Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 521-550
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

A Comparison of Hierarchical Bayesian Models for Small Area Estimation of Counts

Matilde Trevisani, Nicola Torelli

Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste, Trieste, Italy
Email: matildet@deams.units.it, nicolat@deams.units.it

How to cite this paper: Trevisani, M. and Torelli, N. (2017) A Comparison of Hierarchical Bayesian Models for Small Area Estimation of Counts. *Open Journal of Statistics*, **7**, 521-550.
https://doi.org/10.4236/ojs.2017.73036

Received: April 1, 2017
Accepted: June 26, 2017
Published: June 29, 2017

Abstract

Small area estimation (SAE) tackles the problem of providing reliable estimates for small areas, *i.e.*, subsets of the population for which sample information is not sufficient to warrant the use of a direct estimator. Hierarchical Bayesian approach to SAE problems offers several advantages over traditional SAE models including the ability of appropriately accounting for the type of surveyed variable. In this paper, a number of model specifications for estimating small area counts are discussed and their relative merits are illustrated. We conducted a simulation study by reproducing in a simplified form the Italian Labour Force Survey and taking the Local Labor Markets as target areas. Simulated data were generated by assuming population characteristics of interest as well as survey sampling design as known. In one set of experiments, numbers of employment/unemployment from census data were utilized, in others population characteristics were varied. Results show persistent model failures for some standard Fay-Herriot specifications and for generalized linear Poisson models with (log-)normal sampling stage, whilst either unmatched or nonnormal sampling stage models get the best performance in terms of bias, accuracy and reliability. Though, the study also found that any model noticeably improves on its performance by letting sampling variances be stochastically determined rather than assumed as known as is the general practice. Moreover, we address the issue of model determination to point out limits and possible deceptions of commonly used criteria for model selection and checking in SAE context.

Keywords

Small Area Estimation, Hierarchical Bayesian Models, Non-Normal Sampling Stage, Unmatched Models

1. Introduction

In recent years, small area estimation (SAE) has emerged as an important area of

statistics as private and public agencies try to extract the maximum information from sample survey data. Sample surveys are generally designed to provide estimates of characteristics of interest for large areas or domains. However, governments are more and more interested in obtaining statistics for smaller geographical areas such as counties, districts or census divisions, or smaller demographic subsets such as specific age-sex-race subgroups. These domains are called small areas. SAE concerns statistical techniques aimed at producing estimates of characteristics of interest for small areas or domains. A review of SAE methods is in [1] [2] [3]. The simplest approach is to consider direct estimators, that is estimating the variable of interest using the domain-specific sample data. However, it is well known that the domain sample sizes are rarely large enough to support reliable and accurate direct estimators since budget and other constraints usually prevent drawing adequate samples from each of the small areas. When direct estimates are unreliable (or even non computable), a major direction considers the use of explicit small area models that "borrow strength" from related areas across space and/or time or through auxiliary information which is supposed to be correlated to the variable of interest. Explicit models can be classified into two categories: 1) area level models and 2) unit level models. They can be estimated by adopting several alternative approaches and one of these has been the hierarchical Bayesian (HB) paradigm. However, applications of HB models to SAE, though growing [1], still are quite a few. Moreover, they have mainly focused on continuous variables. To date, there is no thorough discussion on what is the most appropriate nonlinear specification of area level models when small area estimates are needed for discrete or categorical variables.

In this paper, we focus on HB area level models for producing small area estimates of counts. In the literature, Bayesian specifications commonly derive from classical models for SAE, *e.g.* the Fay-Harriot model [4], or more properly consider either a generalized linear Poisson model [5] [6] or a multinomial logit model [7]. [8] presented a Normal-logNormal model within the class of the so called unmatched models. Following the HB way of thinking, we independently proposed a Normal-Poisson-logNormal model arguing that this unmatched form could be more appropriate for taking explicitly into account the nature of the variable of interest [9] [10]. An application of this model, originally extended to enable the use of multiple data sources possibly misaligned with small areas, is in [11]. Moreover, we suggested a Gamma-Poisson-logNormal model, that introduces a nonnormal sampling error stage, and advocated a natural extension of the several above specifications by letting sampling variances be stochastically determined rather than fixed to design estimates as is the general practice [12].

For completeness, we mention [13] who compare four HB small area models for producing state estimates of proportions: the original proposal consists of a Beta sampling stage with a logit linking model. Still in a Bayesian context, [14] [15] [16] handle the problem of unknown sampling variances.

Under appropriate conditions each of these models may have some merits and whether it is appropriate depends on various circumstances like size of the areas,

availability of good explanatory variables at area level, accuracy of sampling variance estimates, etc. Practical use of HB models has been boosted by the availability of software that implements Markov chain Monte Carlo (MCMC) simulations so that model estimation can be straightforward and relatively easy. Room is left for investigating the peculiarity of different specifications and for identifying criteria and guidelines for choosing among alternative Bayesian specifications.

Purpose of the present work is comparing alternative HB area level models for SAE of counts. Comparison is made first on a theoretical side and then by a simulation study. This last is aimed to reproduce one of the most relevant instances where SAE has proven its potential, *i.e.* estimation of labour force statistics at a local level finer than the survey planned domains. The specific framework for the simulation is estimation of the number of unemployed (employed) within Local Labor Markets (LLMs, *i.e.* areas including a group of municipalities which share the same labor market conditions). In most developed countries, the major source of information on the labor market is a Labor Force Survey (LFS). In Italy, LFS design has been planned so that reliable (design-based) estimates of given precision can be obtained for regional and provincial quantities, quarterly and yearly respectively. LLMs are a finer regional partition and the sample sizes associated with such minor domains result inadequate to allow for stable (design-based) estimates [12]. Simulated data were generated by assuming population characteristics of interest as well as sampling survey design as known. In one set of experiments, the actual LLM unemployment (employment) figures from census data were utilized, in others population characteristics were varied (by changing the type of distribution symmetry). Still, LLM survey sample sizes were either maintained fixed at actual LFS values or given different values. The sampling design was kept quite simple across all studies, moreover, synthetic estimates comprise the sole source of auxiliary information incorporated into model framework. Although the core of models is quite basic, it is worth noting that it is the framework actually used in Italy to produce totals of unemployed for LLMs since late nineties.

In summary, this paper, through a number of HB area level models for SAE of totals, compares three broad classes: matched, unmatched and nonnormal sampling stage models. A first comparison, on the basis of a design-based simulation from census data, is made by assuming known sampling variances. Secondarily, once detected specific model failures in terms of bias, accuracy and reliability, this hypothesis is abandoned and minor ameliorations are furtherly carried out to models. The comparison is repeated also by varying the finite-population simulation. Moreover, we address the issue of model determination to point out limits and possible deceptions of commonly used criteria for model selection and checking in SAE context, namely, the *deviance information criterion* (DIC) and the posterior predictive p-value (PPp). In the sequel, Section 2 presents the alternative HB models at comparison specifying motivations behind their introduction, Section 3 describes the simulation study, discusses the results and pro-

poses a number of model refinements, finally, Section 4 contains some concluding remarks.

2. HB Models for Small Area Estimation with Count Data

2.1. General Framework

The core of classical small area models consists of linear mixed models. The basic small area model for area level data is the Fay-Herriot model [4] which consists of an area-linking model, e.g. $\theta_i = \boldsymbol{x}_i^{\mathrm{T}}\beta + \nu_i$, $\nu_i \sim N(0,\tau)$ (hereafter $\sim$ stays for "independently distributed as"), coupled with a sampling error model, e.g. $\hat{\theta}_i = \theta_i + \epsilon_i$, $\epsilon_i \mid \theta_i \sim N(0,\sigma_i)$, with θ_i, $\hat{\theta}_i$ and $\boldsymbol{x}_i$ denoting respectively characteristic of interest, survey estimate (when available) and possible auxiliary data, for each area *i*. The linking model is merely a linear model with mixed coefficients: fixed coefficients β, accounting for $\boldsymbol{x}$ effects valid for the entire population, and random area-specific effects ν_i. Sampling variances σ_i are usually assumed to be known; parameters β and τ have to be estimated.

Under a HB approach mixed models are stage-wise specified; in particular, the Fay-Herriot model gives rise to the following specification:

$$\hat{\theta}_i \mid \theta_i, \sigma_i \sim N(\theta_i, \sigma_i) \tag{1}$$

$$\theta_i \mid \beta, \tau \sim N\left(\boldsymbol{x}_i^{\mathrm{T}}\beta, \tau\right) \tag{2}$$

$$(\beta, \tau) \sim p(\beta, \tau). \tag{3}$$

Sampling and linking models, (1) and (2), are unchanged, whereas an additional hyperprior stage, (3), is required within a full HB approach.

Notwithstanding a (proper) informative prior distribution on the hyperparameters would be appropriate for a full Bayesian analysis, for ignorance or because we want inference to be driven solely by the data at hand, noninformative priors are often used (this is still mainstream practice in SAE analyses). In this case, to avoid posterior density to be improper, diffuse yet proper (otherwise said, weakly-informative) priors are routinely assumed. Such a choice—which however needs a careful sensitivity analysis especially when models are barely identified—generally ensures a valid inference.

The classical FH specification may be defective either because it: 1) assumes the sampling errors $\epsilon_i = \hat{\theta}_i - \theta_i$ as normal or because 2) sets a linear link $\theta_i = x_i^{\mathrm{T}}\beta + \nu_i$ directly between θ_i and x_i. Indeed, θ_i's are counts (*i.e.* positive-integer valued variates), moreover, a non-identity link $g(\theta_i) = x_i^{\mathrm{T}}\beta + \nu_i$ may be more appropriate when the predicted variable θ_i is non-continuous and/or the covariates x_i are thought to produce a non-additive effect on it. Of course, the Normal-Normal model (1 - 3) owes its popularity to being in general computationally convenient and inferentially tractable by classical estimation methods. On the other hand, in a HB approach inference is straightforward and computationally feasible thanks to MCMC methods, the most popular computing tools in Bayesian practice. The flexibility inherent to HB modeling and its computational tractability allow the choice of more realistic models for SAE

problems than alternative approaches could never envision.

2.2. Alternative HB Models

We define alternative HB area level models for a SAE problem with count data. In the following, ${}_S\hat{\theta}_i$ indicates a synthetic estimate for small area i, while θ_i, $\hat{\theta}_i$ and x_i have the same meaning as in Section 2.1.

Six model specifications are theoretically conceivable when the parameter of interest is the small area total (such as the number of unemployed or employed in our application). They are:

the Normal-Normal model (NN)

$$\hat{\theta}_i \mid \theta_i, \sigma_i \sim N(\theta_i, \sigma_i) \tag{4}$$

$$\theta_i \mid \beta, \tau \sim N({}_S\hat{\theta}_i + \alpha + \beta x_i, \tau), \tag{5}$$

the log-Normal-Normal model (FH)

$$\log(\hat{\theta}_i) \mid \theta_i, \tilde{\sigma}_i \sim N(\log(\theta_i), \tilde{\sigma}_i) \tag{6}$$

$$\log(\theta_i) \mid \beta, \tau \sim N(\log({}_S\hat{\theta}_i) + \alpha + \beta x_i, \tau), \tag{7}$$

the Normal-logNormal model (YR)

$$\hat{\theta}_i \mid \theta_i, \sigma_i \sim N(\theta_i, \sigma_i) \tag{8}$$

$$\log(\theta_i) \mid \beta, \tau \sim N(\log({}_S\hat{\theta}_i) + \alpha + \beta x_i, \tau), \tag{9}$$

the Normal-Poisson-logNormal model (NPlN)

$$\hat{\theta}_i \mid \theta_i, \sigma_i \sim N(\theta_i, \sigma_i) \tag{10}$$

$$\theta_i \mid \mu_i \sim Poisson(\mu_i) \tag{11}$$

$$\log(\mu_i) \mid \beta, \tau \sim N(\log({}_S\hat{\theta}_i) + \alpha + \beta x_i, \tau), \tag{12}$$

the Poisson-logNormal model (PlN)

$$\hat{\theta}_i \mid \mu_i \sim Poisson(\mu_i) \tag{13}$$

$$\log(\mu_i) \mid \theta_i, \tilde{\sigma}_i \sim N(\log(\theta_i), \tilde{\sigma}_i) \tag{14}$$

$$\log(\theta_i) \mid \beta, \tau \sim N(\log({}_S\hat{\theta}_i) + \alpha + \beta x_i, \tau), \tag{15}$$

and the Gamma-Poisson-logNormal model (GPlN)

$$\hat{\theta}_i \mid \mu_i \sim Poisson(\mu_i) \tag{16}$$

$$\mu_i \mid \theta_i, a_i \sim Gamma(a_i, a_i/\theta_i) \tag{17}$$

$$\log(\theta_i) \mid \beta, \tau \sim N(\log({}_S\hat{\theta}_i) + \alpha + \beta x_i, \tau). \tag{18}$$

For a fully Bayesian specification, the basis of every model hierarchy consists of an hyperprior stage, like as (3), which we leave generically expressed since the discussion is focused on analysing how closer-to-data stages can be variously specified.

Either the NN (4 - 5) or the FH (6 - 7) specifications consist in the Fay-Herriot model with $\hat{\theta}_i$ and θ_i both untransformed (NN) or both log-transformed (FH). NN and FH are both members of the so-called *matched* models [7] in the sense that sampling and linking models can be combined to produce a linear mixed model. Namely, once a suitable function $g(\cdot)$ is chosen to relate the parameter of interest to auxiliary variables through a Normal model ($g(\theta_i) = \mathbf{x}_i\beta + v_i$, with v_i normal variates), also direct estimates are accordingly transformed in the sampling model ($g(\hat{\theta}_i) = g(\theta_i) + e_i$, again with normal errors e_i) in order to combine the two equations into a single linear model (from the foregoing equations, $g(\hat{\theta}_i) = \mathbf{x}_i\beta + v_i + e_i$). Small area estimates are then obtained by inverting $g(\cdot)$.

On the other side, YR (8 - 9) and NPlN (10 - 12) forms are both *unmatched* models in the sense that stage-1 and stage-2 models cannot be combined into a single equation model. You and Rao [8] proposed the YR form when *g* is specifically the log-function, but their arguments can be generalized to any nonlinear function of θ_i. They warned that customary hypotheses on sampling errors e_i may be quite questionable when *g* is nonlinear and area sample size is small (in particular, they refer to the unbiasedness assumption $E(e_i \mid \theta_i) = 0$ and the Taylor approximation ordinarily set for the variance, *i.e.* $var(e_i \mid \theta_i) \approx \{g'(\theta_i)\}^2 \sigma_i$ with $\sigma_i = var(\epsilon_i \mid \theta_i)$). Thus their advice is to let sampling model $\hat{\theta}_i = \theta_i + \epsilon_i$ be unaltered so that condition $E(\epsilon_i \mid \theta_i) = 0$ (*i.e.* design-unbiasedness of $\hat{\theta}_i$) holds and, moreover, the design-variance $v_p(\hat{\theta}_i)$, which is taken as known, can be imputed to the sampling variance σ_i. (Note that we use ϵ_i instead of e_i whenever direct estimates $\hat{\theta}_i$ are left untransformed in the sampling model.) Finally, they choose the HB approach since inference on non-standard specifications may not be feasible by means of classical estimation methods. We note also that with specific reference to FH (6 - 7) model, survey information may be partly wasted, in that transformed direct estimates $\log(\hat{\theta}_i)$ are not defined when $\hat{\theta}_i = 0$. Thus, missing data originate both from areas with null direct estimates (which may not be so rare when area sample size is small) and, as usually, from non-sampled small areas.

Trevisani and Torelli [9] [10] proposed the NPlN model which derives from building, stage by stage, a HB structure suited for SAE problems with count variables. Thereby, first stage (10) is modeled, similarly to YR (8), in terms of untransformed direct estimates $\hat{\theta}_i$, so that both design-unbiasedness and $\sigma_i = v_p(\hat{\theta}_i)$ assignment hold. Then, noting that response variable θ_i of the linking regression analysis is a count, second (11) and third (12) stages consist of a standard Poisson-log-Normal model.

Indeed, this choice has been (determined not only by the type of θ_i variable but also) borrowed from the extensive literature on disease mapping. In such an area, disease counts are modelled as Poisson variates with mean $\mu_i = \rho_i \times E_i$ where ρ_i is the relative risk in area *i* and E_i the expected count. A regression equation is then usually set on a logarithmic scale, $\log(\rho_i) = \mathbf{x}_i\beta + v_i$, to opportunely accommodate for a linear predictor $\mathbf{x}_i\beta$ and any random effect v_i

(which is customarily assumed to be normally distributed). Note that E_i may be taken as random but is usually taken as known and (for including any information prior to inference,) enters the regression equation as an offset. In SAE context, E_i can be set at known synthetic estimate ${}_S\hat{\theta}_i$ and a model-like-disease mapping is then fitted to counts θ_i.

To some readers the specification of $\hat{\theta}_i$ might appear inconsistent with that of θ_i: $\hat{\theta}_i$ are generated from a continuous distribution over the real line while θ_i are drawn from a discrete distribution like the Poisson one. Yet, at a second insight, one realizes that it is the sampling model for $\hat{\theta}_i$'s to be really inconsistent with the integer type of the variable of interest θ_i. Nonetheless, we decided to let sampling model be specified as standardly is in SAE literature (indeed, $\hat{\theta}_i$ might be non-integer since it derives from an estimation process not necessarily constrained to produce integer values, though definitely it cannot be negative) while we originally assumed θ_i to be generated from a Poisson model. It is superfluous to remark that there is no inconsistency in restricting the parameter space of a Normal distribution mean (θ_i) to the sole integers. Moreover, there is no need of discretizing $\hat{\theta}_i$: θ_i naturally arises as an integer for being generated as a Poisson variate. (Incidentally, Bugs software allows specifying a Poisson prior for any continuous quantity.) We apologize to those readers not in need of clarification for such a "byzantine" digression. Even more, if one considers that the core feature for which we turned to a Poisson model was its variance-proportional-to-mean property; though nonnegativity and discreteness of its sampling space are undoubted advantages, they are not so urgent as to require the replacement of the standardly assumed Normal model.

Finally, the two last models, PlN (13 - 15) and GPlN (16 - 18), are characterized by non-normal first stage specifications. The characteristic of interest is a count, thus a canonical Poisson model is set from the very first stage. The PlN specification is a standard generalized linear mixed model for count variates—the $\hat{\theta}_i$s—here written in a form suitable to SAE problems. In particular, the log-Normal stage (14) depends on two sources of random variability: the sampling error, e_i, and the random effect, ν_i. Again, the sampling error variance is set according to the Taylor approximation defined above, *i.e.* $\tilde{\sigma}_i = \theta_i^{-2}\sigma_i$. In order to remedy possible failures implied by the Taylor approximation, GPlN model sets a Gamma distribution for inflating the Poisson variance to the extent of the sampling variability, *i.e.* $\mu_i = \theta_i\gamma_i$ where $\gamma_i \sim \Gamma(a_i, a_i/\theta_i)$ is conveniently specified so that both design unbiasedness $\left(E\left(\hat{\theta}_i \mid \theta_i\right) = \theta_i\right)$ and design variance imputation $\left(var\left(\hat{\theta}_i \mid \theta_i\right) = v_p\left(\hat{\theta}_i\right)\right)$ hold. An ordinary log-Normal model follows at the linking stage (18).

3. A Simulation Study

3.1. Simulation Plan and Performance Measures

To compare the performance of the six HB models, we carried out a simulation study based on reproducing a simplified LFS in the "world" of 1991 Veneto

census data. That is, we generate each LLM population with employment and unemployment rates (% over population N_i) fixed at p_i^e and p_i^u values as derived from census data. Hence, a sample of size $n_i = r_i N_i/1000$, with sampling fraction r_i (over population) fixed at 1999 LFS value, was repeatedly selected by simple random sampling without replacement (SRSWOR) from each LLM. (Table 7 in Section 5 presents census, sampling design as well as one sample of simulated LFS data that were used in our study.)

To strengthen the results of our study, a further series of simulations was carried out by generating over the small areas set considered above three synthetic populations having, in particular, a positively asymmetric (Table 8), an essentially symmetric (Table 9) and a negatively asymmetric (Table 10) distribution of numbers of unemployed (in all cases keeping fixed the mean to the historical value of 3.35% of the LLM unemployment rates). Moreover, as regards the LFS simulation, non-sampled areas were randomly selected (keeping fixed the number to 14 over the total 51 small areas) and sample sizes were varied from $n_i = 100$ to 300 up to 500, $\forall i$ (*i.e.* balanced designs), yet with sampling fraction r_i kept fixed to the realized historical value of 4/5‰.

From each simulated survey sample, the following estimators were calculated: (a) (poststratified) direct estimator $\hat{\theta}_i = N_i \hat{p}_i$, with $\hat{p}_i = y_i/n_i$, y_i being the observed number of employed/unemployed over n_i sampled units; (b) synthetic estimator ${}_S\hat{\theta}_i = N_i \hat{p}$, where $\hat{p} = y/n$, $y = \sum_i y_i$ and $n = \sum_i n_i$; (c) coefficient of variation (*cv*) of direct estimator was estimated by $\widehat{cv}_i = \sqrt{\hat{v}_p\left(\hat{\theta}_i\right)}/\hat{\theta}_i$ with $\hat{v}_p\left(\hat{\theta}_i\right) = N_i\left(N_i - n_i\right)\hat{p}_i\left(1-\hat{p}_i\right)/\left(n_i - 1\right)$ being the sampling design variance estimated according to a SRSWOR scheme; lastly; (d) *cv* of synthetic estimator was estimated by ${}_S\widehat{cv}_i = \sqrt{\widehat{mse}\left({}_S\hat{\theta}_i\right)}/{}_S\hat{\theta}_i$ with mean squared error of synthetic estimator, $mse\left({}_S\hat{\theta}_i\right)$, computed by the method of Marker [17]. (In Table 7, the suffix e/u denotes whether the quantity is related to employment or unemployment.)

Variances σ_i of sampling models (1)—when assumed as known—are set to $\hat{\sigma}_i = \hat{v}_p\left(\hat{\theta}_i\right)$ for sampled areas with non-null direct estimates $\hat{\theta}_i$. Whilst, for those areas having $y_i = 0$ with $n_i < 0$, we impute the synthetic proportion of employed/unemployed, that is we replace $\hat{p}_i$ with $\hat{p}$, in the formula for $\hat{v}_p\left(\hat{\theta}_i\right)$. Lastly, for those areas having $n_i = 0$, missing data $\hat{\theta}_i$ are considered as latent variables according to a Bayesian approach. Various alternatives of value-imputation for the associated σ_i's are then sensible: e.g. in the trials where synthetic estimates are given as initial values to MCMC chains of latent (missing) $\hat{\theta}_i$, the associated σ_i's were consistently fixed at Marker's estimate of $mse\left({}_S\hat{\theta}_i\right)$.

To compare the performance of $\hat{\theta}_i^{HB}$ estimators, based on the formerly introduced HB models, we compute a series of measures out of $R = 100$ simulated samples. (We came to this number after found out that results kept fairly stable even stopping at 50 replications.) Let $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \cdots\right)$ be the vector of the overall direct estimates, then $\hat{\theta}_i^{HB}$ is defined as the posterior mean $E\left(\theta_i \mid \hat{\boldsymbol{\theta}}\right)$ of small area parameter θ_i in the considered HB model. For both FH and PlN

models, where θ_i is modelled on the log-scale, a definition of $\hat{\theta}_i^{HB}$ has to be properly settled. The one (ones) that we adopt are described in Section 3.2. Computations have been made by means of arm package which allows running Bugs, the best-known Bayesian inference software, from within the general statistical package R. Same standard (proper) non-informative hyperpriors are chosen for every model under comparison.

The quantities calculated, for each area *i*, are:

relative bias (*rb*) and the absolute relative bias (*arb*),

$$rb_i = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{ir}^{HB}/true_i - 1\right), \quad arb_i = \frac{1}{R}\left|\sum_{r=1}^{R}\left(\hat{\theta}_{ir}^{HB}/true_i - 1\right)\right|,$$

both measuring the bias of the estimator ($\hat{\theta}_{ir}^{HB}$ denotes the value of the considered estimator of θ_i for the *r*th simulation, $true_i$ is the census number of employed/unemployed);

absolute relative error (*are*) and the relative root mean squared error (*rE*)

$$are_i = \frac{1}{R}\sum_{r=1}^{R}\left|\hat{\theta}_{ir}^{HB}/true_i - 1\right|, \quad rE_i = \sqrt{\widehat{mse}\left(\hat{\theta}_i^{HB}\right)}\Big/true_i$$

with $\widehat{mse}\left(\hat{\theta}_i^{HB}\right) = \sum_{r=1}^{R}\left(\hat{\theta}_{ir}^{HB} - true_i\right)^2/R$, both relating to estimator accuracy;

efficiency (*eff*)

$$eff_i = \frac{1}{R}\sum_{r=1}^{R}\widehat{cv}\left(\hat{\theta}_{ir}^{HB}\right)$$

where expressions for $\widehat{cv}\left(\hat{\theta}_i\right) = \widehat{cv}_i$ and $\widehat{cv}\left({}_S\hat{\theta}_i\right) = {}_S\widehat{cv}_i$ have already been given whereas $\widehat{cv}\left(\hat{\theta}_i^{HB}\right)$ is measured by $\widehat{sd}\left(\theta_i \mid \hat{\boldsymbol{\theta}}\right)/\hat{\theta}_i^{HB}$ with $sd\left(\theta_i \mid \hat{\boldsymbol{\theta}}\right)$ denoting the posterior standard deviation of θ_i; lastly

reliability (*rel*)

$$rel_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\widehat{cv}^2\left(\hat{\theta}_{ir}^{HB}\right)}\Big/rE_i$$

which is intended to measure how much reliable is a standardly used indicator of estimator efficiency (*cv*) when related to a comparable measure (*rE*) yet based on known *true* values.

The degree of *cv* reduction with respect to direct estimators is largely used for selecting the best (model- or design-based) estimator. Nevertheless, it consists of one aspect (essentially depending on the shrinkage degree of direct estimates to the mean level ${}_S\hat{\theta}_i + \alpha + \beta x_i$ or ${}_S\hat{\theta}_i \exp\left(\alpha + \beta x_i\right)$) which is not necessarily the preferential one; furtherly, estimation goal is not unique (is triple according to [1] and [17]). Regardless, we need to know to what extent such an indicator is reliable for measuring the degree of uncertainty about the provided estimates.

Comparison among different HB models is finally completed by looking at some standard criteria for model selection in a Bayesian framework, namely we consider: 1) a likelihood based criterion, the DIC and 2) a predictive distribution-based method, the PP*p*. The DIC is based upon posterior expectation of the deviance which is defined as $D\left(\boldsymbol{\theta}\right) = -2\log L\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}\right) + 2\log f\left(\hat{\boldsymbol{\theta}}\right)$ for a chosen

likelihood $L\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}\right)$ and for some standardizing function f, where $\hat{\boldsymbol{\theta}}$ serve as data in small area context. In Table 4 and similar ones, $\bar{D}=E\left(D\left(\boldsymbol{\theta}\right)|\hat{\boldsymbol{\theta}}\right)$, $D\left(\bar{\boldsymbol{\theta}}\right)=D\left(E\left(\theta|\hat{\boldsymbol{\theta}}\right)\right)$, $p_D=\bar{D}-D\left(\bar{\boldsymbol{\theta}}\right)$, $\mathrm{DIC}=\bar{D}+p_D$, according to the definition given by [18] in their article first proposing this criterion (and the notation therein). PPp is defined, for normal sampling stage models (NN, FH, YR, NPlN), as

$$\mathrm{PP}p=P\left\{\sum_i\left(y_{i,rep}-\theta_i\right)^2/\sigma_i>\sum_i\left(\hat{\theta}_i-\theta_i\right)^2/\sigma_i\,|\,\hat{\boldsymbol{\theta}}\right\} \tag{19}$$

whereas, for the remaining ones (PlN, GPlN), as

$$\mathrm{PP}p=P\left\{\sum_i\left(y_{i,rep}-\mu_i\right)^2/\mu_i>\sum_i\left(\hat{\theta}_i-\mu_i\right)^2/\mu_i\,|\,\hat{\boldsymbol{\theta}}\right\}, \tag{20}$$

with $y_{i,rep}$ indicating hypothetical replicated data under the assumed model and where the reference distribution is derived from the posterior distribution of $\left(\boldsymbol{y}^{rep},\boldsymbol{\theta}\right)$ or, in the last case, $\left(\boldsymbol{y}^{rep},\boldsymbol{\mu}\right)$. According to these criteria, the smaller the DIC or $\left|\mathrm{PP}p-0.5\right|$ the better the model.

3.2. First Findings

Table 1 and Table 2 display averages, over the I small areas, of *rb*, *arb*, *are*, *rE*, *eff* and *rel* (%), for design-based as well as HB model-based estimators, from simulated data (based on the real population) on employment and unemployment respectively. The "average" is measured in terms of the mean and the median (first and second columns respectively for each measure). In Table 2 and related ones which follow, averages were computed by excluding non-sampled areas as well (rows named as *non-na*).

In Table 1 and similar ones, three ways have been adopted to transform $\eta_i=\log\left(\theta_i\right)$ in the FH model back to the original scale. The "classical way" (FH) provides the required estimator by exponentiating $\hat{\eta}_i^{HB}=E\left(\eta_i\,|\,\hat{\boldsymbol{\theta}}\right)$ *i.e.* as

Table 1. Comparison between design-based and HB model-based estimators (simulated data on employment).

	bias				accuracy				efficiency		reliability	
estimator	$\overline{rb}$	rb^M	$\overline{arb}$	arb^M	$\overline{are}$	are^M	$\overline{rE}$	rE^M	$\overline{eff}$	eff^M	$\overline{rel}$	rel^M
Synthetic	0.3	−0.1	4.1	3.4	4.2	3.4	4.4	3.6	4.3	4.3	175	141
Direct	−0.2	0.1	0.9	0.8	6.2	5.5	7.8	6.8	7.8	7.1	102	102
NN	−0.3	−1.0	3.9	3.4	4.7	3.7	5.2	3.9	3.5	1.8	69	55
FH	1.0	0.7	3.4	2.6	3.8	2.8	4.2	3.3	0.3	0.3	9	9
FH(2)	1.1	0.7	3.4	2.7	3.8	2.8	4.2	3.3	3.0	3.6	93	93
FH(3)	1.1	0.7	3.4	2.7	3.8	2.8	4.2	3.3	3.1	3.6	94	93
YR	0.3	0.0	3.2	2.4	3.7	2.8	4.2	3.4	3.4	4.0	99	96
NPlN	0.3	0.0	3.1	2.4	3.7	2.8	4.2	3.5	3.6	4.1	100	101
PlN	1.1	0.7	3.2	2.2	3.7	2.8	4.2	3.4	3.7	4.3	106	111
GPlN	0.0	0.0	3	2.4	3.7	2.8	4.2	3.4	3.6	4.1	100	100

Table 2. Comparison between design-based and HB model-based estimators (simulated data on unemployment).

	bias				accuracy				efficiency		reliability	
estimator	$\overline{rb}$	rb^M	$\overline{arb}$	arb^M	$\overline{are}$	are^M	$\overline{rE}$	rE^M	$\overline{eff}$	eff^M	$\overline{rel}$	rel^M
Synthetic	10	14	25	24	26	24	26	25	26	26	179	114
non-na	11	14	27	29	27	29	27	30	26	26	169	93
Direct	−2	−1	3	2	29	28	36	34	39	36	110	110
NN	−4	−4	21	17	25	25	28	28	28	20	106	82
non-na	1	0	18	16	24	26	27	28	23	20	99	86
FH	10	10	19	16	21	18	24	20	2	2	12	10
non-na	10	10	18	16	21	18	24	20	3	3	13	10
FH(2)	11	12	20	17	22	17	24	20	17	19	85	81
non-na	12	12	19	17	22	17	25	21	20	21	98	91
FH(3)	11	12	20	17	22	17	24	20	17	19	86	80
non-na	12	12	19	17	22	17	25	21	20	22	98	89
YR	−8	−6	14	11	18	16	21	20	20	22	105	109
non-na	−7	−6	12	10	18	17	21	21	24	24	118	110
NPlN	−7	−7	14	11	18	16	21	20	20	23	109	109
non-na	−7	−7	12	9	18	17	21	21	24	24	117	110
PlN	8	9	18	14	22	17	25	21	19	21	89	87
non-na	10	9	17	14	22	17	25	22	22	23	102	92
GPlN	−6	−4	13	11	19	17	22	21	20	22	96	102
non-na	−5	−4	11	9	19	17	23	22	24	23	108	108

$\hat{\theta}_i^{HB} = \exp\left(\hat{\eta}_i^{HB}\right)$; the derived *cv* is necessarily given by multiplying $cv\left(\hat{\eta}_i^{HB}\right) = sd\left(\eta_i \mid \hat{\boldsymbol{\theta}}\right) \Big/ \hat{\eta}_i^{HB}$ by $\hat{\theta}_i^{HB}$, which, though, shows to be clearly inadequate as a measure of efficiency (in Table 2, $\overline{rel} = rel^M = 9\%$, in Table 1, $\overline{rel} = 12\%$, $rel^M = 10\%$, and the size of *mse* coverage keeps being around 10% also in the subsequent unemployment simulation experiments). The "Bayesian way" (referred to as FH(2)) provides $\hat{\theta}_i^{HB}$ as posterior expectation of the back-transformed parameter $\theta_i = \exp\left(\eta_i\right)$. Lastly, FH(3) estimator stems from using the classical formulas for deriving expectation and standard deviation of a lognormal variable *i.e.* by computing $\hat{\theta}_i^{HB} = \exp\left(\hat{\eta}_i^{HB} + var\left(\eta_i \mid \hat{\boldsymbol{\theta}}\right)\Big/2\right)$ and $var\left(\theta_i \mid \hat{\boldsymbol{\theta}}\right) = \exp\left(2\left(\hat{\eta}_i^{HB} + var\left(\eta_i \mid \hat{\boldsymbol{\theta}}\right)\right)\right) - \exp\left(2\hat{\eta}_i^{HB} + var\left(\eta_i \mid \hat{\boldsymbol{\theta}}\right)\right)$. (In each case, ordinary $cv\left(\hat{\theta}_i^{HB}\right)$ formula is then used for *cv* computation.) PlN $\hat{\theta}_i^{HB}$ is obtained solely by the Bayesian way.

It is worth noting that any discrepancy in percent numbers of Table 1 can be appreciated just at one decimal digit. In fact, employment rates range from 34.7% to 44.5% (Table 7) whence cv_i's range from 13.8/4.3% to 11.2/3.5% with n_i = 100/1000 (Figure 1). Then, to assess model performance we will focus especially

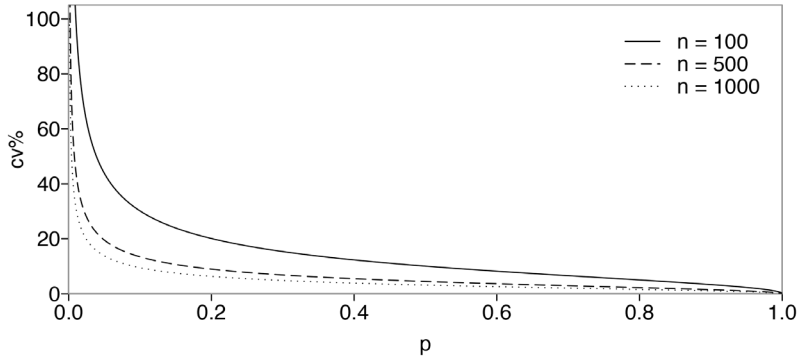

Figure 1. Percentual *cv*'s with respect to proportion *p* and varying *n*.

on simulation experiments on unemployment whose rates range from 2.1% to 8.4% whence cv_i's are much more relevant (from 68.6/21.6% to 33.2/10.4% with $n_i = 100/1000$).

Outcomes pointed out in **Table 2** will be basically repeated in subsequent simulation experiments, thus revealing somehow a typical performance for each model at comparison. Ranking of model-based estimators is to be interpreted as follows: best performance values are boxed while the others are gray/green-colored with intensity growing with worsening performance (the range of expectation/median values, for each measure, has been divided into six equal intervals—being seven the models at comparison; in *rel* columns, only values less than 100% have been shadowed). FH row has been excluded from shadowing since FH estimator is clearly unusable (as will be clarified below). With regard to columns related to efficiency, it is of interest only detecting whether efficiency measures (as they are ordinarily computed) are instead a "lark-mirror". Thus, low *eff* values (*i.e.* low *cv* whence high efficiency) coupled with low *rel* values (*i.e.* low reliability of *cv* in measuring actual estimator accuracy) are boxed and gray-filled.

Direct estimates are expected to be, on average, unbiased, increasingly accurate with growing sample size, as well as their $\widehat{cv}$ (since it is analitically derived from a formula consistent with the sampling design) ought to be a reliable measure of accuracy. **Table 1** and **Table 2** (as well as 8 - 10) prove these expectations, though showing a tendency of such $\widehat{cv}$'s to overestimate direct estimates' uncertainty. On the contrary, synthetic estimates are typically biased and bias (and/or accuracy, their variance being irrelevant) gets worse for data tending to negative asymmetry (see the worsening of bias across **Tables 8-10**). Moreover, it turns out that *mse* of synthetic estimates is badly estimated by Marker's method which tends to heavily overestimate it.

NN gets the worst scores in terms of (absolute) bias and accuracy (**Table 2**). Looking at single small areas situation, $\hat{\theta}_i^{HB}$s suffer from a terrific underestimation especially for the least populated/sampled areas (see the reduced *bias* for *non-na* areas and **Figure 2**, second column of panels). Moreover, the estimated *cv* is, in median terms, scarcely reliable (**Figure 2**, fifth row) except for sampled areas with little information (it is above 100 just for *non-na* areas with the lowest

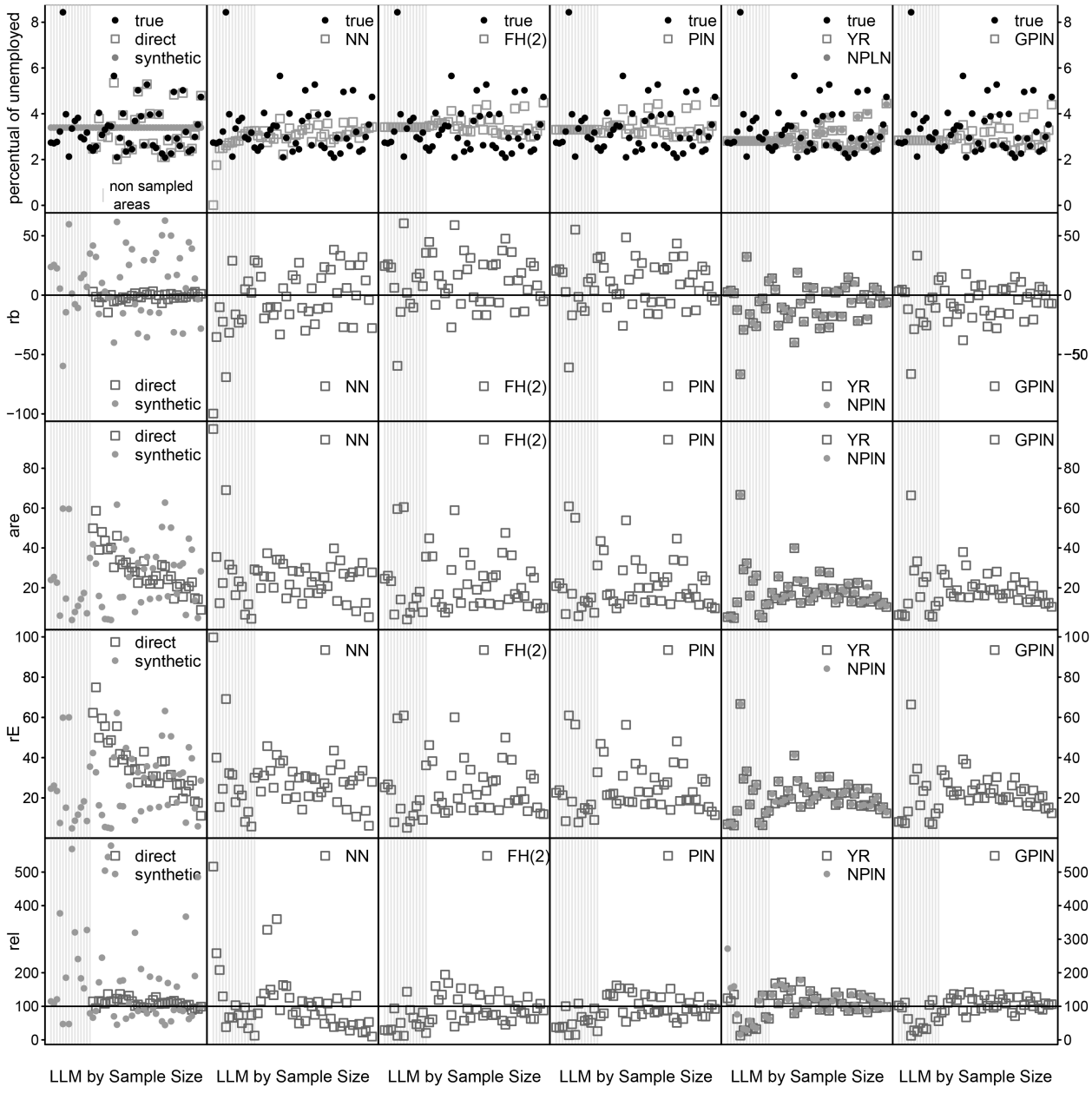

Figure 2. Design-based and HB model-based estimators at comparison (simulated data on unemployment, Table 2). Areas are ordered by increasing sample size; non sampled areas are placed at the extreme left.

sample size). The adoption of a linear form for the linking model is likely inadequate, but, in this study, where the regression model is merely a null model (with the synthetic estimate offset consisting in the sole input of auxiliary information), we may try to handle its elementary components in an effective way before definitely dropping the linear link. In Section 3.3 we will see what are the refinements suited to improve NN performance.

FH and PlN exhibit a similar performance, though defects are exacerbated for FH. They tend to overestimate the θ_i's (positive bias, Table 2; see also Figure 2, second row, third and fourth panels) and their accuracy is low in terms of expectation (high $\overline{are}$ and $\overline{rE}$) again depending on the biased estimation for

some areas (Figure 2, third and fouth rows), though improve in accuracy with respect to NN and both the design-based estimators. Estimated *cv* is scarcely reliable for both of them, yet more markedly for FH and for non-sampled areas. This is likely due to a noticeable shrinkage of the estimates toward the mean (yet a biased mean), as it results also from a low DIC (in particular, a small effective number of parameters, pD; see Table 3, Table 4 which will be discussed later).

Right now, it is worth comparing the three aforementioned FH estimators: FH(2) and FH(3) are practically indistinguishable; FH is less biased than FH(2) (on the other hand, this is well expected since $E\{\exp(\eta_i)\} \geq \exp\{E(\eta_i)\}$ by Jensen inequality) yet its *cv* is totally unreliable. Therefore, without loss of information, only FH(2) and FH(3) will be reported from now on. Since now, a minor difference is detectable between these two just as to reliability (this keeps through all the simulation experiments).

The unmatched models, YR and NPlN, and the non-normal sampling model GPlN show to have the best performance in terms of (absolute) bias, accuracy and reliability (see the highlighted values and regions of light or no shading in Table 2). The unmatched models show to be the most accurate while GPlN seems to be the most "well-centered" and to give the most "well-calibrated" efficiency measure (the last two columns of Figure 2 seem almost indistinguishable; at a careful reading, difference between the unmatched models and GPlN relates

Table 3. Model determination diagnostics (simulated data on employment).

	$\bar{D}$	$D(\hat{\theta})$	p_D	DIC	PPp
NN	50.1	46.3	3.8	53.8	0.18
FH	38.5	27.5	11.0	49.5	0.43[a]
YR	36.8	23.8	13.0	49.8	0.51
NPlN	35.8	21.9	13.9	49.7	0.54
PlN	36.7	1.4	35.2	71.9	0.50
GPlN	36.6	1.0	35.6	72.2	0.50

[a]PPp = 0.43 results by imputing σ_i fixed at $\hat{\sigma}_i = (\hat{cv}_i\hat{\theta}_i)^2$ in (19); otherwise, PPp = 0.47 by using $\hat{\sigma}_i = (\hat{cv}_i\theta_i)^2$ as it is generated by the model.

Table 4. Model determination diagnostics (simulated data on unemployment).

	$\bar{D}$	$D(\hat{\theta})$	p_D	DIC	PPp
NN	56.6	42.5	14.1	70.6	0.11
FH	38.6	23.9	14.6	53.2	0.07[a]
YR	37.3	19.6	17.7	55.0	0.49
NPlN	37.2	19.4	17.7	54.9	0.49
PlN	41.5	6.1	35.4	76.9	0.43
GPlN	37.0	0.5	36.4	73.4	0.49

[a]PPp = 0.40 by using $\hat{\sigma}_i = (\hat{cv}_i\theta_i)^2$ in (19) (see Table 3 for explanation).

to *non-na* areas with lowest sample size: GPlN is somewhat less accurate though the estimated *cv* associated to such areas are fully reliable). However, a tendency to underestimation (negative bias) is a non negligible weak point: as we will see in the next section, such a deficiency can be promptly remedied by a (obvious) model refinement.

Finally, a poor estimation for non sampled areas (placed at the far left of *x*-axis in **Figure 2**) is detectable across all models. A reason for it is definitely the lack of direct information. Thereby, any improvement of model-based estimators over the synthetic ones crucially depends on the model ability of *borrowing strength* from all of the available information (thus producing a consistent estimate of the functional form $\alpha + \beta x_i$ which solely constitutes non sampled areas estimates). In the experiments under study which consider a null model (no x_i) the borrowing is clearly not sufficient.

Table 3 and **Table 4** show $\bar{D}$, $D\left(\hat{\theta}\right)$, p_D, DIC and PPp values averaged over the R replications, for employment and unemployment simulation data respectively. We are essentially interested to detect whether commonly used model selection and validation criteria serve their purpose.

We start with the predictive criterion, the PPp, since it seems to pass the examination. We know "lights and shades" of the models under comparison from the foregoing external validation (*i.e.* by knowing the *true* values of the parameter of interest): PPp effectively chooses the unmatched models and GPlN while rejects NN and FH. (PlN is clearly picked out only for the employment application, yet we stress that focus is on the unemployment study where models' performance differs significantly.) However, PPp is based on a particular measure of discrepancy (the "event" under probability in (19)/(20)), which carries out a model validation relatively to a single aspect of (global) model performance. In particular, it answers the question whether the sampling stage of the examined model might be an adequate mechanism for generating our data. It is immediately evident that either NN or FH sampling model are quite unsuited (the associated PPp's are far below 0.50), and this supports the thesis against either linear models for non-normal variables (NN) or questionable sampling models (the dubious hypotheses under the first stage of FH specification). Nevertheless, this particular PPp measure cannot inform us on many other possible model failures [19] [20] [21].

On the other side, comparison between likelihoods is even more delicate [22], especially with hierarchical models which allow the possibility of marginalization across levels in different ways. FH model that is the second-worst model in terms of performance (**Table 1**, **Table 2**) should be chosen according to the principle "lower the DIC better the model". Is the DIC misleading?

First, the DIC calculated from (NN, FH, YR, NPlN) group cannot be compared to the one from the (PlN, GPlN) pair: the deviance in models with normal sampling stage is focused on $\boldsymbol{\theta}$s whereas is on $\boldsymbol{\mu}$'s for poisson sampling stage specifications. In fact, the DIC has been calculated—as routinely is—relatively to the first stage unobservable variables (or parameters, in classical terminology) of

the HB models. Instead, comparison would be made more consistent by marginalizing the likelihood of all the candidate HB models to the same parameters of interest (in SAE studies, the $\boldsymbol{\theta}$s or small area quantities for they are the primary object). Even doing so, care must be taken in making comparisons of DIC's, since, for instance, normalizing constants vary with the parametric family of distribution (thus all constants must be retained when making comparisons), or, the support for a model can vary with the possible (marginal) likelihoods [23].

Second, the comparison of FH versus all the remaining normal sampling stage models is questionable as well, since data are the log-transformed direct estimates for FH while are the untransformed direct estimates for the other specifications. As we have already noted in Section 2.2, all the null direct estimates are excluded from the usable data for FH, hence the posterior expected deviance, $\bar{D}$, is accordingly decreased [18].

However, despite the critical comments above, can we draw any information from the DIC output for our study? Posterior expected deviance $\bar{D}(\boldsymbol{\theta})$ has been standardized by the maximized log-likelihood, then, if the model is "true", its expectation is approximately the number of the free parameters in $\boldsymbol{\theta}$, that is the total size of unit sample: in our case, it is the number of sampled areas (51 − 14 = 37, see **Table 7**). Thereby, computation of the standardized posterior deviance might be appropriate for checking the overall goodness-of-fit of the model. With this respect, NN clearly does not fit the data, FH and PlN show to be somewhat inconsistent with the data, whereas GPlN, NPlN and YR show, in order, to have the best fit. Such conclusion matches the one derived on the basis of PPp values. As for p_D/DIC, they tell of the model complexity. In our study (where a null model is considered) the number of effective parameters actually lacks in interest (neither mentioning Poisson sampling stage models for which p_D is practically pointless).

A further discussion on model selection and checking is beyond the scope of the paper. Yet, the problem of small area model diagnostics constitutes an important, and still largely unexplored, direction of SAE research.

3.3. Some Refinements

In this section we consider some refinements of the HB models previously introduced, which involve relevant improvements of their performance. The major development consists in letting sampling variances (σ_i or $\tilde{\sigma}_i$ when a logarithmic transformation occurs or a_i in GPlN) be stochastic, whereas, so far, they have been assumed as known and fixed to off-set estimates of sampling design variance $v_p\left(\hat{\theta}_i\right)$.

This extension—that we refer to as *model-based sampling variance function*—arises naturally from a model-based approach to SAE problems. If we assume that a certain model is valid for a certain phenomenon θ, then all the unknown quantities that depend on θ should be made model-generated instead of being imputed as fixed to the model. In our context, the design variance is assumed to be a function of the unknown quantity of interest, *i.e.* $v_p\left(\hat{\theta}_i\right) = f\left(\theta_i\right)$, with f

known from sampling design characteristics. Then, according to the outlined strategy, sampling variances will be obtained as model-based estimates, $f(\theta_i)$, through θ_i.

By the way, this general consideration allows us to explain a specific fault occurring with the customary FH model. Recall that for a FH model with link function g, it is assumed that (according to the Taylor approximation) $\tilde{\sigma}_i = \{g'(\theta_i)\}^2 \sigma_i$, and, in the standard version of the model, sampling error variance σ_i is fixed at the design variance $v_p(\hat{\theta}_i)$ value. In particular, if g is the log function then $\tilde{\sigma}_i = \theta_i^{-2}\sigma_i = cv_i^2$, and $cv_i = \sqrt{\sigma_i/\theta_i}$ is likewise fixed at $cv_p(\hat{\theta}_i) = v_p^{1/2}(\hat{\theta}_i)/\theta_i$. But, the design quantity to be imputed (whatever is: $v_p(\hat{\theta}_i)$, $cv_p(\hat{\theta}_i)$, etc.) is usually function of the unknown θ_i, e.g. $cv_p(\hat{\theta}_i) \approx \sqrt{(1-p_i)/(p_i n_i)} = \sqrt{(N_i - \theta_i)/(\theta_i n_i)}$ under the SRSWR hypothesis as for our simulation. Common practice essentially consists in replacing $\hat{\theta}_i$ to θ_i in the design quantity function, e.g., in our example, $\hat{cv}_p(\hat{\theta}_i) \approx \sqrt{(1-\hat{p}_i)/(\hat{p}_i(n_i-1))}$ with $\hat{p}_i = \hat{\theta}_i/N_i$.

It is now easy to detect why the standard FH estimator $\hat{\theta}_i^{HB}$ is so much positively biased. If, indeed, $\theta_i < \hat{\theta}_i$ then $\hat{cv}_p(\hat{\theta}_i)$ is smaller than $cv_p(\hat{\theta}_i)$ (recall Figure 1), hence model estimator $\hat{\theta}_i^{HB}$ ($< \hat{\theta}_i$ if the model is well posited), in order to adequately fit $\hat{\theta}_i$, is pushed upward *i.e.* tends to overestimate θ_i ($\hat{\theta}_i^{HB} > \theta_i$). Vice versa, if $\theta_i > \hat{\theta}_i$, the greater $\hat{cv}_p(\hat{\theta}_i)$ (than $cv_p(\hat{\theta}_i)$) combined with $\hat{\theta}_i^{HB} > \hat{\theta}_i$ (if the model is well posited) does not pull $\hat{\theta}_i^{HB}$ down (rather, $\hat{\theta}_i^{HB}$ is let to rise, in any case does not underestimate θ_i). To prove such an insight we include in the following tables the results obtained from fitting a standard FH model yet with $cv_p(\hat{\theta}_i)$ fixed to its true value (rows named as 'true cv').

According to the foregoing comments, in all of the HB models so far considered, we "plug-in" the unobservable θ_i into the sampling design variance function $v_p(\hat{\theta}_i) = f(\theta_i)$ by which sampling variance (σ_i or $\tilde{\sigma}_i$ or a_i) is modelled. Incidentally, we note that the implementation of such a strategy is relatively straightforward within the Bayesian approach (not as such in the frequentist one).

Before going through simulation results, we also mention some ameliorations to NN model. The linearity assumed for the linking model is likely inadequate. Yet, since here the linear predictor ($\theta_i = {}_S\hat{\theta}_i + \alpha + \nu_i$) is quite nave (a null one), we may try to adjust it somehow before dropping the linear link definitely. In fact, the bad performance is probably due to mis-calibrated random effects ν_i's: in log-normal linking models, θ_i depends in a multiplicative way on ν_i ($\theta_i = {}_S\hat{\theta}_i e^{\alpha+\beta x_i+\nu_i}$), whence ν_i effect is weighted by ${}_S\hat{\theta}_i$ (*i.e.* by population N_i). Thereby, a simple remedy to NN model deficiency might consist in weighting ν_i by population N_i. By doing that, indeed, a surprising improvement of model performance is immediately obtained (see row NN ($\nu_i \cdot N_i$) in Table 5 and Table 6).

Model performance greatly improves for all of the models thanks to the aforementioned stochastic extension. Table 5 and Tables 8-10 show how all of

Table 5. Comparison between design-based and HB model-based estimators (simulated data on unemployment).

	bias				accuracy				efficiency		reliability	
estimator	$\overline{rb}$	rb^M	$\overline{arb}$	arb^M	$\overline{are}$	are^M	$\overline{rE}$	rE^M	$\overline{eff}$	eff^M	$\overline{rel}$	rel^M
NN($v_i \cdot N_i$)	−5	−4	13	10	19	18	22	21	21	23	104	106
non-na	−5	−4	12	8	19	18	23	22	25	25	116	114
NN($v_i \cdot N_i$) *sv*	4	5	15	14	19	16	22	19	20	22	102	104
non-na	4	5	13	12	18	16	22	20	24	24	117	114
FH(2) true *cv*	3	4	15	13	19	16	22	19	19	21	98	98
non-na	4	4	13	13	19	16	22	20	23	24	111	104
FH(3) true *cv*	3	4	15	13	19	16	22	19	19	21	99	98
non-na	4	4	13	13	19	16	22	20	23	25	112	105
FH(2) *sv*	−2	0	14	12	18	16	21	19	18	21	99	96
non-na	−1	0	13	13	18	17	21	20	22	22	112	107
FH(3) *sv*	−2	0	14	12	18	16	21	19	19	22	102	100
non-na	−1	0	13	13	18	17	21	20	23	23	117	112
YR *sv*	3	4	15	14	18	16	21	19	18	20	98	99
non-na	4	4	14	13	18	17	22	21	22	23	110	104
NPLN *sv*	3	4	15	14	18	16	21	19	19	21	101	97
non-na	4	4	14	14	18	17	22	20	22	23	111	102
GPlN *sv*	2	3	14	13	19	16	22	20	19	22	93	97
non-na	3	3	12	12	19	17	23	22	23	23	105	99

Table 6. Model selection diagnostics (simulated data on unemployment).

	$\bar{D}$	$D(\hat{\theta})$	p_D	DIC	PPp
NN ($v_i \cdot N_i$)	35.8	16.2	19.6	55.4	0.56
($v_i \cdot N_i$)*sv*	34.9	17.1	17.8	52.7	0.58
FH true *cv*	38.2	19.4	18.8	57.1	0.17[a]
FH *sv*	37.9	20.8	17.1	55.0	0.26(0.36)[b]
YR *sv*	36.6	19.7	16.9	53.6	0.51
NPlN *sv*	36.2	19.7	16.5	52.7	0.53
GPlN *sv*	37.5	1.0	36.4	73.9	0.49

[a]PPp = 0.44 by using $\hat{\sigma}_i = (\hat{c}v_i\theta_i)^2$ in (19) (see Table 3). [b]PPp = 0.26/0.36 results from including/excluding $\{\hat{\theta}_i = 0$ with $n_i > 0\}$ cases from the computation.

them—denoted by *sv* *i.e.* stochastic variance—reach comparable good scores (while maintaining performance characteristics detected in Section 3.2).

Results shown in Table 5 and Table 8 were obtained using, respectively, the simulated data from census and from a synthetic positively skewed population (thus maintaining the same type of asymmetry of the original unemployment dataset; see details in Section 3.1), hence both comparable to results in Table 2.

The most noticeable change in Table 5 concerns *bias*. FH *sv* estimator is now almost unbiased, moreover, YR, NPlN and GPlN are no more negatively biased: a stochastic variance helps model centering (the explanation for that has been given above). Such an improvement affects the rest of performance measures making all the *sv* models almost competitive.

In Table 6 results on $\bar{D}$, DIC and PPp show to be concordant in choosing the YR, NPlN and GPlN as the best tern of models. Further, note that the DIC is relatively larger for FH-true *cv* and *sv* models, thus the partial explanation given in Section 3.2 for the tendency of customary FH models to have lower DIC's (the exclusion of $\hat{\theta}_i = 0$ values from the DIC computation) turns out not to be a sufficient reason by alone. Another reason why the DIC tends to be lower for fixed variance FH model is probably the over-shrinkage (thus a lower p_D) of $\hat{\theta}_i^{HB}$ estimators. Finally, note that PPp gets farer from the ideal 0.5 value both for FH and NN *sv* specifications. Although we might have thought to remedy the marked deficiencies of these two models by merely making the sampling variance stochastic, other faults show to be still alive (at least, sampling model is not fully adequate to reproduce survey data). One more time, the YR, NPlN and GPlN tern proves to be the most natural response to model-based SAE of count variables.

Two further synthetic populations were considered: one with an essentially symmetric distribution (Table 9) and the other with negative skewness (Table 10), in order to explore (the expected worsening of) performance (with negative asymmetry) of standard FH models, and, at the same time, detect the parallel reaction of the models that have so far shown to perform better. We show the results only for canonical and extended FH, YR, NPlN models as well as for GPlN, this last just in the *sv* version (being the most naturally conceivable for it), to enlighten the difference in performance of the first one (that worsens with negative asymmetry, mostly in terms of expectation) compared to the competitive outcomes of the stochastic variance models. (Results on "true *cv*" FH have also been reported and written in italic to stress they are only theoretically conceivable.) Commenting on Tables from 8 onward would give rise to observations analogous to the ones already given above. We merely note that the advantage in using stochastic variance models diminishes with increasing sample size (letting fixed variance models regain ground in the rankings), up to the situation where direct estimators result better than model-based ones (see *are* and *rE* in the bottom panel of Table 9 and middle panel of Table 10).

4. Concluding Remarks

The benefits of using HB models for SAE problems have been largely recognized. They include the availability of a wider set of tools to handle complex and more realistic models and to get reliable measures of variability. When estimating counts it might not be clear which specification is more appropriate but relevant quantities should be properly modeled. Object of the study was to show that different model specifications are possible from the ones customarily used in non-

normal/non-linear cases. The purpose was definitely not to show the general superiority of one framework over the others, being the range of situations one has to face with in real analyses practically unlimited. Moreover, there are (at least) two kinds of reasons why the generalization of our results is rather restricted. First, the type of simulation study we have performed is not the one ordinarily carried out in statistical studies of model comparison. In fact, we have not simulated from a posited model, rather we have considered design-based simulations. That is, given a real phenomenon (in our case, LLMs unemployment in a given region and year either known from census data or simulated), a sample survey has been (repeatedly) simulated according to a fixed sampling design. The idea we follow is "all models are wrong but some are useful" [24], and, this concept goes quite naturally with the demand to a SAE analyst for providing estimates of "real world" quantities. A design-based simulation, then, meets the requirement of positing a "true" population instead of a "true" model. Unfortunately, this concept of simulation makes us totally ignorant about the possible presence of any useful/good model among the ones at comparison. It can be the case that all of them are inadequate to fit the phenomenon under study. Then, any comparison quite unlikely would lead to clear-cut results (or select the not-present "best" candidate). Second, in our application the model structure is kept as simple as possible to ease comparison. On the other hand, a null model, *i.e.* with only synthetic estimates entering the model and no additional auxiliary information, can be compared to traditional estimators and are free from further complications derived from specific characteristics of auxiliary information (like as variables measured with errors; type of relationship existing with small areas quantities of interest: e.g. multiplicative or additive effect? on which scale? etc.). On this regard, it is worth noting that, in presence of other sources of information to account for, comparison would have become even more complicated if the data are not simulated on the ground of a posited model, but are considered as real world observations and so generated by a "black box" mechanism.

However, the results from simulation study show persistent model failures for some standard Fay-Herriot specifications and for generalized linear Poisson models with (log-)normal sampling stage in SAE problems with count data, and how even minor model modifications can noticeably improve on their performance. In particular, we advocate the extension of model specification from assuming sampling variances fixed to some off-set estimates (typically design sampling variance estimates) to letting them be (stochastically) generated by the model (at least in their components depending from latent variables estimated within the model itself). On the other hand, unmatched and non-normal sampling stage models show definitely a better performance in terms of bias, accuracy and reliability even in the fixed sampling variance version. Notice that the extension to stochastic sampling variances is straightforward and relatively easy to implement within a Bayesian framework as well as the non-standard specifications are practically feasible only by means of HB models.

Moreover, the study brings out some limits and possible deceptions of com-

monly used criteria for model determination in the context of SAE problems, namely DIC and PPp. The version of PPp here addressed is the one commonly used in Bayesian analyses—within SAE context yet also statistic-worldwide—which tests the (global) validity of model first stage, *i.e.*, in our case, the sampling model. Nevertheless, this particular PPp measure cannot inform us on many other possible model failures. On the other hand, the DIC is not comparable across the models, the main reason being that it is calculated—as routinely is—relatively to the first stage unobservable variables which vary in essence and number across the HB specifications.

In conclusion, future research should definitely focus on defining proper devices for model determination in the field of SAE. Besides, in order to magnify differences between models at comparison in simulation studies, model structure has to be complicated for considering more realistic situations (adding auxiliary information, taking spatial structure into account, etc.), more complex sampling designs are to be experienced, as well as different real population phenomena are to be explored (small areas at different level of territorial aggregation, small area population with different distributional characteristics, etc.).

References

[1] Rao, J.N.K. and Molina, I. (2015) Small Area Estimation. 2nd Edition, Wiley, New York. https://doi.org/10.1002/9781118735855

[2] Pfeffermann, D. (2002) Small Area Estimation—New Developments and Directions. *International Statistical Review*, **70**, 125-143.

[3] Jiang, J. and Lahiri, P. (2006) Mixed Model Prediction and Small Area Estimation. *Test*, **15**, 1-96. https://doi.org/10.1007/BF02595419

[4] Fay, R.E. and Herriot, R.A. (1979) Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **85**, 398-409. https://doi.org/10.1080/01621459.1979.10482505

[5] Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998) Generalized Linear Models for Small-Area Estimation. *Journal of the American Statistical Association*, **93**, 273-282. https://doi.org/10.1080/01621459.1998.10474108

[6] Lu, L. and Larsen, M. (2007) Small Area Estimation in a Survey of High School Students in iowa. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 2627-2634.

[7] Molina, I., Saei, A. and Lombardia, M.J. (2007) Small Area Estimates of Labour Force Participation under a Multinomial Logit Mixed Model. *Journal of the Royal Statistical Society A*, **170**, 975-1000. https://doi.org/10.1111/j.1467-985X.2007.00493.x

[8] You, Y. and Rao, J.N.K. (2002) Small Area Estimation Using Unmatched Sampling and Linking Models. *Canadian Journal of Statistics*, **30**, 3-15. https://doi.org/10.2307/3315862

[9] Trevisani, M. and Torelli, N. (2004) Small Area Estimation by Hierarchical Bayesian Models: Some Practical and Theoretical Issues. *Atti della XLII Riunione Scientifica della Societ Italiana di Statistica*, 273-276.

[10] Trevisani, M. and Torelli, N. (2006) Comparing Hierarchical Bayesian Models for Small Area Estimation. In: Liseo, Montanari, Torelli, Eds., *Metodi statistici per l'in-*

tegrazione di basi di dati da fonti diverse, Franco Angeli, Milano, 17-36.

[11] Trevisani, M. and Gelfand, A. (2013) Spatial Misalignment Models for Small Area Estimation: A Simulation Study. In: *Advances in Theoretical and Applied Statistics*, Springer-Verlag, Berlin Heidelberg, 269-279.

[12] Torelli, N. and Trevisani, M. (2008) Labour Force Estimates for Small Geographical Domains in Italy: Problems, Data and Models. *International Review of Social Sciences*, **4**, 443-464.

[13] Liu, B., Lahiri, P. and Kalton, G. (2014) Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions. *Survey Methodology*, **40**, 1-13.

[14] You, Y. and Chapman, B. (2006) Small Area Estimation Using Area Level Models and Estimated Sampling Variances. *Survey Methodology*, **32**, 97-103.

[15] You, Y. (2008) An Integrated Modeling Approach to Unemployment Rate Estimation for Subprovincial Areas of Canada. *Survey Methodology*, **34**, 19-27.

[16] Maples, J., Bell, W.R. and Huang, E.T. (2009) Small Area Variance Modeling with Application to County Poverty Estimates from the American Community Survey. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 5056-5067.

[17] Shen, A.C. and Louis, T.A. (1998) Triple-Goal Estimates in Two-Stage Hierarchical Models. *Journal of the Royal Statistical Society Series B*, **60**, 455-471. https://doi.org/10.1111/1467-9868.00135

[18] Spiegelhalter, D.K., Best, N., Carlin, B.P. and van der Linde, A. (2002) Bayesian Measures of Model Complexity and Fit (with Discussion). *Journal of the Royal Statistical Society Series B*, **64**, 583-639. https://doi.org/10.1111/1467-9868.00353

[19] Gelman, A., Meng, X.-L. and Stern, H. (1996) Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, **6**, 733-807.

[20] Bayarri, M.J. and Castellanos, M.E. (2007) Bayesian checking of the second levels of hierarchical models. *Statistical Science*, **22**, 322-343. https://doi.org/10.1214/07-STS235

[21] Dey, D.K., Gelfand, A.E., Swartz, T.B. and Vlachos, P.K. (1998) A Simulation-Intensive Approach for Checking Hierarchical Models. *Test*, **7**, 325-346. https://doi.org/10.1007/BF02565116

[22] Dempster, A.P. (1997) The Direct Use of Likelihood for Significance Testing. *Statistics and Computing*, **7**, 247-252. https://doi.org/10.1023/A:1018598421607

[23] Trevisani, M. and Gelfand, A.E. (2003) Inequalities between Expected Marginal log-Likelihoods, with Implications for Likelihood-Based Model Complexity and Comparison Measures. *Canadian Journal of Statistics*, **31**, 239-250. https://doi.org/10.2307/3316084

[24] Box, J.E.P. (1976) Science and Statistics. *Journal of the American Statistical Association*, **71**, 791-799. https://doi.org/10.1080/01621459.1976.10480949

Appendix

Some tables and figures referred to in the text are below listed.

Table 7. Census and one sample of simulated data reported for each LLM: (census data) employment and unemployment rates (% over population), p_i^e and p_i^u; (sampling design data) sampling fraction (over population), r_i, sample size, n_i; (simulated data) direct and synthetic estimates of employment and unemployment rates (%), $\hat{p}_i^e$, $\hat{p}^e$, $\hat{p}_i^u$ and $\hat{p}^u$, as well as the associated cv (%) estimates, $\hat{cv}_i^e$, ${}_S\hat{cv}_i^e$, $\hat{cv}_i^u$ and ${}_S\hat{cv}_i^u$. LLMs have been ordered according to increasing population count.

	census data		sampling design data		one sample of simulated data						
LLM	p_i^e	p_i^u	r_i	n_i	$\hat{p}_i^e$	$\hat{cv}_i^e$	$\hat{p}^e$	${}_S\hat{cv}_i^e$	$\hat{p}_i^u$	$\hat{cv}_i^u$	$\hat{p}_i^u$	${}_S\hat{cv}_i^u$
1	39.9	2.7	0.0	0		5.7	40.4	5.7		29.7	3.5	29.7
2	40.1	2.7	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
3	40.0	3.4	11.6	113	43.4	10.7	40.4	5.7	2.6	57.0	3.5	29.9
4	38.2	3.1	9.7	99	34.3	13.9	40.4	5.7	2.0	70.0	3.5	29.9
5	40.9	2.8	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
6	42.3	3.2	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
7	34.7	8.4	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
8	37.7	4.0	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
9	41.8	2.1	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
10	39.4	3.4	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
11	40.0	3.7	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
12	42.4	4.0	17.6	308	37.3	7.3	40.4	5.7	4.2	26.9	3.5	29.9
13	43.3	2.1	6.4	127	29.9	13.6	40.4	5.7	2.4	57.1	3.5	29.9
14	38.9	3.3	4.7	105	37.1	12.7	40.4	5.7	3.8	49.1	3.5	29.9
15	40.9	3.8	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
16	41.4	4.0	3.8	98	46.9	10.8	40.4	5.7	3.1	57.1	3.5	29.9
17	39.6	4.0	4.5	139	37.4	11.0	40.4	5.7	1.4	70.3	3.5	29.9
18	36.3	5.7	3.8	118	29.7	14.2	40.4	5.7	4.2	43.9	3.5	29.9
19	44.2	2.6	8.3	272	43.4	6.9	40.4	5.7	1.1	57.3	3.5	29.9
20	42.1	3.0	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
21	41.2	4.0	7.3	288	37.5	7.6	40.4	5.7	4.9	26.0	3.5	29.9
22	40.2	3.5	2.7	107	43.0	11.2	40.4	5.7	2.8	57.1	3.5	29.9

Continued

23	40.9	5.3	6.3	274	43.4	6.9	40.4	5.7	6.9	22.1	3.5	29.9
24	42.6	3.0	3.1	134	44.0	9.8	40.4	5.7	2.2	57.2	3.5	29.9
25	40.0	3.9	5.7	256	43.0	7.2	40.4	5.7	2.7	37.2	3.5	29.9
26	38.7	2.9	6.5	337	35.9	7.3	40.4	5.7	3.0	31.1	3.5	29.9
27	39.5	2.9	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
28	41.9	2.4	4.0	230	42.2	7.7	40.4	5.7	1.3	57.3	3.5	29.9
29	44.5	3.2	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
30	41.9	2.4	1.4	88	43.2	12.3	40.4	5.7	0.0	56.1	3.5	29.9
31	40.6	2.6	4.6	298	40.9	7.0	40.4	5.7	2.4	37.3	3.5	29.9
32	41.9	2.5	0.0	0		5.7	40.4	5.7		29.9	3.5	29.9
33	43.4	2.6	1.4	93	43.0	12.0	40.4	5.7	3.2	57.0	3.5	29.9
34	45.0	2.3	4.3	320	44.1	6.3	40.4	5.7	2.2	37.4	3.5	29.9
35	37.7	5.0	3.2	251	38.6	8.0	40.4	5.7	5.6	26.0	3.5	29.9
36	40.2	3.2	7.1	581	41.3	4.9	40.4	5.7	4.0	20.4	3.5	29.9
37	41.9	2.9	5.2	445	39.1	5.9	40.4	5.7	2.7	28.4	3.5	29.9
38	40.1	3.7	2.8	247	38.9	8.0	40.4	5.7	5.3	27.0	3.5	29.9
39	39.2	5.0	4.9	437	35.2	6.5	40.4	5.7	6.9	17.6	3.5	29.9
40	43.1	2.7	2.4	234	43.2	7.5	40.4	5.7	1.7	49.6	3.5	29.9
41	43.7	2.1	3.2	324	39.5	6.9	40.4	5.7	2.5	34.9	3.5	29.9
42	42.5	2.4	5.7	588	45.2	4.5	40.4	5.7	2.5	25.4	3.5	29.9
43	38.7	5.0	5.1	542	40.2	5.2	40.4	5.7	5.7	17.4	3.5	29.9
44	42.5	2.4	2.2	242	42.1	7.5	40.4	5.7	2.5	40.4	3.5	29.9
45	42.4	2.3	3.3	372	49.2	5.3	40.4	5.7	1.9	37.4	3.5	29.9
46	41.6	2.5	2.3	298	38.9	7.3	40.4	5.7	2.7	34.9	3.5	29.9
47	43.0	2.6	2.0	472	39.6	5.7	40.4	5.7	3.6	23.8	3.5	29.9
48	42.5	2.4	2.7	661	42.2	4.5	40.4	5.7	1.7	29.9	3.5	29.9
49	40.8	3.0	2.3	1099	40.3	3.7	40.4	5.7	3.5	15.9	3.5	29.9
50	42.0	3.5	2.2	1100	42.2	3.5	40.4	5.7	3.1	16.9	3.5	29.9
51	38.4	4.7	2.7	1653	37.7	3.2	40.4	5.7	5.1	10.6	3.5	29.9

Table 8. Simulated data on unemployment: $\overline{p} = 3.35$ and positive asymmetry (real data); from top to bottom: $n = 100/N = 25000$, $n = 300/N = 75000$, $n = 500/N = 100000$ (hence $r = 4, 4, 5‰$).

	bias				accuracy				efficiency		reliability	
estimator	$\overline{rb}$	rb^M	$\overline{arb}$	arb^M	$\overline{are}$	are^M	$\overline{rE}$	rE^M	$\overline{eff}$	eff^M	$\overline{rel}$	rel^M
Synthetic	9	12	25	22	26	23	28	25	27	27	153	124
Direct	0	0	7	7	45	45	56	56	60	61	113	112
FH(2) true cv	7	10	21	18	24	21	28	24	21	21	96	97
non-na	8	9	20	19	24	22	28	26	25	25	109	106
FH(3) true cv	7	10	22	18	24	21	28	24	21	21	96	97
non-na	8	9	20	19	24	22	28	26	25	25	108	107
FH(2)	32	35	38	37	39	37	41	39	16	16	60	47
non-na	32	35	38	36	39	36	41	39	18	18	66	53
FH(3)	32	35	38	37	39	37	42	39	16	16	62	49
non-na	32	35	38	36	39	36	41	39	19	19	67	53
YR	−20	−18	22	18	25	19	27	22	22	21	101	95
non-na	−20	−18	22	18	25	19	27	24	25	25	113	116
NPlN	−20	−18	23	18	25	19	27	22	22	21	102	98
non-na	−20	−18	22	18	24	19	27	23	25	25	114	114
FH(2) sv	−2	1	20	18	22	19	25	21	17	17	85	80
non-na	−2	0	19	17	22	18	25	21	20	19	96	93
FH(3) sv	−2	1	20	18	22	19	25	21	17	17	84	81
non-na	−2	0	19	17	22	18	25	21	19	19	95	90
YR sv	5	7	19	16	24	21	28	27	23	24	100	103
non-na	6	6	17	15	24	22	28	27	28	28	112	116
NPlN sv	4	7	18	16	24	21	28	26	24	25	104	106
non-na	5	6	17	15	23	21	28	26	29	29	116	121
GPlN sv	−1	3	13	13	29	27	36	36	31	31	97	100
non-na	2	5	12	11	30	29	38	36	38	38	106	108
Synthetic	11	14	25	24	26	24	27	25	33	33	203	136
Direct	1	1	4	3	25	25	32	31	34	34	108	108
FH(2) true cv	5	5	15	14	21	20	24	23	20	20	94	92
non-na	5	5	12	11	19	18	23	22	24	24	109	107
FH(3) true cv	5	5	15	14	21	20	24	23	20	20	95	93
non-na	5	5	12	11	19	18	23	22	24	24	110	108
FH(2)	13	14	20	18	23	21	26	23	17	18	88	76
non-na	13	13	18	17	21	19	25	23	21	21	100	99

Continued

FH(3)	13	14	20	18	23	21	26	23	18	18	89	77
non-na	13	13	18	17	21	19	25	23	22	22	102	99
YR	−5	−4	14	13	19	17	22	21	19	19	101	97
non-na	−5	−4	11	9	18	17	21	20	19	19	101	104
NPlN	−5	−4	14	13	19	17	22	21	19	19	100	96
non-na	−5	−4	11	10	18	17	21	20	19	19	101	100
FH(2) sv	0	0	15	13	19	17	22	20	18	18	96	93
non-na	0	0	12	10	18	17	21	21	22	22	112	107
FH(3) sv	0	0	15	13	19	17	22	20	19	19	101	97
non-na	0	0	12	10	18	17	21	21	23	23	117	111
YR sv	5	5	16	13	20	19	23	22	19	19	95	92
non-na	5	5	13	12	18	17	22	21	23	23	110	108
NPlN sv	5	5	16	13	20	19	23	22	19	19	95	91
non-na	5	5	13	12	18	17	22	21	23	23	110	107
GPlN sv	5	6	16	13	20	20	24	23	18	18	91	89
non-na	5	7	13	11	19	18	22	23	22	22	106	96
Synthetic	10	13	25	23	25	23	26	24	32	32	205	136
Direct	0	0	4	3	20	20	25	24	25	26	105	104
FH(2)/(3) true cv	5	5	13	11	18	16	21	19	16	16	91	92
non-na (2)	4	3	9	8	16	14	19	18	20	20	108	108
non-na (3)	4	3	9	8	16	14	19	18	20	20	108	107
FH(2)/(3)	8	9	15	15	19	17	22	20	16	16	89	87
non-na (2)	8	7	12	11	16	14	20	18	19	19	106	105
non-na (3)	8	7	12	11	16	14	20	18	19	19	107	106
YR	−2	−2	11	10	17	15	20	18	16	16	97	101
non-na	−2	−3	8	7	15	13	18	17	20	20	114	115
NPlN	−2	−2	12	10	17	15	20	18	16	16	97	99
non-na	−2	−3	8	7	15	13	18	17	19	19	114	117
FH(2) sv	1	1	12	10	17	15	20	18	16	16	94	94
non-na	1	0	9	7	15	14	18	17	19	19	110	108
FH(3) sv	1	1	12	10	17	15	20	18	16	16	97	98
non-na	1	0	9	7	15	14	18	17	20	20	115	112
YR sv	4	4	13	10	17	15	21	19	16	16	93	92
non-na	4	2	9	7	15	14	18	18	19	19	112	110
NPlN sv	4	4	13	10	17	15	21	18	16	16	93	93
non-na	4	2	9	7	15	14	18	18	19	19	112	111
GPlN sv	4	4	13	11	17	16	21	17	16	16	89	89
non-na	4	3	9	8	16	14	19	18	20	19	105	102

Table 9. Simulated data on unemployment: $\bar{p} = 3.35$ and symmetric data; from top to bottom: $n = 100/N = 25000$, $n = 300/N = 75000$, $n = 500/N = 100000$ (hence $r = 4, 4, 5‰$).

	bias				accuracy				efficiency		reliability	
estimator	$\overline{rb}$	rb^{M}	$\overline{arb}$	arb^{M}	$\overline{are}$	are^{M}	$\overline{rE}$	rE^{M}	$\overline{eff}$	eff^{M}	$\overline{rel}$	rel^{M}
Synthetic	15	−2	34	23	35	23	36	24	25	25	162	116
Direct	1	2	7	5	44	41	55	51	58	56	114	114
FH(2) *true cv*	14	−3	32	20	33	21	36	23	18	18	95	83
non-na	15	−3	31	20	33	21	35	23	21	21	108	95
FH(3) *true cv*	14	−3	32	20	33	21	36	23	18	18	94	82
non-na	15	−3	31	20	33	21	35	23	21	21	107	96
FH(2)	40	19	45	20	46	20	47	23	14	14	78	67
non-na	40	19	45	19	46	19	47	22	16	16	85	74
FH(3)	40	19	45	20	46	20	47	23	14	14	79	70
non-na	41	19	45	19	46	19	48	22	16	16	86	76
YR	−16	−27	32	31	33	31	34	32	22	22	94	76
non-na	−15	−26	31	31	32	31	34	32	26	26	106	88
NPlN	−16	−27	33	32	33	32	35	32	22	22	93	76
non-na	−16	−27	32	32	33	32	34	32	26	26	105	89
FH(2) *sv*	4	−11	30	22	31	22	33	23	18	17	88	82
non-na	4	−11	30	20	31	21	33	23	20	21	100	100
FH(3) *sv*	4	−11	30	22	31	22	33	23	17	17	86	80
non-na	4	−11	30	20	31	21	33	23	20	20	98	96
YR *sv*	11	−4	28	18	31	20	34	24	23	23	104	100
non-na	12	−3	26	17	31	19	34	24	27	27	116	115
NPlN *sv*	12	−4	28	19	31	20	34	23	22	22	104	100
non-na	12	−3	27	17	31	20	34	24	27	26	115	113
GPlN *sv*	3	−5	19	15	31	25	37	30	31	31	102	105
non-na	6	−1	15	9	31	26	39	31	38	37	115	117
Synthetic	15	−3	34	23	35	23	35	24	30	30	190	130
Direct	0	1	4	4	26	25	32	32	34	32	106	106
FH(2) *true cv*	9	−3	23	13	28	18	31	21	18	18	87	89
non-na	9	−3	19	11	25	17	29	20	22	22	101	108
FH(3) *true cv*	9	−3	23	13	28	18	31	21	19	18	87	88
non-na	9	−3	19	11	25	17	29	20	23	22	101	107
FH(2)	20	4	30	16	32	18	34	20	15	15	82	80
non-na	19	4	27	13	30	16	33	18	18	18	94	95

Continued

FH(3)	20	4	30	16	32	18	34	20	15	15	84	81
non-na	19	4	27	13	30	16	33	18	19	18	95	94
YR	−4	−13	19	16	25	21	29	24	21	20	91	92
non-na	−4	−12	15	13	21	19	26	23	25	25	107	110
NPlN	−4	−13	20	16	25	21	29	24	20	20	90	91
non-na	−4	−12	15	14	21	20	26	23	25	25	106	109
FH(2) *sv*	1	−9	20	15	25	19	29	22	19	19	90	92
non-na	1	−7	15	11	21	18	26	21	24	24	106	112
FH(3) *sv*	1	−9	20	15	25	19	29	22	21	20	96	98
non-na	1	−7	15	11	21	18	26	21	25	25	113	120
YR *sv*	10	−3	23	15	27	18	31	21	18	17	87	88
NPlN *sv*	10	−3	23	15	27	18	31	21	18	17	87	87
non-na YR/NPlN *sv*	9	−3	20	12	25	17	28	20	22	21	101	104
GPlN *sv*	7	−4	21	14	26	18	30	22	18	18	87	88
non-na	6	−3	16	11	23	17	27	20	22	22	102	109
Synthetic	14	−3	34	23	34	23	35	24	29	29	206	126
Direct	0	−1	3	3	20	19	25	23	26	24	105	104
FH(2)/(3) *true cv*	8	−3	20	12	25	17	28	20	16	15	83	81
non-na (2)	8	−1	15	8	21	14	25	17	19	18	100	105
non-na (3)	8	−1	15	8	21	14	25	17	19	18	100	104
FH(2)	13	0	24	14	27	17	29	19	15	14	83	77
non-na	13	1	20	10	24	14	27	17	18	17	98	102
FH(3)	13	0	24	14	27	17	29	19	15	14	84	80
non-na	13	1	20	10	24	14	27	17	18	17	99	105
YR	−1	−8	17	14	22	18	26	20	17	17	87	88
non-na	−2	−7	11	9	18	16	22	19	21	21	106	110
NPlN	−1	−9	17	14	22	18	26	20	17	17	87	88
non-na	−2	−7	12	10	18	16	22	19	20	20	106	110
FH(2) *sv*	2	−6	17	12	22	17	26	20	17	17	87	86
non-na	1	−5	11	7	18	15	22	18	20	20	106	110
FH(3) *sv*	2	−6	17	12	22	17	26	20	17	17	91	90
non-na	1	−5	11	7	18	15	22	18	21	21	111	116
YR/NPlN *sv*	7	−3	20	12	24	18	27	20	16	15	84	81
non-na (YR)	7	−2	16	9	21	14	24	18	19	18	101	106
non-na (NPlN)	7	−2	16	9	21	14	24	18	19	18	101	103
GPlN *sv*	5	−3	19	11	23	18	27	20	16	16	83	80
non-na	5	−1	13	9	19	13	23	18	19	19	97	108

Table 10. Simulated data on unemployment: $\bar{p} = 3.35$ and negative asymmetry; from top to bottom: $n = 100/N = 25000$, $n = 300/N = 75000$, $n = 500/N = 100000$ (hence $r = 4, 4, 5‰$).

	bias				accuracy				efficiency		reliability	
estimator	$\overline{rb}$	rb^M	$\overline{arb}$	arb^M	$\overline{are}$	are^M	$\overline{rE}$	rE^M	$\overline{eff}$	eff^M	$\overline{rel}$	rel^M
Synthetic	25	−10	48	25	49	25	51	27	26	26	115	105
Direct	−1	−1	8	7	46	41	56	50	60	57	115	115
FH(2) *true cv*	26	−7	46	23	48	24	50	26	17	17	77	76
non-na	26	−8	46	22	47	22	50	25	20	20	89	86
FH(3) *true cv*	26	−7	46	23	48	24	50	26	17	17	75	72
non-na	27	−8	46	22	47	22	50	25	20	20	87	83
FH(2)	55	13	60	13	62	14	64	19	14	14	71	80
non-na	56	11	59	11	62	13	64	19	16	16	78	80
FH(3)	56	13	60	13	62	14	64	19	14	14	72	79
non-na	56	11	59	11	62	13	64	19	16	16	81	79
YR	−10	−32	41	36	43	36	45	37	23	23	78	66
non-na	−10	−32	40	33	42	35	44	36	27	27	88	75
FH(2) *sv*	14	−15	43	27	44	27	47	29	18	18	73	68
non-na	14	−16	42	26	43	26	46	27	20	21	83	81
FH(3) *sv*	14	−15	43	27	44	27	47	29	17	17	72	67
non-na	14	−16	42	26	43	26	46	27	20	20	81	78
YR/NPlN *sv*	20	−8	39	21	43	23	47	27	24	24	92	93
non-na YR	20	−5	36	19	41	23	46	27	28	24	106	106
non-na NPlN	20	−6	36	19	41	22	46	26	28	28	105	106
GPlN *sv*	11	−6	28	19	39	25	46	31	30	30	97	98
non-na	11	0	21	15	35	24	44	30	35	35	111	115
Synthetic	27	−9	48	25	49	25	49	25	32	32	163	129
Direct	1	1	4	3	26	24	33	30	35	30	109	109
true cv (2)	22	−6	36	15	40	18	44	21	18	17	82	86
non-na	22	−3	32	12	37	15	41	19	22	21	99	111
true cv (3)	22	−6	37	15	40	18	44	21	18	17	82	85
non-na	22	−3	32	12	37	15	41	19	22	21	99	109
FH(2)	34	1	45	15	46	17	49	19	15	14	78	80
non-na	34	3	42	12	44	16	47	18	17	17	90	100
FH(3)	34	1	45	15	46	17	49	19	15	14	79	80
non-na	34	3	42	12	45	16	47	18	18	17	91	97
YR	0	−15	25	20	30	23	36	27	22	22	90	89

Continued

non-na	−1	−10	18	14	25	20	30	24	27	27	111	116
NPlN	0	−15	25	20	30	23	36	27	22	22	89	87
non-na	−1	−10	18	14	25	20	30	24	27	27	109	115
FH *sv* (2)	7	−10	26	17	32	21	37	25	21	21	89	93
non-na	6	−5	20	12	26	18	32	21	25	25	110	121
FH *sv* (3)	7	−10	26	17	32	21	37	25	22	22	96	100
non-na	6	−5	20	12	26	18	32	21	27	27	119	132
YR *sv*	18	−6	33	16	37	20	40	22	19	18	83	82
non-na	17	−3	28	12	33	17	37	20	23	21	99	107
NPlN *sv*	18	−6	33	16	37	20	40	22	18	18	83	82
non-na	17	−3	28	12	33	17	37	20	22	21	99	105
GPlN *sv*	10	−7	25	14	31	20	37	23	21	21	86	90
non-na	9	−2	17	9	26	17	31	21	25	24	105	116
Synthetic	27	−9	49	25	49	25	50	25	31	31	157	126
Direct	0	0	4	3	20	19	25	23	27	23	109	106
true cv (2)	20	−6	32	14	36	17	40	19	16	15	77	81
non-na	17	−3	26	9	31	14	35	17	19	18	96	105
true cv (3)	20	−6	32	14	36	17	40	19	16	15	78	81
non-na	18	−3	26	9	31	14	35	17	20	18	96	104
FH(2)	26	−4	37	13	40	16	43	18	15	14	75	80
non-na	24	−1	32	10	35	14	39	17	18	17	91	101
FH(3)	26	−4	37	13	40	16	43	18	15	14	76	79
non-na	24	−1	32	10	35	14	39	17	18	17	92	101
YR	3	−11	22	16	27	20	34	24	18	18	82	81
non-na	0	−8	13	10	20	16	24	20	22	22	106	109
NPlN	4	−11	23	16	28	20	34	24	18	18	82	81
non-na	0	−8	13	10	20	17	24	20	22	22	106	108
FH(2) *sv*	7	−9	23	14	28	19	35	22	18	18	82	85
non-na	4	−5	14	9	20	15	25	19	21	21	105	113
FH(3) *sv*	7	−9	23	14	28	19	35	22	19	18	87	90
non-na	4	−5	14	9	20	15	25	18	22	22	112	120
YR/NPlN *sv*	16	−8	30	14	33	18	38	20	16	15	78	79
non-na (YR)	13	−4	22	9	27	15	31	18	20	18	97	104
non-na (NPlN)	13	−4	22	9	27	15	31	17	20	19	97	104
GPlN *sv*	10	−6	23	13	29	19	35	22	17	16	78	85
non-na	7	−4	16	9	23	15	28	18	21	21	101	103