Semi-Parametric Empirical Best Prediction for small area estimation of unemployment indicators

M.F. Marino[*]   M.G. Ranalli [†]   N. Salvati [‡]   M. Alfò [§]

Abstract

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 render 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 non trivial task. In this paper, we propose a Semi-Parametric Empirical Best Predictor for a (possibly) non-linear 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.

Key Words: Binary data; Exponential Family; Finite Mixture; General parameters; Mixed logistic model; Unit-level model.

1   Introduction

The Italian National Institute for Statistics (ISTAT) regularly provides estimates of unemployment indicators based on data obtained through the Italian Labor Force Survey (ILFS). The ILFS allows to obtain quarterly estimates of the main aggregates regarding the labor market; these are particularly important both at the local and the central government levels for the development of labor market policies. These estimates are planned

---

[*]Department of Statistics, Computer Science, Applications, Università degli Studi di Firenze, Italy. `mariafrancesca.marino@unifi.it`

[†]Department of Political Science, Università degli Studi di Perugia, Italy. `giovanna.ranalli@unipg.it`

[‡]Department of Economics and Management, Università di Pisa, Italy. `nicola.salvati@unipi.it`

[§]Department of Statistical Science, Sapienza Università di Roma, Italy. `marco.alfo@uniroma1.it`

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, 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 to 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'Aló et al., 2012, 2017).

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 Liner 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 et al. (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'Aló et al. (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 collects, every year, 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 et al. (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) and 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 et al. (2018).

In all of these approaches, the area-specific random effects are assumed to be iid 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 e.g. on

2

(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. For this reason, ad-hoc alternatives, mainly based on plug-in predictors and Taylor linearizations, were proposed and are currently largely applied (Saei and Chambers, 2003; González-Manteiga et al., 2007; Molina et al., 2007; López-Vizcaíno et al., 2013).

In this paper, we describe a further alternative and develop a Semi-Parametric 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 - Simar, 1976; Laird, 1978; Lindsay, 1983a,b). This estimate known to be a discrete distribution defined over a finite number of locations leading to a (semi-parametric) 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 and, later, focusing on the relevant case of binary data. We compare our proposal to the EBP (Jiang, 2003) and to the plug-in estimator (e.g. Saei and Chambers, 2003; González-Manteiga et al., 2007) 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 LLMAs 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.
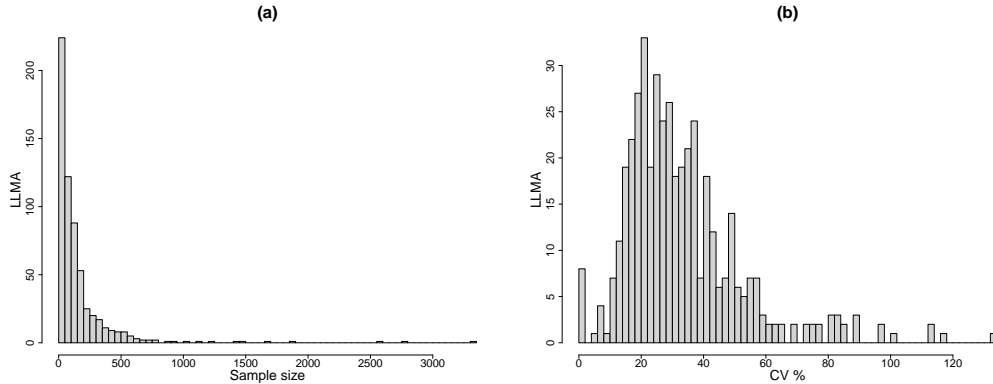
2 The ILFS data

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 LLMAs. LLMAs refer to 611 unplanned domains obtained as clusters of municipalities where the bulk of the labour force lives and works, and where establishments can find the largest amount of the labour force necessary to occupy the offered jobs. They respond to the need for meaningfully, comparable, sub-regional labour market areas for the reporting and the analysis of statistics. LLMAs 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, LLMAs were re-defined by the analysis of daily working commuting flows using a new allocation process, an evolution of the previous algorithm. Nearly half of the LLMAs stands in the size class from $10,000$ up to $50,000$ inhabitants, whereas the highest proportion of the population (35.0%) lives in LLMAs with a dimension between $100,000$ and $500,000$ inhabitants. In 332 LLMAs (over 70% of the national population), more than three quarter of the labour force lives and works in the same LLMA, that is self-containment is more than 75%.

Figure 1a shows the distribution of the LLMAs by sample size. This plot does not include the 158 areas with zero sample size. Among the observed LLMAs, the 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 LLMAs are characterized by a very small sample size that hinders reliability of direct estimates. Figure 1b 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% that 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. *Sex-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,

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

and university degree or beyond; *U-count*: a discrete variable measuring the number of unemployed in a given sex-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 *Sex-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 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 for the Italian LLMAs cannot be based on direct survey estimation as direct estimates cannot be computed and/or published for most of the LLMAs. For these reason, 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 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 semi-parametric approach.

5

Table 1: Sample percentage distribution with standard errors (S.E.) of *Unemployed status* by *Sex-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

Table 2: Sample percentage distribution with standard errors (S.E.) of *Unemployed status* by *Educationial 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

3   The Empirical Best Prediction

Let U denote a finite population of size N, which can be partitioned into m non-overlapping small areas/domains, with U_i denoting the i-th small area with size $N_i, i = 1, \ldots, m$. 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 $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$, with $j = 1, \ldots, N_i$. Also, let $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m$ be iid, q-dimensional, vectors of area-specific random effects ($q \leq p$) with density $f_\alpha(\cdot)$, $E_\alpha(\boldsymbol{\alpha}_i) = 0$, and $E_\alpha(\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i') = \boldsymbol{\Sigma}$ for all $i = 1, \ldots, m$. Last, let $\boldsymbol{w}_{ij}$ denote a q-dimensional subset of $\boldsymbol{x}_{ij}$ associated to $\boldsymbol{\alpha}_i$. 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 non-sampled 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 ($i = 1, \ldots, m, j \in s_i$), while the values of $\boldsymbol{x}_{ij}$ and $\boldsymbol{w}_{ij}$, are known for all units in the population ($i = 1, \ldots, m, j = 1, \ldots, N_i$). 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. We will discuss this issue in more details in the application to the ILFS data. Last, we assume that sampling is non-informative for the small area distribution of $Y_{ij} \mid \boldsymbol{x}_{ij}$, 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 $\boldsymbol{\alpha}_i$, responses Y_{ij} from the same small area i are independent with density in the Exponential Family

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

for $i = 1, \ldots, m$ and $j = 1, \ldots, N_i$. 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 $\boldsymbol{\beta}$ denote a p-dimensional vector of fixed regression coefficients and let us assume that θ_{ij} is modeled via the following regression model

$$\theta_{ij} = \eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{w}'_{ij}\boldsymbol{\alpha}_i.$$

The joint distribution of $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iN_i})'$ for the i-th small area, conditional on the vector of area-specific random effects $\boldsymbol{\alpha}_i$, is obtained by exploiting conditional independence

$$f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i; \boldsymbol{X}_i) = \prod_{j=1}^{N_i} f_{y|\alpha}(y_{ij} \mid \boldsymbol{\alpha}_i; \boldsymbol{x}_{ij}) = \exp\left\{\sum_{j=1}^{N_i} \frac{y_{ij}\,\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right\},$$

where $\boldsymbol{X}_i$ denotes the matrix of covariates associated to units in the i-th area. The marginal distribution of the area-specific sequence $\boldsymbol{y}_i$ is obtained by integrating out $\boldsymbol{\alpha}_i$:

$$f_y(\boldsymbol{y}_i; \boldsymbol{X}_i) = \int_{\mathbb{R}^q} f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i; \boldsymbol{X}_i) f_\alpha(\boldsymbol{\alpha}_i) d\boldsymbol{\alpha}_i.$$

Typically, a parametric specification for $f_\alpha(\boldsymbol{\alpha}_i)$ 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 $\boldsymbol{x}_{ij}$ is taken, that is $f_\alpha(\boldsymbol{\alpha}_i \mid \boldsymbol{X}_i) = f_\alpha(\boldsymbol{\alpha}_i)$ or $E(\boldsymbol{\alpha}_i \mid \boldsymbol{X}_i) = E(\boldsymbol{\alpha}_i) = \boldsymbol{0}$. 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 $\boldsymbol{X}_i$, see Neuhaus and McCulloch (2006). In the following, we will assume that, if needed, such an approach is applied and that $f_\alpha(\boldsymbol{\alpha}_i \mid \boldsymbol{X}_i) = f_\alpha(\boldsymbol{\alpha}_i)$.

3.2 EBP and MSE approximation

We are interested in using sample data on responses Y_{ij} $(i = 1, \ldots, m, j \in s_i)$ and population data on covariates $\boldsymbol{x}_{ij}$ $(i = 1, \ldots, m, j = 1, \ldots, N_i)$ to predict a (possibly) non-linear function of fixed and random effects, say $\zeta(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$, with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m)$. According

to Jiang (2003), the Best Predictor (BP) of ζ in terms of minimum MSE is given by

$$\tilde{\zeta}^{BP} = E_{\alpha|y}\left[\zeta(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma}) \mid \boldsymbol{y}\right] = \int_{\mathbb{R}^{m \times q}} \zeta(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma}) f_{\alpha|y}(\boldsymbol{\alpha} \mid \boldsymbol{y}) d\boldsymbol{\alpha}, \tag{1}$$

where

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

and $f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i; \boldsymbol{X}_i) = \prod_{j \in s_i} f_{y|\alpha}(y_{ij} \mid \boldsymbol{\alpha}_i; \boldsymbol{x}_{ij})$. 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:

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

where, as before, $f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i; \boldsymbol{X}_i)$ 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 in few cases, for instance when $f_{y|\alpha}(\cdot \mid \cdot)$ 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 non trivial 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 $\hat{\zeta}^{EBP} = \tilde{\zeta}^{BP}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Sigma}})$. 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:

$$\text{MSE}(\hat{\zeta}^{EBP}) = E[(\hat{\zeta}^{EBP} - \zeta)^2] = \frac{1}{m} e(\boldsymbol{\Phi}) + d(\boldsymbol{\Phi}) + o_p(1/m), \tag{3}$$

where

$$e(\boldsymbol{\Phi}) = E_y\left[\left(\frac{\partial \tilde{\zeta}^{BP}}{\partial \boldsymbol{\Phi}}\right)' m V(\hat{\boldsymbol{\Phi}}) \left(\frac{\partial \tilde{\zeta}^{BP}}{\partial \boldsymbol{\Phi}}\right)\right], \tag{4}$$

$$d(\boldsymbol{\Phi}) = E_\alpha[(\zeta)^2] - E_y[(\tilde{\zeta}^{BP})^2]$$
$$= \int_{\mathbb{R}^{m \times q}} \zeta(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma})^2 f_\alpha(\boldsymbol{\alpha}) d\boldsymbol{\alpha} - E_y\left[\left(\int_{\mathbb{R}^{m \times q}} \zeta(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma}) f_{\alpha|y}(\boldsymbol{\alpha} \mid \boldsymbol{y}) d\boldsymbol{\alpha}\right)^2\right], \tag{5}$$

and $f_\alpha(\boldsymbol{\alpha}) = \prod_{i=1}^{m} f_\alpha(\boldsymbol{\alpha}_i)$ denotes the joint density of the random effects $\boldsymbol{\alpha}_i, i = 1, \ldots, m$. An estimator of $\text{MSE}(\hat{\zeta}^{EBP})$ can be obtained by replacing $\boldsymbol{\Phi}$ in equation (3) by a consistent

estimator, that is

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

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

As it is clear, computing the MSE requires the solution of (multiple) integrals that may not admit a closed form expression. As stated before, Monte Carlo approximations or numerical integration techniques are required and this makes the computation extremely time-consuming. Bootstrap may represent a further alternative, 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 non-optimal 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 Semi-Parametric 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 non trivial 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 Semi-Parametric 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 multi-dimensional) integrals that do not admit a closed form expression. An alternative may be based on leaving the distribution of $\boldsymbol{\alpha}_i$ completely unspecified and follow the approach detailed by Aitkin (1996, 1999). The area-specific random effects are treated as nuisance parameters and a NonParametric Maximum Likelihood (NPML) estimate of their distribution is derived. Different contributions to the theory of NPML can be found in the literature (Simar, 1976; Laird, 1978; Böhning, 1982; Lindsay, 1983a,b). Results by Lindsay (1983a,b) 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 $\boldsymbol{\xi}_g = (\xi_{g1}, \ldots, \xi_{gq})'$, $g = 1, \ldots, G$, 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 $(\boldsymbol{y}_i, \boldsymbol{X}_i)$ in the sample. That is, in the presence of categorical

covariates, the number of locations does not necessarily grow with m.

Let $\boldsymbol{\Phi}$ denote the global vector of model parameters, $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \phi, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_G, \pi_1, \ldots, \pi_G)'$; the observed data likelihood is approximated by

$$L(\boldsymbol{\Phi}) = \prod_{i=1}^{m} \int_{\mathbb{R}^q} f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i; \boldsymbol{X}_i) \, d\boldsymbol{\alpha}_i \simeq \prod_{i=1}^{m} \sum_{g=1}^{G} f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\xi}_g; \boldsymbol{X}_i)\pi_g, \qquad (6)$$

where $f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\xi}_g; \boldsymbol{X}_i) = \prod_{j \in s_i} f_{y|\alpha}(y_{ij} \mid \boldsymbol{\alpha}_i = \boldsymbol{\xi}_g; \boldsymbol{x}_{ij})$ denotes the product of densities in the Exponential Family with canonical parameter θ_{ijg} defined by the following (mixed) model:

$$\theta_{ijg} = \eta_{ijg} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{w}'_{ij}\boldsymbol{\xi}_g.$$

As it is clear, expression (6) resembles the likelihood of a finite mixture of distributions, with weights $\pi_g = \Pr(\boldsymbol{\alpha}_i = \boldsymbol{\xi}_g)$. That is, $\boldsymbol{\alpha}_i \sim \sum_{g=1}^{G} \pi_g \delta(\boldsymbol{\xi}_g)$, 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 (e.g. based on Gaussian quadrature or Monte Carlo approaches) applied when a parametric specification is considered, are exactly of the type in equation (6). The only substantial difference is that locations $\boldsymbol{\xi}_g$ 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 et al., 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), as detailed in Sections 1 and 2 of the on-line Supplementary Material, where the EM algorithm is described. A crucial point in the proposed approach is the choice of the number of mixture components in (6). 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 $\boldsymbol{\alpha}_i \sim \sum_{g=1}^{G} \pi_g \delta(\boldsymbol{\xi}_g)$ is the true mixing, with G known, the usual ML asymptotics apply.

4.2   Semi-Parametric 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 non-overlapping domains or small areas. Furthermore, let $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_G)'$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)'$ 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, $\zeta(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\pi})$ by exploiting sample data on responses Y_{ij} and populations data on covariates $\boldsymbol{x}_{ij}$. Under the proposed approach, the Semi-Parametric Best Predictor (sp-BP) of ζ is defined according to the following expression:

$$\tilde{\zeta}^{\text{sp-BP}} = E_{\alpha|y}\left[\zeta(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\pi}) \mid \boldsymbol{y}\right] = \sum_{g_1 \cdots g_m} \zeta(\boldsymbol{\beta}, \boldsymbol{\xi}_{g_1,\ldots,g_m}, \boldsymbol{\pi}) \prod_{i=1}^{m} \tau_{ig_i}, \tag{7}$$

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

$$\tau_{ig} = \Pr\left(z_{ig} = 1 \mid \boldsymbol{y}_i\right) = \frac{\pi_g \, f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\xi}_g; \boldsymbol{X}_i)}{\sum_{l=1}^{G} \pi_l \, f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\xi}_l; \boldsymbol{X}_i)}, \tag{8}$$

where, as before, $f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\xi}_g; \boldsymbol{X}_i)$ refers to sample data only. As it is clear, expression (7) denotes the expected value of $\zeta(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\pi})$, with respect to the posterior distribution of the random effects $\boldsymbol{\alpha}$. 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 $\boldsymbol{\beta}, \boldsymbol{\xi}$, and $\boldsymbol{\pi}$ by consistent estimates. Here, we consider the estimates derived by the EM algorithm described in Section 1 of the on-line Supplementary Material. In the following, we will refer to such a quantity as the Semi-Parametric Empirical Best Predictor (sp-EBP) of ζ, denoted by $\hat{\zeta}^{\text{sp-EBP}} = \tilde{\zeta}^{\text{sp-BP}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\pi}})$.

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

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

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

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

The computational burden to obtain the above quantities is substantially lower than that

11

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 $\boldsymbol{\Phi}$ in (9) by a consistent estimator such as that obtained by maximizing the observed data likelihood in equation (6). That is,

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

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 $\hat{\boldsymbol{\Phi}}$ in $d^{\text{sp}}(\boldsymbol{\Phi})$, 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 directly not hold but, rather, need to be adapted.

Let $\boldsymbol{\Phi}_0$ denote the "true" vector of model parameters and let us consider a second-order Taylor expansion of $d^{\text{sp}}(\boldsymbol{\Phi})$ around $\boldsymbol{\Phi}_0$ evaluated at $\hat{\boldsymbol{\Phi}}$:

$$d^{\text{sp}}(\hat{\boldsymbol{\Phi}}) = d^{\text{sp}}(\boldsymbol{\Phi}_0) + \left(\frac{\partial d^{\text{sp}}}{\partial \boldsymbol{\Phi}}\right)'\bigg|_{\boldsymbol{\Phi}_0}(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0) + \frac{1}{2}(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)'\left(\frac{\partial^2 d^{\text{sp}}}{\partial \boldsymbol{\Phi}\boldsymbol{\Phi}'}\right)(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0) + o_p(m^{-1}), \tag{11}$$

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

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

where $b^{\text{sp}}(\boldsymbol{\Phi})$ denotes a bias correction defined as

$$b^{\text{sp}}(\boldsymbol{\Phi}) = \left(\frac{\partial d^{\text{sp}}}{\partial \boldsymbol{\Phi}}\right)'\bigg|_{\boldsymbol{\Phi}_0} mE(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0) + \frac{m}{2}E\left[(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)'\left(\frac{\partial^2 d^{\text{sp}}}{\partial \boldsymbol{\Phi}\boldsymbol{\Phi}'}\right)\bigg|_{\boldsymbol{\Phi}_0}(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)\right]$$
$$= b_1^{\text{sp}}(\hat{\boldsymbol{\Phi}}) + b_2^{\text{sp}}(\hat{\boldsymbol{\Phi}}). \tag{12}$$

As it is shown in Section 3 of the on-line Supplementary Material, the former term on the right hand side of equation (12) is given by

$$b_1^{\text{sp}}(\hat{\boldsymbol{\Phi}}) = \left(\frac{\partial d^{\text{sp}}}{\partial \boldsymbol{\Phi}}\right)'\bigg|_{\boldsymbol{\Phi}_0} \frac{m}{2}I_e(\boldsymbol{\Phi}_0)^{-1}E\left\{tr\left[I_e(\boldsymbol{\Phi}_0)^{-1}\left[\frac{\partial^2 S^k(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}\partial \boldsymbol{\Phi}'}\right]\bigg|_{\boldsymbol{\Phi}_0}I_e(\boldsymbol{\Phi}_0)^{-1}I_e(\boldsymbol{\Phi}_0)\right]_{1\leq k\leq K}\right\}$$
$$- \left(\frac{\partial d^{\text{sp}}}{\partial \boldsymbol{\Phi}}\right)'\bigg|_{\boldsymbol{\Phi}_0} \frac{m}{2}I_e(\boldsymbol{\Phi}_0)^{-1}tr\left\{I_e(\boldsymbol{\Phi}_0)^{-1}\left[\frac{\partial I_e^k(\boldsymbol{\Phi}_0)}{\partial \boldsymbol{\Phi}'}\right]_{1\leq k\leq K}\right\}.$$

Here, $I_e(\boldsymbol{\Phi}_0)$ denotes the expected information matrix, while $S^k(\boldsymbol{\Phi})$ and $I_e^k(\boldsymbol{\Phi}_0)$ denote the k-th element of the score function $S(\boldsymbol{\Phi})$ and the k-th row of the expected information matrix $I_e(\boldsymbol{\Phi}_0)$, respectively.

On the other hand, it can be shown that the latter term in equation (12) , $b_2^{\text{sp}}(\hat{\boldsymbol{\Phi}})$,

can be computed as

$$b_2^{\mathrm{sp}}(\hat{\boldsymbol{\Phi}}) = \frac{m}{2} tr \left\{ \left(\frac{\partial^2 d^{\mathrm{sp}}}{\partial \boldsymbol{\Phi} \partial \boldsymbol{\Phi}'} \right) \bigg|_{\boldsymbol{\Phi}_0} V(\hat{\boldsymbol{\Phi}}) \right\}.$$

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

$$\widehat{\mathrm{MSE}}^*(\hat{\zeta}^{\mathrm{sp\text{-}EBP}}) = d^{\mathrm{sp}}(\hat{\boldsymbol{\Phi}}) + \frac{1}{m} \left[e^{\mathrm{sp}}(\hat{\boldsymbol{\Phi}}) - b^{\mathrm{sp}}(\hat{\boldsymbol{\Phi}}) \right]. \tag{13}$$

We report the computational details required to derive $b_1^{\mathrm{sp}}(\hat{\boldsymbol{\Phi}})$ and $b_2^{\mathrm{sp}}(\hat{\boldsymbol{\Phi}})$ in Section 3 of the on-line 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 ($i = 1, \ldots, m, j = 1, \ldots, N_i$), and let α_i denote an area-specific random effect. Again, let $\boldsymbol{x}_{ij}$ denote a p-dimensional vector of covariates, and $\boldsymbol{X}_i$ 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:

$$\theta_{ij} = \log \frac{p_{ij}}{1 - p_{ij}} = \eta_{ij} = \alpha_i + \boldsymbol{x}_{ij}' \boldsymbol{\beta}. \tag{14}$$

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 will use the EBP for the quantity

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

In fact, since N_i is usually very large in most applications, as it is the case in the one at hand, 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 ($i = 1, \ldots, m, j \in s_i$), while covariates $\boldsymbol{x}_{ij}$ are available at the population level ($i = 1, \ldots, m, j = 1, \ldots, N_i$). 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 $\xi_1, \ldots, \xi_G$, with $\sum_{g=1}^G \pi_g = 1$. By adopting a canonical

13

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

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

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

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

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

$$\tilde{p}_i^{\text{sp-BP}} = \sum_{g=1}^{G} p_{ig} \frac{\exp\left[\sum_{j \in s_i} y_{ij}\eta_{ijg} - \sum_{j \in s_i} \log\left(1 + e^{\eta_{ijg}}\right)\right]\pi_g}{\sum_{l=1}^{G} \exp\left[\sum_{j \in s_i} y_{ij}\eta_{ijl} - \sum_{j \in s_i} \log\left(1 + e^{\eta_{ijl}}\right)\right]\pi_l}$$

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

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

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

the sp-BP of p_i is given by

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

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

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

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

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

14

where

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

and

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

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 $\Pr\left(Y_{i\cdot} = h; \boldsymbol{X}_i\right)$ is obtained as

$$\Pr\left(Y_{i\cdot} = h; \boldsymbol{X}_i\right) = \sum_{g=1}^{G} \Pr\left(Y_{i\cdot} = h \mid \xi_g; \boldsymbol{X}_i\right) \pi_g,$$

where $\Pr\left(Y_{i\cdot} = h \mid \xi_g; \boldsymbol{X}_i\right)$ represents the probability of observing h successes in n_i independent, but non identically 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 $(p_{i1g}, \ldots, p_{in_ig})$. The last term in equation (18) is obtained as

$$E_\alpha[(\hat{p}_i^{\text{sp-EBP}} - \tilde{p}_i^{\text{sp-BP}})^2] = \sum_{h=0}^{n_i} \left[\left(\frac{\partial \tilde{p}_{i(h)}^{\text{sp-BP}}}{\partial \boldsymbol{\Phi}}\right)' mV(\hat{\boldsymbol{\Phi}}) \left(\frac{\partial \tilde{p}_{i(h)}^{\text{sp-BP}}}{\partial \boldsymbol{\Phi}}\right) \right] \Pr\left(Y_{i\cdot} = h; \boldsymbol{X}_i\right), \quad (19)$$

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

The second-order, bias-corrected, estimator of $\text{MSE}(\hat{p}_i^{\text{sp-EBP}})$, that is $\widehat{\text{MSE}}^*(\hat{p}_i^{\text{sp-EBP}})$, 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 are generated under some model assumptions and sample data are selected from the simulated population. In particular, population data are generated considering $m = 100, 200, 500$ small areas; then, samples are selected by simple random sampling

without replacement within each area. The population and the sample sizes are 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 generate the target variable $Y_{ij}, i = 1, \ldots, m, j = 1, \ldots, N_i$, from a Bernoulli distribution with success probability defined by

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

with $\beta = 1$, $x_{ij} \sim \mathrm{Unif}(-1, b_i)$, 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 $\Pr(\nu = 1) = 0.7$, $\mu_1 = 0$, $\mu_2 = 3$, 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 e.g. for properly representing non-homogeneous unemployment rates typically observed in the North/South of Italy, as we will see in Section 7. Clearly, the chosen scenarios represent two extreme situations; we expect that, in real data 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 ($G = 2, \ldots, 5$) 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 with the parametric EBP by Jiang and Lahiri (2001) and the Naive predictor considered in González-Manteiga et al. (2007), both based on the assumption of Gaussian random effects. For the EBP, parameter estimates were derived via the 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

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	0.858	0.132	0.001	0.000	0.943	0.056	0.001	0.000

the intractable integrals, we followed the approach suggested by Boubeta et al. (2016). That is, we started by generating $B = 2,500$ replicates of the area-specific random effects $\alpha_i^{(b)}$ 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 will denote EBP estimates of small area proportions by $\hat{p}_i^{\text{EBP}}$. 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 the PQL approach via the `glmmPQL` function from the R `MASS` package (Venables and Ripley, 2002). To get predictions, parameter estimates were directly plugged into the expression for the area-specific proportions:

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

The performance of the small area estimators were evaluated by computing, for each area $i = 1, \ldots, m$, the bias and the Root Mean Squared Error (RMSE), defined as follows:

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

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

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 ($\hat{p}_i^{\text{EBP}}$), the sp-EBP ($\hat{p}_i^{\text{sp-EBP}}$), or the Naive ($\hat{p}_i^{\text{Naive}}$) approach. For completeness, we also report in Section 5 of the on-line Supplementary Material the distribution of the Mean Absolute Error (MAE) across small areas for the EBP, the sp-EBP, and the Naive predictor under different experimental scenarios. Together with the bias, MAE is frequently used to evaluate the quality of

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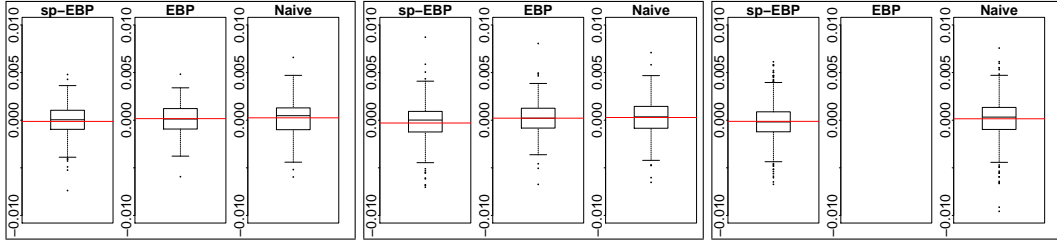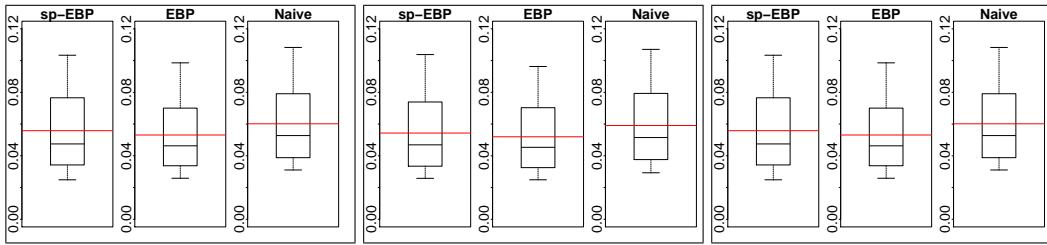

Figure 3: Scenario 1: Distribution of the RMSE over areas for $\hat{p}_i^{\text{sp-EBP}}$, $\hat{p}_i^{\text{EBP}}$, and $\hat{p}_i^{\text{Naive}}$, for $m = 100$ (left panel), $m = 200$ (central panel), and $m = 500$ (right panel).

predictions 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 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 and RMSE. When $m = 500$, performance values of EBP are not showed 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 showed 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 seem to produce predictions with a reduced quality than those obtained via the semi-parametric alternative we propose. In particular, we may notice that $\hat{p}_i^{\text{sp-EBP}}$ clearly outperforms the two competitors in terms of both bias and RMSE. Also, results for $\hat{p}_i^{\text{Naive}}$ and $\hat{p}_i^{\text{EBP}}$ seem to 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

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

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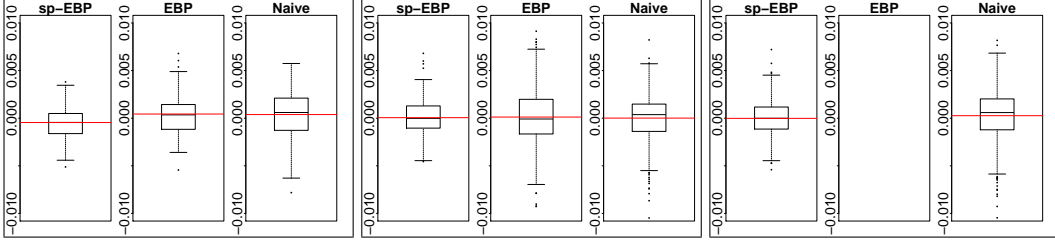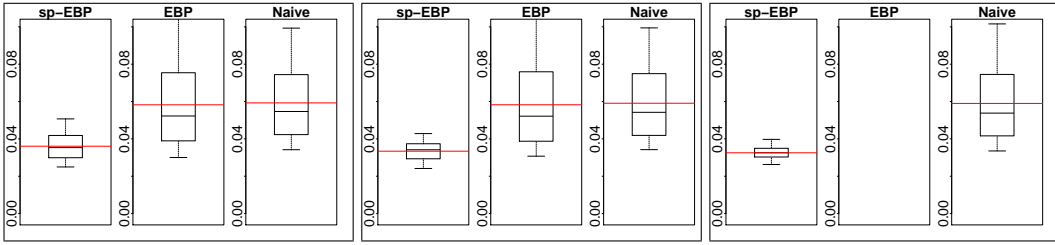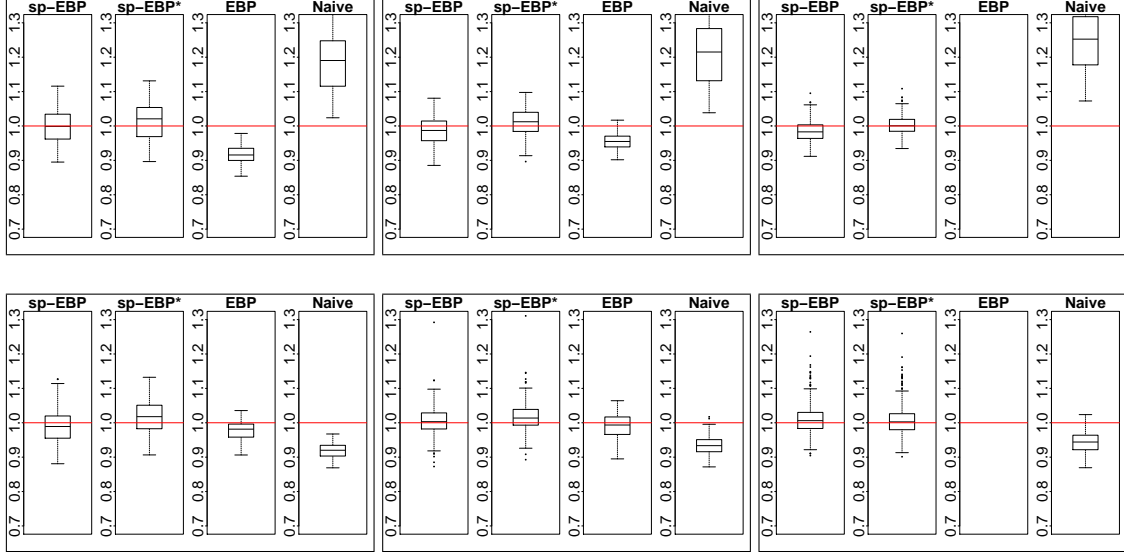

MSE estimators to evaluate the accuracy of the predictions we discussed so far. In particular, for $\hat{p}_i^{\text{sp-EBP}}$, we considered the proposed MSE estimator reported in equations (10) and (13); we will refer to the square root of these quantities as $\widehat{\text{RMSE}}(\hat{p}_i^{\text{sp-EBP}})$ and $\widehat{\text{RMSE}}^*(\hat{p}_i^{\text{sp-EBP}})$, respectively. For the estimator $\hat{p}_i^{\text{EBP}}$, we used the approximate MSE estimator proposed by Hobza and Morales (2016); the corresponding square root will be denoted by $\widehat{\text{RMSE}}(\hat{p}_i^{\text{EBP}})$. Last, for the Naive predictor $\hat{p}_i^{\text{Naive}}$, 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 will be 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:

$$\text{R}_i = \frac{\sum_{t=1}^{T} \widehat{\text{RMSE}}(\hat{p}_{it}^{\text{Model}})}{\sqrt{\sum_{t=1}^{T}(\hat{p}_{it}^{\text{Model}} - p_{it})^2}}, \quad i = 1, \dots, m.$$

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 $\widehat{\text{RMSE}}^*(\hat{p}_i^{\text{sp-EBP}})$ seem to 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

Figure 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 other hand, $\widehat{\mathrm{RMSE}}^*(\hat{p}_i^{\mathrm{sp\text{-}EBP}})$ 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.

The estimator $\widehat{\mathrm{RMSE}}(\hat{p}_i^{\mathrm{EBP}})$ underestimate 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 et al. (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 $V(\hat{\boldsymbol{\Phi}})$ 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, $\widehat{\mathrm{RMSE}}(\hat{p}_i^{\mathrm{sp\text{-}EBP}})$ allows to estimate the actual RMSE with a higher precision than the

20

corresponding bias-corrected version $\widehat{\text{RMSE}}^*(\hat{p}_i^{\text{sp-EBP}})$. 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.

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

$$\text{CR}_i = T^{-1} \sum_{t=1}^{T} \mathbb{1}\left(|\hat{p}_{it} - p_{it}| \leqslant 1.96 \times \widehat{\text{RMSE}}(p_{it}^{\text{Model}}) \right), \quad i = 1, \ldots, m.$$

As it is clear from the 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.

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 $m = 100, 200, 500$.

	Coverage			Computational Time		
m	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

To conclude, we also compared MSE estimators in terms of computational complexity. The last column of Tables 4 reports the computational time (averaged 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 $\widehat{\mathrm{RMSE}}^*(\hat{p}_i^{\mathrm{sp\text{-}EBP}})$ requires a higher computational effort than that required for $\widehat{\mathrm{RMSE}}(\hat{p}_i^{\mathrm{sp\text{-}EBP}})$: 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 LLMAs in Italy

In this section, we use ILFS data to estimate unemployment incidence for 611 LLMAs 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; (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 authours. This is part of the on-line Supplementary Material at the publisher's web-site, 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 et al. (2014) and considered the variables introduced in Section 2 and their transformations in the linear predictor, that is *Sex-Age* (reference = 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 ($G = 2, \ldots, 6$) 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. 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

22

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.

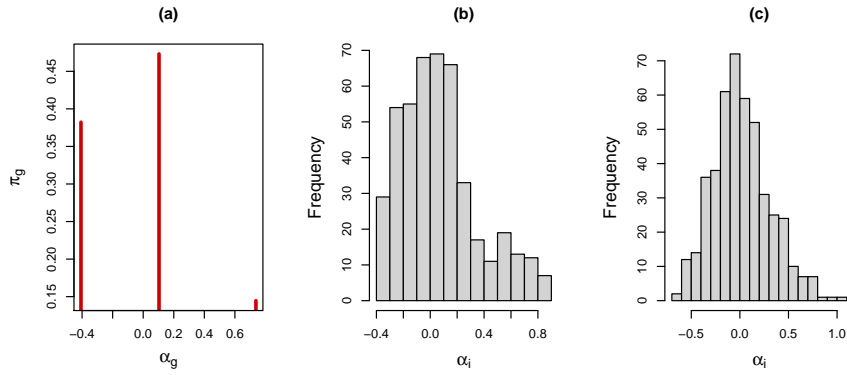|	Unspecific			Gaussian		
	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	-3.052	0.176	<0.001	-3.002	0.150	<0.001
M:25-34	0.118	0.057	0.038	0.122	0.054	0.024
M:35-65	-0.787	0.054	<0.001	-0.778	0.048	<0.001
F:15-24	-0.222	0.062	<0.001	-0.221	0.058	<0.001
F:25-34	0.113	0.05	0.023	0.118	0.054	0.029
F:35-65	-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
$\log(U\text{-}count)$	0.113	0.012	<0.001	0.104	0.022	<0.001
ℓ		-21833.599			-21837.357	
AIC		43695.199			43696.714	

based on an unspecified random effect distribution with respect to its parametric counterpart, even though differences in terms of parameter estimates are rather negligible. In particular, looking at the estimates for *Sex-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 those of males aged 15-24. On the other hand, for a male who is aged 35-65 years, the odds are $e^{-0.787} = 0.455$, that is 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 diploma is higher than that 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.

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

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

where $\hat{\bar{\xi}} = \sum_g \hat{\xi}_g \hat{\pi}_g$ is the overall intercept estimate reported in Table 5. By focusing the attention on Figure 1a, we may clearly observe 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 1b and 1c, 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.

Figure 7: Semi-parametric approach: estimated prior (a) and posterior (b) distribution for α_i's; parametric approach: posterior distribution for α_i's (c).

7.2 Small area predictions

As highlighted in Section 3, we need the covariate values, $\boldsymbol{x}_{ij}$, to be known for all units in the population to predict the target variable. This would require access e.g. to census micro-data. However, in the important and special case where the components of $\boldsymbol{x}_{ij}$ 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 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 prediction for the 611 LLMAs obtained using direct estimation, the proposed sp-EBP approach, the parametric EBP, and the Naive approach. Direct estimates are computed using Hájek-type estimators with adjusted weights that account for nonresponse 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. As expected, model-based maps are smoother when compared to direct estimates; 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

Figure 8: Maps of the estimated unemployment incidences for LLMAs in Italy in 2012: direct estimates, sp-EBP, EBP, and Naive estimates.
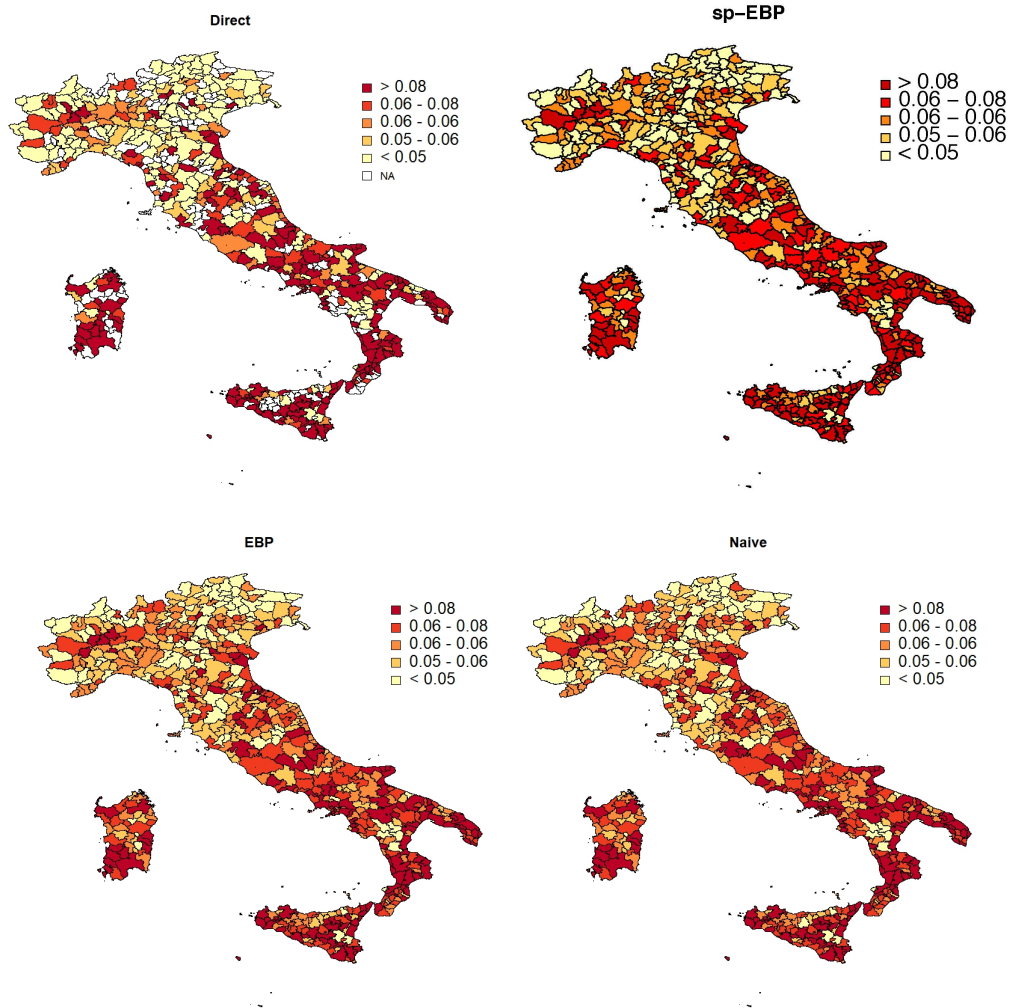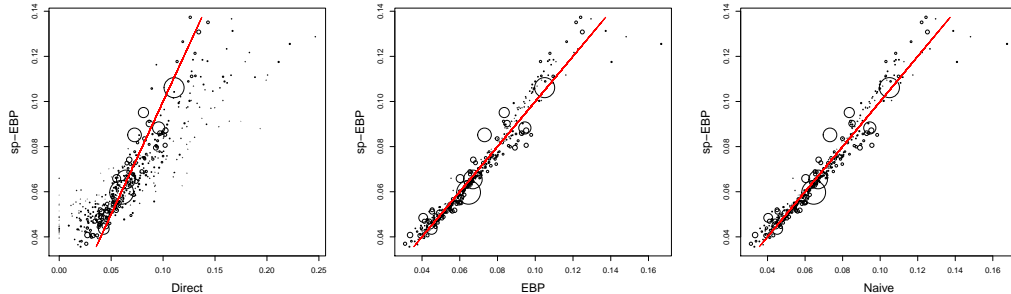
Figure 9: sp-EBP estimates of small area proportions versus the corresponding direct (left), EBP (centre), and Naive estimates (right). Dots' size is proportional to the sample size.

requirement that model-based small area estimates should be coherent with, in the sense of being close to, the corresponding unbiased direct estimates, albeit 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 to 0.977 (sp-EBP vs. Naive).
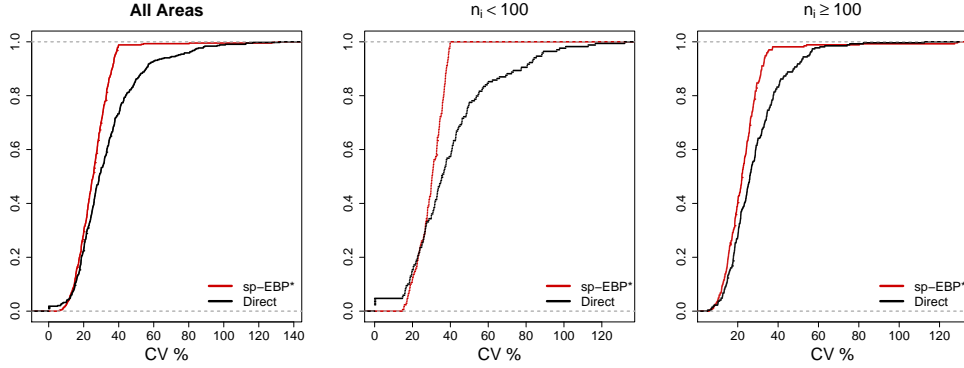
Coherence of direct and sp-EBP estimates can be also evaluated by computing a goodness-of-fit diagnostic (Brown et al., 2001), which is obtained from the following Wald-test statistic:

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

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 could 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 95-th percentile of a χ^2 distribution with 452 d.f., $\chi^2_{452,0.95} = 502.56$. 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 (ecdfs) of the vcoefficients of variation (CV) of both estimators. The first panel uses CV's from all areas, while the second (third) one focuses on small areas with sample size smaller (higher) than 100. As it is clear, by looking at the first panel, the

Figure 10: CV's empirical cumulative density functions for the sp-EBP and the direct estimator.

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: CV's 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, about the 60% of the Italian LLMAs, CVs associated to the direct estimator are above the standard 33% threshold which is typically considered for reliability in the SAE context. Such a percentage reduces to about 20% 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 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 semi-parametric 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 LLMAs 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 semi-parametric 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 the corresponding MSE. The simulation study, where different sample sizes were considered, showed that the semi-parametric approach is always reliable, especially for large m. Such a gain comes from the discrete nature of the mixing distribution estimate which substantially simplifies the calculations to get the EBPs and the corresponding MSEs.

We illustrated the benefits of our proposal discussing the estimation of unemployment incidence for Italian LLMAs in Italy in 2012. In this context, direct estimates cannot be published for most of the LLMAs 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 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 estimates, still being comparable with alternative model-based estimates.

Although the approach we propose is presented for responses with density in the Exponential Family, we did not explore the behavior of the small area sp-EBPs 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 components membership. Last, developing design-consistent small area estimators under the proposed methodology represent 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.

Acknowledgements

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6:251–262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55:117–128.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 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. *Journal of the American Statistical Association*, 83:28–36.

Böhning, D. (1982). Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound poisson process. *The Annals of Statistics*, 10:1006–1008.

Boubeta, M., Lombardía, M. J., and Morales, D. (2016). Empirical best prediction under area-level Poisson mixed models. *Test*, 25(3):548–569.

Boubeta, M., Lombardía, M. J., and Morales, D. (2017). Poisson mixed models for studying the poverty in small areas. *Computational Statistics & Data Analysis*, 107:32–47.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 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. (1995). Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23:221–233.

Chen, S. X. and Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7:875–892.

D'Aló, 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. *Journal of the Indian Society of Agricultural Statistics*, 66(1):43–53.

D'Aló, M., Falorsi, S., and Solari, F. (2017). Space-time unit-level eblup for large data sets. *Journal of Official Statistics*, 33(1):61–77.

Dempster, A., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

González-Manteiga, W., Lombardía, M., 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. *Computational Statistics and Data Analysis*, 51:2720–2733.

Hobza, T. and Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. *Journal of Official Statistics*, 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.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, 93:720–729.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111:117–127.

Jiang, J. and Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53:217–243.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805–811.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Lindsay, B. and Lesperance, M. (1995). A review of semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47:29–39.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, 11:86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, Part II: the exponential family. *The Annals of Statistics*, 11:783–792.

López-Vizcaíno, E., Lombardía, M. J., and Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical modelling*, 13:153–178.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92:162–170.

Molina, I., Nandram, B., Rao, J., et al. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. *The Annals of Applied Statistics*, 8(2):852–885.

Molina, I., Saei, A., and Lombardía, M. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society: Series A*, 70:265–283.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46:69–85.

Neuhaus, J. M. and McCulloch, C. E. (2006). Separating between-and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series B*, 68:859–872.

Oakes, D. (1999). Direct calculation of the information matrix via the em. *Journal of the Royal Statistical Society: Series B*, 61:479–482.

Pinheiro, J. and Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12–35.

Prasad, N. and Rao, J. (1990). The estimation of mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85:163–171.

Rao, J. N. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Redner, R. and Walker, H. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26:198–239.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(1):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*, pages 1–35. Southampton Statistical Sciences Research Institute, Southampton.

Simar, L. (1976). Maximum likelihood estimation of a compound poisson process. *The Annals of Statistics*, 4:1200–1209.

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

Supplementary Material to "Semi-Parametric Empirical Best Prediction for small area estimation of unemployment indicators"

M.F. Marino[*]   M.G. Ranalli [†]   N. Salvati [‡]   M. Alfò [§]

1   EM algorithm for parameter estimation

In this section, we provide computational details of the the EM algorithm required to get model parameter estimates. Let us start from the observed data likelihood in equation (6) of the manuscript:

$$L(\boldsymbol{\Phi}) = \prod_{i=1}^{m} \sum_{g=1}^{G} f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\xi}_g; \boldsymbol{X}_i)\pi_g,$$

where $f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\xi}_g; \boldsymbol{X}_i) = \prod_{j \in s_i} f_{y|\alpha}(y_{ij} \mid \boldsymbol{\alpha}_i = \boldsymbol{\xi}_g; \boldsymbol{x}_{ij})$. In the following, we will denote this quantity as f_{ig} to simplify notation.

As we highlighted before, even if a direct maximization of the above expression is always affordable, an indirect approach, based on the use of an EM algorithm, is frequently adopted. For this purpose, let us define the binary indicator variable z_{ig} which is equal to one if area i belongs to the g-th component of the finite mixture. The EM algorithm starts from the definition of the complete-data log-likelihood:

$$\ell_c(\boldsymbol{\Phi}) = \sum_{i=1}^{m} \sum_{g=1}^{G} z_{ig} \left[\log f_{ig} + \log(\pi_g)\right], \tag{S.1}$$

In the E-step of the algorithm, we compute the expected value of the complete data log-likelihood in equation (S.1), conditional on the observed data and the current parameter estimates, say $\hat{\boldsymbol{\Phi}}^{(t)}$. Due to linearity in the indicator variables z_{ig}, at the generic $(t+1)$-th iteration of the algorithm, we need to derive the following quantities:

$$\tau_{ig}^{(t+1)} = E_{\hat{\boldsymbol{\Phi}}^{(t)}}\left[z_{ig} \mid \boldsymbol{y}_i, \hat{\boldsymbol{\Phi}}^{(t)}\right] = \Pr\left(z_{ig} = 1 \mid \boldsymbol{y}_i, \hat{\boldsymbol{\Phi}}^{(t)}\right) = \frac{\pi_g^{(t)} f_{ig}^{(t)}}{\sum_{l=1}^{G} \pi_l^{(t)} f_{il}^{(t)}}, \tag{S.2}$$

[*]Department of Statistics, Computer Science, Applications, Università degli Studi di Firenze, Italy. `mariafrancesca.marino@unifi.it`

[†]Department of Political Science, Università degli Studi di Perugia, Italy. `giovanna.ranalli@unipg.it`

[‡]Department of Economics and Management, Università di Pisa, Italy. `nicola.salvati@unipi.it`

[§]Department of Statistical Science, Sapienza Università di Roma, Italy. `marco.alfo@uniroma1.it`

where $f_{ig}^{(t)}$ denotes the joint conditional distribution of the i-th small area in the g-th component of the finite mixture, under the current estimate of model parameters $\mathbf{\Phi}^{(t)}$. The (conditional) expected value of the complete data log-likelihood is given by

$$Q\left(\mathbf{\Phi} \mid \hat{\mathbf{\Phi}}^{(t)}\right) = \sum_{i=1}^{m}\sum_{g=1}^{G} \tau_{ig}^{(t+1)}\left[\log(f_{ig}) + \log(\pi_g)\right]. \tag{S.3}$$

In the M-step, parameter estimates are obtained by maximizing (S.3) with respect to model parameters $\mathbf{\Phi}$. A closed form expression is available for the mixture component probabilities:

$$\pi_g^{(t+1)} = \sum_{i=1}^{m} \frac{\tau_{ig}^{(t+1)}}{m}.$$

For the parameters in the regression model, updates depend on the nature of the conditional distribution for Y_{ij}. Generally, the problem reduces to a weighted ML estimation problem for generalized linear models, which can be solved by using standard Newton-type recursions.

The E- and the M-step of the algorithm are iterated until convergence, which can be specified in terms of log-likelihood or parameter values, using relative or absolute norms. We report in the following an example of the algorithm structure in a programmable form. A crucial point in this framework is represented by the initialization of model

begin
    Initialize $\mathbf{\Phi}$ and $\tau_{ig}, i = 1, \ldots, n, g = 1, G$ **repeat**
        update τ_{ig} Expectation step
        update $\mathbf{\Phi}$ Maximization step
    until $\ell(\mathbf{\Phi}^{(t)}) - \ell(\mathbf{\Phi}^{(t-1)}) > \varepsilon;$
end
    Algorithm 1: Pseudo-code of the EM algorithm for parameter estimation

parameters. To avoid local maxima and/or spurious solutions, a multi-start strategy is frequently adopted by based both on a deterministic and a random starting rule. For the former, fixed model parameters can be initialized at the corresponding estimates from a homogeneous generalized linear model, while random parameters can be set equal to the G locations of a standard Gaussian quadrature approximation. On the other hand, random starting solutions can be obtained by randomly perturbing the deterministic ones. Overall, for a given G, the solution that at convergence of the algorithm corresponds to the highest log-likelihood value is retained as the optimal one.

2   Computing the standard errors for parameter estimates

A disadvantage of the EM algorithm is that it does not directly provide estimates for the covariance matrix of model parameter estimates. However, we may rely on the approach

discussed by Louis (1982) and, for practical purposes, on the formula described by Oakes (1999). Let

$$J(\hat{\boldsymbol{\Phi}}) = - \left[\frac{\partial^2 \ell(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi} \partial \boldsymbol{\Phi}'} \right] \bigg|_{\boldsymbol{\Phi} = \hat{\boldsymbol{\Phi}}}$$

denote the observed information matrix. According to Oakes (1999), we may write

$$J(\hat{\boldsymbol{\Phi}}) = - \left\{ \frac{\partial^2 Q(\boldsymbol{\Phi} \mid \hat{\boldsymbol{\Phi}})}{\partial \boldsymbol{\Phi} \partial \boldsymbol{\Phi}'} \bigg|_{\boldsymbol{\Phi} = \hat{\boldsymbol{\Phi}}} + \frac{\partial^2 Q(\boldsymbol{\Phi} \mid \hat{\boldsymbol{\Phi}})}{\partial \boldsymbol{\Phi} \partial \hat{\boldsymbol{\Phi}}'} \bigg|_{\boldsymbol{\Phi} = \hat{\boldsymbol{\Phi}}} \right\}, \tag{S.4}$$

where the first term denotes the expected complete-data Hessian matrix conditional on the observed data and the parameter estimates. To derive such a quantity, we proceed as follows. Let us denote by $h[\cdot]$ the inverse link function. As a first step, we need to compute the score functions associated with each element of $\boldsymbol{\Phi}$, based on the sample data only; these are given by

$$S(\boldsymbol{\alpha}_g) = \frac{\partial Q(\boldsymbol{\Phi} \mid \hat{\boldsymbol{\Phi}})}{\partial \boldsymbol{\alpha}_g} = \sum_{i=1}^m \tau_{ig} S_{ig}(\boldsymbol{\alpha}_g) = \sum_{i=1}^m \tau_{ig} \sum_{j \in s_i} \{y_{ij} - h[\eta_{ijg}]\} \, \boldsymbol{w}_{ij}, \quad g = 1, \dots, G,$$

$$S(\boldsymbol{\beta}) = \frac{\partial Q(\boldsymbol{\Phi} \mid \hat{\boldsymbol{\Phi}})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^m \sum_{g=1}^G \tau_{ig} S_{ig}(\boldsymbol{\beta}) = \sum_{i=1}^m \sum_{g=1}^G \tau_{ig} \sum_{j \in s_i} \{y_{ij} - h[\eta_{ijg}]\} \, \boldsymbol{x}_{ij},$$

$$S(\pi_g) = \frac{\partial Q(\boldsymbol{\Phi} \mid \hat{\boldsymbol{\Phi}})}{\partial \pi_g} = \sum_{i=1}^m \left[\frac{\tau_{ig}}{\pi_g} - \frac{\tau_{iG}}{\pi_G} \right], \quad g = 1, \dots, G-1, \tag{S.5}$$

where τ_{ig} represents the posterior probability for the i-th small area to belong to the g-th component of the finite mixture. Given the score functions above, the expectation of the complete data Hessian matrix can be obtained from the following equations:

$$H(\boldsymbol{\alpha}_g, \boldsymbol{\alpha}_g) = \frac{\partial S(\boldsymbol{\alpha}_g)}{\partial \boldsymbol{\alpha}_g'} = - \sum_{i=1}^m \tau_{ig} \sum_{j \in s_i} \left\{ \frac{\partial h[\eta_{ijg}]}{\partial \eta_{ijg}} \right\} \boldsymbol{w}_{ij} \boldsymbol{w}_{ij}', \quad g = 1, \dots, G,$$

$$H(\boldsymbol{\beta}, \boldsymbol{\alpha}_g) = \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\alpha}_g'} = - \sum_{i=1}^m \tau_{ig} \sum_{j \in s_i} \left\{ \frac{\partial h[\eta_{ijg}]}{\partial \eta_{ijg}} \right\} \boldsymbol{x}_{ij} \boldsymbol{w}_{ij}', \quad g = 1, \dots, G,$$

$$H(\boldsymbol{\beta}, \boldsymbol{\beta}) = \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = - \sum_{i=1}^m \sum_{g=1}^G \tau_{ig} \sum_{j \in s_i} \left\{ \frac{\partial h(\eta_{ijg})}{\partial \eta_{ijg}} \right\} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}',$$

$$H(\pi_g, \pi_l) = \frac{\partial S(\pi_g)}{\partial \pi_g} = - \sum_{i=1}^m \left[\frac{\tau_{ig}}{\pi_g^2} \mathbb{1}(g = l) + \frac{\tau_{iG}}{\pi_G^2} \right], \quad g = 1, \dots, G-1.$$

The remaining (non-redundant) elements of the expected complete data Hessian matrix, that is $H(\boldsymbol{\alpha}_g, \boldsymbol{\alpha}_k)$, $H(\boldsymbol{\beta}, \pi_g)$, and $H(\boldsymbol{\alpha}_g, \pi_g)$, are all null due to parameter distinctiveness.

The second component involved in expression (S.4) is the first derivative of the observed data score with respect to model parameter estimates (at convergence). For simplicity, we compute this quantity by numerical differentiation.

According to Friedl and Kauermann (2000), the covariance matrix of parameter estimates may be based on the standard sandwich formula (Huber, 1967; White, 1980):

$$V(\hat{\boldsymbol{\Phi}}) = J(\hat{\boldsymbol{\Phi}})^{-1} V^{\star}(\hat{\boldsymbol{\Phi}}) J(\hat{\boldsymbol{\Phi}})^{-1}, \tag{S.6}$$

where $V^{\star}(\hat{\boldsymbol{\Phi}}) = \sum_{i=1}^{m} S_i(\hat{\boldsymbol{\Phi}}) S_i(\hat{\boldsymbol{\Phi}})'$ denotes the estimate of the covariance matrix of the score $S(\boldsymbol{\Phi})$ and $S_i(\hat{\boldsymbol{\Phi}})$ is the individual score vector. Such an approach helps stabilize the estimate of the observed information and ensures robustness to potential model misspecification.

3 Analytic computation of the bias correction term for the MSE estimation

We provide computational details for the bias correction term required to get a second-order, bias-corrected, MSE estimator of the proposed np-EBP. Let $\boldsymbol{\Phi}_0$ denote the "true" vector of model parameters and let us consider a second-order Taylor expansion of $d^{\mathrm{np}}(\boldsymbol{\Phi})$ around $\boldsymbol{\Phi}_0$ evaluated at $\hat{\boldsymbol{\Phi}}$:

$$d^{\mathrm{np}}(\hat{\boldsymbol{\Phi}}) = d^{\mathrm{np}}(\boldsymbol{\Phi}_0) + \left(\frac{\partial d^{\mathrm{np}}}{\partial \boldsymbol{\Phi}}\right)' \bigg|_{\boldsymbol{\Phi}_0} (\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0) + \frac{1}{2}(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)' \left(\frac{\partial^2 d^{\mathrm{np}}}{\partial \boldsymbol{\Phi}\boldsymbol{\Phi}'}\right)(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0) + o_p(m^{-1}),$$

where d^{np} is a short hand for $d^{\mathrm{np}}(\boldsymbol{\Phi})$. From the expression above, we get

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

where $b^{\mathrm{np}}(\boldsymbol{\Phi})$ denotes a bias correction term which is given by:

$$b^{\mathrm{np}}(\boldsymbol{\Phi}) = \left(\frac{\partial d^{\mathrm{np}}}{\partial \boldsymbol{\Phi}}\right)' \bigg|_{\boldsymbol{\Phi}_0} m E(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0) + \frac{m}{2} E\left[(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)' \left(\frac{\partial^2 d^{\mathrm{np}}}{\partial \boldsymbol{\Phi}\boldsymbol{\Phi}'}\right)\bigg|_{\boldsymbol{\Phi}_0} (\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)\right]$$

$$= b_1^{\mathrm{np}}(\hat{\boldsymbol{\Phi}}) + b_2^{\mathrm{np}}(\hat{\boldsymbol{\Phi}}). \tag{S.7}$$

To derive $b_2^{\mathrm{np}}(\hat{\boldsymbol{\Phi}})$, we proceed as follows:

$$b_2^{\mathrm{np}}(\hat{\boldsymbol{\Phi}}) = \frac{m}{2} E\left[(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)' \left(\frac{\partial^2 d^{\mathrm{np}}}{\partial \boldsymbol{\Phi}\partial \boldsymbol{\Phi}'}\right)\bigg|_{\boldsymbol{\Phi}_0} (\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)\right]$$

$$= \frac{m}{2} E\left\{tr\left[\left(\frac{\partial^2 d^{\mathrm{np}}}{\partial \boldsymbol{\Phi}\partial \boldsymbol{\Phi}'}\right)\bigg|_{\boldsymbol{\Phi}_0} (\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_o)'\right]\right\}$$

$$= \frac{m}{2} tr\left\{E\left[\left(\frac{\partial^2 d^{\mathrm{np}}}{\partial \boldsymbol{\Phi}\partial \boldsymbol{\Phi}'}\right)\bigg|_{\boldsymbol{\Phi}_0} (\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)'\right]\right\}$$

$$= \frac{m}{2} tr \left\{ \left(\frac{\partial^2 d^{\mathrm{mp}}}{\partial \mathbf{\Phi} \partial \mathbf{\Phi}'} \right) \Bigg|_{\mathbf{\Phi}_0} E \left[(\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0)(\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0)' \right] \right\}$$

$$= \frac{m}{2} tr \left\{ \left(\frac{\partial^2 d^{\mathrm{mp}}}{\partial \mathbf{\Phi} \partial \mathbf{\Phi}'} \right) \Bigg|_{\mathbf{\Phi}_0} V(\hat{\mathbf{\Phi}}) \right\}.$$

The first term in the right hand side of expression (S.7) requires a more complex computation. Let us consider a first-order Taylor expansion of the score function $S(\mathbf{\Phi}) = \partial \log L(\mathbf{\Phi})/\partial \mathbf{\Phi}$ around $\mathbf{\Phi}_0$ evaluated at $\hat{\mathbf{\Phi}}$:

$$0 = S(\hat{\mathbf{\Phi}}) = S(\mathbf{\Phi}_0) + \left[\frac{\partial S(\mathbf{\Phi})}{\partial \mathbf{\Phi}} \right] \Bigg|_{\mathbf{\Phi}_0} (\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0) + o_p(||\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0||).$$

Based on the above expression, we have

$$\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0 = J(\mathbf{\Phi}_0)^{-1} S(\mathbf{\Phi}_0) + o_p(||\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0||), \tag{S.8}$$

where $J(\mathbf{\Phi}_0) = -[\partial S(\mathbf{\Phi})/\partial \mathbf{\Phi}]_{\mathbf{\Phi}_0}$ denotes the Fisher information matrix. Then, let us notice that

$$J(\mathbf{\Phi}_0) = I_e(\mathbf{\Phi}_0) + o_p(1),$$

where $I_e(\mathbf{\Phi}_0)$ corresponds to the expected information matrix, so that

$$\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0 = I_e(\mathbf{\Phi}_0)^{-1} S(\mathbf{\Phi}_0) + o_p(||\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0||), \tag{S.9}$$

Let us now consider a second-order Taylor expansion of the k-th element of the score function, $S^k(\mathbf{\Phi})$, around $\mathbf{\Phi}_0$ evaluated at $\hat{\mathbf{\Phi}}$, with $k = 1, \ldots, K$, and K being the number of free model parameters:

$$0 = S^k(\hat{\mathbf{\Phi}}) = S^k(\mathbf{\Phi}_0) + \left[\frac{\partial S^k(\mathbf{\Phi})}{\partial \mathbf{\Phi}} \right] \Bigg|_{\mathbf{\Phi}_0} (\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0)$$

$$+ \frac{1}{2} \left[(\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0)' \left[\frac{\partial^2 S^k(\mathbf{\Phi})}{\partial \mathbf{\Phi} \partial \mathbf{\Phi}'} \right] \Bigg|_{\mathbf{\Phi}_0} (\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0) \right] + o_p(||\hat{\mathbf{\Phi}} - \mathbf{\Phi}||^2),$$

which, according to expression (S.9), may be written as

$$0 = S^k(\hat{\mathbf{\Phi}}) = S^k(\mathbf{\Phi}_0) + \left[\frac{\partial S^k(\mathbf{\Phi}_0)}{\partial \mathbf{\Phi}} \right] \Bigg|_{\mathbf{\Phi}_0} (\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0)$$

$$+ \frac{1}{2} \left[S(\mathbf{\Phi}_0)' I_e(\mathbf{\Phi}_0)^{-1} \left[\frac{\partial^2 S^k(\mathbf{\Phi}_0)}{\partial \mathbf{\Phi} \partial \mathbf{\Phi}'} \right] \Bigg|_{\mathbf{\Phi}_0} I_e(\mathbf{\Phi}_0)^{-1} S(\mathbf{\Phi}_0) \right] + o_p(||\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0||^2).$$

The corresponding second order multivariate Taylor expansion of $S(\mathbf{\Phi})$ around $\mathbf{\Phi}_0$ evaluated at $\hat{\mathbf{\Phi}}$ may be obtained by stacking those for $S^k(\mathbf{\Phi})$, with for $k = 1, \ldots, K$, and is given by

$$0 = S(\hat{\mathbf{\Phi}}) = S(\mathbf{\Phi}_0) + \left[\frac{\partial S(\mathbf{\Phi})}{\partial \mathbf{\Phi}} \right] \Bigg|_{\mathbf{\Phi}_0} (\hat{\mathbf{\Phi}} - \mathbf{\Phi}_0)$$

$$+ \frac{1}{2}\left[S(\boldsymbol{\Phi}_0)'I_e(\boldsymbol{\Phi}_0)^{-1}\left[\frac{\partial^2 S^k(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right]\Bigg|_{\boldsymbol{\Phi}_0} I_e(\boldsymbol{\Phi}_0)^{-1}S(\boldsymbol{\Phi}_0)\right]_{1\leq k\leq K} + o_p(||\hat{\boldsymbol{\Phi}}-\boldsymbol{\Phi}_0||^2).$$

Solving for $(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}_0)$, we obtain

$$\hat{\boldsymbol{\Phi}}-\boldsymbol{\Phi}_0 = I_e(\boldsymbol{\Phi}_0)^{-1}\left\{ S(\boldsymbol{\Phi}_0) + \frac{1}{2}\left[S(\boldsymbol{\Phi}_0)'I_e(\boldsymbol{\Phi}_0)^{-1}\left[\frac{\partial^2 S^k(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right]\Bigg|_{\boldsymbol{\Phi}_0} I_e(\boldsymbol{\Phi}_0)^{-1}S(\boldsymbol{\Phi}_0)\right]_{1\leq k\leq K}\right\} + o_p(||\hat{\boldsymbol{\Phi}}-\boldsymbol{\Phi}_0||^2).$$

Going back to the computation of the bias correction term, we may substitute the above expression into $b_1^{\mathrm{np}}(\hat{\boldsymbol{\Phi}})$ and obtain the following approximation:

$$b_1^{\mathrm{np}}(\hat{\boldsymbol{\Phi}}) = \left(\frac{\partial d^{\mathrm{np}}}{\partial\boldsymbol{\Phi}}\right)'\Bigg|_{\boldsymbol{\Phi}_0} mE(\hat{\boldsymbol{\Phi}}-\boldsymbol{\Phi})$$

$$= \left(\frac{\partial d^{\mathrm{np}}}{\partial\boldsymbol{\Phi}}\right)'\Bigg|_{\boldsymbol{\Phi}_0} m\bigg\{ I_e(\boldsymbol{\Phi}_0)^{-1}E[S(\boldsymbol{\Phi}_0)]$$

$$+ \frac{1}{2}I_e(\boldsymbol{\Phi}_0)^{-1}E\left[S(\boldsymbol{\Phi}_0)'I_e(\boldsymbol{\Phi}_0)^{-1}\left[\frac{\partial^2 S^k(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right]\Bigg|_{\boldsymbol{\Phi}_0} I_e(\boldsymbol{\Phi}_0)^{-1}S(\boldsymbol{\Phi}_0)\right]_{1\leq k\leq K}\bigg\}$$

Clearly, due to $E[S(\boldsymbol{\Phi}_0)] = 0$, the first term in the curly brackets vanishes. The remaining terms are computed as follows

$$b_1^{\mathrm{np}}(\hat{\boldsymbol{\Phi}}) = \left(\frac{\partial d^{\mathrm{np}}}{\partial\boldsymbol{\Phi}}\right)'\Bigg|_{\boldsymbol{\Phi}_0} \frac{m}{2}I_e(\boldsymbol{\Phi}_0)^{-1}E\left\{ tr\left[S(\boldsymbol{\Phi}_0)'I_e(\boldsymbol{\Phi}_0)^{-1}\left[\frac{\partial^2 S^k(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right]\Bigg|_{\boldsymbol{\Phi}_0} I_e(\boldsymbol{\Phi}_0)^{-1}S(\boldsymbol{\Phi}_0)\right]_{1\leq k\leq K}\right\}$$

$$= \left(\frac{\partial d^{\mathrm{np}}}{\partial\boldsymbol{\Phi}}\right)'\Bigg|_{\boldsymbol{\Phi}_0} \frac{m}{2}I_e(\boldsymbol{\Phi}_0)^{-1}E\left\{ tr\left[I_e(\boldsymbol{\Phi}_0)^{-1}\left[\frac{\partial^2 S^k(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right]\Bigg|_{\boldsymbol{\Phi}_0} I_e(\boldsymbol{\Phi})^{-1}S(\boldsymbol{\Phi}_0)S(\boldsymbol{\Phi}_0)'\right]_{1\leq k\leq K}\right\}$$

$$= \left(\frac{\partial d^{\mathrm{np}}}{\partial\boldsymbol{\Phi}}\right)'\Bigg|_{\boldsymbol{\Phi}_0} \frac{m}{2}I_e(\boldsymbol{\Phi}_0)^{-1}E\left\{ tr\left[I_e(\boldsymbol{\Phi}_0)^{-1}\left[\frac{\partial^2 S^k(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}'}\right]\Bigg|_{\boldsymbol{\Phi}_0} I_e(\boldsymbol{\Phi}_0)^{-1}I_e(\boldsymbol{\Phi}_0)\right]_{1\leq k\leq K}\right\}$$

$$- \left(\frac{\partial d^{\mathrm{np}}}{\partial\boldsymbol{\Phi}}\right)'\Bigg|_{\boldsymbol{\Phi}_0} \frac{m}{2}I_e(\boldsymbol{\Phi}_0)^{-1}tr\left\{ I_e(\boldsymbol{\Phi}_0)^{-1}\left[\frac{\partial I_e^k(\boldsymbol{\Phi}_0)}{\partial\boldsymbol{\Phi}'}\right]_{1\leq k\leq K}\right\},$$

where $I_e^k(\boldsymbol{\Phi}_0)$ denotes the k-th row of the expected information matrix and $S(\boldsymbol{\Phi}_0)S(\boldsymbol{\Phi}_0)'$ is approximated by $I_e(\boldsymbol{\Phi}_0)$.

4   Model derivatives for Bernoulli responses

In this section, we provide explicit formulas for computing model derivatives in the case of binary data. As before, let $\tilde{p}_{i(h)}$ denote the np-BP of p_i, conditional on $y_{i\cdot} = h$, with

$y_{i.} = \sum_{j \in s_i} y_{ij}$ and let $\tau_{ig(h)}$ be the posterior probability for the i-th small area to belong to the g-th component of the finite mixture, conditional on $y_{i.} = h$. This latter quantity is defined as

$$\tau_{ig(h)} = \frac{\exp\left[\alpha_g h - \sum_{j \in s_i} \log\left(1 + e^{\eta_{ijg}}\right)\right] \pi_g}{\sum_{l=1}^{G} \exp\left[\alpha_l h - \sum_{j \in s_i} \log\left(1 + e^{\eta_{ijl}}\right)\right] \pi_l}.$$

The first derivative vector of $\tilde{p}_{i(h)}$ with respect to model parameters has elements:

$$\frac{\partial \tilde{p}_{i(h)}^{\text{np-BP}}}{\partial \alpha_g} = \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \frac{\partial}{\partial \alpha_g} \left(\sum_{l=1}^{G} p_{ijl}\, \tau_{il(h)} \right) \right\}$$

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \left[\tau_{ig(h)} \frac{\partial p_{ijg}}{\partial \alpha_g} \right] + \left[\sum_{l=1}^{G} p_{ijl} \frac{\partial \tau_{il(h)}}{\partial \alpha_g} \right] \right\}$$

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \tau_{ig(h)} p_{ijg}(1 - p_{ijg}) + \right.$$

$$\left. + \sum_{l=1}^{G} p_{ijl} \left[\frac{\pi_l f_{il(h)} S_{il(h)}(\alpha_l)}{\sum_{k=1}^{G} \pi_k f_{ik(h)}} \mathbb{1}(l = g) - \frac{\pi_l f_{il(h)} \pi_g f_{ig(h)} S_{ig(h)}(\alpha_g)}{\left[\sum_{k=1}^{G} \pi_k f_{ik(h)} \right]^2} \right] \right\}$$

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \tau_{ig(h)} p_{ijg}(1 - p_{ijg}) + \sum_{l=1}^{G} p_{ijl}\, \tau_{il(h)} S_{il(h)}(\alpha_l) \mathbb{1}(l = g) - \sum_{l=1}^{G} p_{ijl} \tau_{il(h)} \tau_{ig(h)} S_{ig(h)}(\alpha_g), \right\}.$$

with $f_{ig(h)}$ and $S_{ig(h)}(\alpha_g) = h - \sum_{r \in s_i} p_{irg}$ denoting the joint conditional distribution and the score function for the i-th small area, respectively, both conditional on the g-th component of the finite mixture and $y_{i.} = h$.

For the fixed parameters $\boldsymbol{\beta}$, model derivatives are obtained as follows:

$$\frac{\partial \tilde{p}_{i(h)}^{\text{np-BP}}}{\partial \boldsymbol{\beta}} = \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \left(\sum_{l=1}^{G} p_{ijl} \tau_{il(h)} \right) \right\}$$

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \left[\sum_{l=1}^{G} \tau_{il(h)} \frac{\partial p_{ijl}}{\partial \boldsymbol{\beta}} \right] + \left[\sum_{l=1}^{G} p_{ijl} \frac{\partial \tau_{il(h)}}{\partial \boldsymbol{\beta}} \right] \right\}$$

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \left[\sum_{l=1}^{G} \tau_{il(h)} p_{ijl}(1 - p_{ijl}) \boldsymbol{x}_{ij} \right] + \right.$$

$$\left. + \sum_{l=1}^{G} p_{ijl} \left[\frac{\pi_l f_{il(h)} S_{il(h)}(\boldsymbol{\beta})}{\sum_{k=1}^{G} \pi_k f_{ik(h)}} - \frac{\pi_l f_{il(h)} \sum_{u=1}^{G} \pi_u f_{iu(h)} S_{iu(h)}(\boldsymbol{\beta})}{\left[\sum_{k=1}^{G} \pi_k f_{ik(h)} \right]^2} \right] \right\}$$

7

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \sum_{l=1}^{G} \tau_{il(h)} p_{ijl} (1 - p_{ijl}) \boldsymbol{x}'_{ij} + \sum_{l=1}^{G} p_{ijl} \tau_{il(h)} S_{il(h)}(\boldsymbol{\beta}) + \right.$$

$$\left. - \sum_{l=1}^{G} p_{ijl} \tau_{il(h)} \sum_{k=1}^{G} \tau_{ik(h)} S_{ik(h)}(\boldsymbol{\beta}) \right\},$$

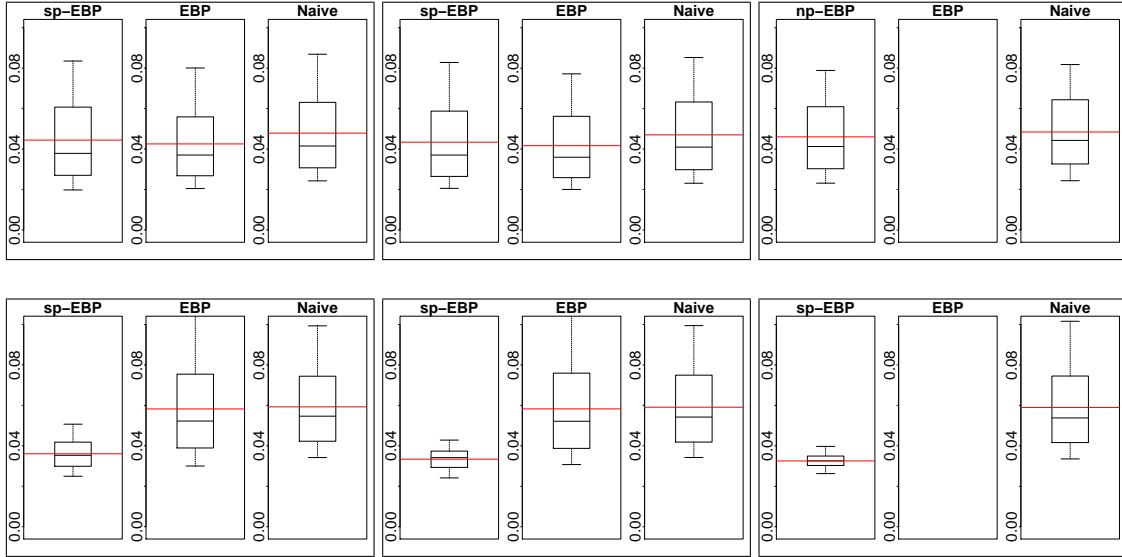with $S_{ig(h)}(\boldsymbol{\beta}) = - \sum_{r \in s_i} p_{irg} \boldsymbol{x}_{ir}$. Last, for the component probabilities π_g, we have:

$$\frac{\partial \tilde{p}_{i(h)}^{\text{np-BP}}}{\partial \pi_g} = \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \sum_{l=1}^{G} p_{ijl} \left[\frac{f_{il(h)}}{\sum_{k=1}^{G} \pi_k f_{ik(h)}} \mathbb{1}(g = l) - \frac{f_{il(h)} \pi_l f_{ig(h)}}{(\sum_{k=1}^{G} \pi_k f_{ik(h)})^2} \right] \right\}$$

$$= \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \sum_{l=1}^{G} p_{ijl} \left[\frac{\tau_{il(h)}}{\pi_l} \mathbb{1}(g = l) - \frac{\tau_{il(h)} \tau_{ig(h)}}{\pi_g} \right] \right\}.$$

5 Simulation results: Mean Absolute Error of the estimators

In this section, we report the distribution of the Mean Absolute Bias (MAE) across small areas for the three estimators under comparison in the simulation study. For each area, the MAE index is computed as follows:

$$\text{MAE}_i = T^{-1} \sum_{t=1}^{T} |\hat{p}_{it}^{\text{Model}} - p_{it}|, \quad i = 1, \dots, m.$$

Figure 1: Distribution of the MAE over areas for the np-EBP, the the EBP, and the Naive approach, for $m = 100$ (left panel), $m = 200$ (central panel), and $m = 500$ (right panel), under Scenario G (upper panel) and Scenario M (lower panel).

References

Friedl, H. and Kauermann, G. (2000). Standard errors for EM estimates in generalized linear models with random effects. *Biometrics*, 56:761–767.

Huber, P. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the 5th Berkeley Symposium*, volume 1, pages 221–233.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44:226–233.

Oakes, D. (1999). Direct calculation of the information matrix via the em. *Journal of the Royal Statistical Society: Series B*, 61:479–482.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and direct test for heteroskedasticity. *Econometrica*, 48:817–838.