Small area methods and big data sources

Caterina Giusti

Department of Economics and Management, University of Pisa

Workshop
“Small Area Methods and living conditions indicators in European poverty studies in the era of data deluge and Big data”

Pisa, 8-10 May 2018

IN SUPREMÆ DIGNITATIS
1343

Centro Camilo Dagum
Tuscan Interuniversity Centre
Advanced Statistics for Equitable and Sustainable Developmen

SMALL AREA METHODS
FOR MONITORING OF POVERTY
AND LIVING CONDITIONS IN EU
Jean Monnet Chair

Small Area Estimation (SAE) and big data

- In resource-constrained environments - where censuses and household surveys are rare - the use of big data may create an option for gathering localized and timely information at a fraction of the cost of traditional methods
- In countries where official surveys are regularly conducted, big data represent a valuable resource also because they can be used to **improve the accuracy of local estimates**
- In the last few years our research team in Pisa focused on **model based methods** for estimation of **local poverty indicators**
- The work done has been developed within the **SAMPLE**, **e-Frame** and **InGRID** FP7 projects → now **InGRID-2** (www.inclusivegrowth.eu) and **MAKSWELL** (www.makswell.eu) H2020 projects

InGRID
Supporting expertise in inclusive growth

MAKSWELL
MAKing Sustainable development and WELL-being frameworks work for policy analysis

Small Area Estimation (SAE) and big data

- Estimation of poverty indicators for small areas is a crucial issue for **policy making**
- Small areas are **geographical areas or domains** fro which the survey sample size is not large enough to obtain reliable estimates
- Poverty is also crucial in the framework of SDGs indicators
- In the multidimensional definition of poverty, monetary poverty indicators such as the At-Risk-Of-Poverty Rate (ARPR) still play a crucial role
- Estimation of **monetary poverty indicators using data from official surveys** such as the EU-SILC requires the use of **SAE methodologies** since the sample size is usually to small to compute direct estimates at local level (e.g. below the NUTS2 level in Italy)

Small Area Estimation (SAE) and Big Data

- Marchetti et al. (2015) identified **three possible approaches** for the use of big data sources together with SAE methods:
  1. Use big data to validate small area estimates
  2. Use big data as covariate in small area models
  3. Use survey data to remove the bias from estimates obtained using big data

1. Use Big Data to validate small area estimates

- Socio-economic measures obtained from big data can be **compared** with similar measures obtained from survey data, e.g. poverty indicators
- If there is accordance between big data estimates and survey data estimates then there is a double checked evidence of the socio-economic measures of interest
- If there is discrepancy there is need of further investigation

1. Use Big Data to validate small area estimates: example

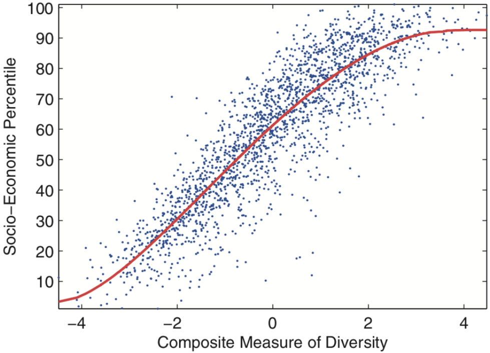

Figure: The relation between social network diversity and socioeconomic rank (Eaglet at al., 2010.

1. Use Big Data to validate small area estimates: example

- Generalising the approach of Eagle et al. (2010), we studied the possible **agreement between the level of poverty and the diversity of its inhabitants' mobility** in the provinces of the Tuscany region, Italy
- We considered SAE estimates of the **ARPR** for the **10 Tuscany provinces** computed by applying an **unit level M-quantile model** to **EU-SILC** 2008 and population census data
- The ARPR was defined as the share of households with income below the poverty line (60% of the Italian median equivalised household income)
- As covariate information for the households living in Tuscany we used data referring to the household (e.g. number of components) and to the head of the household (e.g. occupational status)

1. Use Big Data to validate small area estimates: example

- We considered an indicator of mobility defined using a large dataset of private vehicles in central Italy, tracked with a GPS device
- The travels were tracked using the GPS by the OCTOTelematics s.p.a., a data collection service for insurance companies
- The dataset is comprised of information on approximately **ten million different car journeys made by 150,000 vehicles tracked during May 2011**
- Focusing on Tuscany, the dataset refers to **37,326 vehicles**, which correspond to 1.5 percent of the total vehicles registered in Tuscany in 2011
- As 'big data' indicator we computed for each of the 10 provinces the **standard deviation of the mobility**, M_d, $d = 1, \ldots, 10$, indicated by S_{M_d}
- The mobility of area d, M_d, is defined as a mean measure of entropy of all the vehicles "resident" in area d

1. Use Big Data to validate small area estimates: example

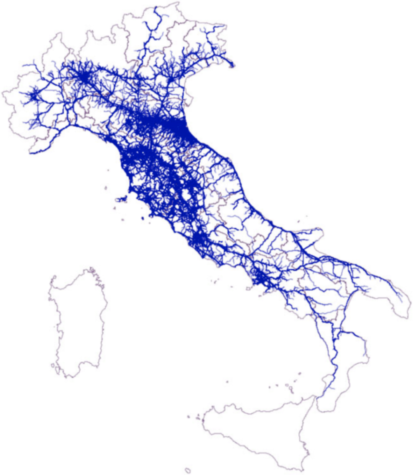

Figure: Spatial distribution of GPS trajectories in the dataset. The trajectories correspond to car travels performed by vehicle passed through an area corresponding to central Italy in May 2011. (Pappalardo at al., 2013.)

1. Use Big Data to validate small area estimates: example

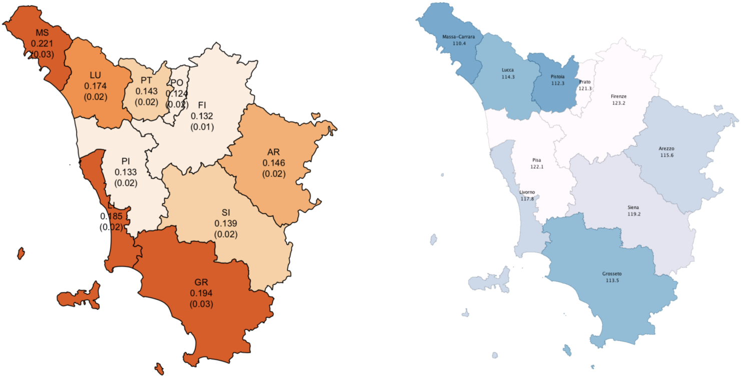

Figure: Model-based ARPRs with corresponding s.e. (left) and S_{M_d} (right) for the 10 provinces of the Tuscany region.

1. Use Big Data to validate small area estimates: example

- We analysed the relation between the two measures
- We decided to refine the analysis and to consider the second approach to the joint use of SAE methods and big data

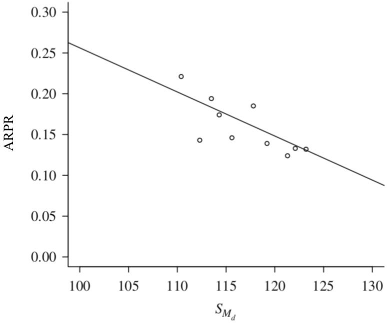

Figure: Model-based ARPRs versus S_{M_d} for the 10 provinces of the Tuscany region.

2. Use Big Data as covariate in small area models

- Big data often provide **unit level data**
- The outcome variable should be linked to the auxiliary variables in order to use unit level data in a small area model
- However, the reference population of survey and big data is often different
- Moreover, to technical challenges and law restrictions it is often impossible to have unit level big data that can be linked with administrative archives or census or survey data
- **Big data can be aggregated at area level** and then used in an area level model

$$\hat{\theta}_d = \boldsymbol{d}_d^T \boldsymbol{\beta} + u_d + \varepsilon_{\boldsymbol{d}}$$

where d is the area index, and $\boldsymbol{d}_d$ is a vector of p variables including big data sources

2. Use Big Data as covariate in small area models

- Due to the availability of direct estimates from the EU-SILC 2011 survey, we refined the previous analysis
- As areas of interest we considered the **57 Local Labour Systems** (LLSs) of the Tuscany region
- LLSs sample sizes: from 10 to 246 (mean 39.52, median 25)
- 24 out of the 57 LLSs are "out-of-sample areas" with a zero sample size in the EU-SILC 2011
- The target parameters are the **ARPR** and the **mean of the household equivalised income** for each LLS
- As covariate information available for all 57 LLSs we considered data from the EU-SILC survey and big data on individuals' mobility
- The hypothesis is that **mobility data are predictive of poverty measures**

Small area estimation: Fay-Harriot model

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

- θ_d target parameter in area d; $\hat{\theta}_d$ its direct estimate, $d = 1, \ldots, D$
- $\boldsymbol{x}_d$ vector of p auxiliary variables for area d

$$\begin{aligned}
\hat{\theta}_d &= \theta_d + \varepsilon_d \quad \varepsilon_d \sim N(0, \psi_d) \\
\theta_d &= \boldsymbol{x}_d \boldsymbol{\beta} + u_d \quad u_d \sim N(0, \sigma_u^2) \\
\hat{\theta}_d &= \boldsymbol{x}_d^T \boldsymbol{\beta} + u_d + \varepsilon_d
\end{aligned}$$

- Model parameters can be estimated by maximum likelihood methods (ψ_d are considered known)

By using auxiliary information the accuracy of the estimates can be improved

Measurement error in the covariates

- FH model hypothesis: auxiliary data are measured without error
- When this is not the case, there is the need to **account for the measurement error in the covariates**, otherwise:
  - 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): Fay-Herriot model extension that accounts for measurement error in covariates, e.g. when information comes from surveys

The FH model with measurement error in the covariates

- x_d is the true value of the auxiliary variable x in small area d
- If x_d is unknown than it can be estimated from survey data, $\hat{x}_d$ (that is a measure of x_d with sampling error)
- The FH-measurement-error model is as follows

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

  - $r_d(\hat{x}_d, x_d) = u_d + (x_d - \hat{x}_d)^T \beta$
  - $u_d \sim N(0, \sigma_u^2)$, $\varepsilon_d \sim N(0, \psi_d)$, $u_d \perp \varepsilon_d$
  - $u_d \perp \hat{x}_d, \hat{\theta}_d$
  - As in Ybarra and Lohr (2008), Marchetti et al. (2015) suppose that $\hat{\theta}_d \perp \hat{x}_d$

The FH model with measurement error in the covariates

- The resulting **EBLUP** (Empirical Best Linear Unbiased Predictor) is:

$$\hat{Y}_{dME} = \hat{\gamma}_d y_d + (1 - \hat{\gamma}_d)\hat{x}_d^T \hat{\beta}$$

  where $\hat{\gamma}_d = (\hat{\sigma}_u^2 + \hat{\beta}^T C_d \hat{\beta})/(\hat{\sigma}_u^2 + \hat{\beta}^T C_d \hat{\beta} + \psi_u^2)$ and $C_d = MSE(\hat{x}_d)$

- Ybarra and Lohr (2008) propose a Jackknife estimator for the $MSE(\hat{Y}_{dME})$ in the case of $\hat{x}_d \perp \hat{\theta}_d$
- Marchetti et al (2015) propose an alternative parametric bootstrap to estimate the $MSE(\hat{Y}_{dME})$

Our hypotheses on big data on mobility

- Big Data can be analysed from two alternative perspectives: as collected on a **self-selected sample** from the population - that is, under a **survey design perspective** - or not
- In Marchetti et al. (2015) we chose to follow the first perspective
- The **self-selection bias** is related to (Bethlehem, 2002):
  - the correlation between the target variable and the response behavior;
  - the variance of the response behavior;
  - the variance of the target variable;
  - to the average of the response behavior
- The **weighting adjustment** and **response propensity** are two methods that can be used to reduce the self-selection bias
- The main issue is data availability to apply these methods to big data

Self-selection bias of big data auxiliary variables

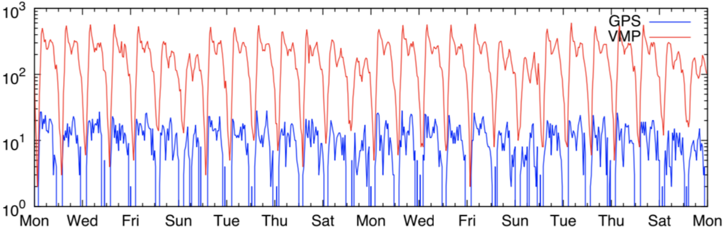

Figure: Traffic sensed by a Variable Message Panels device and GPS traffic volume in one of the twelve entry gates of Pisa.

Self-selection bias of big data auxiliary variables

- Pappalardo et al. (2013) show that the mobility measures based on big data are coherent with the mobility measures registered for all the vehicles in the municipality of Pisa (derived from traffic sensors)
- This result suggests that the mobility measures and the event ‘having a GPS’ should be independent
- The correlation coefficient between the mobility measure and the event ‘having a GPS’ should tend to 0
- Given this, according to the work of Bethlehem (2002), **the bias due to the self-selection process should be negligible in our application**
- Thus, we handled these data as if they were a **simple random sample** from the population of vehicles

2. Use Big Data as covariate in small area models: example

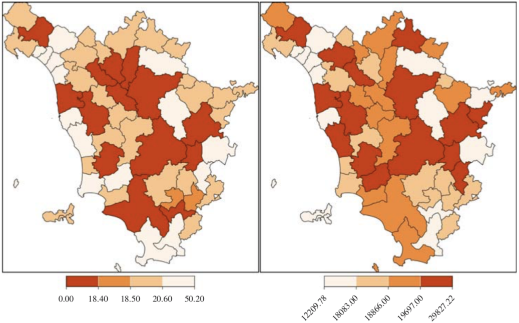

Figure: Estimates of the mean equivalised income in Euros (right) and of the HCR (left) for the Local Labour Systems of Tuscany region. Small area estimates based on EU-SILC 2011 and Mobility Data 2011. **Out-of-sample areas estimated using a synthetic estimator**.

2. Use Big Data as covariate in small area models: example 2

- Following again the second approach suggested by Marchetti at al. (2015), in a second application we used and indicator defined with **emotional data coming from Twitter** as auxiliary variable in an area level SAE model to estimate **Italian households' share of food consumption expenditure** in the 110 Italian provinces in 2012
- The **share of total expenditure that an household dedicate to food items** is an important indicator of the **household living conditions** (Deaton, 2003)
- as auxiliary variables in the area level models we used data coming from the **Population and Housing Census** 2011, from the **Survey on Social Actions and Services on Single and Associates Municipalities** 2012 and an indicator computed using **big data from Twitter**

Twitter data: the iHappy indicator

- We consider as potential covariate for our SAE working model the **iHappy indicator**
- The iHappy indicator referring to the year 2012 was computed by collecting and coding more than 43 millions of tweets posted on a daily basis in all the Italian provinces
- The words and emoticons of the tweets were classified using a training set in two categories: "happy" and "unhappy", together with a residual class "other"
- Then, Curini et al. (2015) derived the frequency distribution of the happy and unhappy tweets in the entire population
- The iHappy indicator was then computed for each Italian province as **the percentage ratio of the number of happy tweets to the sum of happy and unhappy tweets**

Results: correlation coefficients

Table: Linear correlation (ρ) between the selected auxiliary variables for the FH model and the SFCE variable.

	ρ
iHappy	-0.350
Share of owners of the house	-0.258
Share of household lead by female	-0.497
Expenses for household with children	-0.500
Expenses for old-aged persons	-0.332
Expenses for immigrants	-0.335
Expenses for at risk of poverty persons	-0.130
Expenses for services to families	-0.509

Results: map of the SAE estimates

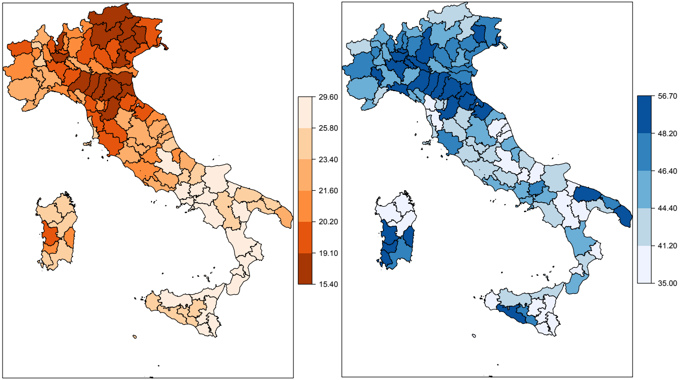

Figure: Map of the FH estimated of the SFCE (left) and map of the iHappy indicator (right) for the 110 provinces in Italy. In both the maps a darker colour corresponds to a better situation.

3. Use survey data to remove the bias from estimates obtained using Big Data

- An option is to use big data directly to measure the socio-economic indicators of interest
- It is realistic to think the **big data are not representative of the whole target population** (*coverage/self-selection* problems)
- How can we deal with these problems?
- There are many statistical methods that could be use to allow statistical inference for big data!
- For example, using a quality survey we could check difference in the distribution of **common variables between big data and survey data**
- A crucial point however is the availability of **identifiable big data** (Shlomo and Goldstein, 2015) that can be **linked** to survey data

Open questions

- Can we have **identifiable big data** to be use to compute local estimates of socio-economic indicators?
- Are **unit-level SAE models** an option with big data?
- Can **record linkage methodologies** be used to link survey and big data?
- Should we consider alternative approaches, such as ‘learning samples’ in official surveys to **derive the ‘big data profile’ of sampled units**?
- More generally (also for area level SAE models), can we **derive the ‘error profile’ of big data**?

References

Bethlehem, J.G. (2002)
Weighting Nonresponse Adjustments Based on Auxiliary Information
Survey Nonresponse (Groves, Dillman, Eltinge, and Little Eds.) New York: John Wiley and Sons

Curini, L. and Iacus, S. and Canova, L. (2015)
Measuring idiosyncratic happiness through the analysis of twitter: An application to the Italian case
Social Indicators Research 121, 525-542

Eagle, N. and Macy, M. and Claxton, R. (2010)
Network diversity and economic development
Science 328, 1029-1031.

Deaton, A. (2003)
Household surveys, consumption, and the measurement of poverty
Economic Systems Research 15, 135?159.

Marchetti, S. and Giusti, C. and Pratesi, M. and Salvati, N. and Giannotti, F. and Pedreschi, D. and Rinzivillo, S. and Pappalardo, L. and Gabrielli, L. (2015)
Small Area Model-Based Estimators Using Big Data Sources
Journal of Official Statistics 31, 263 – 281.

Shlomo, N. and Goldstein, H. (2015)
Editorial: Big data in social research
Journal of the Royal Statistical Society - Series A 178, 787?790

Ybarra, L.M.R., and S.L. Lohr. (2008)
Small Area Estimation When Auxiliary Information is Measured With Error
Biometrika 95, 919-931.