STATISTICA, anno LXXV, n. 1, 2015

SMALL AREA ESTIMATION WITH COVARIATES PERTURBED FOR DISCLOSURE LIMITATION

Silvia Polettini [1]

Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, Roma, Italia

Serena Arima

Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, Roma, Italia

1. Introduction

In recent years, small area estimation has emerged as an important area of statistics as private companies and public agencies need to extract the maximum information from sample survey data. Sample surveys are generally designed to produce reliable estimates of totals and means of variables of interest for given domains. However, governments and general users are more and more interested in obtaining statistical summaries for smaller domains, called small areas, created by cross classifying demographic and geographical variables. Due to budget constraints, the samples in these subdomains are often too small and direct survey estimates may be unreliable, with exceedingly high standard errors. In order to obtain improved estimates, model-based approaches, usually based on mixed effects regression models, are introduced to link the small areas and borrow strength from similar domains (see Rao, 2003; Datta, 2009; Pfefferman, 2013 for a review). The model proposed by Fay and Herriot (1979) is the most popular small area model when data are available at area-level. It borrows strength from data available from all areas by assuming a hierarchical structure and incorporates auxiliary information from other data sources such as administrative records or censuses. The frequentist predictor of small area mean based on the Fay-Herriot model, which is also known as empirical best linear unbiased predictor (EBLUP), results in a convex combination of the direct estimator and the synthetic estimator from the model. Properties of the predictors of small area means, such as bias and mean squared error, are derived conditionally on the auxiliary information and under the assumption that auxiliary data are measured without error. When auxiliary information is measured with error, an estimator, accounting for the measurement error in the covariates, has been proposed in Ybarra and Lohr (2008). They

[1] Corresponding Author. E-mail: silvia.polettini@uniroma1.it

suggested a suitable modification to the Fay-Herriot estimator that accounts for sampling variability in the auxiliary information, and derive its properties, in particular showing that it is approximately unbiased. In a subsequent paper Arima *et al.* (2012a) rewrite the measurement error model as a hierarchical Bayesian model. Their predictors exhibit smaller empirical mean squared errors than the frequentist predictors of Ybarra and Lohr (2008) and are more stable in terms of variability and bias.

To grant the request for useful statistical data without disclosing confidential information about respondents, Statistical Offices routinely apply statistical disclosure control (SDC) methods (for a comprehensive presentation of the topic, see Willenborg and de Waal, 2001). The risk of disclosure can be lowered by aggregating or suppressing data, referred to as non perturbative methods, or by applying perturbative methods, that mask the data purposely. Quoting Little (1993), the paradigm is that "masking is primarily concerned with identification of *individual* records, whereas statistical analysis is concerned with making inference about *aggregates*", so that the information loss can in principle be confined to the individual level. Examples of perturbative methods are noise addition (Kim, 1986; Fuller, 1993; Brand, 2002), designed for continuous data, and Post Randomization Method (PRAM, Gouweleeuw *et al.*, 1998) and data swapping (Dalenius and Reiss, 1982), designed for categorical data. The implementation of disclosure limitation protocols prevents users from using the observed survey data to estimate the small area aggregates of interest. Area-level models partially overcome the problem by relying on external sources; however the perturbed sample may still provide valuable information for the small area model. The focus of this paper will be the prediction of small area quantities when the available covariates arise from data treated for disclosure limitation. We consider random data perturbation techniques, focusing in particular on PRAM and noise addition. The protection methods just mentioned perturb the data through ad-hoc probabilistic models that in fact introduce measurement error. We can therefore exploit the analogy between random data perturbation schemes and measurement/misclassification errors to cast the small area estimation problem within the framework of Ybarra and Lohr (2008) and Arima *et al.* (2012a). Their models however only considered the particular and somehow unrealistic situation in which the auxiliary covariates measured with error are all continuous variables. We propose a Bayesian area-level model analogous to the one in Arima *et al.* (2012a) and explicitly introduce the error distributions induced by protection of both categorical and continuous variables for confidentiality purposes. In this case the information available on the perturbation scheme allows us to model the misclassification process without having to rely on strong assumptions about the measurement error, whose validity is crucial when dealing with categorical data.

The paper is organized as follows: in Section 2 we give a brief overview of the masking techniques with focus on the Post Randomization Method. In Section 3 we introduce the measurement error models as the starting point of the proposed models described in Section 4. In Section 5 the performance of the proposed approach is investigated through simulated data. We conclude with a brief discussion in Section 6.

2. Masking techniques: Post Randomization and Noise Addition

The Post Randomization Method was proposed in Gouweleeuw *et al.* (1998) to protect a microdata file of categorical variables by allowing the original scores of certain variables on all records to change to possibly different scores according to a prescribed probability mechanism. PRAM is defined through a transition matrix that specifies the probability that each record's categories are transformed into each of the other categories.

Consider for simplicity applying PRAM to a single categorical variable Z, with categories $\{1, \ldots, K\}$. Let Z^* denote the corresponding perturbed variable. Let $p_{lh} = Pr(Z^* = h|Z = l)$, be the probability of transition from category l of Z to category h of Z^*. PRAM consists of the following: given $Z_j = l$ for subject j in the original microdata file, the score on Z^* for this record, namely Z^*_j, is determined by sampling from a discrete probability distribution with masses $p_{l1}, ..., p_{lK}$ at scores $1, \ldots, K$. All records in the original data set are protected, with the procedure being applied independently to each unit. The $K \times K$ matrix of transition probabilities $P = \{p_{lh}\}_{l,h=1,\ldots,K}$ is referred to as the PRAM matrix. PRAM can also be applied to $p > 1$ categorical variables $Z_1, \ldots, Z_p$, independently or simultaneously. In the latter case, PRAM can be defined by specifying the transition matrix for the compounded variable whose $K_1 \times K_2 \cdots \times K_p$ categories are formed by combination of all $K_1, K_2, \ldots, K_p$ scores for all variables considered; in the former case, the PRAM transition matrix P is the Kronecker product of the transition matrices $P_1, \ldots P_p$ for the p variables considered: $P = P_p \otimes P_{p-1} \otimes \cdots \otimes P_1$. Clearly the application of independent PRAM is likely to destroy observed information about the structure of association among variables. Under the second instance of PRAM, certain dependencies between the variables used to form the compounded variable can be taken into account. Also, if there are variables that must be kept as in the original data file, the compounded variable may include those variables that must remain unchanged. Choice of the entries of the transition matrix P depends on the extent of protection that is deemed adequate and on particular data features that are to be maintained in the released data. Indeed PRAM may be designed to preserve approximately the distribution of the protected variables (see Gouweleeuw *et al.*, 1998) as it will be discussed next. Moreover, including variables that must be kept unchanged and specifying block-diagonal PRAM matrices make it possible to preserve the distribution in pre-specified subpopulations.

Denote by T (T^*, respectively) the vector of frequencies of the original (perturbed) variable Z (Z^*); let T_l and T^*_l represent the l-th element of the vectors T, T^* respectively, that is, the frequency on the l-th category of the original and perturbed variables Z, Z^*. Also, denote by z the vector of scores on Z observed on all n units of the original microdata file. Since $p_{lh} = Pr(Z^* = h|Z = l) = E(I(Z^* = h)|Z = l)$, where $I(\cdot)$ is the indicator function, $V(I(Z^* = h)|Z = l) = p_{lh}(1 - p_{lh}) = \{V_l\}_{h,h}$, and $Cov(I(Z^* = h), I(Z^* = j)|Z = l) = -p_{lh}p_{lj} = \{V_l\}_{h,j}$, it is easy to verify that

$$E(T^*|z) = P'T.$$

Therefore, if the transition matrix P is invertible,

$$\hat{T}^P = (P^{-1})'T^* \tag{1}$$

is an unbiased estimator for T (Gouweleeuw *et al.*, 1998).

The conditional variance of the estimator is $V(\hat{T}^P|z) = (P^{-1})'V(T^*|z)P^{-1}$, where $V(T^*|z) = \sum_l T_l V_l$; here V_l is the covariance matrix of the transition process from original score l $(l = 1, \ldots, K)$, with elements

$$\{V_l\}_{h,j} = \begin{cases} p_{lh}(1 - p_{lh}) & \text{if } h = j \\ -p_{lh}p_{lj} & \text{if } h \neq j \end{cases} \qquad h, j = 1, \ldots, K \tag{2}$$

as already discussed. Gouweleeuw *et al.* (1998) propose the following plug-in estimator:

$$\hat{V}(T^*|z) = \sum_l \hat{T}_l^P V_l \tag{3}$$

to quantify the uncertainty introduced by the noise process.

The so-called invariant PRAM amounts to choosing the transition matrix P such that $P'T = T$. One simple choice of invariant P is a matrix with entries

$$p_{hh} = 1 - \vartheta * T^{min}/T_k; \qquad p_{lh} = \frac{1 - p_{hh}}{K - 1}, \qquad l, h = 1, \ldots, K, \quad l \neq h, \tag{4}$$

where T^{min} is the minimum observed frequency among categories of Z and the initial probability $\vartheta \in (0, 1)$ of changing each of the original categories is adjusted to satisfy the invariance requirement

$$P'T = T. \tag{5}$$

A different, two-stage, solution is reported e.g. in Willenborg and de Waal (2001). The property (5) defines a frequency-invariant transformation of the original variable Z; in this case the perturbed frequency table is itself an unbiased estimator for the original table. As long as point estimation is concerned, knowledge of the transition matrix is not needed. Full knowledge of the PRAM matrix is needed in (3) for variance estimation even under invariant post randomization.

The simplest form of noise addition (Kim, 1986; Fuller, 1993; Brand, 2002) amounts to adding independent random noise η to the observed variables. Let Σ denote the variance-covariance matrix of the continuous variables of interest. The random noise η is given a known distribution (usually, the Normal) with mean zero and covariance matrix $\Sigma_\eta = \alpha\Sigma, \alpha > 0$. The level of noise introduced clearly depends on the parameter α. Denoting by μ and Σ the mean and covariance of X, choice of $\Sigma_\eta = \alpha\Sigma$ implies that the perturbed variable $X^* = X + \eta$ has the same mean as the original variables and covariance matrix $V(X^*) = (1 + \alpha)\Sigma$. Correlations are exactly preserved, and Kim (1986) shows that the covariance matrix of the original data can be consistently estimated from the masked data as long as α is known.

3. Measurement error models

Measurement error is still an open issue in statistical practice. The use of covariates which are affected by measurement error has three main effects. First, it causes bias in parameter estimation for statistical models; second, it leads to a loss of power for detecting relationship among variables and third it masks the features of the data, making graphical model analysis difficult (Carroll *et al.*, 2006). Although measurement error models have been mainly developed for the analysis of experimental data, their role in official statistics is crucial. For example, most of the unit-level small area models make use of covariates which are often measured imprecisely either because they arise from poor quality administrative data, or due to lack of memory, rounding, and other obvious mechanisms related to respondents. Also, measurement error may be artificially induced for confidentiality reasons, which is the focus of this paper. To prevent identification of respondents and protect their privacy, prior to data release National Statistical Offices may perturb survey data according to an appropriate random mechanism (so called data masking). Work by Fuller (1993) and Little (1993) shows how to account for some SDC treatments using standard statistical methods for the analysis of incomplete data when the values of the masking parameters are disclosed.

Several different small area estimators have been proposed to correct for biases of estimation caused by measurement error. Ybarra and Lohr (2008) have considered a Fay-Herriot area-level model with the auxiliary information used in the covariates measured with error. Let Y_i be an area-level summary of the response variable for area i. Along with Y_i, auxiliary data w_i are available as supplementary information for the ith area, where w_i is a $q \times 1$ vector. The Fay-Herriot model is defined as

$$Y_i = \theta_i + \epsilon_i, \quad \theta_i = w_i'\delta + u_i, \quad i = 1, \cdots, m, \tag{6}$$

where the sampling errors ϵ_i's and the model errors u_i's are all independently distributed. It is assumed that $\epsilon_i \sim N(0, \psi_i)$ and $u_i \sim N(0, \sigma_u^2)$, $i = 1, \cdots, m$. The sampling variances ψ_i's are assumed known. When the model parameters and the auxiliary variables are *known*, the best unbiased predictor of θ_i, that minimizes the MSE of prediction is given by

$$\tilde{\theta}_i = \gamma_i Y_i + (1 - \gamma_i)\{w_i'\delta\}, \tag{7}$$

where $\gamma_i = \sigma_u^2/(\psi_i + \sigma_u^2)$, and the MSE of $\tilde{\theta}_i$ is given by $\psi_i\gamma_i$. Ybarra and Lohr (2008) considered the situation in which some of the covariates are measured with error: let w_i be the auxiliary variables measured without error and let X_i be the $p \times 1$ vector of covariates measured with error. The true values X_i of the covariates are not perfectly observable and they are estimated by a set of estimators $X_i^* = \hat{X}_i$ such that $MSE(X_i^*) = C_i$, where C_i are known quantities. They showed that if the unknown X_i in the best predictor $\tilde{\theta}_i$ is replaced by its estimator X_i^*, which is subject to error, the resulting naive predictor

$$\tilde{\theta}_{i,N} = \gamma_i Y_i + (1 - \gamma_i)\{w_i'\delta + X_i^{*}{}'\beta\}, \tag{8}$$

has MSE that is larger than the MSE of the direct estimator Y_i.

Hence Ybarra and Lohr (2008) proposed the following estimator

$$\tilde{\theta}_{i,BP} = \gamma_i^{BP} Y_i + (1 - \gamma_i^{BP})\{w_i'\delta + X_i^{*\prime}\beta\}, \tag{9}$$

where $\gamma_i^{BP} = \frac{\sigma_u^2 + \beta' C_i \beta}{\sigma_u^2 + \beta' C_i \beta + \psi_i}$. They proved that $\tilde{\theta}_{i,BP}$ has minimum mean squared error among all linear combinations of Y_i and $w_i'\delta + X_i^{*\prime}\beta$ (Theorem 1 in Ybarra and Lohr, 2008). When the model parameters are unknown, they suggested to plug in formula (9) the estimates of the parameters obtained with the modified Prasad-Rao's method of moments described in Ybarra and Lohr (2008).
In a subsequent paper, Arima *et al.* (2012a) provided a Bayesian solution for the measurement error problem for the same small area setup. They reformulated the frequentist model described above into a multi-stage hierarchical Bayesian model as follows:

Stage 1. $Y_i | X_i^*, \theta_i, \beta, \delta, \sigma_u^2, w_i \overset{ind}{\sim} N(\theta_i, \psi_i)$, $i = 1, \ldots, m$;

Stage 2. $X_i^* | \theta_i, X_i \overset{ind}{\sim} N(X_i, C_i)$, $\quad i = 1, \ldots, m$;

Stage 3. $\theta_i | X_i, \beta, \delta, \sigma_u^2 \overset{ind}{\sim} N(X_i'\beta + w_i'\delta, \sigma_u^2)$, $i = 1, \ldots, m$;

Stage 4. Prior distribution: $\pi(X_1, ..., X_m, \beta, \delta, \sigma_u^2) \propto 1$.

They showed that the above posterior density is proper under the very mild condition that $m > p + q + 2$. They also found that the Bayesian estimator of θ has smaller empirical mean squared error than the frequentist predictors in Ybarra and Lohr (2008) and is more stable in terms of variability and bias.

4. The proposed model

Exploiting the analogy between random data perturbation schemes and measurement/misclassification error, we consider measurement error models with covariates perturbed for disclosure limitation. We consider two random data perturbation methods, namely PRAM for categorical variables, and noise addition for continuous variables. In particular, in order to obtain area-level covariates from record-level data, we propose to aggregate unit-level categorical variables to the area level as the number of sampled records in all categories of each variable. As discussed below, we propose to rephrase the problem of perturbed categorical variables in terms of perturbed continuous variables, so that the results in Ybarra and Lohr (2008) and Arima *et al.* (2012a) can be easily adapted.

Suppose there are m areas labelled $1, \ldots, m$, each with $n_1, \ldots, n_m$ individual observations. We denote by y_i the response of the $i-$th area ($i = 1, ..., m$). Let w_i be a continuous or discrete area-specific covariate measured without error. Let X_i^* be the observed continuous area-specific covariate measured with error from an external source or resulting from perturbation of the unit level data X_{ij}. Let $Z_{ij}^*, j = 1, \ldots, n_i$ be the score of the discrete auxiliary variable observed for record j within area i, and obtained by post randomization of the original variable Z_{ij}. We assume that the categorical variables Z, Z^* have K possible categories and that PRAM is performed uniformly over all sampled units. Let T_i be vector of frequencies of the unperturbed categorical variable Z within area i, and let T_i^* be the corresponding vector after PRAM of variable Z. If the area sizes are not too small, the conditional distribution of the vector of frequencies T_i^* in the perturbed sample given the original score vector z can be approximated by a K-variate degenerate normal distribution with mean $P'T_i$ and covariance matrix depending on both the PRAM matrix P and the unperturbed sample frequencies T_i, according to formula (3). As a consequence, having denoted by $T_{i,l}$ and $T_{i,l}^*$ the l-th element of T_i and T_i^*, respectively, the covariates $(T_{i,1}^*, \dots T_{i,K-1}^*)$, given the original score vector z can be modelled by a multivariate normal distribution with mean μ_i^Z equal to the first $K-1$ elements of $P'T_i$ and covariance matrix

$$\Sigma_i^Z = \sum_{l=1}^{K-1} T_{i,l} V_{i,l}^-,$$

where $V_{i,l}^-$, $l = 1, \dots, K$ is the $K-1 \times K-1$ submatrix obtained by dropping the K-th row and column of $V_{i,l}$ in (2).

Since we work at the area level, the quality of the approximation will depend on both the area size and the number of categories. In our problem, when the invariant PRAM is selected, the multinomial variate $(T_{i,1}^*, \dots T_{i,K}^*)$ has mean vector equal to the observed frequencies T_i; therefore the approximation is good provided the categories of the post randomized variable have large enough within area frequencies. This is indeed a stringent requirement. In Section 5 we investigate the robustness to departures from this assumption, finding a certain stability in the model performances under investigation (see Table 2). If the above assumption is not met, however, the model can be adapted to include, rather than the normal approximation, the exact multinomial distribution, whose parameters we estimate using the information on PRAM.

In light of the previous considerations, we refer to the following model:

Stage 1 $Y_i | X_i^*, Z_i^*, \theta_i, \beta, \delta \overset{ind}{\sim} N(\theta_i, \psi_i)$;

Stage 2.1 $X_i^* | \theta_i, X_i \overset{ind}{\sim} N(X_i, C_i)$;

Stage 2.2 $(T_{i,1}^*, \dots, T_{i,K-1}^*) | \theta_i, Z \overset{ind}{\sim} N_{(K-1)}(\mu_i^Z, \Sigma_i^Z)$;

Stage 3 $\theta_i | X_i, T_i, \boldsymbol{\beta}, \delta \overset{ind}{\sim} N(\beta_1 T_{i,1} + \dots + \beta_{K-1} T_{i,K-1} + \beta_K X_i + \delta w_i, \sigma_u^2)$, $i = 1, \dots, m$;

Stage 4 Prior distribution $\pi(X_1, \dots, X_m, T_1, \dots, T_m, \boldsymbol{\beta}, \delta, \sigma_u^2) \propto 1$.

Stage 1 defines the model's likelihood. Stage 2.1 and Stage 2.2 define the measurement error model for the continuous and the discrete covariates, respectively. Since μ_i^Z and Σ_i^Z in Stage 2.2 depend on the unknown frequency distribution T_i of Z in the $i-$th area, we use the plug-in estimators of formula (1) and (3). Notice that estimation of μ_i^Z is required only under the non invariant PRAM. In Stage 3 the parameter of interest θ_i is modelled as a function of the covariates. Following

Arima *et al.* (2012a), we specify in Stage 4 a uniform improper prior that yields to a proper posterior density under the mild condition that $m > p + q + 2$.

5. Simulation study

In this section, we use simulated data to illustrate the performance of the proposed approach in estimating model parameters in different PRAM scenarios. We take $m = 60$ and generate $Z_{ij} \sim Multinomial(1, p = (0.25, 0.40, 0.35))$, $i = 1, ..., m$ and $j = 1, ..., n_i$ where n_i ranges from 50 to 200. For each area i, we denote $T_{i,1}, T_{i,2}, T_{i,3}$ the absolute frequencies of the categories of Z over the n_i units. We also generate $w_i \sim N(0, 1)$. We set

$$\begin{array}{rcl} \theta_i & \sim & N(10 + 2w_i + 2T_{i,1} + 3T_{i,2}, 1) \\ y_i & \sim & N(\theta_i, 3) \end{array}$$

We consider both invariant and non invariant PRAM, and three different levels of perturbation (note that the diagonal elements of the PRAM matrices represent the probability of not changing the score for a single record). For the invariant PRAM, at each iteration, we generate the perturbed Z^*_{ij} covariates according to the following perturbation scenarios:

$$P_1 = \begin{pmatrix} 0.900 & 0.050 & 0.050 \\ 0.031 & 0.938 & 0.031 \\ 0.034 & 0.034 & 0.932 \end{pmatrix} \; P_2 = \begin{pmatrix} 0.800 & 0.100 & 0.100 \\ 0.062 & 0.876 & 0062 \\ 0.068 & 0.068 & 0.864 \end{pmatrix} \; P_3 = \begin{pmatrix} 0.700 & 0.150 & 0.150 \\ 0.093 & 0.814 & 0.093 \\ 0.102 & 0.102 & 0.796 \end{pmatrix}$$

The P matrices above are generated by selecting the values $(0.1, 0.2, 0.3)$, respectively, for the initial probability ϑ of changing each of the observed scores prior to the adjustment (see formula (4)) needed to achieve the invariance property.

For each scenario, we estimate the proposed model, labelled M_1, and compare it to two alternative models, M_2 and M_3. For sake of completeness, we also introduce the benchmark model M_0 in which we use the area-level frequencies of the first $K - 1$ categories of the unperturbed Z as auxiliary variables.
Model M_2 is defined as

Stage 1 $Y_i | X^*_i, \theta_i, \beta, \delta \overset{ind}{\sim} N(\theta_i, \psi_i)$;

Stage 2 $\theta_i | X_i, \boldsymbol{\beta}, \delta, \sigma^2_u \overset{ind}{\sim} N(\beta_1 \hat{T}^P_{i,1} + ... + \beta_{K-1} \hat{T}^P_{i,K-1} + \beta_K X_i + \delta w_i, \sigma^2_u)$, $i = 1, \ldots, m$;

Stage 3 Prior distribution $\pi(X_1, ..., X_m, \boldsymbol{\beta}, \delta, \sigma^2_u) \propto 1$,

where $\hat{T}^P_i = (\hat{T}^P_{i,1}, \hat{T}^P_{i,2}, \ldots \hat{T}^P_{i,K})$ is obtained as $\hat{T}^P_i = (P')^{-1} T^*_i$.
Model M_2 accounts for measurement error in the categorical covariates by considering the first $K - 1$ elements of the PRAM estimator $\hat{T}^P_i$ instead of the observed covariates $T^*_{i,1}, \ldots, T^*_{i,K-1}$. Indeed, the latter are not treated as random variables as in the proposed model; rather, the PRAM estimates of the unperturbed covariates $T_{i,1}, \ldots, T_{i,K-1}$ are plugged into the model equation and used as covariates

measured without error. Making a parallel with the estimators proposed for the measurement error in continuous auxiliary variables, the idea behind the formulation of model M_2 is akin to the plug-in estimator defined in (8).

Model M_3 ignores the measurement error in the categorical covariates and uses the frequencies of the perturbed variable Z^* as auxiliary variables measured without error, that is

Stage 1 $Y_i|X_i^*, \theta_i, \sigma_u^2, \beta, \delta \overset{ind}{\sim} N(\theta_i, \psi_i)$;

Stage 2 $\theta_i|X_i, \boldsymbol{\beta}, \delta, \sigma_u^2 \overset{ind}{\sim} N(\beta_1 T_{i,1}^* + ... + \beta_{K-1} T_{i,K-1}^* + \beta_K X_i + \delta w_i, \sigma_u^2)$, $i = 1, \ldots, m$;

Stage 3 Prior distribution $\pi(X_1, ..., X_m, \boldsymbol{\beta}, \delta, \sigma_u^2) \propto 1$.

We performed a total of 100 iterations. Given the characteristics of the problem under study, that amounts to a random perturbation model, the results exhibit a great stability, that enables us to limit the number of simulations. Table 1 shows the results of the simulation study for the invariant PRAM formulation. For each model, each row shows the posterior means of the unknown parameters under the different scenarios. In parentheses, we show the actual coverage percentage (ACP) of 95% credibility intervals for the model parameters and the root mean squared errors (RMSE).

All models yield very similar and trustworthy estimates under the PRAM matrix P_1, although with quite different RMSE. When the probability of misclassification increases, model M_2 performs very poorly. This is in line with the results in Ybarra and Lohr (2008) who show that using the estimator of the covariate affected by error as in Equation (8) may induce bias and increase MSEs of model estimates. On the other hand, M_1 and M_3 show similar results when $P = P_2$ but when $P = P_3$, the proposed model outperforms M_3 in terms of both ACP and RMSE.

The validity of the normal approximation has been questioned by a referee. Indeed the within-area frequencies of the categories of Z might be small thus invalidating the normal approximation. To investigate the effect of the violation of such an assumption we conducted a simulation study, letting n_i range from 5 to 20. The results are reported in Table 2. It seems that the sample size does not dramatically affect the estimates whose actual coverage probabilities of the 95% credible intervals are very similar to those obtained previously.

Let us now consider the non invariant PRAM scenario: at each iteration, we generate the perturbed Z_{ij}^* covariates according to the following perturbation matrices:

$$\tilde{P_1} = \begin{pmatrix} 0.900 & 0.070 & 0.030 \\ 0.010 & 0.900 & 0.090 \\ 0.050 & 0.050 & 0.900 \end{pmatrix} \quad \tilde{P_2} = \begin{pmatrix} 0.800 & 0.150 & 0.050 \\ 0.010 & 0.800 & 0.190 \\ 0.100 & 0.100 & 0.800 \end{pmatrix} \quad \tilde{P_3} = \begin{pmatrix} 0.700 & 0.250 & 0.050 \\ 0.100 & 0.700 & 0.200 \\ 0.150 & 0.150 & 0.700 \end{pmatrix}$$

Notice that we chose the same initial probabilities ϑ of staying with the original score as in the invariant setup. Due to the adjustment required to achieve the invariance property (5), the effective protection level may differ among the two types of post-randomization. For ease of comparison, we computed the posterior

TABLE 1
Invariant PRAM: posterior mean by misclassification probability P for different models. In parentheses ACP and RMSE, respectively. In the last line, we show the average ACP and RMSE of the small area means.

		P_1	P_2	P_3
M_0	δ_0	10.008 (1.00; 0.067)	10.007 (1.00; 0.072)	10.001 (1.00; 0.067)
	δ_1	2.322 (1.00; 0.322)	2.223 (1.00; 0.322)	2.322 (1.00; 0.322)
	β_1	1.943 (1.00; 0.057)	1.933 (1.00; 0.057)	1.943 (1.00; 0.057)
	β_2	3.044 (1.00; 1.044)	3.041 (1.00; 1.041)	3.044 (1.00; 1.044)
	θ	(0.927; 3.492)	(0.930; 3.540)	(0.932; 3.553)
M_1	δ_0	10.563 (0.995; 2.014)	11.671 (0.980; 3.269)	11.625 (0.960; 3.704)
	δ_1	1.964 (0.990; 0.752)	1.766 (0.990; 1.145)	1.515 (0.990; 1.338)
	β_1	2.103 (1.00; 0.156)	2.147 (0.980; 0.219)	2.210 (0.950; 0.273)
	β_2	2.935 (0.998; 0.939)	2.885 (0.970; 0.891)	2.845 (0.925; 0.853)
	θ	(0.930; 3.4922)	(0.933; 3.536)	(0.932; 3.546)
M_2	δ_0	11.307 (0.960; 2.571)	14.460 (0.860; 5.786)	17.295 (0.800; 8.691)
	δ_1	2.228 (0.980; 0.867)	2.116 (0.990; 1.392)	1.877 (0.970; 1.870)
	β_1	2.053 (0.970; 0.146)	2.037 (0.910; 0.193)	1.985 (0.920; 0.189)
	β_2	2.954 (0.920; 0.958)	2.902 (0.924; 3.547)	2.869 (0926; 0.880)
	θ	(0.926; 3.495)	(0.931; 3.551)	(0.932; 3.566)
M_3	δ_0	10.208 (0.980; 2.011)	11.386 (0.980; 3.833)	11.425 (0.960; 3.671)
	δ_1	2.091 (0.990; 0.761)	1.851 (0.990; 1.249)	1.537 (0.960; 1.405)
	β_1	2.068 (0.960; 0.155)	2.107 (0.930; 0.955)	2.164 (0.900; 0.268)
	β_2	2.954 (0.930; 0.971)	2.920 (0.890; 0.930)	2.884 (0.880; 0.896)
	θ	(0.927; 3.492)	(0.930; 3.540)	(0.932; 3.553)

probability that a given score in the perturbed file corresponds to the same score in the original variable Z (see e.g. Shlomo and Skinner, 2010), namely

$$R(h) = Pr(Z = h|Z^* = h) = \frac{p_{hh}Pr(Z = h)}{\sum_l p_{lh}Pr(Z = l)}$$

and the posterior odds ratios (Gouweleeuw *et al.*, 1998)

$$PO(h) = \frac{Pr(Z = h|Z^* = h)}{Pr(Z \neq h|Z^* = h)} = \frac{p_{hh}Pr(Z = h)}{\sum_{l \neq h} p_{lh}Pr(Z = l)}$$

of the observed score h, $h = 1, \ldots, K$ under all scenarios. $R(h)$ has been proposed as a measure of risk, which is reasonable when the diagonal is the leading term in the misclassification matrix; we consider it here as an indicator of the proximity between the original and perturbed samples. $OR(h)$ is the ratio between the expected number of units whose score h is not changed and the expected number of units whose score is modified into h. For each given score h this quantity measures the degree of confusion (and protection) induced by PRAM on each category of the perturbed variable Z^*, with small values indicating higher protection. Both measures are reported in Table 3 for all the scenarios selected for our simulation experiment; we see in particular that the non invariant PRAM with matrix $\tilde{P}_3$ is more perturbative than the corresponding invariant matrix P_3.

Table 4 shows the ACP and the RMSE of the competing models for each PRAM scenario under the non invariant setup. For the estimation of model parameters, the proposed model outperforms the other models in all scenarios in terms of both ACP and RMSE. The higher the misclassification probability, the better the relative performance of model M_1. We note on passing the particularly poor performance of model M_2 under non invariant matrices $\tilde{P}_2$ and $\tilde{P}_3$. Overall, the advantage of introducing the measurement error mechanism in the proposed model is here more evident, compared to the invariant PRAM scenarios; as expected, the estimates obtained under the non invariant PRAM are, in general, more variable than those obtained under the invariant PRAM. With the exception of the non invariant PRAM with misclassification matrix $\tilde{P}_3$, under which the extent of perturbation is large, estimates obtained under model M_1 are close to those that would have been produced were the original data available, with a good coverage and generally larger RMSE than under the benchmark M_0, as expected.

Considering the prediction of small area means, all models perform similarly under the non invariant setup.

Considering inference on small area means, all models yield very accurate predictions: the ACPs are very similar to each other and the perturbation seems not to largely affect the estimation of the small area means. Compared to the alternative models, M_1 exhibits larger ACPs with smaller RMSEs, especially in the invariant setup.

The similar performances of all methods in predicting small area means may be due to the limited amount of perturbation implied by the scenarios considered (see Table 3). Also, this can be partially ascribed to the fact that the small area means do not depend on the units but they only depend on the perturbed covariates

TABLE 2

Invariant PRAM: posterior mean by misclassification probability P for different models and n_i ranging from 5 to 20. In parentheses ACP and RMSE, respectively. In the last line, we show the average ACP and RMSE of the small area means.

		P_1	P_2	P_3
M_0	δ_0	10.457 (1.00; 0.461)	10.457 (1.00; 0.462)	10.457 (1.00; 0.461)
	δ_1	2.188 (1.00; 0.189)	2.188 (1.00; 0.189)	2.188 (1.00; 0.189)
	β_1	2.020 (1.00; 0.021)	2.020 (1.00; 0.022)	2.020 (1.00; 0.022)
	β_2	2.921 (1.00; 0.921)	2.921 (1.00; 0.921)	2.921 (1.00; 0.921)
	θ	(0.922; 2.366)	(0.934; 2.551)	(0.939; 2.641)
M_1	δ_0	11.255 (0.98; 1.373)	11.766 (0.920; 1.893)	12.310 (0.800; 2.451)
	δ_1	2.224 (1.00; 0.360)	2.181 (1.00; 0.424)	2.122 (1.00; 0.462)
	β_1	2.053 (1.00; 0.141)	2.093 (0.977; 0.195)	2.084 (1.00; 0.172)
	β_2	2.773 (0.86; 0.777)	2.674 (0.679; 0.684)	2.589 (0.560; 0.600)
	θ	(0.946; 2.332)	(0.948; 2.500)	(0.951; 2.590)
M_2	δ_0	11.365 (0.760; 1.527)	12.599 (0.540; 2.768)	14.3711 (0.130; 4.540)
	δ_1	2.134 (0.970, 0.369)	2.023 (0.970; 0.522)	1.842 (0.960; 0.685)
	β_1	1.993 (0.920; 0.161)	1.943 (0.910; 0.226)	1.784 (0.880; 0.282)
	β_2	2.800 (0.640; 0.807)	2.664 (0.370; 0.682)	2.450 (0.140; 0.521)
	θ	(0.925; 2.407)	(0.940; 2.606)	(0.947; 2.701)
M_3	δ_0	10.845 (0.920; 1.049)	11.203 (0.890; 1.457)	11.686 (0.830; 1.942)
	δ_1	2.187 (0.970; 0.366)	2.168 (0.980; 0.466)	2.128 (0.990; 0.520)
	β_1	2.048 (0.930; 0.177)	2.108 (0.910; 0.267)	2.115 (0.950; 0.250)
	β_2	2.842 (0.780; 0.849)	2.765 (0.640; 0.781)	2.687 (0.580; 0.704)
	θ	(0.946; 2.532)	(0.935; 2.551)	(0.939; 2.641)

TABLE 3

Posterior probabilities $R(h) = Pr(Z = h|Z^* = h)$ and posterior odds ratios $OR(h)$ for each category of the perturbed variable Z^* under all scenarios.
P: invariant PRAM; $\tilde{P}$: non invariant PRAM.

	$R(1)$	$R(2)$	$R(3)$	$OR(1)$	$OR(2)$	$OR(3)$
P_1	0.90	0.94	0.93	9.27	15.36	13.13
$\tilde{P}_1$	0.91	0.91	0.88	10.47	10.29	7.24
P_2	0.80	0.88	0.86	4.12	7.18	6.07
$\tilde{P}_2$	0.84	0.82	0.76	5.13	4.41	3.16
P_3	0.71	0.82	0.79	2.40	4.45	3.73
$\tilde{P}_3$	0.65	0.71	0.73	1.89	2.43	2.65

TABLE 4

Non Invariant PRAM: posterior mean by misclassification probability P for different models. In parentheses ACP and RMSE, respectively. In the last line, we show the average ACP and RMSE of the small area means.

		P_1	P_2	P_3
M_0	δ_0	10.008 (1.00; 0.067)	10.008 (1.00; 0.067)	10.008 (1.00; 0.067)
	δ_1	2.322 (1.00; 0.322)	2.322 (1.00; 0.322)	2.322 (1.00; 0.322)
	β_1	1.943 (1.00; 0.057)	1.943 (1.00; 0.057)	1.943 (1.00; 0.057)
	β_2	3.044 (1.00; 1.044)	3.044 (1.00; 1.044)	3.044 (1.00; 1.044)
	θ	(0.927; 3.492)	(0.930; 3.540)	(0.932; 3.553)
M_1	δ_0	11.408 (1.00; 2.978)	12.272 (0.970; 4.129)	11.079 (0.980; 4.031)
	δ_1	1.641 (0.99; 1.043)	1.142 (0.980; 1.614)	0.742 (0.950; 2.064)
	β_1	2.170 (0.96; 0.230)	2.314 (0.930; 0.361)	2.348 (0.780; 0.407)
	β_2	2.924 (1.00; 0.929)	2.871 (0.990; 0.877)	2.678 (0.766; 0.710)
	θ	(0.927; 3.492)	(0.933; 3.536)	(0.932; 3.546)
M_2	δ_0	13.105 (0.940; 4.545)	18.700 (0.750; 10.041)	23.507 (0.590; 14.198)
	δ_1	1.929 (0.990; 1.138)	1.420 (0.960; 2.156)	0.745 (0.930; 3.185)
	β_1	2.170 (0.880; 0.253)	2.354 (0.750; 0.423)	2.562 (0.350; 0.617)
	β_2	2.847 (0.760; 0.855)	2.623 (0.440; 0.639)	2.396 (0.030; 0.430)
	θ	(0.926; 3.495)	(0.931; 3.551)	(0.932; 3.566)
M_3	δ_0	11.081 (0.980; 3.064)	12.148 (0.970; 4.285)	11.397 (0.980; 4.450)
	δ_1	1.767 (0.990; 1.008)	1.208 (0.940; 1.674)	0.754 (0.930; 2.121)
	β_1	2.101 (0.940; 0.209)	2.221 (0.40; 0.309)	2.290 (0.790; 0.386)
	β_2	2.978 (0.960; 0.985)	2.937 (0.990; 0.946)	2.718 (0.760; 0.741)
	θ	(0.927; 3.492)	(0.931; 3.540)	(0.932; 3.553)

averaged over all units. Moreover, the mechanism of "borrowing strength" from the related areas improves the estimate of the small area means, thus mitigating the effect of perturbation.

6. Discussion

In this paper, we have exploited the connections between measurement error and data perturbation for disclosure limitation in the context of small area estimation. We have primarily focused on the model proposed in Arima *et al.* (2012a), where some of the covariates (all continuous) are measured with error. We have extended the aforementioned model in order to account for categorical covariates perturbed for disclosure limitation. We have considered random data perturbation techniques, focusing in particular on PRAM perturbation mechanism. Under this method the data are perturbed according to a specific probabilistic mechanism and therefore are subjected to misclassification error; for this reason we can embed post randomized data in the measurement error model. In order to obtain area-level covariates from categorical record-level data, we propose to aggregate unit-level variables to the area level as the number of sampled records in all categories of each variable. These quantities can be considered as approximately normal variables. In this way, we rephrase the problem of perturbed categorical variables in terms of perturbed continuous variables for which the results in Arima *et al.* (2012a) can be easily adapted. In case of small area sizes and rare categories, the normal approximation may fail. Although the simulations indicate a robustness of the proposed procedure, the model can be easily modified to include the exact Multinomial distribution in Stage 2.2. In both formulations, we can model the misclassification process without having to rely on strong assumptions about the measurement error, whose validity is crucial when dealing with categorical data.

The proposed model has been compared to a model that ignores the presence of perturbation and a model in which PRAM estimates of the perturbed covariates are used. We have conducted two simulation studies investigating the effect on parameter estimates of invariant and non invariant PRAM perturbation mechanism. Under both the invariant and non invariant scenarios, simulations showed that using the PRAM estimates of the perturbed covariates is worse than ignoring the perturbation, in terms of accuracy and variability of parameters estimates. On the other hand, the proposed model provided more accurate and more stable estimates with respect to both the competing models, while accounting for uncertainty about the underlying covariates. With respect to non invariant PRAM, the variability of the estimates are larger than those obtained with the invariant PRAM.

In this paper, we have focused on area-level models to keep the parallel with the previous literature mentioned in this paper; for this reason we have aggregated unit-level covariates to the area level. However, unit-level models can also be fruitfully employed for prediction of small area quantities. In a disclosure limitation context, they arise more naturally since the perturbation is usually performed at the unit-level; they also allow for interesting investigations of disclosure risk assessment supported by the estimated regression model. Measurement error approach to unit-level small area estimation models has also been explored, for example, in Ghosh *et al.* (2006) and Arima *et al.* (2012b). In these models also, the measurement error affects only the continuous variables. An investigation of the effect of PRAM-ed covariates in logistic regression is presented in Woo and Slavković (2012). We are actually working on extending the unit-level model in order to consider perturbed categorical variables and to predict the true value of the perturbed variables for each unit.

References

S. Arima, G. Datta, B. Liseo (2012a). *Bayesian estimators for small area models when auxiliary information is measured with error.* Scandinavian Journal of Statistics, in press.

S. Arima, G. Datta, B. Liseo (2012b). *Objective Bayesian analysis of a measurement error small area model.* Bayesian Analysis, 7(2), pp. 363–384.

R. Brand (2002). *Microdata protection through noise addition.* In J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*, Springer, vol. 2316 of *Lecture Notes in Computer Science*, pp. 97–116.

R. J. Carroll, D. Ruppert, L. Stefanski, C. Crainiceanu (2006). *Measurement Error in Nonlinear Models: a Modern Perspective.* Chapman & Hall, CRC, 2nd ed.

T. Dalenius, S. P. Reiss (1982). *Data-swapping: A technique for disclosure control.* Journal of Statistical Planning and Inference, 6, no. 1, pp. 73 – 85.

G. Datta (2009). *Model-based approach to small area estimation.* Handbook of Statistics: Sample Surveys: Inference and Analysis, Volume 29B, Eds.: D. Pfeffermann and C.R. Rao. The Netherlands: North-Holland, pp. 251–288.

R. Fay, R. Herriot (1979). *Estimates of income for small places: an application of James-Stein procedures to census data.* Journal of the American Statistical Association, 74, pp. 269–277.

W. A. Fuller (1993). *Masking procedures for microdata disclosure limitation.* Journal of Official Statistics, 9, pp. 383–406.

M. Ghosh, K. Sinha, D. Kim (2006). *Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error model.* Scandinavian Journal of Statistics, 33, pp. 591–568.

J. Gouweleeuw, P. Kooiman, L. Willenborg, P.-P. De Wolf (1998). *Post randomisation for statistical disclosure control: Theory and implementation.* Journal of Official Statistics, 14, pp. 463–478.

J. Kim (1986). *A method for limiting disclosure of microdata based on random noise and transformation.* In *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 370–374.

R. J. A. Little (1993). *Statistical analysis of masked data.* Journal of Official Statistics, 9, pp. 407–426.

D. Pfefferman (2013). *New important developments in small area estimation.* Statistical Science, 28, pp. 40–68.

J. N. K. Rao (2003). *Small Area Estimation.* Wiley series in survey methodology. John Wiley and Sons, New York.

N. Shlomo, C. Skinner (2010). *Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata.* Ann. Appl. Stat., 4, no. 3, pp. 1291–1310.

L. Willenborg, T. de Waal (2001). *Elements of Statistical disclosure control.* Springer, New York.

Y. Woo, A. Slavković (2012). *Logistic regression with variables subject to post randomization method.* In J. Domingo-Ferrer, I. Tinnirello (eds.), *Privacy in Statistical Databases*, Springer Berlin Heidelberg, vol. 7556 of *Lecture Notes in Computer Science*, pp. 116–130.

L. Ybarra, S. Lohr (2008). *Small area estimation when auxiliary information is measured with error.* Biometrika, 95, pp. 919–931.

Summary

We exploit the connections between measurement error and data perturbation for disclosure limitation in the context of small area estimation. Our starting point is the model in Ybarra and Lohr (2008), where some of the covariates (all continuous) are measured with error. Using a fully Bayesian approach, we extend the aforementioned model including continuous and categorical auxiliary variables, both possibily perturbed by disclosure limitation methods, with masking distributions fixed according to the assumed protection mechanism. In order to investigate the feasibility of the proposed method, we conduct a simulation study exploring the effect of different post-randomization scenarios on the small area model.

Keywords: Disclosure limitation; Hierarchical Bayesian models; measurement error; PRAM; small area.