Area level SAE models with measurement errors in covariates: an application to sample surveys and Big Data sources

Monica Pratesi, Caterina Giusti, Stefano Marchetti, Nicola Salvati, Fosca Giannotti, Dino Pedreschi

Dept of Economics and Management, Dept of Informatics, University of Pisa

SAE 2014
Small Area Estimation 2014 Poznan, 3-5 September 2014
Part I

Using Big Data in Small Area Estimation
Our aim is to use the huge source of data coming from human activities - the big data - to make accurate inference at a small area level.

We identified three possible approaches:

1. Use big data as covariate in small area models
2. Use survey data to remove self-selection bias from estimates obtained using big data
3. Use big data to validate small area estimates
Use Big Data as Covariate in Small Area Models

- Big data often provide unit level data
- Outcome variable have to be linked to auxiliary variable in order to use unit level data in a small area model
- Due to technical challenges and law restriction it is unfeasible at this stage to have unit level big data that can be linked with administrative archive or census or survey data
- Big data can be aggregate at area level and then used in an area level model

\[ \hat{\theta}_i = d_i^T \beta + u_i + \varepsilon_i \]

\( d_i \) is a vector of \( p \) variables gathered from big data sources
Use Survey Data to Remove Self-Selection Bias from Estimates Obtained Using Big Data

- An option is to use big data directly to measure poverty and social exclusion.
- It is realistic to think the big data are not representative of the whole population of interest (self-selection problem).
- Using a quality survey we can check difference in the distribution of common variables between big data and survey data.
- If there aren’t common variables we can use known correlated data to check difference in the distribution.
- Given this difference we can compute weights that allow the reduction of bias due to the self-selection of the big data.
Use Big Data to Validate Small Area Estimates

- Poverty and deprivation measures obtained from big data can be compared with similar measures obtained from survey data.
- If there is accordance between big data estimates and survey data estimates then there is a double checked evidence of the level of poverty and deprivation.
- If there is discrepancy there is need of further investigation.
Part II

Use Big Data as Covariate in Small Area Models
To use small area methods there is the need of survey and population data.

If available, the Census give plenty information on all population units.

However, Census info is only available every 10 years and the time lag with survey can affect the analyses.

One possible alternative: use area-level small area models with auxiliary information coming from alternative sources:

- Surveys
- Population registers
- Big Data
Small Area Estimation: Fay-Harriot Model

Based on mixed models, the Fay and Harriot model relates direct estimates with area level auxiliary variables

- Let \( \theta_i \) be a target parameter and \( \hat{\theta}_i \) its direct estimate in area \( i \)
- Let \( x_i \) be the auxiliary vector of \( p \) auxiliary variables for area \( i \)

\[
\hat{\theta}_i = \theta_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma_{\varepsilon_i}^2)
\]
\[
\theta_i = x_i \beta + u_i \quad u_i \sim N(0, \sigma_u^2)
\]

- The Fay and Harriot model is then

\[
\hat{\theta}_i = x_i^T \beta + u_i + \varepsilon_i
\]

- Model parameters can be estimated by maximum likelihood methods (\( \sigma_{\varepsilon_i}^2 \) are considered known)

By the use of the auxiliary information the accuracy of the estimates is improved
Measurement error in the covariates

- The proprieties of the small area estimators derived under the FH model are based on the hypothesis that the auxiliary data are available for all the areas and that they are measured without error.

- When this is not the case there is the need to account for the measurement error in the covariates, otherwise:
  - traditional FH estimators can be worst of the corresponding direct estimators in terms of precision;
  - the estimated MSEs of FH estimators can give a misleading notion of precision.

- Ybarra and Lohr (2008) proposed an extension of the Fay-Herriot model that can be used in these cases, e.g. when covariate information comes from surveys.
The FH model with measurement error in the covariates

- Suppose $x_d$ is the true value of the auxiliary variable in small area $d$ available for small area estimation
- When $x_d$ is measured with error, we substitute an estimator $\hat{x}_d$ for $x_d$ and use the following model:

$$\theta_d = \hat{x}_d^T \beta + r_d(\hat{x}_d, x_d) + \epsilon_d$$

where $r_d(\hat{x}_d, x_d) = u_d + (x_d - \hat{x}_d)^T \beta$, with $u_d$ and $\epsilon_d$ subject to the same hypothesis of the FH model
- Additional hypothesis: $u_d$ and $\theta_d$ are independent from $\hat{x}_d$, $\forall d$
- The last condition is satisfied when the auxiliary variables affected by measurement error come from a different source with respect to the target variables
- However, this problem can be solved changing the model according to Ybarra PhD thesis (Arizona State University, 2003)
The FH model with measurement error in the covariates

- The resulting EBLUP estimator derived by Ybarra and Lohr (2008) is

$$\hat{\theta}^{FHME}_d = \hat{\gamma}_d \hat{\theta}_d + (1 - \hat{\gamma}_d) \hat{x}_d^T \hat{\beta}$$

(1)

where $\hat{\gamma}_d = (\hat{\sigma}_u^2 + \hat{\beta}^T C_d \hat{\beta}) / (\hat{\sigma}_u^2 + \hat{\beta}^T C_d \hat{\beta} + \sigma_{\varepsilon}^2)$, with $C_d = \text{MSE}(\hat{x}_d)$ and $\hat{\beta}$ and $\sigma_{\varepsilon}^2$ are estimated according to an iterative procedure for the modified least squares as in Cheng and Van Ness (1999)

- In this modified least squares method $w_d, d = 1, \ldots, D$ are a set of finite weights bounded away from 0.

- The regression parameters $\tilde{\beta}$ satisfy the equation

$$\sum_{d=1}^{D} w_d (\hat{x}_d \hat{x}_d^T - C_d) \tilde{\beta} = \sum_{d=1}^{D} \hat{x}_d \hat{\theta}_d$$
The FH model with measurement error in the covariates

- When it is no possible to obtain estimates of $\beta$ from the previous equation, Ybarra and Lohr (2008) suggest to modify the estimator as follows
- Define $G = \sum_{d=1}^{D} w_d \hat{x}_d \hat{x}_d^T$
- Write $G^{-1/2}(\sum_{d=1}^{D} w_d C_d) G^{-1/2} = P\Lambda P^T$, where $P$ is orthogonal and $\Lambda$ is diagonal
- Let $Q$ be the diagonal matrix with $j$th diagonal entry $Q_{ij} = (1 - \Lambda_{ij})^{-1}$ if $1 - \Lambda > 1/D$ and $Q_{jj} = 0$ otherwise
- Estimates of $\beta$ and $\sigma_u^2$ are obtained setting the weights $w_d$ equal to 1 and computing the two first estimates as follows:

$$
\hat{\beta}_w = G^{-1/2}PDP^T G^{-1/2} \sum_{d=1}^{D} w_d \hat{x}_d \hat{\theta}_d
$$

$$
\hat{\sigma}_u^2 = (D - p)^{-1} \sum_{d=1}^{D} \left\{ (\hat{\theta}_d - \hat{x}_d^T \hat{\beta}_w)^2 - \psi_d^2 - \hat{\beta}_w^T C_d \hat{\beta}_w \right\}
$$
The FH model with measurement error in the covariates

- Then, the weights are set equal to \( w_d = 1/(\hat{\sigma}_u^2 + \psi_d^2 + \hat{\beta}_w^t C_d \hat{\beta}_w) \) and equations (2) and (3) are computed again.
- If desired, this process can be iterated.
- Ybarra and Lohr (2008) proof the consistency of these estimators and propose an analytic and a jackknife estimator of \( \text{MSE}(\hat{\theta}_{d}^{FHME}) \).
- Since the jackknife approach described in Ybarra and Lohr (2008) was too unstable with our data, in the application the \( \text{MSE}(\hat{\theta}_{d}^{FHME}) \) for both sampled and out of sample areas is estimated using a parametric bootstrap approach.
MSE estimation for the FHME estimator

Steps of the parametric bootstrap:

- Estimate $\beta$ and $\sigma_u^2$ following equation (2) and (3)
- Generate parametrically the errors $u_d^* \sim N(0, \hat{\sigma}_u^2)$ and $e_d^* \sim N(0, \hat{\psi}_d^2)$
- Generate the bootstrap true values $\theta_d^* \hat{x}_d^T \hat{\beta} + u_d^*$
- Generate the bootstrap direct estimates $\hat{\theta}_d^* = \theta_d^* + e_d^*$
- Generate a bootstrap matrix of auxiliary variables with errors $\hat{x}_d^* = \hat{x}_d + \varepsilon_d$, where $\varepsilon_d \sim N_p(0, C_d)$
- Using (1) obtain a bootstrap estimate $\hat{\theta}_d^{FHME*}$ of $\theta_d^*$
- Repeat the process B times obtaining B values of $\hat{\theta}_d^{FHME*b}$ and of $\theta_d^{*b}$
- The bootstrap MSE estimator of $\hat{\theta}_d^{FHME}$ is $\sum_{b=1}^{B} (\hat{\theta}_d^{FHME*b} - \theta_d^{*b})^2$
In the application we used area-level data coming from the EU-SILC survey 2011 and, as covariate information, data coming from the EU-SILC survey itself and from Big Data on mobility.

The areas of interest are the 57 Local Labour Systems (LLSs) of the Tuscany region. Of these areas, 24 are “out of sample areas” for the EU-SILC survey.

The Local Labour Systems (LLSs) are the areas in which most of the daily activity of the people who live and work in them takes place. They are intermediate between LAU 1 and LAU 2 levels.

Covariate information available to be used in the SAE models for the mean income and for the HCR?
The Census 2001 data are too old: the use of lagged Census info may lead to biased small area estimators.

We decided to use as covariate information available at the area level data coming from the EU-SILC survey itself and Big Data on individuals’ mobility.

The hypothesis is that data on mobility can be predictive of well-being measures, generalizing the approach of Eagle et al. (2010).

The Big Data refer to 10 million car travels tracked using the GPS by the OCTOTelematics s.p.a., a data collection service for insurance companies.

We used as covariates the mobility entropy, measuring the number of locations visited by each vehicle, and the radius of gyration, measuring how spread out are the visited locations from the vehicle center of mass.
Results of the application: we were able to map all the LLSs, even the out of sample ones

Figure: Estimates of the mean income (left) and of the HCR (right) for the LLSs of the Tuscany region
Moreover, using the SAE models for the mean income and HCR at the LLSs level we were able to reduced the Mean Squared Error of the direct estimates.

We also produced an estimation of the estimates’ variability using a resampling technique (bootstrap).

The gain of precision was observed for all the areas:

Figure: Ratio between the MSE of the SAE and of the direct estimates
Concluding remarks

- In time of spending review and budget restrictions auxiliary information that is relatively inexpensive and readily available, but that is still representative of the population under consideration - as the Big Data may be - is of substantial interest.

- The disadvantage is that these data require a more complex small area estimation model.

- Our evidence of correlation between poverty local estimates obtained by surveys and independent estimates from Big Data sources is encouraging and suggest to go on in this direction.