A SEMI-PARAMETRIC FAY-HERRIOT-TYPE MODEL
WITH UNKNOWN SAMPLING VARIANCES

Silvia Polettini

1 Dipartimento di Metodi e Modelli per il Territorio, l’Economia e la Finanza, Sapienza
Università di Roma, (e-mail: silvia.polettini@uniroma1.it)

KEYWORDS: Bayesian hierarchical model, Dirichet process, small area, unknown
sampling variances

1 Introduction

Availability of survey data allows users to obtain estimates for a whole variety
of subpopulations, called small areas, obviously not restricted to the planned
domains. Often the sample sizes for such domains are too small to provide
sufficiently accurate design-based estimate of the domain parameters. To
improve such estimates, indirect estimators introduce a linking model that relies
on auxiliary information to connect small areas thus borrowing strength and in-
creasing the effective sample size. Modern methods for small area estimation
(SAE) heavily rely on mixed effects modelling; see Rao, 2003.

This contribution focuses on area level models, that rely on aggregated,
area-specific, quantities and allow to take into account the sampling design
through the direct survey estimates and their corresponding (design-based)
variance estimates. The most popular area-level model is the Fay-Herriot
model (Fay & Herriot, 1979). It prescribes a sampling model for the direct
survey estimates, supplemented by a linking model for the small area param-
eters of interest. Let \( m \) be the number of sampled small areas and denote
by \( \hat{\theta}_i \), the design unbiased, direct survey estimators of the small area parame-
ters \( \theta_i \), \( i = 1, \ldots, m \). Let \( \varepsilon_i \) be the associated sampling error and let \( \psi_i \) be the
corresponding variance under the sampling model. Assuming \( \varepsilon_i \sim N(0, \psi_i) \)
indipendently, we get \( \hat{\theta}_i \sim N(\theta_i, \psi_i), i = 1, \ldots, m \). To borrow information
across areas, a linking model for \( \theta_i \) is introduced, namely \( \theta_i = x'_i \beta + v_i \), where
\( x_i = (x_{i1}, \ldots, x_{ip})' \) is a vector of auxiliary variables, \( \beta \) is a vector of regression
coefficients, and finally the \( v_i \)'s are area-specific random effects accounting for
heterogeneity and lack of fit. Normality of the \( v_i \)'s is usually assumed, so that
\[
\theta_i | \beta, \sigma^2_v \sim N(x'_i \beta, \sigma^2_v), \quad i = 1, \ldots, m.
\]

(1)

Combining the previous equations, one obtains a mixed effects linear re-
gression model with normal random components, \( \hat{\theta}_i = x'_i \beta + \varepsilon_i + v_i \). For known
\(\beta, \sigma^2, \psi_i\), the Best Linear Unbiased Predictors (BLUP) of the small area parameters can be obtained. A large part of the literature has focused on procedures for estimating model parameters to obtain the so called EBLUP and propose measures of uncertainty for this predictor.

In the Fay-Herriot setup, accuracy and precision of small area predictors depend on the validity of the model. In this respect, two specific major underlying assumptions are investigated in this contribution, namely the normality of random effects leading to (1) and the assumption of known sampling variances. While the distributional assumption on \(\varepsilon_i\) may be justified by the properties of the direct estimators \(\hat{\theta}_i\), the normality assumption for the random effects \(\nu_i\) has no justification other than computational convenience and is difficult to detect in practice, since it involves unobservable quantities. In applications, examples abound of skewed distributions that require suitable (often, logarithmic) transformations of the data, as in estimation of economic aggregates; treating overdispersion may also be an issue, as in estimation of poverty; moreover, multimodalities may result from omission of important covariates. Finally, the impact of outliers is an ubiquitous problem. Some of these issues have been addressed via general distributions and finite mixture models. Our proposal does not rely on ad hoc models and avoids issues typical of mixture models, trying to exploit the structure in the data to gain efficiency without incurring in a consistency-efficiency trade-off.

As to the second aspect, most of the literature has focused on reflecting uncertainty in estimating \(\sigma^2\), while assuming known \(\psi_i\) (see Prasad & Rao, 1990; Datta & Lahiri, 2000). In practice however these quantities are estimated from the sample; in some applications, smoothed estimators are introduced, and then treated as known, with the consequence that inferences neglect the uncertainty associated with this step. Sensitivity of inferences to uncertainty about sampling variances has been investigated by Bell, 2008. See Maiti et al., 2014 for a detailed discussion of the literature.

2 Proposed model

We extend the Fay Herriot model within a Bayesian approach in two directions. First, uncertainty on variances is introduced in the model, so as to reflect the fact that they are actually estimated from survey data. Assuming that (independent) information is available about sampling variances, we adopt the approach of You & Chapman, 2006 assuming a common distribution generating the variance parameters (Dass et al., 2012). Second, a Bayesian semiparametric approach is pursued: the default normality assumption for random
effects is replaced by a nonparametric specification, namely using a Dirichlet process (Ferguson, 1973; Antoniak, 1974). Under this setup, different levels of shrinkage of the small area means are possible and outliers can be accounted for. The complete model specification reads as follows:

$$\hat{\theta}_i = \theta_i + e_i, \quad e_i \sim N(0, \psi_i), \quad \text{independently, } i = 1, \ldots, m \quad (2)$$

$$\theta_i = x_i'\beta + \nu_i, \quad \nu_i \sim G(\cdot), \quad \text{independently, } i = 1, \ldots, m \quad (3)$$

$$G \sim DP(M, N(0, \sigma^2)) \quad (4)$$

$$d_i S^2_i \sim \psi_i \chi^2_{d_i}, \quad \text{independently and independent on } \hat{\theta}_i, \quad i = 1, \ldots, m \quad (5)$$

$$\psi_i \sim IG(a_0, b_0), \quad \text{independently, } i = 1, \ldots, m \quad (6)$$

$$\sigma^2_i \sim IG(a_1, b_1) \quad (7)$$

$$\beta \sim N(0, dI) \quad \text{where } I \text{ is the identity matrix} \quad (8)$$

$$M \sim Gamma(a_2, b_2) \quad (9)$$

where $d_i$ represent the degrees of freedom for estimating the sampling variance, $a_0, b_0, a_1, b_1, d, a_2, b_2$ are known constants and $IG(\cdot, \cdot)$ denotes the Inverse Gamma distribution. Under (2)–(9) estimation of area means and variances is performed at the same time and smoothing of variances is performed within the same model. The modelling assumption in (5) is fully justified in a normal setting, with simple random sampling; in our case, it can be taken as an attempt to achieve a more comprehensive quantification of uncertainty in estimating domain means that also accounts for variability of sampling variances. See Maiti, 2003 for comments. The proposed hierarchical model differs from the one analysed in Maiti et al., 2014 for the nonparametric modelling of the random effects.

It is of particular interest to understand the performance of the model in predicting small area quantities under the extended Fay Herriot model above, primarily the domain means $\theta_i$ in (3). Besides measures of prediction error, for which a natural quantification of uncertainty under the proposed approach is the posterior variance, an aspect that has received less attention in the literature is the construction of confidence intervals for the area means. Recently Diao et al., 2014 focus on deriving accurate CIs for small area means using the EBLUPs and estimators of MSPE of EBLUPs based on various methods of estimation of model parameters. Under the proposed Bayesian model, a natural approach to the former issue is to produce posterior credible intervals. We consider here both problems and compare the proposed method to existing solutions. The variable shrinkage property allows us to reduce prediction error and although the proposed model accounts for all sources of uncertainty, applications show that the resulting credibility intervals tend to be shorter than the
corresponding frequentist solutions. The frequentist coverage of such intervals is also analysed through simulation, comparing the proposed approach to other solutions presented in the literature.

References


