SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Peer Review Process
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board
A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

SOCIETÀ ITALIANA DI STATISTICA
Sede: Salita de' Crescenzi 26 - 00186 Roma
Tel +39-06-6869845 - Fax +39-06-68806742
email: sis@caspur.it web:http://www.sis-statistica.it

La Società Italiana di Statistica (SIS), fondata nel 1939, è una società scientifica eretta ad Ente morale ed inclusa tra gli Enti di particolare rilevanza scientifica. La SIS promuove lo sviluppo delle scienze statistiche e la loro applicazione in campo economico, sociale, sanitario, demografico, produttivo ed in molti altri settori di ricerca.

Organi della società:

Presidente:
- Prof.ssa Monica Pratesi, Università di Pisa

Segretario Generale:
- Prof.ssa Filomena Racioppi, Sapienza Università di Roma

Tesoriere:
- Prof.ssa Maria Felice Arezzo, Sapienza Università di Roma

Consiglieri:
- Prof. Giuseppe Arbia, Università Cattolica del Sacro Cuore
- Prof.ssa Maria Maddalena Barbieri, Università Roma Tre
- Prof.ssa Francesca Bassi, Università di Padova
- Prof. Eugenio Brentari, Università di Brescia
- Dott. Stefano Falorsi, ISTAT
- Prof. Alessio Pollice, Università di Bari
- Prof.ssa Rosanna Verde, Seconda Università di Napoli
- Prof. Daniele Vignoli, Università di Firenze

Collegio dei Revisori dei Conti:
- Prof. Francesco Campobasso, Prof. Michele Gallo, Prof. Francesco Sanna, Prof. Umberto Salinas (supplente)

SIS2017 Committees

Scientific Program Committee:
Rosanna Verde (chair), Università della Campania "Luigi Vanvitelli"
Maria Felice Arezzo, Sapienza Università di Roma
Antonino Mazzeo, Università di Napoli Federico II
Emanuele Baldacci, Eurostat
Pierpaolo Brutti, Sapienza Università di Roma
Marcello Chiodi, Università di Palermo
Corrado Crocetta, Università di Foggia
Giovanni De Luca, Università di Napoli Parthenope
Viviana Egidi, Sapienza Università di Roma
Giulio Ghellini, Università degli Studi di Siena
Ippoliti Luigi, Università di Chieti-Pescara "G. D'Annunzio"
Matteo Mazziotta, ISTAT
Lucia Paci, Università Cattolica del Sacro Cuore
Alessandra Petrucci, Università degli Studi di Firenze
Filomena Racioppi, Sapienza Università di Roma
Laura M. Sangalli, Politecnico di Milano
Bruno Scarpa, Università degli Studi di Padova
Cinzia Viroli, Università di Bologna

Local Organizing Committee:
Alessandra Petrucci (chair), Università degli Studi di Firenze
Gianni Betti, Università degli Studi di Siena
Fabrizio Cipollini, Università degli Studi di Firenze
Emanuela Dreassi, Università degli Studi di Firenze
Caterina Giusti, Università di Pisa
Leonardo Grilli, Università degli Studi di Firenze
Alessandra Mattei, Università degli Studi di Firenze
Elena Pirani, Università degli Studi di Firenze
Emilia Rocco, Università degli Studi di Firenze
Maria Cecilia Verri, Università degli Studi di Firenze

Supported by:
Università degli Studi di Firenze
Università di Pisa
Università degli Studi di Siena
ISTAT
Regione Toscana
Comune di Firenze
BITBANG srl

Index

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Preface

The 2017 SIS Conference aims to highlight the crucial role of the Statistics in Data Science. In this new domain of "meaning" extracted from the data, the increasing amount of produced and available data in databases, nowadays, has brought new challenges. That involves different fields of statistics, machine learning, information and computer science, optimization, pattern recognition. These afford together a considerable contribute in the analysis of "Big data", open data, relational and complex data, structured and no-structured. The interest is to collect the contributes which provide from the different domains of Statistics, in the high dimensional data quality validation, sampling extraction, dimensional reduction, pattern selection, data modelling, testing hypotheses and confirming conclusions drawn from the data. In the mention that statistics is the "grammar of data science", statistics has become a basic skill in data science: it gives right meaning to the data. Still, it isn't replaced by newer techniques from machine learning and other disciplines but it complements them. The Conference is also addressed to the new challenges of the new generations: the native digital generations, who are called to develop professional skills as "data analyst", one of the more request professionality of the 21st Century, crossing the rigid disciplinary domains of competence. In this perspective, all the traditional statistical topics are admitted with an extension to the related machine learning and computer science ones. The present volume includes the short papers of the contributions that will be presented in the 4 invited speaker sessions; in the 19 specialized sessions; in the 11 solicited sessions; in the 6 foreign societies sessions and in the 17 contributed sessions as well as, in the panel session.

Rosanna Verde
President of the Scientific Programme Committee

Alessandra Petrucci
President of the Local Organizing Committee

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Accounting for measurement error in small area models: a study on generosity.

Modelli per piccola area con errore di misurazione: uno studio sulla generosità

Silvia Polettini and Serena Arima

Abstract In this paper we focus on a recently documented effect of economic inequality, namely that higher income individuals tend to be less generous than poorer individuals, *but only in contexts where macro-level economic inequality is high*, or is perceived as high. We consider data from the Measuring Morality study, a nationally representative survey of United States residents, that contains a validated behavioural measure of generosity (the dictator game) along with the household income of respondents. We fit a small area model to this data with the aim of investigating the role of economic inequality on generosity in the US. We observe that model covariates (reported income and Gini index) are subject to measurement error and investigate the effect of introducing the measurement error in this model.

Abstract *Il lavoro considera il ruolo della disuguaglianza economica sulla generosità, a partire da uno studio recente secondo cui gli individui con redditi più elevati tendono ad essere meno generosi degli individui meno abbienti, ma solo in contesti di grande disuguaglianza economica. I dati analizzati provengono dal Measuring Morality study, un'indagine effettuata negli USA in cui viene rilevato il reddito e una misura validata di generosità (dictator game). Per ogni area di residenza è stato anche ricavato l'indice di Gini, come misura di disuguaglianza economica. In questo lavoro si stima la generosità mediante un modello per piccole aree con reddito e disuguaglianza come variabili ausiliarie. Il modello viene esteso al fine di considerare l'errore di misurazione nelle variabili ausiliarie, sia continue che discrete.*

Key words: small area estimation, measurement error, misclassification, Bayesian inference.

Silvia Polettini
Dip. di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, via del Castro Laurenziano, 9, e-mail: silvia.polettini@uniroma1.it

Serena Arima
Dip. di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, via del Castro Laurenziano, 9, e-mail: serena.arima@uniroma1.it

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

1 Introduction

There is an increasing interest in understanding the implications of income for behaviour, in particular generosity toward others. Well grounded literature on this topic has portrayed a picture of higher-income individuals as consistently more selfish than poorer individuals [13]. A different perspective is reported in a recent paper [6], where the relationship between economic inequality, income, and generosity is tested. Analysing data from the Measuring Morality study (a nationally representative survey of United States residents), as well as a follow-up experiment, the authors identify a previously undocumented effect of economic inequality, namely that higher income individuals in the US tend to be less generous than poorer individuals, *but only in contexts where macro-level economic inequality is high*, or is perceived as high. The Authors comment that the results obtained challenge the prevailing view in the literature that higher income individuals are necessarily less generous and conclude that "inequitable resource distributions undermine collective welfare" and that redistributive policies may "attenuate, or even reverse, the negative relationship between income and generosity, in turn increasing the generosity of those individuals who have the most to give".

The Measuring Morality study data contain a validated behavioural measure of generosity (the dictator game) along with the household income of respondents; moreover, Gini indices were available from the American Community Survey. The authors fit a mixed effects model to these data, where significant, negative, interaction between income and inequality is found. Using a Bayesian approach, we consider the same model, in a small area context and speculate on the fact that both income and the Gini index are subject to measurement error for different reasons: indeed income is self reported and the Gini index is estimated from another survey. As stressed in the literature, ignoring the measurement error in the covariates may lead to inconsistent estimates and can severely invalidate inferences.

The paper is organized as follows: in Section 2 we introduce the problem of measurement error in small area estimation and propose a small area model accounting for measurement error in covariates and present. In Section 3 we present and discuss the results obtained when the model is applied to the generosity data.

2 A measurement error small area model for generosity data

In this paper, we focus on unit level small area models, whithin a Bayesian framework. Unit level small area models relate the unit values of the study variable to unit-specific auxiliary variables with known area means. See [11] for an up-to-date review.

Suppose there are m areas and let N_i be the known population size of area i. We denote by Y_{ij} the response of the $j-$th unit in the $i-$th area ($i = 1,...,m$; $j = 1,...,N_i$). A random sample of size n_i is drawn from the $i-$th area. The goal is to predict the small area means $\bar{Y}_i = \frac{1}{N_i}\sum_{i=1}^{N_i} y_{ij}$, $i = 1,\ldots,m$, based on the available

data. To develop reliable estimates, auxiliary information is introduced as covariates and usually a mixed effects model is specified as

$$Y_{ij} = \alpha + \beta w_{ij} + u_i + \varepsilon_{ij} \qquad i = 1,...,m; \quad j = 1,...,N_i \tag{1}$$

with ε_{ij} and u_i independent, $\varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma_e^2)$ and $u_i \overset{iid}{\sim} N(0,\sigma_u^2)$. [8] and [9] were the first to consider the problem of measurement error in small area models for unit-level data. They assume that the true, area-level, covariate, w_i, is measured with error as

$$S_{ij} = w_i + \eta_{ij}, \qquad \eta_{ij} \overset{iid}{\sim} N(0,\sigma_\eta^2) \qquad i = 1,...,m; \quad j = 1,...,n_i \tag{2}$$

where ε_{ij}, u_i and η_{ij} are taken mutually independent. [8] also assumed that $w_i \overset{iid}{\sim} N(\mu_w,\sigma_w^2)$, defining the structural measurement error model. They considered both an empirical Bayes and a hierarchical Bayes approach to derive predictors of small area means θ_i. [12] extended the approach in [8] including sample information on the covariate values. [8] also proposed a fully Bayesian approach, by specifying a hierarchical model, with vague prior distributions for all the model parameters, whose posterior distributions are estimated via Gibbs sampling. [1, 3] extended the above approach, proposing to use the Jeffreys' prior on the model parameters. The aforementioned literature considers the case in which the measurement error only affects continuous variables, according to the measurement error model of equation (1). For discrete covariates, measurement error means misclassification. To allow for auxiliary discrete covariates measured with error, [4] propose to model the misclassification mechanism through an unknown transition matrix P and estimate all the unknown parameters in a fully Bayesian framework. Following [4], for each unit in each area, we consider the following covariates: t_{ij} – the vector of p continuous or discrete covariates measured without error, w_i and x_{ij} – respectively, a vector of q continuous covariates and h discrete variables (with a total of K categories), both measured with error. Denote by s_{ij} and z_{ij} the observed values of the latent w_i and x_{ij}, respectively. Without loss of generality, in what follows we assume $h = 1$.

Following the notation in [8], the proposed measurement error model can be written in the usual multi-stage way: for $j = 1,\ldots,n_i$, $i = 1,\ldots,m$ and for $k,k' = 1,\ldots,K$

- Stage 1. $y_{ij} = \theta_{ij} + e_{ij} \quad e_{ij} \overset{iid}{\sim} N(0,\sigma_e^2)$
- Stage 2. $\theta_{ij} = t_{ij}^{'}\delta + w_i^{'}\gamma + \sum_{k=1}^{K} I(x_{ij}=k)\beta_k + u_i \quad u_i \overset{iid}{\sim} N(0,\sigma_u^2)$
- Stage 3. $S_{ij}|w_i \overset{iid}{\sim} N(w_i, \Sigma_s = diag(\sigma_{s_1}^2,...,\sigma_{s_q}^2)) \quad W_i \overset{iid}{\sim} N(0, \Sigma_w = diag(\sigma_{w_1}^2,...,\sigma_{w_q}^2))$
  $Pr(Z_{ij} = k|X_{ij} = k') = p_{k'k}, \quad p_{k'.} \sim Dir(\alpha_{k',1},\ldots,\alpha_{k',K})\ Pr(X_{ij} = k') = \frac{1}{K}$
- Stage 4. $\beta,\delta,\gamma,\sigma_e^2,\sigma_u^2,\sigma_{s_1}^2,...,\sigma_{s_p}^2$ are, loosely speaking, a-priori mutually independent.

Stage 3 defines the measurement error model for both continuous and discrete covariates. For the discrete covariates, the misclassification mechanism is specified according to the $K \times K$ matrix P, whose (k',k) element, $p_{k'k}$, denotes the probability

that the observable variable Z_{ij} takes the $k-$th category when the true unobservable variable X_{ij} takes the $k'-$th category. We also assume that the misclassification probabilities are the same across subjects and that all the categories have the same prior probability $\frac{1}{K}$ to occur. Over each row of P, we place a Dirichlet $Dir(\alpha_{k',1},\ldots,\alpha_{k',K})$ prior distribution, with known $\alpha_{k',1},\ldots,\alpha_{k',K}$. In Stage 4 we assume Normal priors for β,δ, and γ and inverse gamma distributions for σ_e^2 and σ_u^2 and σ_s^2. Hyperparameters have been chosen to have flat priors. Finally, we fix Σ_w and $(\alpha_{k',1},\ldots,\alpha_{k',K})$. According to the above assumptions, we can estimate the transition matrix P and the measurement error variance σ_s^2 jointly with all the other model parameters. As the posterior distribution cannot be derived analytically in closed form, we obtain samples from the posterior distribution using Gibbs sampling.

3 Results and conclusions

We fit a unit level small area model with measurement error in covariates, which also allows us to evaluate the relationship between economic inequality, income and generosity. We use data from the Measuring Morality study, a nationally representative survey of United States residents consisting of a sample of 1498 respondents in the US. For each respondent, income and some personal and demographic variables (such as age, gender, education, ...) have been collected. Respondents completed a validated behavioural measure of generosity: the dictator game. Respondents learned that they had been randomly assigned the role of *decider* and had received 10 tickets, each worth one entry in a raffle to win a monetary prize of either 10 or 500. They could transfer any number of tickets to the next participant, a *receiver* who did not have any tickets. By giving tickets, respondents could benefit another person at a cost to themselves in a zero-sum opportunity to win money. This measure of generosity was administered to individuals with different incomes residing in areas (US states plus the District of Columbia) that vary in levels of inequality, measured according to the Gini's coefficient. The number of respondents in each area ($m = 9$ divisions) ranges from 72 to 286. In the proposed model we take generosity as the response variable and income, standardized Gini coefficients and their interaction as auxiliary variables. According to the survey design, household income was collected as a 19-classes variable; for ease of interpretation in the application we recoded it into five classes (C_1 : less than 12500; $C_2 : [12500, 30000)$, $C_3 : (30000, 60000]$, $C_4 : (60000, 125000]$, C_5 : over 125000). Since income is self reported and the Gini index is estimated using data from the 2012 American Community survey, we can suspect that both auxiliary variables are subject to measurement error. In order to evaluate the impact of accounting for this source of error, we fit both the standard model that ignores the measurement error and the model proposed in Section 2. Figure 1 shows the posterior distribution of the model parameters. The left panel reports the posterior distribution of the regression parameters under the proposed measurement error model: income is the only factor that significantly impacts on the response variable, since for all the other pa-

rameters the 95% credible intervals contain the zero value ($CI_{Gini} : [-0.207, 0.349]$, $CI_{C1*Gini} : [-0.632, 0.241]$, $CI_{C2*Gini} : [-0.542, 0.217]$, $CI_{C3*Gini} : [-0.533, 0.189]$, $CI_{C4*Gini} : [-0.827, -0.028]$). With respect to the income, it is apparent that generosity increases with income, with the exception of the last class, in which the effect on generosity is comparable to that of the second one. This actually means that the richest are less generous with respect to the others, which is line with findings in the mainstream literature on the subject. On the other hand, when one ignores the measurement error, all the covariates and their interactions seem to be significant (Figure 1, right panel). In particular, income exhibits a positive effect on generosity, with no distinctions between income classes, which contradicts the economic theories; moreover, an unexpectedly positive effect of inequality is found. With respect to the measurement error for income, the posterior distribution of $P_{1,1}$ is concentrated around 0.5 and almost uniformly distributed over the other categories. This is an empirical evidence that income is often underreported by the respondents. The distributions of the other diagonal elements of P are concentrated around 0.9 and credibile intervals do not contain 1. We conclude that measurement error has a significant impact on income. The small area estimates produced under the model with and without measurement error are reported in Table 1. As can be seen, allowing for measurement error in both continuous and categorical covariates also impacts on estimation of the small area means in both point estimates (in particular for the first division, which is one of the smallest ones) and measures of uncertainty. Also, although the posterior means are not very different for the large areas, the ranking of the divisions varies. As can be seen, allowing for measurement error in both continuous and categorical covariates also impacts on estimation of the small area means. Although the posterior means are not very different, the ranking of the divisions varies. In conclusion, our application reveals that ignoring the measurement error in covariates may drive inferences and yeld misleading conclusions.

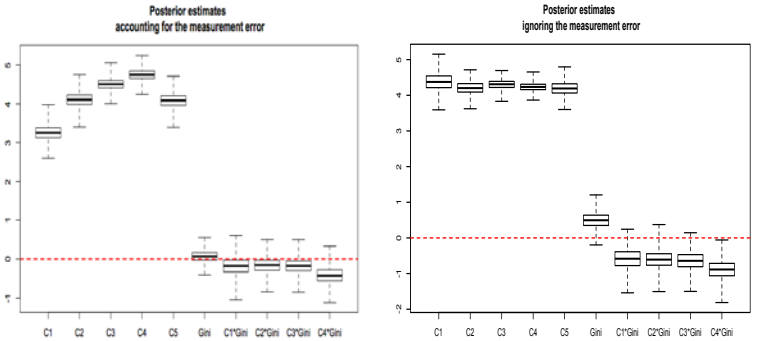
Table 1 Small area estimates: posterior means of the small area means obtained with the model that does not account for the measurement error (first row) and the model that accounts for it (second row). Standard deviations in brackets.

Division	1	2	3	4	5	6	7	8	9
θ_{NoErr}	4.17	4.11	4.25	4.44	4.19	4.28	4.25	4.37	4.22
	(0.27)	(0.33)	(0.18)	(0.20)	(0.24)	(0.10)	(0.14)	(0.16)	(0.23)
θ_{Err}	4.27	4.09	4.26	4.43	4.17	4.30	4.25	4.38	4.23
	(0.36)	(0.41)	(0.38)	(0.37)	(0.40)	(0.33)	(0.34)	(0.32)	(0.40)

References

1. Arima, S., Datta, G.S., Liseo, B.: Objective Bayesian analysis of a measurement error small area model. *Bayesian Analysis*, **72 (2)**,363–384 , (2012)

Fig. 1 Posterior distribution of the model parameters. Left panel: posterior distributions obtained from the proposed model. Right panel: posterior distributions from the model that ignores the measurement error.



2. Arima, S., Datta, G.S., Liseo, B.: Bayesian Estimators for Small Area Models when Auxiliary Information is Measured with Error. Scandinavian Journal of Statistics, **42 (2)**,518–529, (2014)
3. Arima, S., Datta, G.S., Liseo, B.: Models in Small Area Estimation when Covariates are Measured with Error, in Analysis of Poverty Data by Small Area Estimation, 151–170, (2015)
4. Arima, S., Polettini, S.: A unit-level small area model with misclassified covariates, arXiv:1611.02845 [stat.ME],(2016)
5. Carroll, R.J., Ruppert, D., Stefanski, L., Crainiceanu, C.: Measurement error in nonlinear models: a modern perspective. 2nd edn. Chapman & Hall, CRC, (2006)
6. Côtè, S., House, J., Willer, R.: High economic inequality leads higher-income individuals to be less generous . Ann. PNAS, **112**, 52, 15838–15843 (2015)
7. Engel, C.: Dictator Games: A Meta Study. Experimental Economics **14(4)**, 583?-610, (2011)
8. Ghosh, M., Sinha, K. and Kim, D.: Empirical and Hierarchical Bayesian estimation in finite population sampling under structural measurement error model. Scandinavian Journal of Statistics, **33(3)**, (2006)
9. Ghosh, M., Sinha, K.: Empirical Bayes estimation in finite population sampling under functional measurement error models. Journal of Statistical Planning Inference,**137**, 2759–2773,(2007)
10. Polettini, S., Arima, S.: Small area estimation with covariates perturbed for disclosure limitation. Statistica, **25 (1)**, 57–72, (2015)
11. Rao, J.N.K. and Molina, I.: Small Area Estimation, 2nd Edition, Wiley, Hoboken, New Jersey, (2015).
12. Torabi, M., Datta, G.S. and Rao, J.N.K. Empirical Bayes estimation of small area means under nested error linear regression model with measurement error in the covariates, Scandinavian Journal of Statistics, **36**, 355–368, (2009).
13. Trautmann, S.T., van de Kuilen, G. and Zeckhauser, R.J.: Social class and (un)ethical behavior: A framework, with evidence from a large population sample Perspectives on Psychological Science **8(5)**:487–497, (2013).
14. Ybarra, L.M.R., Lohr, S.L.: Small area estimation when auxiliary information is measured with error. Biometrika, **95(4)**, 919–931, (2008).