The Survey Statistician, 2024, Vol. 89, 59–67.

Social big data to enhance small area estimates

Stefano Marchetti[1] **and Francesco Schirripa Spagnolo**[2]

[1]University of Pisa, Italy, stefano.marchetti@unipi.it
[2]University of Pisa, Italy, francesco.schirripa@unipi.it

Abstract

The paper discusses the challenge and the opportunity of the use of big data in small area estimation applications. Big data come in an unstructured way, and they are self-selected. Therefore, they have to be handled with care, and they need adequate statistical methods to produce sound statistics. Together with the discussion, we present an application where data mined from Twitter (now changed to X) are used to improve small area estimates of consumption expenditure for leisure at the local level in Italy.

Keywords: Twitter, area-level models, leisure consumption

1 Introduction

In recent decades, there has been a growing demand for small area official statistics for decision-making at the local level. By small area, we mean subdomains of a population (such as geographical areas or socio-economic groups) where direct estimates – based on area units only – from surveys typically do not offer accurate estimations. To overcome this problem, survey statisticians rely on a model-based approach and use statistical models to *borrow strength* across areas. Rao and Molina (2015) and Pratesi (2016) provide a comprehensive account of model-based approaches for Small Area Estimation (SAE).

At the same time, as a result of technological innovations and the growth of the Internet and the Web, the availability of new kinds of unstructured and heterogeneous data originating from ICT systems, the so-called big data, is increasing at an unprecedented rate. Examples of big data sources are GPS data, mobile phone data, internet searches, and social networking. Many of these data can be viewed as proxies of social behaviour along various dimensions. For instance, data coming from social networks, blogs, or web search keywords can trace desires, opinions, and feelings; records of mobile phone calls and GPS trajectories can trace the movement of individuals (Marchetti et al., 2015). Consequently, a growing number of analysts and academics have looked into the benefits of utilizing big data in socioeconomic studies.

For what concerns the use of big data in SAE there is a need to point out the dimension of big data in terms of the number of units in the small areas. We start by recalling the definitions of big data, which are many. A popular definition is: "*big data is the information asset characterized by such a high Volume, Velocity and Variety to require specific technology and analytical methods for its transformation into Value*" (De Mauro et al., 2016). It is also known as the four V's definition (Volume, Velocity, Variety, Value). Other definitions add more V's, like Veracity, Variability, etc., up to 42 V's in Farooqi et al. (2019).

Assuming that a high volume of data is fundamental to dealing with big data, this does not automatically imply that we have the availability of a large sample size at the small area level. Indeed, we can distinguish two kinds of big data: i) "horizontal" and ii) "vertical".

The first kind of big data is characterized by many observations for each unit, e.g. vehicles with GPS track can produce data for longitude, latitude and altitude variables every 5", for about 520 thousand observations in one month for each variable for each unit. Nevertheless, having many observations for each unit does not imply having observed many units, e.g. we can have a few vehicles with GPS track at the small area level, like in Marchetti et al. (2015).

The latter are big data characterized by a large number of units at the small area level. Of course, we can have big data that are both vertical and horizontal.

Moreover, the high spatial granularity of big data and their availability at an unprecedented temporal detail may enable us to use them to infer in near real-time socio-economic characteristics for an entire nation as well as for disaggregated geographical domains, which are important for timely, evidence-based policies. The integration of big data in the SAE framework represents a valuable resource to improve the accuracy of local estimates.

In the last decade, the use of big data in the SAE framework has been increasingly explored by researchers, with the aim of estimating well-being and other socioeconomic indicators, such as poverty indicators, for unplanned domains such as provinces in Italy (NUTS 3 in the Eurostat nomenclature), as their knowledge may be useful in better planning local policies and distributing welfare resources.

Model-based SAE can be divided into two approaches: area-level and unit-level models (Rao and Molina, 2015). The first uses aggregated area-level data in the regression model. The latter uses unit-level data (microdata), where a model fitted on sample data is then projected on the population data. Area-level models represent a natural way to include big data that can be aggregated at the area-level (Porter et al., 2014; Marchetti et al., 2015, 2016; Schmid et al., 2017); on the contrary, to the best of our knowledge, the use of unit level models by including big data has been only recently investigated (Pratesi et al., 2022).

We can identify three possible approaches for the use of big data in the SAE framework (Marchetti et al., 2015; Pratesi and Schirripa Spagnolo, 2023):

1) Local indicators are created by using big data sources, and they are then compared with the results obtained from SAE techniques.

2) Big data are used to generate new covariates for small area models

3) Survey data are used to check and remove the self-selection bias of the values of the indicators obtained using big data.

The first approach has been shown by Marchetti et al. (2015), where an entropy-based mobility variability index has been compared to poverty incidence at the province (small area) level in Tuscany, showing the potential of big data to catch the direction of poverty incidence. The idea of this first approach is to use big data to estimate a target parameter or an index strictly related to that target,

and then validate the results using traditional high-quality surveys. If results based on big data prove to be reliable over time, then they can be used to anticipate results from surveys, which are typically available several times after the big data.

At the moment, the second approach is the most explored. In particular, under this approach, Porter et al. (2014) used Google Trends data as covariates in a standard spatial Fay–Herriot (FH) model (as in Pratesi et al., 2009) to estimate the relative changes in rates of household Spanish-speaking in the United States. Marchetti et al. (2015), presented an application of the modified FH model (Fay and Herriot, 1979) proposed by Ybarra and Lohr (2008) to estimate poverty indicators for local areas in Tuscany using big data on mobility as covariates. In particular, the authors used mobility indexes based on different car journeys between locations automatically tracked with a GPS device. Similarly, Marchetti et al. (2016) used data coming from the social network Twitter to predict the share of food consumption expenditure of Italian households at the provincial level. In particular, they included as a covariate in the FH model an indicator, called iHappy, obtained from the analysis of Twitter (now named X) data measuring happy tweets to the total of tweets at the provincial level. They showed that this indicator has a good predictive power for food consumption expenditure and can be used as a proxy to measure households' living conditions. Another interesting application developed following the second approach described above was proposed by Schmid et al. (2017), who used mobile phone data to estimate subnational estimates of the share of illiterate individuals by gender at the local (*commune*) level in Senegal.

When we deal with horizontal big data - the big data size at the small area level is small - if we want to use big data as auxiliary variables in the SAE models, we can use the modified version of the FH model proposed by Ybarra and Lohr (2008), which allows for the sampling error in the auxiliary variables.

On the other hand, when we deal with vertical big data, the survey error at the area-level can be negligible and a standard FH model can be appropriate if we want to use them as auxiliary variables in the small area model (Marchetti et al., 2016).

For what concern the third approach, in the last year, many scholars have focused on the issue of selection bias that arises when big data are used. In particular, according to the third approach mentioned above, integrating data from a probability survey and a non-probability source is used to make valid inferences (for a review on this topic see Yang and Kim, 2020; Lohr and Raghunathan, 2017). In the SAE framework, this problem has been addressed by Pratesi et al. (2022), who proposed a method based on the integration of a probability and a nonprobability sample to reduce the selection bias associated with the big data source when the aim is to predict statistics related to enterprises at the local level. In their work, the authors assumed that the variable of interest is only in the non-probability (big data) data source.

In this paper, we show the potential of big data coming from the social network Twitter to improve small area estimates based on an area-level FH model (see details on Rao and Molina, 2015). We use the second approach for this application among the three possible approaches suggested above.

2 Potential of social big data to improve small area estimate, an application

In this section, we use a social index based on text analysis of georeferenced tweets from Twitter to improve the small area estimates of consumption expenditure for leisure at the province level in Italy in 2017.

2.1 Data

In this application, we use the Italian Household Budget Survey (HBS) 2017, aggregated data coming from the tax and population registers and the iHappy index based on big data from Twitter.

In Italy, the HBS is the main source of information concerning consumption expenditure. The HBS has a stratified two-stage design, which allows for reliable estimates at the regional level (NUTS 2). The sample size is about 17000 households and 40000 persons. The sample size at the province level varies between 20 and 1036, with a median of 125. Therefore, in most of the provinces direct estimation leads to unreliable estimates. From the HBS we estimate the consumption expenditure for leisure activities, which is a driver of subjective well-being (Noll and Weick, 2015), and it is our target variable from which we want to obtain a mean estimate at the province level in Italy.

Model-based small area methods require the use of auxiliary variables which are related to the target variable. We found that province-level variables coming from the Italian tax register are suitable for our purpose. From this source, we have available the following variables: per capita tax, per capita income from real estate, per capita income from labour, and proportion of taxpayers. From the population register we obtain the mean age at the province level.

We also considered as a potential source of auxiliary information big data obtained from Twitter in 2017. In particular, we use the iHappy index at the province level available from the Opinion Analytics platform Voices from the Blogs (available here `http://media2.corriere.it/corriere/pdf/2018/cultura/ihappy2017-18-def.pdf?fbclid=IwAR2dUuMDUxXEJICRi6Owpw4ICpFGi3jTLlyOZ-y1YxXj1tAaC6GIAfj2bn4`)[1]. This index referring to the year 2017 was obtained from more than 52 million tweets posted on a daily basis in all the Italian provinces. A text analysis of the tweets classifies them into two categories: "happy" and "unhappy" (Curing et al., 2015). The iHappy index is the percentage ratio of the number of happy tweets to the sum of happy and unhappy tweets. The iHappy index for Italy in 2017 is 55.1%. The happiest province is Genova (north-west of Italy) with 59.5%, while the unhappiest is Aosta with (also in the north-west) 49.9%.

2.2 Small area estimation method

Given the availability of data aggregated at the provincial level we use an area-level model, as described in Rao and Molina (2015). Let θ_i be the target parameter (mean or total of a target variable) for area i, and let $\hat{\theta}_i$ be its direct estimator, then $\hat{\theta}_i = \theta + \varepsilon$, where under random sampling we usually assume $\varepsilon \sim N(0, \psi_i)$, and ψ_i is the variance of $\hat{\theta}_i$. Let $\boldsymbol{x}_i$ be a vector of auxiliary variables for area i, then we assume $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i$, where $\boldsymbol{\beta}$ is a vector of regression coefficients and $u_i \sim N(0, \sigma_u^2)$. The FH model is then

$$\hat{\theta}_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i + \varepsilon_i,$$

where u_i represents the area i random effect, which is independent from ε_i.

The best linear unbiased predictor (BLUP) of θ_i is $\tilde{\theta}_i = \hat{\theta}_i\gamma_i + \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}(1-\gamma_i)$, where $\gamma_i = \sigma_u^2/(\sigma_u^2+\psi_i)$ and $\tilde{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, while σ_u^2 and ψ_i are unknown. Usually, ψ_i is considered known, and the smoothed estimator of ψ_i is treated as if it is the true sampling variance. The BLUP is a convex combination of the direct estimator $\hat{\theta}_i$ and of the predicted value $\boldsymbol{x}_i^T\boldsymbol{\beta}$, with weights γ_i and $1-\gamma_1$, where γ_i is the relative sizes of the model error variance σ_u^2 and the sampling error variance ψ_i.

Using maximum likelihood estimation or restricted maximum likelihood estimation we can obtain es-

[1] Link verified on December 11th 2023

timates of σ_u^2 and β, so to obtain the empirical best linear unbiased predictor (EBLUP):

$$\hat{\theta}_i^{FH} = \psi_i\hat{\gamma}_i + \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_i(1-\hat{\gamma}_i), \quad \hat{\gamma}_i = \frac{\hat{\sigma_u^2}}{\hat{\sigma_u^2}+\psi_i}.$$

The mean squared error (MSE) of the BLUP is $MSE(\tilde{\theta}_i) = E[(\tilde{\theta}_i - \theta_i)^2] = \psi_i\gamma_i$. Therefore γ_i measures the reduction of the variability of the BLUP with respect to the variability of the direct estimator. The MSE of the EBLUP is $MSE(\hat{\theta}_i^{FH}) = \psi_i\hat{\gamma}_i + \boldsymbol{x}_i^T V(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i(1-\hat{\gamma}_i)^2 + \psi_i^2(\psi_i+\sigma_u^2)^{-3}V(\hat{\sigma}_u^2)$. Analytical estimation of the MSE of the EBLUP is possible, details in Rao and Molina (2015).

Estimates in our application have been obtained using the package `emdi` (Kreutzmann et al., 2019) in the `R` environment (R Core Team, 2023).

2.3 Estimate of the mean consumption expenditure for leisure at the provincial level with and without big data covariates

We show how the use of the iHappy index as a covariate in the FH model in addition to other register-based covariates can improve the efficiency of estimates in many areas. Our target is the mean consumption expenditure for leisure at the province level in Italy. As discussed before, the sample size of the HBS does not allow for reliable estimates at such a level of aggregation. Therefore, we resort to SAE models, and in particular to the FH area-level model described above because we have access to aggregated data only.

First, we obtain direct estimates of the mean consumption expenditure at the province level (θ_i) for all the 107 Italian provinces using micro-data from the HBS 2017 edition[2]. Let y_{ij} be the expenditure for leisure for household j in province i, and let w_{ij} be the associated survey weight, adjusted for non-response and measurement error, and calibrated at the provincial level. Then a direct estimator of θ_i is $\hat{\theta}_i = \sum_{j=1}^{n_i} y_{ij}w_{ij} / \sum_{j=1}^{n_i} w_{ij}$. According to Statistics Canada, there are no restrictions to publish estimates with a coefficient of variation (CV) less than 16%, other national statistical offices use different thresholds (Eurostat, 2013). Using a CV less than 16% as a reference, in our case only 3 provinces out of 107 have such a CV, making evident the need to resort to SAE methods.

We adopt two different FH models to improve the efficiency of direct estimates, one without and one with the iHappy index (x_1). As auxiliary variables in the FH model without iHappy, we use the province mean age (x_2) and the per capita tax (x_3).

The two FH models successfully increase the efficiency of direct estimates. In table 1 we show the number of provinces for which the CV is less or equal to 16% and for which is greater than 16%, i.e. the number of provinces for which the estimates are considered reliable or not. As already noted, only three direct estimates can be considered reliable. On the contrary, the small area estimates are reliable for all 107 provinces.

Table 1: Number of provinces by coefficient of variation (CV)

Estimator	CV $\leq 16\%$	CV $> 16\%$
Direct	3	104
FH without iHappy	107	0
FH with iHappy	107	0

Given that both the FH models work well, we observed that the estimates based on the FH model that include the iHappy index are for 87 out of 107 provinces a little bit more efficient than the estimates

[2]These data are made available to the authors under the European Project MAKSWELL, grant agreement 770643

obtained without the iHappy index. In table 2 we show the ratio between the estimated MSE of small area estimates obtained with and without the iHappy index. A value smaller than one means that small area estimates obtained with iHappy are more efficient than those obtained without.

Table 2: Summary of the ratio between estimated MSEs of small area estimates obtained with and without the iHappy index

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.969	0.985	0.990	0.994	0.996	1.068

The gain in efficiency is very limited, but is however present.

The estimates (with iHappy) of the mean consumption expenditure for leisure at the provincial level in Italy in 2017 are mapped in figure 1. The values are expressed in euro per month, and represent the average household expenditure for leisure in a month. The mean consumption expenditure for leisure at national level in Italy is 42.61 euro per month. At province level the expenditure for leisure ranges from 1.76 per month in Nuoro, the central part of Sardinia (the Italian island on the 40°N parallel) to 85.00 euro per month in Milan, north-west of Italy. Milan is an outlier, with a very high expenditure level. The highest one without considering Milan is 67.42 euro per month in Bolzano, in the north-east of Italy, on the border with Austria. A similar level of consumption expenditure (67.22 euro) is in Monza-Brianza, which borders the Milano province. From figure 1 is also evident the socio-economic north-south divide present in Italy, with the highest values of consumption expenditure for leisure in the northern provinces and the lowest values in the southern provinces.

3 Concluding remarks

This article summarises the possible use of big data in small area estimation. Big data are a valuable source of information and knowledge that come in an unstructured way. After appropriate elaboration, they can be used, among others, in the framework of small area estimation, mainly in three ways: i) to create indexes that predict similar indicators from reliable data sources, such as surveys, ii) as auxiliary variables in small area estimation methods, iii) to estimate small area target parameters that are adjusted for self-selection and measurement error using survey sample. All these three methods have been explored in the literature, however, the approach ii) is the most used, and it is used in the application shown in this article. Similarly to previous works, we use an index obtained from Twitter data as a covariate in a small area model to estimate the consumption expenditure for leisure in Italy at the provincial (small area) level in 2017. Even if the application can be considered as a bit more than an example, it shows the potential of the use of big data in SAE.

Big data represents a challenge and an opportunity, in the field of small area estimation and in many other fields of statistics. In the last years, many efforts have been made to use them to obtain sound statistics. However, it is important to remark that it is often difficult to have access to big data, that are mainly available to big tech companies, banks, big retail shops, etc. Nevertheless, the possibility of scraping data from the web makes it possible to have access to a wide range of big data, that can be used by researchers to explore the potential of this new source of information.

Acknowledgment

This paper is supported by the Ministry of University and Research (MUR) as part of the FSE REACT-EU - PON 2014-2020 “Research and Innovation” resources - Innovation Action - DM MUR 1062/2021 - Title of the Research: “Statistical Machine Learning nelle Indagini Campionarie”.

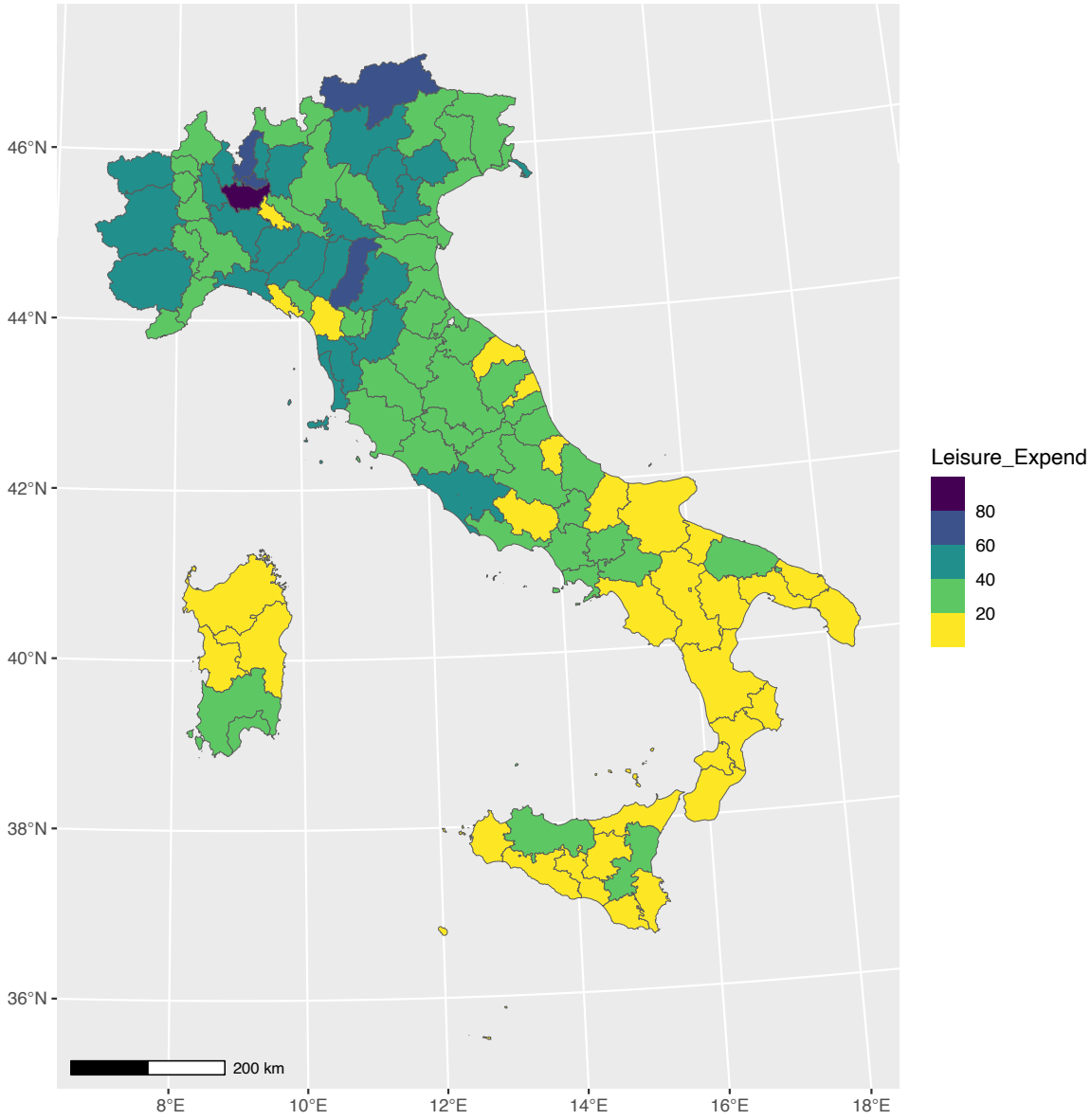

Figure 1: Estimates of consumption expenditure on leisure at provincial level in Italy, 2017

References

Curing, L., S. Iacus, and L. Canova (2015). Measuring idiosyncratic happiness through the analysis of twitter: An application to the italian case. *Social Indicators Research 121*(2), 525–542.

De Mauro, A., M. Greco, and M. Grimaldi (2016, 03). A formal definition of big data based on its essential features. *Library Review 65*, 122–135.

Eurostat (2013). *Handbook on precision requirements and variance estimation for ESS households surveys*. Eurostat.

Farooqi, M. M., M. A. Shah, A. Wahid, A. Akhunzada, F. Khan, N. ul Amin, and I. Ali (2019). Big data in healthcare: A survey. In F. Khan, M. A. Jan, and M. Alam (Eds.), *Applications of Intelligent Technologies in Healthcare*, pp. 143–152. Springer International Publishing.

Fay, R. E. and R. A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association 74*(366a), 269–277.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software 91*(7), 1–33.

Lohr, S. L. and T. E. Raghunathan (2017). Combining survey data with other data sources. *Statistical Science 32*(2), 293–312.

Marchetti, S., C. Giusti, and M. Pratesi (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.

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

Noll, H.-H. and S. Weick (2015). Consumption expenditures and subjective well-being: empirical evidence from germany. *International Review of Economics 62*(2), 101–119.

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

Pratesi, M. (Ed.) (2016). *Analysis of Poverty Data by Small Area Estimation*. New Yok: John Wiley & Sons.

Pratesi, M., N. Salvati, et al. (2009). Small area estimation in the presence of correlated random area effects. *Journal of Official Statistics 25*(1), 37.

Pratesi, M. and F. Schirripa Spagnolo (2023). Small area methodology for measuring poverty at a local level. In J. Silber (Ed.), *Research Handbook on Measuring Poverty and Deprivation*, pp. 129–140. Edward Elgar Publishing.

Pratesi, M., F. Schirripa Spagnolo, G. Bertarelli, S. Marchetti, M. Scannapieco, N. Salvati, and D. Summa (2022). Inference for big data assisted by small area methods: an application to OBEC (on-line based enterprise characteristics). In A. Balzanella, M. Bini, , C. Cavicchia, and R. Verde (Eds.), *Book of short papers SIS 2022*, pp. 305–311. Pearson.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, J. N. and I. Molina (2015). *Small area estimation*. New Yok: John Wiley & Sons.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society Series A: Statistics in Society 180*(4), 1163–1190.

Yang, S. and J. K. Kim (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 1–26.

Ybarra, L. M. and S. L. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika 95*(4), 919–931.