 + e^{\theta_{ijg}})] \pi_g}{\sum_{l=1}^G \exp[\sum_{j \in s_i} y_{ij} \theta_{ijl} - \sum_{j \in s_i} \log(1 + e^{\theta_{ijl}})] \pi_l} \\ &= \sum_{g=1}^G p_{ig} \frac{\exp[\alpha_g y_{i \cdot} - \sum_{j \in s_i} \log(1 + e^{\theta_{ijg}})] \pi_g}{\sum_{l=1}^G \exp[\alpha_l y_{i \cdot} - \sum_{j \in s_i} \log(1 + e^{\theta_{ijl}})] \pi_l}, \end{aligned}$$

where $y_{i \cdot} = \sum_{j \in s_i} y_{ij}$ and $p_{ig} = N_i^{-1} \sum_{j=1}^{N_i} p_{ijg}$. By letting

$$\tau_{ig(y_{i \cdot})} = \frac{\exp[\alpha_g y_{i \cdot} - \sum_{j \in s_i} \log(1 + e^{\theta_{ijg}})] \pi_g}{\sum_{l=1}^G \exp[\alpha_l y_{i \cdot} - \sum_{j \in s_i} \log(1 + e^{\theta_{ijl}})] \pi_l},$$

the sp-BP of p_i is given by

$$(16) \quad \hat{p}_i^{\text{sp-BP}} = \sum_{g=1}^G p_{ig} \tau_{ig(y_{i \cdot})}.$$

The corresponding sp-EBP, denoted by $\hat{p}_i^{\text{sp-EBP}}$, is obtained by substituting ML estimates of model parameters into expression (16):

$$(17) \quad \hat{p}_i^{\text{sp-EBP}} = \sum_{g=1}^G \hat{p}_{ig} \hat{\tau}_{ig(y_{i \cdot})},$$

while the quality of predictions obtained via $\hat{p}_i^{\text{sp-EBP}}$ can be evaluated through the following MSE expression:

$$(18) \quad \text{MSE}(\hat{p}_i^{\text{sp-EBP}}) = E_\alpha[(p_i)^2] - E_y[(\tilde{p}_i^{\text{sp-BP}})^2] + E_\alpha[(\hat{p}_i^{\text{sp-EBP}} - \tilde{p}_i^{\text{sp-BP}})^2],$$

where

$$E_\alpha[(p_i)^2] = \sum_{g=1}^G p_{ig}^2 \pi_g$$

and

$$E_y[(\tilde{p}_i^{\text{sp-BP}})^2] = \sum_{h=0}^{n_i} (\tilde{p}_{i(h)}^{\text{sp-BP}})^2 \Pr(Y_{i \cdot} = h; \mathbf{X}_i).$$

Here, $\tilde{p}_{i(h)}^{\text{sp-BP}}$ denotes the sp-BP of p_i conditional on $y_{i \cdot} = h$, that is,

$$\tilde{p}_{i(h)}^{\text{sp-BP}} = \sum_{g=1}^G p_{ig} \left[\frac{\exp[\xi_g h - \sum_{j \in s_i} \log(1 + e^{\theta_{ijg}})] \pi_g}{\sum_{l=1}^G \exp[\xi_l h - \sum_{j \in s_i} \log(1 + e^{\theta_{ijl}})] \pi_l} \right] = \sum_{g=1}^G p_{ig} \tau_{ig(h)}.$$

The term $\Pr(Y_i. = h; \mathbf{X}_i)$ is obtained as

$$\Pr(Y_i. = h; \mathbf{X}_i) = \sum_{g=1}^G \Pr(Y_i. = h \mid \xi_g; \mathbf{X}_i) \pi_g,$$

where $\Pr(Y_i. = h \mid \xi_g; \mathbf{X}_i)$ represents the probability of observing h successes in n_i independent but nonidentically distributed Bernoulli trials. This quantity can be obtained using the probability mass function of a Poisson-Binomial random variable [see [Chen and Liu \(1997\)](#)] with parameter $(p_{i1g}, \dots, p_{in_i g})'$. The last term in equation (18) is obtained as

$$\begin{aligned} & E_{\alpha}[(\hat{p}_i^{\text{sp-EBP}} - \tilde{p}_i^{\text{sp-BP}})^2] \\ (19) \quad & = \sum_{h=0}^{n_i} \left[\left(\frac{\partial \tilde{p}_{i(h)}^{\text{sp-BP}}}{\partial \Phi} \right)' m V(\hat{\Phi}) \left(\frac{\partial \tilde{p}_{i(h)}^{\text{sp-BP}}}{\partial \Phi} \right) \right] \Pr(Y_i. = h; \mathbf{X}_i), \end{aligned}$$

where, as before, $V(\hat{\Phi})$ is the covariance matrix of model parameter estimates and $\partial \tilde{p}_{i(h)}^{\text{sp-BP}} / \partial \Phi$ is the vector of model derivatives conditional on $y_i. = h$. Explicit formulas for these latter quantities are provided in Section 4 of the online Supplementary Material.

The second-order, bias-corrected estimator of $\text{MSE}(\hat{p}_i^{\text{sp-EBP}})$, denoted by $\widehat{\text{MSE}}^*(\hat{p}_i^{\text{sp-EBP}})$, is obtained according to expression (13) after adapting the bias correction term to the binary case.

6. Model-based simulation study. In this section, we evaluate the empirical properties of the proposed approach via a large scale (model-based) simulation study. This consists of $T = 1,000$ samples, where binary population data were generated under some model assumptions and sample data were selected from the simulated population. In particular, population data entailed $m = 100, 200, 500$ small areas; samples were selected by simple random sampling without replacement within each area. The population and the sample sizes were constant across areas and are fixed to $N_i = 100$ and $n_i = 10$, respectively. According to the simulation study discussed by [González-Manteiga et al. \(2007\)](#), for each unit j in small area i we generated the target variable $Y_{ij}, i = 1, \dots, m, j = 1, \dots, N_i$, from a Bernoulli distribution with success probability defined by

$$(20) \quad p_{ij} = \frac{\exp(\alpha_i + x_{ij} \beta)}{1 + \exp(\alpha_i + x_{ij} \beta)},$$

with $\beta = 1, x_{ij} \sim \text{Unif}(-1, b_i)$, and $b_i = i/8, i/16, i/48$ for $m = 100, 200$ and 500 , respectively. To evaluate the impact of parametric assumptions on the distribution of the area-specific random effects, we considered two different scenarios. The first one (Scenario 1) uses area-specific random effects from a zero mean,

Gaussian distribution with standard deviation equal to $\sigma_1 = 0.5$. The second scenario (Scenario 2) involves area-specific random effects generated from a mixture of Gaussian distributions, $\alpha_i \sim \nu N(\mu_1, \sigma_2) + (1 - \nu)N(\mu_2, \sigma_2)$, where ν represents a random draw from a Bernoulli distribution with $\Pr(\nu = 1) = 0.7$, $\mu_1 = 0$, $\mu_2 = 3$ and $\sigma_2 = 0.05$. Based on this latter quantity, it is evident that, under this scenario, the random effect distribution closely resembles that of a discrete distribution putting masses ν and $1 - \nu$ on locations μ_1 and μ_2 . In this framework, the population is made by two separate sets of small areas having different *baseline* levels for the success probabilities. This may be reasonable, for example, for properly representing nonhomogeneous unemployment rates typically observed in the north/south of Italy, as we see in Section 7. Clearly, the chosen scenarios represent two extreme situations; we expect that, in real applications, the random effect distribution lies in between them.

In this simulation study, our aim is that of evaluating the empirical behavior of the proposed approach. For each simulated sample, we estimated model parameters for a varying number of mixture components ($G = 2, \dots, 5$) and selected the optimal G according to the AIC index. We report in Table 3 the distribution of the optimal number of mixture components G across simulations. As it can be observed, in most of the cases the AIC index leads to selecting a model with $G = 2$ components only. This reflects the reduced variability of the random effect distribution considered under both simulation scenarios. However, it is worth to highlight that, for higher sample sizes, the chance of selecting a higher G slightly increases, especially when α_i is a random draw from a Gaussian density. This result is clearly related to the requirement of a higher number of components to properly approximate the “true” continuous distribution of the area-specific effects.

Starting from parameter estimates derived from the proposed approach, the sp-EBP for small area proportions was derived according to equation (17). The proposed predictor was then compared to the parametric EBP by Jiang and Lahiri (2001) and to the naive predictor considered in González-Manteiga et al. (2007), where both are based on the assumption of Gaussian random effects. For the EBP parameter estimates were derived via a ML approach based on a Laplace approximation available in the `glmer` function from the R `lme4` package [Bates et al. (2015)]. Given the estimates, small area proportions and corresponding MSEs were derived by adopting the formulas detailed in Section 3. To evaluate the in-

TABLE 3
Distribution of the optimal number of mixture components across simulations

m / k	Scenario 1				Scenario 2			
	2	3	4	5	2	3	4	5
100	0.988	0.012	0.000	0.000	0.968	0.032	0.000	0.000
200	0.962	0.038	0.000	0.000	0.950	0.049	0.001	0.000
500	0.866	0.133	0.001	0.000	0.943	0.056	0.001	0.000

tractable integrals, we followed the approach suggested by Boubeta, Lombardía and Morales (2016). That is, we started by generating $B = 2,500$ replicates of the area-specific random effects, say $\alpha_i^{(b)}$, from a Gaussian density with zero mean and variance equal to the corresponding ML estimate. Then, we considered their antithetic transform $\alpha_i^{(B+b)} = -\alpha_i^{(b)}$ to obtain $2B$ random effect values. Finally, integrals were approximated by the corresponding empirical means. In the following, we denote EBP estimates of small area proportions by \hat{p}_i^{EBP} . Although this approach is optimal, the computational complexity greatly limits its applicability. Via the current simulation study we aim at understanding whether the sp-EBP approach we propose could represent an effective alternative which is optimal in terms of minimum MSE and simpler from a computational point of view.

For completeness, we also included in the simulation study results from the naive approach. In this case, parameter estimates were obtained using a PQL approach available in the `gllmmPQL` function from the R MASS package [Venables and Ripley (1994)]. To get predictions, parameter estimates were directly plugged into the expression for the area-specific proportions:

$$(21) \quad \hat{p}_i^{\text{Naive}} = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\exp(\hat{\alpha}_i + x_{ij}\hat{\beta})}{1 + \exp(\hat{\alpha}_i + x_{ij}\hat{\beta})}.$$

The performance of the small area estimators were evaluated by computing, for each area $i = 1, \dots, m$, the bias and the root mean squared error (RMSE), defined as follows:

$$\text{BIAS}_i = T^{-1} \sum_{t=1}^T (\hat{p}_{it}^{\text{Model}} - p_{it}),$$

$$\text{RMSE}_i = \sqrt{T^{-1} \sum_{t=1}^T (\hat{p}_{it}^{\text{Model}} - p_{it})^2},$$

where $\hat{p}_{it}^{\text{Model}}$ denotes the model-based proportion estimate for the i th small area in the t th simulated sample obtained via either the EBP (\hat{p}_i^{EBP}), the sp-EBP ($\hat{p}_i^{\text{sp-EBP}}$), or the naive (\hat{p}_i^{Naive}) approach, $i = 1, \dots, m$. For completeness, we also report in Section 5 of the online Supplementary Material the distribution of the mean absolute error (MAE) across small areas for the EBP, the sp-EBP and the naive predictor under the two experimental scenarios. Together with the bias, MAE is frequently used to evaluate the quality of results in the small area literature, even though it may not be a consistent tool to evaluate predictions obtained by posterior means [Gneiting (2011)]. Figures 2 and 3 show the BIAS and the RMSE distribution across small areas for the three estimators under investigation for Scenario 1 and $m = 100, 200, 500$, respectively; the dotted red line denotes the corresponding mean values. As expected, when looking at the first two panels (i.e., $m = 100, 200$), the sp-EBP performs better than the naive estimator and slightly worse than the EBP with a gap that reduces as m increases both in terms of BIAS

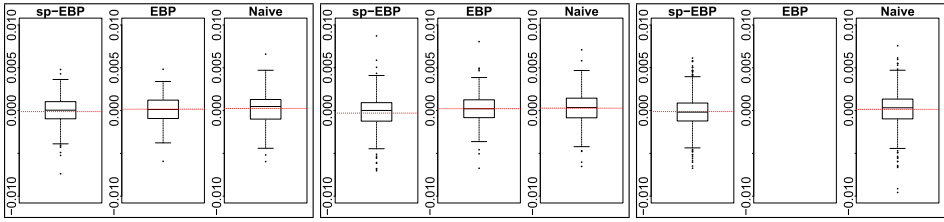


FIG. 2. Scenario 1: distribution of the BIAS over areas for \hat{p}_i^{sp-EBP} , \hat{p}_i^{EBP} and \hat{p}_i^{Naive} , for $m = 100$ (left panel), $m = 200$ (central panel) and $m = 500$ (right panel).

and RMSE. When $m = 500$, performance values of EBP are not shown due to the computational burden required to get the estimates. For one replication we needed 161.612 minutes on an Intel(R) I5-3330 architecture – 3.0 GHz, and, therefore, we couldn't obtain results for $T = 1,000$ replications in a reasonable amount of time.

Figures 4 and 5 show the performance of the estimators under Scenario 2. As before, results for the EBP approach for $m = 500$ are not shown due to computational issues. As it is clear by looking at these plots, when the assumption of Gaussian random effects does not hold, parametric approaches produce predictions with a reduced quality than those obtained via the semiparametric alternative we propose. In particular, we may notice that \hat{p}_i^{sp-EBP} clearly outperforms the two competitors in terms of both bias and RMSE. Also, results for \hat{p}_i^{Naive} and \hat{p}_i^{EBP} slightly worsen as m increases. This may be possibly due to the higher information available and the stronger impact of the random effect distribution on the overall response variability when the number of small areas increases.

A further purpose of this simulation study is to investigate the performance of the MSE estimators to evaluate the accuracy of the predictions we discussed so far. In particular, for \hat{p}_i^{sp-EBP} , we considered the proposed MSE estimator reported in equations (10) and (13); we refer to the square root of these quantities as $\widehat{RMSE}(\hat{p}_i^{sp-EBP})$ and $\widehat{RMSE}^*(\hat{p}_i^{sp-EBP})$, respectively. For the estimator \hat{p}_i^{EBP} , we used the approximate MSE estimator proposed by Hobza and Morales (2016); the

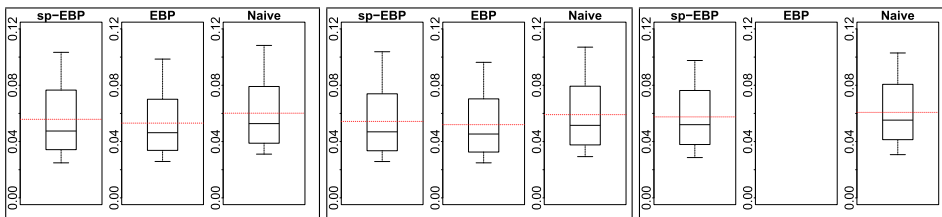


FIG. 3. Scenario 1: distribution of the RMSE over areas for \hat{p}_i^{sp-EBP} , \hat{p}_i^{EBP} and \hat{p}_i^{Naive} , for $m = 100$ (left panel), $m = 200$ (central panel) and $m = 500$ (right panel).

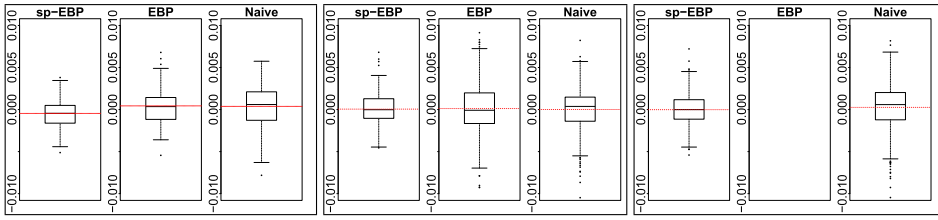


FIG. 4. Scenario 2: distribution of the BIAS over areas for \hat{p}_i^{Naive} , \hat{p}_i^{EBP} and $\hat{p}_i^{\text{sp-EBP}}$ for $m = 100$ (left panel), $m = 200$ (central panel) and $m = 500$ (right panel).

corresponding square root is denoted by $\widehat{\text{RMSE}}(\hat{p}_i^{\text{EBP}})$. Last, for the naive predictor \hat{p}_i^{Naive} , we considered the approach suggested by González-Manteiga et al. (2007), based on linearizing the GLMM in equation (20) and, then, applying the Prasad-Rao MSE approximation for the corresponding linear mixed model; the square root of such an estimator is denoted by $\widehat{\text{RMSE}}(\hat{p}_i^{\text{Naive}})$.

The performance of the RMSE estimators were evaluated by considering the ratio (R) between the estimated RMSE for the model-based estimates and the corresponding actual RMSE for each small area prediction, that is,

$$R_i = \frac{\sum_{t=1}^T \widehat{\text{RMSE}}(\hat{p}_{it}^{\text{Model}})}{\sqrt{\sum_{t=1}^T (\hat{p}_{it}^{\text{Model}} - p_{it})^2}},$$

for $i = 1, \dots, m$. The distribution over areas for such a ratio for varying m and varying random effect distributions is reported in Figure 6. Under Scenario 1 $\widehat{\text{RMSE}}(\hat{p}_i^{\text{sp-EBP}})$ and $\widehat{\text{RMSE}}^*(\hat{p}_i^{\text{sp-EBP}})$ perform generally better than the alternatives. In particular, simulation results suggest that the former estimator is more appropriate when a reduced number of small areas is available ($m = 100, 200$), while its precision decreases in case of larger m . On the other hand, $\widehat{\text{RMSE}}^*(\hat{p}_i^{\text{sp-EBP}})$ shows slight overestimation of the actual Monte Carlo RMSE for $m = 100$ and $m = 200$, but it has to be preferred in the presence of a large number of small areas since the R_i index is strongly concentrated around 1.

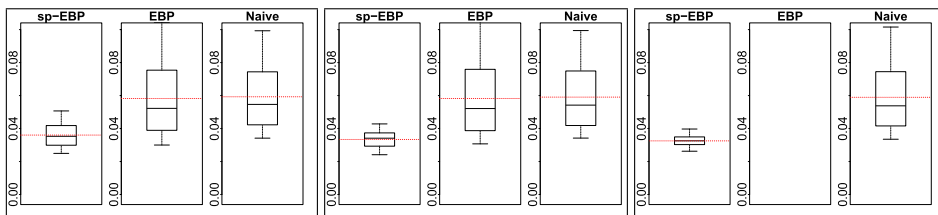


FIG. 5. Scenario 2: distribution of the RMSE over areas for \hat{p}_i^{Naive} , \hat{p}_i^{EBP} and $\hat{p}_i^{\text{sp-EBP}}$ for $m = 100$ (left panel), $m = 200$ (central panel) and $m = 500$ (right panel).

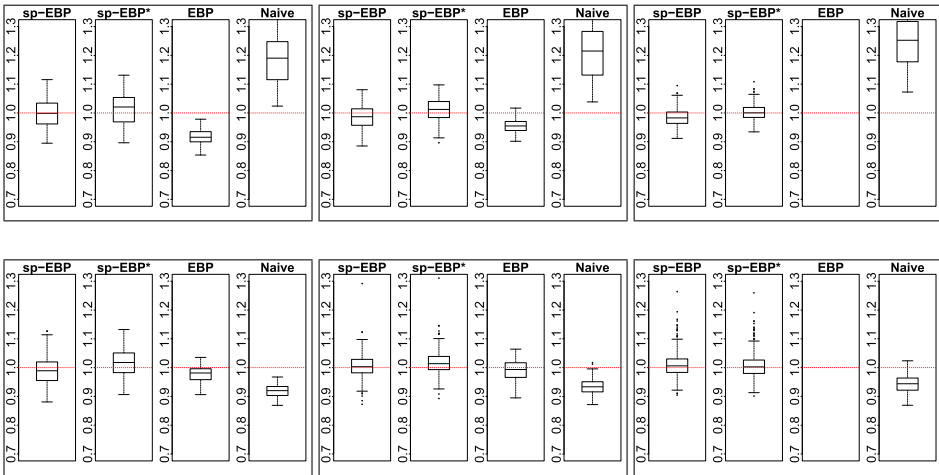


FIG. 6. Distribution of the RMSE ratio over areas for the *sp-EBP* (without bias correction), the *sp-EBP** (with bias correction), the *EBP* and the *Naive* approach, for $m = 100$ (left panel), $m = 200$ (central panel) and $m = 500$ (right panel), under Scenario 1 (upper panel) and Scenario 2 (lower panel).

The estimator $\widehat{\text{RMSE}}(\hat{\rho}_i^{\text{EBP}})$ underestimates the actual Monte Carlo RMSE, with a ratio which is always lower than 1 for $m = 100$. The quality of the results improves with m , even though it is always lower than that provided by the proposed approach. Such a finding may be possibly due to the estimation of the covariance matrix for parameter estimates which is not as accurate as expected with $B^* = 250$ bootstrap resamples. In fact, it is worth noticing that Boubeta, Lombardía and Morales (2016) highlighted the need of a very accurate estimate of the covariance matrix of parameter estimates to ensure high quality of the results. For this reason in their simulation study, the authors suggested to estimate $V(\hat{\Phi})$ by running a Monte Carlo experiment based on 10^4 iterations in advance. In practice, when dealing with large sample sizes, such an approach is computationally very expensive and this is the reason why we considered a bootstrap approach based on $B^* = 250$ iterations only. Last, the estimator suggested by González-Manteiga et al. (2007) overestimates the actual RMSE in all the scenarios we considered in this simulation study.

By looking at the bottom panel in Figure 6, we observe that, when dealing with non-Gaussian random effects, the MSE estimator of the *sp-EBP* has again good performances with an average ratio close to 1 for all values of m . The effect of the bias correction term is less evident than before. When a reduced number of small areas is available, $\widehat{\text{RMSE}}(\hat{\rho}_i^{\text{sp-EBP}})$ allows us to estimate the actual RMSE with a higher precision than the corresponding bias-corrected version $\widehat{\text{RMSE}}^*(\hat{\rho}_i^{\text{sp-EBP}})$. However, when $m = 500$, the two estimators seem to perform similarly. The above results are not that surprising from our perspective. The bias correction term

strongly relies on asymptotic results from ML theory. As a consequence, the quality of the approximation and, in turn of the results, improves only when dealing with large sample sizes that render asymptotics more reliable.

Considering that, in real applications, we expect the random effect distribution to lie in between the two “extreme” settings we considered in this simulation study and, also, that we generally need to deal with a large number of small areas, using $\widehat{\text{RMSE}}^*(\hat{p}_i^{\text{sp-EBP}})$ seems to be generally more appropriate. From Figure 6 we may also notice that, under Scenario 2, the MSE estimator of the EBP works quite well (apart from being computationally prohibitive from large m), whereas that for the naive estimator consistently underestimates the actual RMSE.

Table 4 shows the mean coverage rate (CR) for nominal 95% Wald-type confidence intervals over simulations, that is,

$$\text{CR}_i = T^{-1} \sum_{t=1}^T \mathbb{1}(|\hat{p}_{it} - p_{it}| \leq 1.96 \times \widehat{\text{RMSE}}(p_{it}^{\text{Model}})),$$

for $i = 1, \dots, m$. As it is clear from this table, the proposed estimators show a good performance with an average empirical coverage of approximately 92–94% in all cases, except for $m = 100$ under Scenario 1. On the other hand both the EBP and, particularly, the naive approach show a more erratic behavior. The former approach leads to under coverage for Scenario 1 and to over coverage for Scenario 2. This behavior is reversed for the naive estimator.

To conclude, we also compared MSE estimators in terms of computational complexity. The last columns of Tables 4 report the computational time (averaged

TABLE 4
Average coverage rate over areas and computational time (in minutes) of $\widehat{\text{RMSE}}(\hat{p}_i^{\text{Naive}})$, $\widehat{\text{RMSE}}(\hat{p}_i^{\text{EBP}})$, $\widehat{\text{RMSE}}(\hat{p}_i^{\text{sp-EBP}})$ and $\widehat{\text{RMSE}}^*(\hat{p}_i^{\text{sp-EBP}})$, for $m = 100, 200, 500$

m	Coverage			Computational Time		
	100	200	500	100	200	500
Scenario 1						
sp-EBP	0.888	0.920	0.940	0.059	0.122	0.332
sp-EBP*	0.890	0.923	0.944	0.528	1.042	2.986
EBP	0.864	0.912		16.907	41.385	*161.609
Naive	0.962	0.967	0.976	0.005	0.018	0.206
Scenario 2						
sp-EBP	0.928	0.931	0.927	0.053	0.110	0.303
sp-EBP*	0.933	0.933	0.928	0.481	0.959	2.437
EBP	0.966	0.974		17.229	42.112	*162.528
Naive	0.903	0.906	0.910	0.004	0.018	0.219

*Results refer to a single Monte Carlo draw.

over simulations) required to get the estimates on an Intel(R) I5-3330 architecture (3.0 GHz) under each simulation setting. As it can be seen, the proposed MSE estimators show good performance also in this respect. When compared to the naive approach, they clearly require a higher effort which is, however, always under control. When compared to the EBP approach, the computational burden is considerably reduced. It is important to notice that, due to computational issues, results reported for the EBP approach when $m = 500$ refer to a single Monte Carlo draw in place of being the average of $T = 1,000$ draws as for the other methods. In this respect, it is clear that this approach does not represent an option for empirical applications with large m , as the one we discuss here.

When comparing the two MSE estimators we propose (with and without bias correction), we may observe that deriving $\widehat{\text{RMSE}}^*(\hat{\rho}_i^{\text{sp-EBP}})$ requires a higher computational effort than that required for $\widehat{\text{RMSE}}(\hat{\rho}_i^{\text{sp-EBP}})$. This is clearly due to the computation of model derivatives in equation (12) which does not represent an easy task. However, such an effort is rewarded by the quality improvements we discussed so far, at least for large m .

7. Estimating unemployment incidence for LLMA in Italy. In this section, we use ILFS data to estimate unemployment incidence for the 611 LLMA in Italy. According to the simulation results in Section 6, the sp-EBP is a potentially useful approach as: (i) it performs better than the naive predictor in terms of bias and efficiency; and (ii) it dramatically decreases the computational complexity of the MSE estimator for the parametric EBP which becomes unfeasible for a large number of small areas and/or large sample sizes. The use of the proposed approach is made easy by the availability of a (computationally efficient) algorithm for estimation and inference developed in R language from the authors. This is part of the online Supplementary Material at the publisher's website together with an example data set similar to the real one.

7.1. The model. To predict unemployment incidence in Italy, we considered a response variable Y_{ij} taking value 1 if unit j in small area i is unemployed and 0 otherwise. We followed an approach similar to that used by Molina, Nandram and Rao (2014) and considered the variables introduced in Section 2 and their transformations in the linear predictor, that is, *gender-age* (reference = M:15–24), *educational level* (reference = no education or primary school diploma) and the logarithmic transform of *U-count*. We ran the EM algorithm described in the Supplementary Material (Section 1) for different model specifications and a varying number of components ($G = 2, \dots, 6$) for the random effect distribution. The optimal solution, corresponding to the smallest AIC value, is based on $G = 3$ components and includes in the linear predictor a random intercept and main covariate effects only. We report in Table 5 model parameter estimates, standard errors, and resulting p-values together with the corresponding log-likelihood and AIC index.

TABLE 5

Parameter estimates, standard errors and corresponding *p*-values for the mixed logistic model fitted to the ILFS data based on an unspecific (left) and a Gaussian (right) random effect distribution

	Unspecific			Gaussian		
	Estimate	SE	<i>p</i> -value	Estimate	SE	<i>p</i> -value
Intercept	-3.052	0.176	<0.001	-3.002	0.150	<0.001
M:25-34	0.118	0.057	0.038	0.122	0.054	0.024
M:35-65	-0.787	0.054	<0.001	-0.778	0.048	<0.001
F:15-24	-0.222	0.062	<0.001	-0.221	0.058	<0.001
F:25-34	0.113	0.05	0.023	0.118	0.054	0.029
F:35-65	-1.039	0.052	<0.001	-1.027	0.050	<0.001
Middle School	0.206	0.059	<0.001	0.211	0.054	<0.001
High School	0.235	0.062	<0.001	0.239	0.053	<0.001
University Degree or Beyond	-0.032	0.083	0.682	-0.029	0.064	0.653
log(U-count)	0.113	0.012	<0.001	0.104	0.022	<0.001
ℓ		-21,833.599			-21,837.357	
AIC		43,695.199			43,696.714	

For comparison, we also report such quantities for the corresponding parametric model based on Gaussian random effects. Looking at this table, we may first observe that the AIC index suggests a better fit of the model based on an unspecified random effect distribution with respect to its parametric counterpart, even though differences in terms of parameter estimates are rather negligible. For this reason in the following, we focus on results obtained by means of the semiparametric approach only. In particular, looking at the estimates for *gender-age*, we may notice that, when controlling for the effect of other explanatory variables in the model and for the effect of unobserved heterogeneity, the odds of being unemployed for younger people is higher than that for the older ones. For instance, the odds of being unemployed for a male in the 25-34 group are $e^{0.118} = 1.125$ times higher than those of males aged 15-24. On the other hand, for a male who is aged 35-65 years, the odds are 54.5% lower than those for the baseline category. Such differences are even stronger for females. Turning to the *educational level*, the odds of being unemployed for subjects with middle or high school diplomas are higher than those of low educated subjects (parameter estimates for middle and high school diplomas are all positive). On the other hand, having a university degree or higher education has not a significant effect. These findings are in line with the results reported in the preliminary analysis; low educated females and relatively younger individuals (the reference category) are more frequent in the inactive category. Last, as expected, results reported in Table 5 suggest that the probability of being unemployed increases as the total number of unemployed registered at the 2011 Census increases.

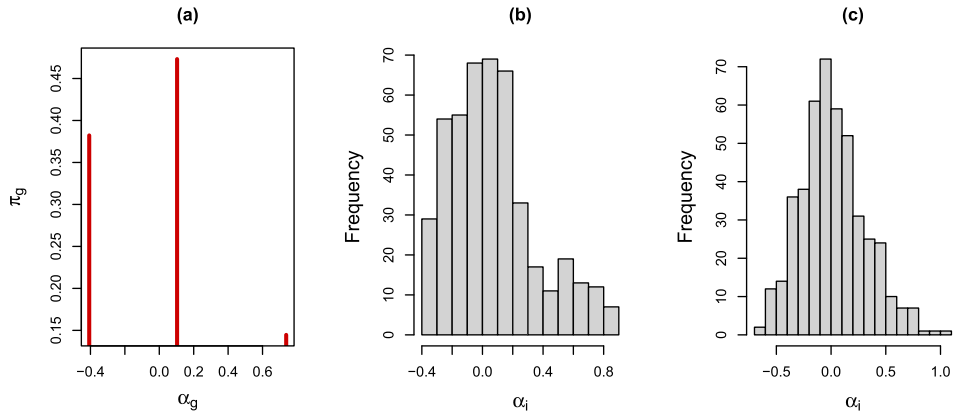


FIG. 7. *Semiparametric approach: estimated prior (a) and posterior (b) distribution for α_i 's; parametric approach: posterior distribution for α_i 's (c).*

Figure 7 shows the estimated prior (Figure 1(a)) and posterior distribution (Figure 1(b)) estimates for the random effects obtained using the proposed (semiparametric) approach together with the estimated posterior distribution deriving from the parametric approach (Figure 1(c)). In particular, in Figure 1(b), we report the posterior mean of the area-specific random intercepts calculated as

$$\hat{\alpha}_i = \sum_{g=1}^G (\hat{\xi}_g - \hat{\xi}) \hat{\tau}_{ig},$$

where $\hat{\xi} = \sum_g \hat{\xi}_g \hat{\pi}_g$ is the overall intercept estimate reported in Table 5. By focusing on Figure 1(a), we may notice that observed data lead to the estimation of a random effect with a clear degree of skewness. If the standard Gaussian assumption had been reasonable, the NPML estimate of the random effect distribution would have been a symmetric distribution centered around zero. As a consequence, we may conclude that such an assumption may not be that adequate for the current application. Furthermore, by comparing Figures 1(b) and 1(c), we may observe that the parametric assumption also affects the posterior mean of the area-specific intercepts, leading to a less skewed distribution than that obtained under the proposed approach.

7.2. Small area predictions. As highlighted in Section 3, we need the covariate values, \mathbf{x}_{ij} , to be known for all units in the population to predict the target variable. This would require access, for example, to Census microdata. However, in the important and special case where the components of \mathbf{x}_{ij} are all categorical or take a finite number of values, the method described in this paper only requires the corresponding area level cross-tabulations to be available. This is the case of the

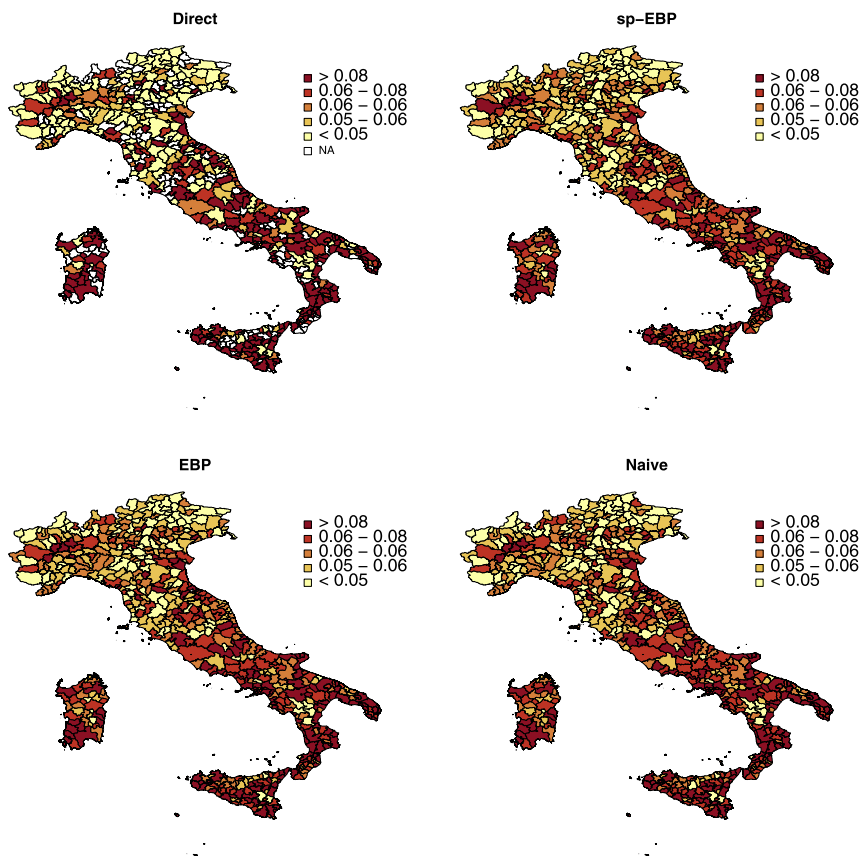


FIG. 8. Maps of the estimated unemployment incidences for LLMAs in Italy in 2012: direct estimates, sp-EBP, EBP and naive estimates.

ILFS data, where information on the covariates in the model are available at an aggregate level for the whole population. Figure 8 shows the map of unemployment incidence predictions for the 611 LLMAs obtained using direct estimation (upper-left panel), the proposed sp-EBP approach (upper-right panel), the parametric EBP (lower-left panel) and the naive approach (lower-right panel). Direct estimates are computed using Hájek-type estimators with adjusted weights that account for non response and calibrate to population level information of demographic variables. The patterns of unemployment produced by the proposed approach are consistent with those obtained by all the other methods; model-based maps are smoother when compared to direct estimates. As expected, relatively larger values for unemployment incidences are mainly located in the south of Italy and in the islands.

To assess the quality of predictions, we used a set of diagnostic tools based on the requirement that model-based small area estimates should be coherent with, in the sense of being close to, the corresponding unbiased direct estimates, albeit

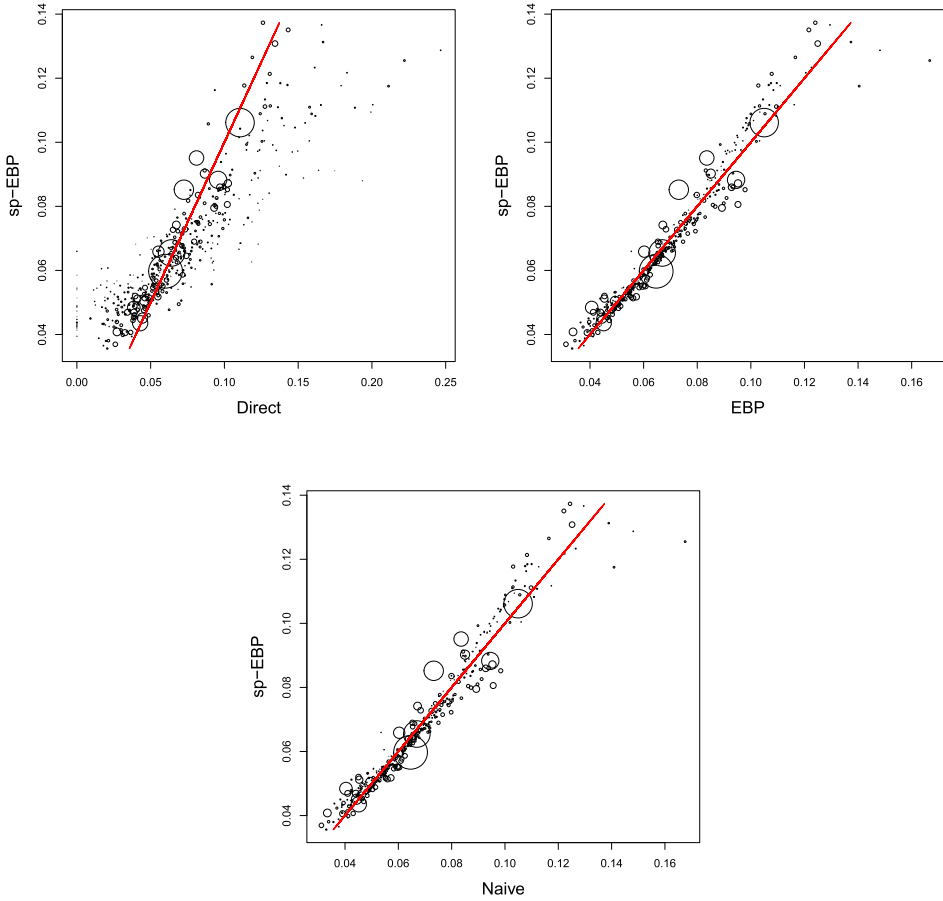


FIG. 9. *sp-EBP estimates of small area proportions versus the corresponding direct (left), EBP (centre) and naive estimates (right). Dots' size is proportional to the sample size.*

more precise. Figure 9 shows the estimates derived from the sp-EBP approach versus the direct, the EBP and the naive estimates, respectively. From this figure (first panel), we may observe that our approach leads to predictions which are close to those provided by a direct approach, with a correlation coefficient equal to 0.881. From the remaining panels in Figure 9, it is evident that model-based estimates for unemployment incidence are all very close to each other, with correlation coefficients equal to 0.978 (sp-EBP vs. EBP) and 0.977 (sp-EBP vs. naive).

Coherence of direct and sp-EBP estimates can be also evaluated by computing a goodness-of-fit diagnostic [Brown et al. (2001)] which is obtained from the following Wald-test statistic:

$$(22) \quad W = \sum_{i=1}^m \frac{(\hat{p}_i^{\text{Direct}} - \hat{p}_i^{\text{sp-EBP}})^2}{\widehat{\text{Var}}(\hat{p}_i^{\text{Direct}}) + \widehat{\text{MSE}}^*(\hat{p}_i^{\text{sp-EBP}})},$$

where the estimated MSE of the sp-EBP is calculated by using formulas in Section 5. Considering the results of the simulation experiments with $m = 500$, we decided to consider the bias-corrected MSE reported in equation (13). The above test is based on the idea that, should model-based estimates be close to the “true” small area parameters of interest, the unbiased direct estimates can be considered as random variables with expected value equal to the value of the corresponding model-based estimates. Here, $W = 360.56$, and such a value needs to be compared to the 95th percentile of a χ^2 distribution with 452 d.f., that is, $\chi_{452,0.95}^2 = 502.56$. In this respect, we may conclude that model-based estimates are not significantly different from direct estimates.

To assess the potential gain in precision we obtain by using the proposed sp-EBP approach in place of the direct one, we compare in Figure 10 the empirical cumulative density functions (ECDs) of the coefficients of variation (CV) of both estimators. The first panel uses CVs from all areas, while the second (third) one focuses on small areas with sample size smaller (larger) than 100. As it is clear by looking at the first panel, the ECDF corresponding to sp-EBP almost always dominates the one for the direct estimates, highlighting that CV values for the former approach are lower than those estimated with the latter. Only for very small CV values, the ECDFs show an inverse relation; CVs for direct predictions are smaller than those for sp-EBP. This is more evident in the second panel and is related to the presence of some areas with a small sample size for which $\hat{p}_i^{\text{Direct}}$ is zero or is very close to zero, and so is $\widehat{\text{RMSE}}(\hat{p}_i^{\text{Direct}})$. However, also in this case, more than 60% of Italian LLMA have CV% associated to the direct estimator above the standard 33% threshold which is typically considered for reliability in the SAE context. Such a percentage reduces to about 50% when considering the proposed sp-EBP approach; in addition, less than 5% of the estimates have CV% larger than 40%. When we move to higher CVs, the sp-EBP approach always provides smaller CV values when compared to the direct approach, and such CVs are always smaller than 40%. When focusing on the third panel in Figure 10, it

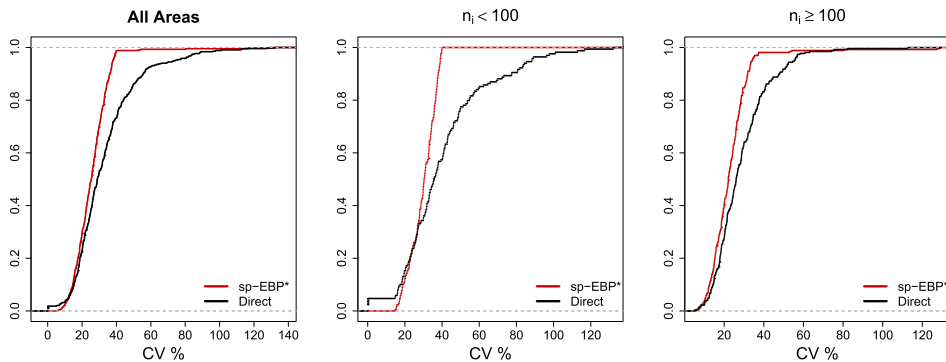


FIG. 10. CVs empirical cumulative density functions for the sp-EBP and the direct estimator.

is evident that, as the sample size gets larger, direct and model-based estimates tend to have quite similar CV values, although those associated with model-based estimates are still consistently smaller.

8. Conclusions. The paper described some tools to derive best predictions for responses with distribution in the Exponential Family in the presence of clustered data. In particular, we proposed a semiparametric version of the EBP and the corresponding second-order, bias-corrected MSE approximation using a NPML approach and leaving the distribution of the random effects unspecified. Motivated by a real application to data on unemployment incidence in LLMA in Italy, we focused on a binary response modeled via a mixed logistic model with random intercepts which represents a relevant case in the SAE framework.

Simulation experiments showed that the proposed estimator performs equally or better than the competitors. In particular, when moving far from the assumption of Gaussian distributed random effects, the proposed semiparametric approach performed better than the corresponding parametric versions. Also, when compared to the parametric EBP, simulation results highlighted better performance of the proposed approach in terms of computational load required to get predictions and corresponding MSE. The simulation study, where different sample sizes were considered, showed that the semiparametric approach is always reliable, especially for large m . Such a gain comes from the discrete nature of the mixing distribution estimate which substantially simplifies calculations.

We illustrated the benefits of our proposal discussing the estimation of unemployment incidence for Italian LLMA in 2012. In this context, direct estimates cannot be published for most of the LLMA given the unacceptable large value of the coefficient of variation for those areas with a small sample size. In this respect, model-based approaches represent a necessary strategy. Since the sample size and the number of small areas are particularly large in this application, the implementation of the EBP turns out to be particularly cumbersome and the evaluation of its precision prohibitive. This application indicated that the proposed methodology leads to estimates which are coherent with, but more efficient than, the direct ones still being comparable with alternative model-based estimates.

Although the approach we proposed is presented for responses with density in the Exponential Family, we did not explore the behavior of the small area sEBPs for counts or multinomial responses. However, a possible extension to multi-category outcomes is quite straightforward. Also, we notice that suitable extensions of the proposed approach to allow for spatial correlation could be envisioned by properly modeling, for each small area, the prior mixture probabilities as a function of neighborhood component memberships. Last, developing design-consistent small area estimators under the proposed methodology represents a topic of interest, especially for those researchers working in survey sampling from a design-based or a model-assisted perspective. More specifically, we could adopt a model-assisted approach, thereby the model is used only to motivate the predictors, but

their properties are evaluated only with respect to the randomization distribution induced by the sampling design.

SUPPLEMENTARY MATERIAL

Supplement to “Semiparametric empirical best prediction for small area estimation of unemployment indicators” (DOI: [10.1214/18-AOAS1226SUPP](https://doi.org/10.1214/18-AOAS1226SUPP); .pdf). The online Supplementary Material describes the EM algorithm for parameter estimation and the procedure for estimating the covariance matrix of model parameters. Also, computational details for deriving the bias correction term for the MSE estimator of the proposed sp-EBP, as well as explicit formulas for computing model derivatives in the case of binary data are reported. Some additional simulation results are also presented. Last, a computationally efficient algorithm for estimation and inference developed in R language from the authors, together with an example data set, is made available as part of the online Supplementary Material.

REFERENCES

- AITKIN, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Stat. Comput.* **6** 251–262.
- AITKIN, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55** 117–128. [MR1705676](#)
- BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48.
- BATTESE, G., HARTER, R. and FULLER, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.
- BÖHNING, D. (1982). Convergence of Simar’s algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Ann. Statist.* **10** 1006–1008. [MR0663451](#)
- BOUBETA, M., LOMBARDÍA, M. J. and MORALES, D. (2016). Empirical best prediction under area-level Poisson mixed models. *TEST* **25** 548–569. [MR3531842](#)
- BOUBETA, M., LOMBARDÍA, M. J. and MORALES, D. (2017). Poisson mixed models for studying the poverty in small areas. *Comput. Statist. Data Anal.* **107** 32–47. [MR3575057](#)
- BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.
- BROWN, G., CHAMBERS, R., HEADY, P. and HEASMAN, D. (2001). Evaluation of small area estimation methods: an application to unemployment estimates from the UK LFS. In *Proc. Statistics Canada Symp. Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Hull: Statistics Canada.
- CHEN, J. H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. [MR1331665](#)
- CHEN, S. X. and LIU, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statist. Sinica* **7** 875–892. [MR1488647](#)
- D’ALÒ, M., FALORSI, S. and SOLARI, F. (2017). Space-time unit-level EBLUP for large data sets. *J. Off. Stat.* **33** 61–77.
- D’ALÒ, M., DI CONSIGLIO, L., FALORSI, S., RANALLI, M. G. and SOLARI, F. (2012). Use of spatial information in small area models for unemployment rate estimation at sub-provincial areas in Italy. *J. Indian Soc. Agricultural Statist.* **66** 43–53. [MR2953458](#)

- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. [MR0501537](#)
- GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. [MR2847988](#)
- GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D. and SANTAMARÍA, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Comput. Statist. Data Anal.* **51** 2720–2733. [MR2338998](#)
- HOBZA, T. and MORALES, D. (2016). Empirical best prediction under unit-level logit mixed models. *J. Off. Stat.* **32** 661–692.
- HOBZA, T., MORALES, D. and SANTAMARÍA, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST* **27** 270–294. [MR3799140](#)
- JIANG, J. (1998). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93** 720–729. [MR1631373](#)
- JIANG, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *J. Statist. Plann. Inference* **111** 117–127. [MR1955876](#)
- JIANG, J. and LAHIRI, P. (2001). Empirical best prediction for small area inference with binary data. *Ann. Inst. Statist. Math.* **53** 217–243. [MR1841133](#)
- LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. [MR0521328](#)
- LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.
- LINDSAY, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11** 86–94. [MR0684866](#)
- LINDSAY, B. G. (1983b). The geometry of mixture likelihoods. II. The Exponential Family. *Ann. Statist.* **11** 783–792. [MR0707929](#)
- LINDSAY, B. G. and LESPERANCE, M. L. (1995). A review of semiparametric mixture models. *J. Statist. Plann. Inference* **47** 29–39. [MR1360957](#)
- LÓPEZ-VIZCAÍNO, E., LOMBARDÍA, M. J. and MORALES, D. (2013). Multinomial-based small area estimation of labour force indicators. *Stat. Model.* **13** 153–178. [MR3179522](#)
- MARINO, M. F., RANALLI, M. G., SALVATI, N. and ALFÒ, M. (2019). Supplement to “Semi-Parametric Empirical Best Prediction for small area estimation of unemployment indicators.” DOI:10.1214/18-AOAS1226SUPP.
- MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170. [MR1436105](#)
- MOLINA, I., NANDRAM, B. and RAO, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. *Ann. Appl. Stat.* **8** 852–885. [MR3262537](#)
- MOLINA, I., SAEI, A. and LOMBARDÍA, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *J. Roy. Statist. Soc. Ser. A* **170** 975–1000. [MR2408988](#)
- MUNDLAK, Y. (1978). On the pooling of time series and cross section data. *Econometrica* **46** 69–85. [MR0478489](#)
- NEUHAUS, J. M. and MCCULLOCH, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 859–872. [MR2301298](#)
- OAKES, D. (1999). Direct calculation of the information matrix via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 479–482. [MR1680298](#)
- PINHEIRO, J. and BATES, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comput. Graph. Statist.* **4** 12–35.
- PRASAD, N. G. N. and RAO, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.* **85** 163–171. [MR1137362](#)

- RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. *Wiley Series in Survey Methodology*. Wiley, Hoboken, NJ. [MR3380626](#)
- REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. [MR0738930](#)
- RODRIGUEZ, G. and GOLDMAN, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser. A* **158** 73–89.
- SAEI, A. and CHAMBERS, R. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. In *S3RI Methodology Working Papers* 1–35. Southampton Statistical Sciences Research Institute, Southampton.
- SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4** 1200–1209. [MR0428582](#)
- VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern Applied Statistics with S-Plus. Statistics and Computing*. Springer, New York. [MR1337030](#)

M. F. MARINO
DEPARTMENT OF STATISTICS
COMPUTER SCIENCE, APPLICATIONS “G. PARENTI”
UNIVERSITY OF FLORENCE
VIALE MORGAGNI 59
FLORENCE
ITALY
E-MAIL: mariafrancesca.marino@unifi.it

N. SALVATI
DEPARTMENT OF ECONOMICS AND MANAGEMENT
UNIVERSITY OF PISA
VIA RIDOLFI, 10
PISA 56124
ITALY
E-MAIL: nicola.salvati@unipi.it

M. G. RANALLI
DEPARTMENT OF POLITICAL SCIENCE
UNIVERSITY OF PERUGIA
VIA PASCOLI 20
PERUGIA 06123
ITALY
E-MAIL: giovanna.ranalli@unipg.it

M. ALFÒ
DEPARTMENT OF STATISTICS
“SAPIENZA” UNIVERSITY OF ROME
PIAZZALE A. MORO 5
ROME 00185
ITALY
E-MAIL: marco.alfò@uniroma1.it