2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8302

A proposal to deal with sampling bias in social network big data

Iacus, Stefano Maria [a]; Porro, Giuseepe [b]; Salini, Silvia [a] and Siletti, Elena [a]
[a] Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Italy, [b] Department of Law, Economics and Culture, Università degli Studi dell'Insubria, Italy

Abstract

Selection bias is the bias introduced by the non random selection of data, it leads to question whether the sample obtained is representative of the target population. Generally there are different types of selection bias, but when one manages web-surveys or data from social network as Twitter or Facebook, one mostly need to focus with sampling and self-selection bias.

In this work we propose to use official statistics to anchor and remove the sampling bias and unreliability of the estimations, due to the use of social network big data, following a weighting method combined with a small area estimations (SAE) approach.

Keywords: *Big data; Well-being; Social indicators; Sentiment analysis; Self-selection bias; Small area estimation.*

1. Introduction

Despite, social media users could be thought of as the world's largest focus group, and be analysed as such (Hofacker et al., 2016), when one deals with such a data it seems that one cannot to take into account the selection bias. Indeed dealing with Twitter data (as for Iacus et al. (2017)), it is obvious that the sample is done by people that have Internet access, that have decided to open an account on Twitter and that are active users.

In statistical literature, studies largely addresses this bias using the propensity score (PS) approach (Rosenbaum & Rubin, 1983) or the Heckman approach (Heckman, 1979). Both methods attempt to match the self-selected intervention group with a control group that has the same propensity to select the intervention, but they both relay on information that dealing with data such as Twitter data are not disposable. In web-survey context, these issues have been addressed by some strategies based on weighting procedures and model-based approach (Bethlehem & Biffignandi, 2012), nevertheless, also these proposals relies on the availability of unit level information from big data sources, that nowadays are still a mirage, and that, dealing with aggregated big data, are always impossible to achieve.

2. Our proposal

We propose to manage sampling bias, due to the use of aggregated data from social networks, combining a weighting method with a small area estimation model. Our proposal start from this consideration: SAE models have been traditionally used to check and remove unreliability from direct estimations, because if we use direct estimations from Twitter data, those can suffer of selection bias as introduced above, first of all we use a weighting method and then we check and remove their unreliability using SAE models.

In big data context SAE models have been recently used, employing this new kind of data as a covariates when official statistics are missing or they are poor. Porter et al. (2014) use Google trends searches as covariates in a spatial FH model, while in Falorsi et al. (2017) the time series query share extracted always from Google Trends is used as covariate to improve the SAE model estimates for Italian regional young unemployed. Marchetti et al. (2015) use big data on mobility as covariates in a FH model to estimate poverty indicators; where accounting for the presence of measurement error they follow the Ybarra & Lohr (2008) approach. Moreover, themselves have proposed the use official data to verify and remove the self-selection bias due to the use of big data, but in addition to the suggestion, no concrete proposals has been made. Finally, Marchetti et al (2016) use data coming from Twitter as covariate to estimate Italian households' share of food consumption expenditure.

In order to proceed in our direction we have to take into account some topics. When we deal with big data, we often have not a really unit level data to use for direct estimations.

To overcome this problem we can consider different hierarchical levels of aggregations. As an example, we can think of Italian provinces as a unit level for regions. In this way, should be clear that the use of small sample techniques are suitable. Going back to the example: also if we manage million of tweet, if we consider provinces as statistics unit, this number will always be very small. A good and desirable property of big data is the high time frequency, however this feature is often disregarded for the official statistics. In this work, we consider data with the same frequency, but the opportunity to use data with a different time frequencies could be an interesting methodological challenge for the future. Lastly, dealing with timely and spatial information, we should take into account both time and space correlations too. Following these addresses, we now present our method step by step, and we propose an application as toy example.

2.1. The method

About SAE model, we consider area level models, because we assume to have area level covariates. Furthermore, because these data are available for several periods of time *T* and for *D* domains, to consider also eventually time and space correlations, we have chosen a spatio-temporal Fay-Herriot (STFH) model, proposed by Marhuenda et al. (2013). Thus, for domain *d* and *t* time periods, let μ_{dt} be the target parameter, the STFH model, as all the FH models, has two stages, where in first stage, the "sampling model" is defined as follow:

$$\hat{\bar{y}}_{dt}^{DIR} = \mu_{dt} + e_{dt}\ , \qquad e_{dt} \sim N(0, V(\hat{\bar{y}}_{dt}^{DIR}))\ , d = 1, \dots, D, t = 1, \dots, T \tag{1}$$

where e_{dt} are the sampling errors that are assumed to be independent and normally distributed, and $V(\hat{\bar{y}}_{dt}^{DIR})$ is the sampling variance of the direct estimator. Especially, we consider as direct estimator the regional sampling mean, weighted by some characteristics to overcome the non-sampling structure of our data

$$\hat{\bar{y}}_{dt}^{DIR} = \frac{1}{\sum_{i=1}^{n_{dt}} w_{idt}} \sum_{i=1}^{n_{dt}} y_{idt} w_{idt} \tag{2}$$

where n_{dt} is the number of provinces in region *d* at time *t*, and w_{idt} are the weights used. For the sampling variance we use the same weights.

While in second stage, the "linking" model is

$$\mu_{dt} = \boldsymbol{x}_{dt}'\beta + u_d + v_{dt}\ , \qquad u_d \sim N(0, \sigma_1^2)\ , v_{dt} \sim N(0, \sigma_2^2) \tag{3}$$

it relates all areas through the regression coefficients, $\boldsymbol{x}_{dt}$ is a column vector containing the aggregated values of *k* covariates for the *d*-th area in *t*-th period, and β is the vector of coefficients. u_d are the area effects, that follow a first order spatial autocorrelation, SAR(1), process with variance σ_1^2, spatial autocorrelation parameter ρ_1 and proximity matrix $\mathbf{W}$ of dimension $d \times d$. Especially, $\mathbf{W}$ is a row-standardized matrix obtained from an initial proximity matrix $\mathbf{W}^{\mathrm{I}}$, whose diagonal elements are equal to zero and the residual entries

equal to one, when the two domains are neighbours, and zero otherwise. Normality for u_d is required for the mean squared error, but not for point estimations. Furthermore v_{dt} represents the area-time random effects that are i.i.d. for each area *d*, following the first order autoregressive, AR(1), process with autocorrelation parameter ρ_2 and variance parameter equal to σ_2^2.

The final model is defined as $$\hat{\bar{y}}_{dt}^{DIR} = \boldsymbol{x}'_{dt}\beta + u_d + v_{dt} + e_{dt} \quad (4)$$

Then, $\boldsymbol{\theta} = (\rho_1, \sigma_1^2, \rho_2, \sigma_2^2)$ is the vector of unknown parameters involved in the STFH model. Marhuenda et al (2013) give the empirical best linear unbiased estimator (EBLUE) of β, and the empirical best linear unbiased predictors (EBLUPs) of u_d and v_{dt}. Both are obtained by replacing a consistent $\widehat{\boldsymbol{\theta}}$ in the respectively BLUE and BLUPs introduced by Henderson (1975). Also due to Marhuenda et al (2013) is the parametric bootstrap procedure for the estimation of the mean squared error (MSE) of the EBLUPs, that for *B* bootstrap replies has the following form $$MSE(\hat{\mu}_{dt}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\mu}_{dt}^{b} - \mu_{dt}^{b}\right)^2 \quad (5)$$

In this way the point estimates $\hat{\mu}_{dt}$ of μ_{dt} can be supplemented with (5) as measure of uncertainty.

3. Application

The assessment of well-being mostly at local level is an important task for policy makers, because they increasingly need to target their policies and actions not to the nation, but at local domains. Unfortunately very often there is a lack of this level data, even more if the interest is for high frequency data too. To fill this gap, the use of data from social networks can be considered a good option to improve well-being knowledge. In this section considering a well-being index from Twitter data and some official statistics, we implement the proposed approach to check and, if necessary, to remove unreliability of estimations.

3.1. A Subjective Well-being Index with the Twitter Data

Since 2012 Iacus et al. (2015) propose to apply iSA (integrated Sentiment Analysis, Ceron et al. (2016)) method, to derive a composite index of subjective well-being that capture different aspects and dimensions of individual and collective life. This index named Social Well Being Index (SWBI) monitor the subjective well-being expressed by the society through the social networks, epsetially, in Iacus et al. (forthcoming) the SWBI index is provided for the Italian provinces from 2012 to 2016 and combined with the "Il Sole 24 Ore Quality of Life index". SWBI is not the result of some aggregation of individual well-being measurements, but it directly measures the aggregate composition of the sentiment throughout the society at province or regional level. For this reason, about the weights, we

can't consider users characteristics as traditionally, but aggregated to area ones. SWBI has been inspired by the definitions introduced by the think-tank NEF (New Economic Foundation), for its Happy Planet Index (New Economics Foundation, 2012), and it is defined as a manifold, dynamic combination of different features, with indicators which look beyond the single item questions and capture more than simply life satisfaction.

The eight SWBI dimensions concern three different well-being areas: personal well-being, social well-being and well-being at work. Data source are tweets written in Italian language and from Italy and data are accessed through Twitter's public API. A small part of these data (around 1to 5% each day) contain geo-reference information which allows to build the SWBI indicator at province level in Italy.

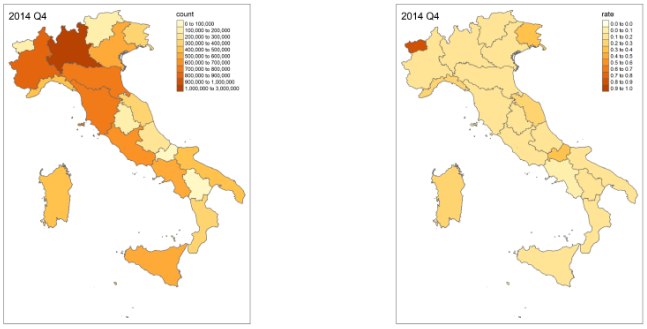

Figure 1. Twitter counts, on the left, and Twitter rates, on the right, in 2014-Q4.

In the application presented here, we consider SWBI quarterly data, with an area aggregation at the Italian provincial and regional levels. Now we shortly describe the dimension of these data. To compute the the SWBI index we consider more than two hundred million tweets (201,496,621) in 22 quarters from 2012 to June 2017. Despite this huge total number, we have to reveal a decrease in the number of tweets for all the four quarters of 2015 and the first quarter of 2016, anyway we stress that also for these quarters the counts were still in the thousands (minimum count equal to 1727 tweets in 2016-Q1 for Basilicata). To have a more realistic view of the situation, we consider a Twitter rate: the ratio between the number of analysed tweets and the number of inhabitants in the region in the same period. Considering the simple counts, we would have seen that most of the SWBI info comes from Lombardia, Piemonte, Emilia and Toscana, while considering also the resident population, we note that information is also substantial in particular or small regions such as Friuli, Sardegna, Valle d'Aosta and Molise, for the last two we remark their large variability during the period too. While the dispersion for big regions like Lazio and Lombardia seems to be smaller. However, we can conclude that during our observational period the average Twitter rate is equal to about 20% of the population, with a mean value always greater than 9% (minimum for Campania), for all regions. Looking at Figure 1, as

example, it is clear that, considering the Twitter rate (on the right), all the Italian regions are homogeneously observed with the exception of the Valle d'Aosta which have a higher but almost anomalous rate.

3.2. The implemented model

To implement the proposed model, as toy example, we consider only the `wor` dimension of the SWBI index, at quarterly time level, from 2012-Q1 to 2017-Q2. We compute regional values following (2) and using as weights: the broadband coverage and the Twitter rate. The broadband coverage is provided by "Il Sole 24 Ore" and Infratel Italia, this coverage can be considered as the opportunity to access internet in the different provinces. While the Twitter rate, computed in each period and province level, can be a proxy of the use of Twitter.

The weighted quality of job has remained stable, with very little variability between the regions, the distributions are very compressed, until the second half of 2015. From 2015-Q3 weighted `wor` grows, and especially from the second half of 2016, this dimension attained values greater than 80. Even the differences between the regions are more evident: the distributions are less crushed. We remark that the shapes of the quality of work weighed or not, computed as simple means, are quite similar. It seems that difference for the weighted `wor` index among regions are small, instead are greater in time. Considering the different rankings obtained by the two indices, in the 29% of the cases there are no differences and only for the 14% of the cases, there is a difference (Δ) in the rankings greater than 5 positions. Regions with the greater Δ are Trentino, Campania, Marche, and Sardegna, for the first two we remark the greatest position improvement, while for the last two we remark the greatest worsening of position. Focusing on time, we recorded the major Δ in 2014-Q1, 2015-Q1 and both the considered 2017 quarters. The mean of the ranking Δ is equal to 2.01(SD = 2.41).

Referring to the model (4) we use as direct estimator of regional quality of job the weighted `wor` and its sampling variance. Because in one Italian region, Valle d'Aosta, there is only one province, we decided to drop this region from our data. In the recked STFH model data are available for $T = 22$ time instants, from the first quarter of 2012 to the second quarter of 2017, and the dominions are the $D = 19$ considered Italian regions. The considered area level auxiliary variables were, before any process of selection, in the job context: the unemployment and the inactivity rate, computed both in relation to the labour force, as traditionally, and to the resident population; while in the socio demographic context: the birth, the mortality and the natural rate. All the covariates come from official statistics distributed by ISTAT (http://istat.it/), as representatives for all the Italian regions at quarters time level. The row-standardized proximity matrix $\mathbf{W}^{c}$ of dimension 19×19 , has been obtained from an initial proximity matrix $\mathbf{W}^{Ic}$, whose diagonal elements are equal to zero

and the residual entries equal to one, when the two regions had some common borders, and zero otherwise. Since in Italy, there are two regions corresponding to two islands, for them we take as neighbours the regions with direct naval connections.

3.3. Results and discussion

After fitting the model the selected covariates were the unemployment rate, calculated traditionally dividing by the labour force and the mortality rate. The coefficients were both negative: regions with larger unemployment and mortality rate have smaller quality of job. The estimated spatial autocorrelation ρ_1 is significant enough with a small and negative value of about -0.02. While the temporal autocorrelation parameter ρ_2 is still significant and greater with a positive value equal about to 0.86. The value equal to zero for $\hat{\sigma}_1^2$ is coherent with analysis of distribution discussed above. Quality of job change in time but less or not at all between regions.

Comparing the resulting EBLUPs obtained by fitting the STFH model and the direct estimates, weighted or not, we can conclude that the direct weighted estimates are approximately design unbiased. Looking at the rankings, what change if we use EBLUPs estimates instead of direct, weighted or not, estimates ? Comparing the rankings obtained with the simple means `wor` and those with EBLUPs estimates, we find that in the 31% of the cases the position is the same and in the 14% of the cases the position Δ is greater than 4. Regions and time situation is the same as when we compared above simple means with weighted means. The mean of the ranking Δ is equal to 1.97 (SD = 2.3). While comparing the rankings obtained with the weighted means `wor` and those with EBLUPs estimates, the situation is very different: in the 88% of the cases the position is the same and in less than 1% of the cases the position Δ is greater than 4. Only in one case we have a great ranking Δ (Marche in 2015-Q3, $\Delta = 7$). The average of the differnces in this case is equal to 0.16 (SD = 0.54). This means that moving to weighted estimates to EBLUPs estimates the ranking are quite the same.

In SAE literature, traditionally use coefficients of variations (CV) to analyse the gain of efficiency of the EBLUPs estimates. National statistical institutes are committed to publish statistics with a minimum level of reliability (<20%). The CVs of our three indices, where those computed for the STFH model have been obtained using boostrap and (5), except few exceptions, are always lower 20%. For 14 regions the CVs are still less than 10%, while the highest CVs values have been obtained only in few quarters for 5 regions: Calabria, Friuli, Lazio, Marche, and Trentino. What is clear is that whenever we observe a peak of CVs, the EBLUPs estimates improve reliability, but also considering only weighted indices this happen. CVs obtained for EBLUPs estimates are quite always lower than both others, but in some cases they are larger, for very small values, than those obtained for weighted

estimations, examples for Trentino in 2014-Q4 and 2015-Q1. Thus, EBLUPs based on STFH model are always more reliable than direct simple mean.

More results and discussion will be introduced during the conference presentation.

References

Bethlehem, J. & Biffignandi, S. (2012) *Handbook of Web Surveys*. John Wiley and Sons, Inc., New York, DOI 10.1002/9781118121757

Ceron, A., Curini, L. & Iacus, S.M. (2016) iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences*, 367-368, 105-124.

Falorsi, S., Fasulo, A., Naccarato, A. & Pratesi, M. (2017) Small area model for Italian regional monthly estimates of young unemployed using Google trends data. ISI2017-Marrakech.

Heckman, J.J. (1979) Sample selection bias as a specification error. Econometrica 47(1), 153-161.

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2), 423-447.

Hofacker. C.F., Malthouse. E.C. & Sultan, F. (2016). Big data and consumer behavior: imminent opportunities. *Journal of Consumer Marketing*, 33(2), 89-97.

Iacus, S.M., Porro, G., Salini, S. & Siletti, E. (2015) Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being. ArXiv e-prints 1512.01569

Iacus, S.M., Porro, G., Salini, S. & Siletti, E. (2017) How to exploit big data from social networks: a subjective well-being indicator via twitter. In: Petrucci, A. & Verde, R. (eds) *SIS 2017. Statistics and data science: new challenges, new generations*. Proceedings of the Conference of the Italian Statistical Society, Firenze University Press, Firenze, 537-542.

Iacus, S.M., Porro, G., Salini, S. & Siletti, E. (forthcoming) Social networks data and subjective well-being: an innovative measurement for italian provinces. Italian Journal of Regional Studies.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L. & Gabrielli, L. (2015) Small area model-based estimators using big data sources. Journal of Official Statistics 31(2), 263-281.

Marchetti, S., Giusti, C. & Pratesi, M. (2016) The use of twitter data to improve small area estimates of households' share of food consumption expenditure in italy. AStA Wirtschafts und Sozialstatistisches Archiv 10(2), 79-93.

Marhuenda, Y., Molina, I. & Morales, D. (2013) Small area estimation with spatio-temporal Fay Herriot models. *Computational Statistics & Data Analysis*, 58, 308-325.

Molina, I. & Marhuenda, Y. (2015) sae: An R Package for Small Area Estimation. *The R Journal*, 7(1), 81-98, https://journal.r-project.org/archive/2015/RJ-2015-007/index.html.

New Economics Foundation (2012) *The happy planet index: 2012 report. a global index of sustainable well-being*. Tech. rep.

Porter, A.T., Holan, S.H., Wikle, C.K. & Cressie, N. (2014) Spatial Fay Herriot models for small area estimation with functional covariates. *Spatial Statistics* 10, 27-42.

Rosenbaum, P.R, Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41-55.