

Small Area Estimation with Covariates Perturbed for Disclosure Limitation

Stima per piccole aree con covariate perturbate per tutelare la riservatezza dei rispondenti

Silvia Poletti and Serena Arima

Abstract We exploit the connections between measurement error and data perturbation for disclosure limitations in the context of small area estimation. Our starting point is an area level model in which some of the covariates (all continuous) are measured with error. Using a fully Bayesian approach, we extend such model including continuous and categorical auxiliary variables, both 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 an extensive simulation study exploring the effect of different protection scenarios on the small area mean predictions. We also perform a comparative analysis of the proposed estimator.

Abstract *In questo lavoro esploriamo le connessioni tra la stima in presenza di errore di misurazione e l'analisi di dati perturbati per motivi di confidenzialità per i modelli di stima per piccole aree. Il punto di partenza è un modello di area in cui alcune covariate, tutte continue, sono misurate con errore. Mediante un approccio bayesiano, includiamo in tale modello variabili sia continue che categoriche, entrambe perturbate per tutelare la riservatezza dei rispondenti secondo uno schema probabilistico noto. Al fine di valutare l'efficacia del metodo proposto, viene effettuato uno studio di simulazione che esplora gli effetti di diversi scenari di protezione sulle previsioni di area. Lo stimatore proposto viene confrontato con i metodi attualmente noti in letteratura.*

Key words: MCMC, measurement error models, small area estimation, statistical disclosure limitation.

Silvia Poletti

Department of Methods and Models for Economics, Territory and Finance, Sapienza Università di Roma, e-mail: silvia.poletti@uniroma1.it

Serena Arima

Department of Methods and Models for Economics, Territory and Finance, Sapienza Università di Roma e-mail: serena.arima@uniroma1.it

1 Introduction

In recent years, small area estimation has emerged as an important area of statistics as private users 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, and exhibit 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 [10], [3] and [9] for a review).

There is a strong and increasing pressure from the research community to extract from survey data the largest amount of information, small area estimation being one instance of the possible analyses of legitimate users. The mission of statistical agencies is supplying society with useful statistical information, under the commitment of protecting confidentiality of the respondents. Statistical Disclosure Control (SDC) aims at protecting statistical data in such a way that these can be released without disclosing confidential information that can be linked to specific individuals or entities (see [11]). SDC can be defined as the set of techniques designed to measure and reduce the risk of disclosure to an acceptable level while releasing as much information as possible. Reduction of the risk of disclosure of individual information can be achieved by aggregating or suppressing data, referred to as non perturbative methods, or by applying perturbative methods, that mask the data purposely. Quoting [8], “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.

The model proposed by [4] 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. 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. [12] considered the situation in which the auxiliary information is measured with error. Suppose there are m small areas of interest and suppose θ_i is the population characteristic of interest in area i . The quantity of interest may be a total, a mean or a proportion. Let y_i be a direct estimator of θ_i for area i . [12] considered the situation in which the p -vector X_i of auxiliary data is measured with error as X_i^* , with known $MSE(X_i^*) = C_i$. [12] proposed to use the auxiliary information X_i^* to improve prediction of θ_i and to incorporate the mean squared error of X_i^* into the error bounds of the small area estimator. [1] rewrite the measurement error model as a hierarchical Bayesian model; they use improper noninformative priors

on the model parameters and show, under a mild condition, that the joint posterior distribution is proper and the marginal posterior distributions of the model parameters have finite variances. The Bayesian predictors show smaller empirical mean squared errors than the frequentist predictors of [12] and are more stable in terms of variability and bias.

As discussed, it is often difficult to obtain unperturbed microdata that allow one to estimate the small area aggregates of interest: indeed, perturbative disclosure limitation techniques are often applied to survey data to prevent reidentification of respondents; examples of such techniques include noise addition ([7, 5, 2]) and PRAM ([6]). The focus of this paper will be the analysis of data treated by such techniques for prediction of small area means.

As far as data analysis is concerned, there is a strong analogy between random data perturbation schemes such as PRAM and noise addition on one hand, and measurement/misclassification errors, on the other. The measurement error models proposed in [12] and [1] only considered the particular and somehow unrealistic situation in which the auxiliary covariates measured with error are all continuous variables. However it may happen, as in the case under analysis, that auxiliary categorical variables are of interest and they are measured with error, that is are subject to misclassification. We exploit the analogy between data perturbation and measurement errors in variables to propose a Bayesian area-level model allowing for measurement errors in both continuous and categorical auxiliary variables and explicitly introducing the error distributions of both categorical and continuous variables induced by the above mentioned SDC techniques. As in [12], we assume that the observed continuous variable X_i^* may be obtained as a perturbation of the original variable X_i : the perturbation mechanism is specified as $X_i^*|X_i \sim N(X_i, C_i)$, with C_i known positive definite, $i = 1, \dots, m$, independently. With respect to the categorical variables, we consider the distribution induced by the Post Randomization Method (PRAM); see Section 2 for the definition of the above mentioned protection techniques. The proposed model is presented in Section 3.

2 Masking techniques: Post Randomization and Noise Addition

The Post Randomization Method was proposed in [6] 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 with which 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 $\{1, \dots, K\}$. 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 $p_{l1}, \dots, p_{lK}$ at scores $1, \dots, K$. 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 $P = \{p_{lh}\}_{l,h=1,\dots,K}$ is referred to as the PRAM matrix. PRAM can also be applied to $p > 1$ categorical variables $Z_1, \dots, Z_p$, 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 $K_1, K_2, \dots, K_p$ scores for all variables considered; in the former case, the PRAM transition matrix P is the Kroneker product of the transition matrices $P_1, \dots, P_p$ 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. On the other hand, 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, a compounded variable may include those variables that must remain unchanged. The characteristics and the specific values of the elements of the transition matrix P depend 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 so as to preserve approximately the distribution of the protected variables (see [6]). Moreover, including variables that must be kept unchanged and specifying block-diagonal PRAM matrices makes it possible to preserve the distribution in pre-specified subpopulations.

Let us denote by T_Z (T_{Z^*} , respectively) the vector of frequencies of the scores on the original (perturbed) variable Z (Z^*), and 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 $\text{Cov}(I(Z^* = h), I(Z^* = j)|Z = l) = -p_{lh}p_{lj} := \{V_l\}_{h,j}$, it is easy to verify that

$$P'T_Z = E(T_{Z^*}|z).$$

Therefore, if the transition matrix P is invertible, we have that $\hat{T}_Z = (P^{-1})'T_{Z^*}$ is an unbiased estimator for T_Z .

The conditional variance of the estimator is $V(\hat{T}_Z|z) = (P^{-1})'V(T_{Z^*}|z)P^{-1}$, where $V(T_{Z^*}|z) = \sum_l T_{Z^*l}V_l$; here T_{Z^*l} denotes the frequency associated with the l -th score on the perturbed variable and V_l is the covariance matrix of the transition process from original score l , whose elements $\{V_l\}_{h,j}$, $h, j = 1, \dots, K$, are reported above. [6] propose the following plug-in variance estimator for $\hat{T}_Z$:

$$\hat{V}(\hat{T}_Z) = \sum_l \hat{T}_{Zl}V_l \quad (1)$$

An interesting application of the post randomization method is the so-called invariant PRAM, that amounts to choosing the transition matrix P such that $P'T_Z = T_Z$. Such a choice for P 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 tran-

sition matrix is not needed. Full knowledge of the PRAM matrix is needed in (1) for variance estimation even under invariant post randomization.

The simplest form of noise addition ([7, 5, 2]) 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 .

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 [7] shows that the covariance matrix of the original data can be consistently estimated from the masked data as long as α is known.

Work by [5] and [8] shows how to account for some SDC methods using standard statistical methods for the analysis of incomplete data when the values of the masking parameters are disclosed.

3 The proposed model

Suppose there are m areas labelled $1, \dots, m$, each with $n_1, \dots, n_m$ individual observations. We denote by y_i the response of the i -th area ($i = 1, \dots, m$). For ease of notation, we consider a single continuous auxiliary variable, but generalization to a vector of covariates, possibly observed exactly, is straightforward and proceeds as in [1]. Let X_i^* be the observed continuous area-specific covariate measured with error or resulting from perturbation of the original variable X_i . Moreover, let Z_{ij}^* , $j = 1, \dots, n_i$ 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.

The proposed measurement error model can be written as the following multi-stage model:

Stage.1 $y_i | \theta_i \sim N(\theta_i, \psi_i)$, ψ_i known, $i = 1, \dots, m$, independently;

Stage.2 $\theta_i | \beta_0, \beta_1, \beta_2, \sigma_v^2 \sim N(\beta_0 + \beta_1 X_i + \beta_2 Z_i, \sigma_v^2)$

Stage.3 $X_i^* | X_i \sim N(X_i, c_i)$, c_i known variance and $X_i \sim N(\mu, \sigma_x^2)$

$$Z_i = (Z_i^{(1)}, \dots, Z_i^{(k)}, \dots, Z_i^{(K-1)}), \quad Z_i^{(k)} = \sum_{j=1}^{n_i} I(Z_{ij} = k), \quad k = 1, \dots, K-1,$$

$$Pr(Z_{ij}^* = k | Z_{ij} = k') = p_{k'k} \text{ known transition probabilities } (k = 1, \dots, K-1)$$

$$Pr(Z_{ij} = k) \sim Unif\left(\frac{1}{K}\right)$$

Stage.4 $\beta_0, \beta_1, \beta_2, \mu, \sigma_x^2, \sigma_v^2$ are mutually independent with $\beta_0, \beta_1, \beta_2, \mu \sim N(0, \tau^2)$, $\sigma_x^2 \sim IG(a_x, b_x)$ and $\sigma_v^2 \sim IG(a_v, b_v)$.

Stage 3 specifies the perturbation mechanism for continuous and discrete auxiliary variables. For ease of notation, we have considered a single continuous auxiliary variable, but generalization to a vector of covariates, possibly observed exactly, is straightforward and proceeds as in [1]. In Stage 4 we define the prior distributions for the unknown parameters: we use flat normal prior ($\tau^2 = 10^5$) for the regression coefficients and for the μ and flat inverse gamma distribution for the variance parameters ($a_x = b_x = a_v = b_v = 0.001$). The posterior distributions cannot be obtained analytically; however, full conditional distributions can be easily derived and samples from the joint posterior distribution can be obtained using standard Markov Chain Monte Carlo algorithm.

The properties of the obtained small area estimator $\hat{\theta}_i$ are investigated using an extensive simulation study: the bias and mean squared error of the estimator are studied under different scenarios, also investigating the tradeoff between disclosure risk and data utility in this specific setting. We evaluate the effect on the estimator of ignoring or misspecifying the perturbation mechanism in both continuous and categorical auxiliary variables for comparison. Moreover we examine the impact of the choice of the prior distributions through a sensitivity analysis.

References

1. Arima, S., Datta, G.S., Liseo, B.: Bayesian estimators for small area models when auxiliary information is measured with error, submitted (2012)
2. Brand, R.: Microdata protection through noise addition. In: Domingo-Ferrer, J. (ed.) *Inference Control in Statistical Databases*, Lecture Notes in Computer Science, vol. 2316, pp. 97–116. Springer (2002)
3. Datta, G.: Model-based approach to small area estimation. In: Pfeiffermann, D. and Rao, C.R. (eds.) *Handbook of Statistics: Sample Surveys: Inference and Analysis*, Volume 29B, The Netherlands: North-Holland pp. 251–288 (2009)
4. Fay, R., Herriot, R.: Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277 (1979)
5. Fuller, W.A.: Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383–406 (1993)
6. Gouweleeuw, J., Kooiman, P., Willenborg, L., De Wolf, P.P.: Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics* **14**, 463–478 (1998)
7. Kim, J.: A method for limiting disclosure of microdata based on random noise and transformation. In: *Proceedings of the Survey Research Methods Section*, pp. 370–374. American Statistical Association (1986)
8. Little, R.J.A.: Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426 (1993)
9. Pfefferman, D.: New important developments in small area estimation. *Statistical Science* **28**, 40–68 (2013)
10. Rao, J.N.K.: *Small Area Estimation*. Wiley series in survey methodology. John Wiley and Sons, New York (2003)
11. Willenborg, L., de Waal, T.: *Elements of Statistical Disclosure Control*. Springer, New York (2001)
12. Ybarra, L., Lohr, S.: Small area estimation when auxiliary information is measured with error. *Biometrika* **95**, 919–931 (2008)