Noname manuscript No.
(will be inserted by the editor)

The use of Twitter data to improve small area estimates of households' share of food consumption expenditure in Italy

Stefano Marchetti · Caterina Giusti · Monica Pratesi

Received: date / Accepted: date

Abstract The use of big data in many socio-economic studies has received a growing interest in the last few years. In this work we use emotional data coming from Twitter as auxiliary variable in a small area model to estimate Italian households' share of food consumption expenditure (the proportion of food consumption expenditure on the total consumption expenditure) at provincial level. We show that the use of Twitter data has a potential in predicting our target variable. Moreover, the use of these data as auxiliary variable in the small area working model reduces the estimated mean squared error in comparison with what obtained by the same working model without the Twitter data.

Keywords Big Data · Area level model · Emotional data

Abstract Die Nutzung von Big Data in vielen sozio-konomischen Studien hat in den vergangenen Jahren zunehmend Interesse geweckt. Im vorliegenden Beitrag verwenden wir aus Twitter stammende Emotionsdaten als Hilfsvariable in einem Small Area-Modell zur Schtzung des proportionalen Anteils der Nahrungsmittelausgaben an den gesamten Konsumausgaben italienischer Haushalte auf Provinzebene. Wir zeigen, dass die Verwendung von Twitter-Daten zu besseren Schtzungen der Zielvariablen beitragen kann. Zudem reduziert die Nutzung dieser Daten als Hilfsvariable im verwendeten Small Area-Modell die geschtzte mittlere quadratische Abweichung im Vergleich zum verwendeten Modell ohne Twitter-Daten.

Keywords Big Data · Area-level-Modelle · Emotionsdaten

S. Marchetti
University of Pisa, Italy
Tel.: +39-0502216320
E-mail: stefano.marchetti@unipi.it

C. Giusti
University of Pisa, Italy

M. Pratesi
University of Pisa, Italy

1 Introduction

In the last years an increasing number of researchers and analysts around the world have investigated the value of using the so-called big data in socio-economic studies. As big data we refer to the huge amounts of digital information about human activities produced by a wide range of high-throughput tools and technologies. Indeed, GPS data, calls from mobile phones, internet searches and social networking are nowadays suggesting new approaches to conduct socio-economic studies. The main advantage of these data is their availability at an unprecedented spatial and temporal detail, which may enable to use them to infer some relevant socio-economic characteristics for entire nations as well as for microregions composed of just a few households (Blumenstock et al, 2015).

Recent studies have shown the value of mobile phone data to tackle problems related to economic development and humanitarian action. For example, Eagle et al (2010) showed that the diversity of individuals relationships - as measured by the entropy of mobile phones' calls - is strongly correlated with the economic development of communities. Blumenstock et al (2015) analyzed the power of anonymized data from mobile phone networks to predict the poverty and wealth of individual subscribers, as well as to create high-resolution maps of the geographic distribution of wealth in Africa. Decuyper et al (2014) assessed the suitability of indicators derived from mobile phone data as a proxy for food security indicators.

In resource-constrained environments where censuses and household surveys are rare, the approach used in the previous studies 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. Marchetti et al (2015) suggested three approaches to use big data in sinergy with small area estimation methods. These methods are currently used by many researchers to produce estimates of several socio-economic target indicators - for example, poverty indicators - for unplanned domains such as Provinces and Municipalities in Italy (LAU 1 and 2 in Eurostat nomenclature), as their knowledge can help in planning local policies and distributing welfare resources. Another approach to use big data in small area estimation was suggested by Porter et al (2014).

In this paper we focus on the use of data coming from the social network Twitter to investigate their potential in predicting the share of food consumption expenditure of Italian households at local level, following the second approach on the use of big data presented in Marchetti et al (2015). We show here that the iHappy indicator derived from Twitter has a good predictive power for the share of expenditure that Italian households devote to the consumption of food and beverages - an indicator that can be used as a proxy to measure households' living conditions, as we better explain in the next section.

The paper has the following structure: the description of the data used in the analysis is in section 2; the small area estimation model is presented in section 3; the results of the application are detailed in section 4. Finally, we draw some concluding remarks in section 5.

2 Description of the data

The Household Consumption Expenditure represents a crucial measure for assessing households' living conditions both at national or at more detailed geographical level (Marchetti and Secondi, 2016). The primary source of data on households' expenditure in Italy is the Household Budget Survey (HBS) carried out annually by ISTAT. In 2012 the sample of the HBS was composed by approximately 28000 households. Data were collected on the basis of a two-stage sample design where the first stage were the municipalities (476 out of approximately 8000 in 2012) and the second stage were the households. The Regions (NUTS 2 level according to Eurostat) are the finest geographical level for which direct estimates of the target indicators are reliable. However, the knowledge of measures able to assess households' living conditions and well-being at a more detailed geographical level is often crucial, since this knowledge can for example enable policy makers in planning local polices aiming at reducing poverty and social exclusion (Giusti et al, 2016).

Using small area methodologies, Marchetti and Secondi (2016) produced reliable estimates of the monthly equivalised[1] Household Consumption Expenditure in Italian provinces in 2012, also taking into account the local differences in the prices by using Purchasing Power Parities (PPPs).

Using the same data, the households' consumption expenditure can be classified into food (and beverages) and non food expenditure. The share of total expenditure that an household dedicate to food items is an important indicator of the household living conditions: at risk of poverty households usually spend an higher share of their total expenditure on food with respect to the other households, with a lower impact of the share of expenditure dedicated to other resources and commodities (Lechene, 2000; Regmi et al, 2001; Deaton, 2003; Meyer and Sullivan, 2003; Barigozzi et al, 2009). As we are interested in the estimation of the share of food consumption at a local level, we resort to the small area methodologies.

Small area methods require auxiliary information able to predict the response variable, as we better explain in section 3. In our case we need auxiliary variables able to predict households' share of food consumption expenditure in 2012 for the 110 Italian provinces. As possible sources of auxiliary variables we use data coming from the Population and Housing Census 2011 and from the Survey[2] on Social Actions and Services on Single and Associates Municipalities 2012.

From the Population Census we collected information at provincial level such as the number of households, the average households' size, the tenure status, the female-headed households quota. As the target variable of our analysis can be considered as a proxy of the households' living conditions, we also considered as valuable source of auxiliary information the expenditure that Italian municipalities made in 2012 for

[1] The equivalence scale is the Carbonaro scale used by ISTAT, according to which the expenditure of a family is divided by a specific coefficient depending on the household size (for example equal to 0.66 for a household with 1 member, 1.33 for a household with 3 members and up to 2.40 for a household with 7 members or more). In this way the expenditures of households of any size can be directly compared with those of households composed by two members.

[2] This survey is a census survey, although some nonresponses can occur. Here we ignore the nonresponses and we use these data as census data.

interventions of social protection. These interventions includes the costs information on local welfare policies, such as services, benefits and transfers directed to households with children, old-age persons, poor and social excluded persons, immigrants.

Besides these sources of official statistics, we also considered as a potential source of auxiliary information big data from Twitter, following the second approach for the joint use of big data and small area methods presented in Marchetti et al (2015). In particular, we considered here as potential covariate for our small area working models the iHappy indicator referring to the year 2012. The iHappy indicator is made available every year since 2012 for all the 110 Italian provinces on the Opinion Analytics platform *Voices from the Blogs*. 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. The overall average of the iHappy indicator in 2012 was equal to 44.5%, with a minimum value of 35.1% for Oristano and a maximum value of 56.6% for Sassari, both provinces of the Sardinia region. Indeed, the spatial variability of the iHappy values was rather high, as it is evident from the "emotional map" of Figure 3b.

Curini et al (2015) also performed some econometric analysis to investigate the determinants of Italians' happiness, as measured by the iHappy indicator - using available data. They considered some static variables such as the overall quality of institutions, that seemed to matter only marginally in affecting the average level of happiness of the Italian provinces. On the contrary, meteorological variables and events related to specific days, such as the variability of the spread between German and Italian Bonds or the payday, resulted to have the largest impact. Thus, in a certain sense the iHappy indicator can be considered as a "thermometer of emotional mood" of Italian Provinces.

In the present work we are not interested in an econometric approach to find out the factors influencing the target variable at local level. We focus instead on the predictive power that the iHappy indicator may have at provincial level as covariate in an operational model to estimate the outcome. Thus, our research question is: can the iHappy indicator be a good covariate in a small area model for the estimation of the share of food consumption expenditure of Italian households? In the next sections we present the methodologies and the results of our analyses aiming at investigating this question.

3 The Fay-Herriot model for small area estimation

Data obtained from surveys are often used to estimate characteristics for subsets of the survey population. If the sample from a subset is small, then a traditional design-based survey estimator can have unacceptably large variance. These subsets has been defined as *small areas* (Rao and Molina, 2015).

In literature a wide range of methods have been used for obtaining reliable small-area estimates (Pfeffermann, 2013; Rao and Molina, 2015), mostly model-based estimators, which can be classified into area- and unit-level models.

Model-based estimators are derived from small area models linking the study variable and the auxiliary information. These models are often based on random area-specific effects that account for between area variation beyond that explained by auxiliary variables included in the model. Area- (or aggregate-) level models relate small area direct estimates to area-specific auxiliary variables, unit-level models relate the unit values of a study variable to unit-specific auxiliary variables. Both area- and unit-level model-based estimators have been extended to account for time-series and spatial information (Rao and Molina, 2015). In the family of unit-level models an alternative (frequentist) approach to the traditional link model – based on random area-specific effects – has been proposed in literature by Chambers and Tzavidis (2006) and Tzavidis et al (2010), and it is based on M-quantile models. Applications of this last approach can be found in Giusti et al (2012) and Pratesi et al (2012). For a review of the small area methods applied to the analysis of poverty data see also Pratesi (2016).

In this study the available data allow us to rely only on area-level models. In addition, we do not have time-series data and the spatial correlation of the target direct estimates is low. So our choice falls on the Fay and Herriot (1979) estimator (FH). In what follows a short description of the method is given.

Let us assume that there are m small areas of interest and that θ_i represents the target parameter of the area i, such as a mean, a proportion or a percentile. A survey provides a direct estimator $\hat{\theta}_i^{dir}$ of θ_i for some or all of the small areas. As usual, we assume that under the sampling design $E[\hat{\theta}_i^{dir}] = \theta_i$. A p-vector $\mathbf{X}_i$ contains the auxiliary data sources of population characteristics for area i.

Let us assume that the auxiliary variables $\mathbf{X}_i$ are known exactly. The FH model is as follows

$$\hat{\theta}_i^{dir} = \mathbf{X}_i^T \beta + u_i + e_i \quad i = 1, \ldots, m, \tag{1}$$

where $u_i \overset{iid}{\sim} N(0, \sigma_u^2), i = 1, \ldots, m$ are the model errors and $e_i \overset{ind}{\sim} N(0, \psi_i^2), i = 1 \ldots, m$ are the design errors, with e_i independent from u_j for all i and j. It is assumed that the quantity of interest in area i is $\theta_i = \mathbf{X}_i^T \beta + u_i$.

Under the assumption of normality of both the errors (model and sampling design), the best linear unbiased predictor of θ_i is

$$\tilde{\theta}_i^{FH} = \gamma_i \hat{\theta}_i^{dir} + (1 - \gamma_i) \mathbf{X}_i^T \tilde{\beta}, \quad \gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \psi_i^2}, \tag{2}$$

where $\tilde{\beta}$ is the Best Linear Unbiased Estimator of β. The predictor $\tilde{\theta}_i^{FH}$ is a convex combination of the direct estimator $\hat{\theta}_i^{dir}$ and of the predicted value $\mathbf{X}_i^T \tilde{\beta}$ from the regression model. The extent to which it depends on the the direct estimator or on the predicted value for the area is determined by γ_i and hence by the relative sizes of the model error variance σ_u^2 and the sampling error variance ψ_i^2.

According to the theory of small area estimation (Rao and Molina, 2015), the parameters β and σ_u^2 are unknown and must be estimated, while ψ_i^2 is assumed to be

known. The estimators of the ψ_i^2s are often smoothed, and the smoothed estimators are treated as if they were the true sampling variances (Datta et al, 2005).

Estimators of β and σ_u^2 can be obtained using the restricted maximum likelihood from the marginal distribution $\hat{\theta}_i^{dir} \sim N(\mathbf{X}_i^T\beta, \sigma_u^2 + \psi_i^2)$ (Rao, 2003, see paragraph 6.2.4 page 100). By plugging in the estimates of β and σ_u^2 into equation (2) we obtain the empirical best linear unbiased predictor

$$\hat{\theta}_i^{FH} = \hat{\gamma}_i\hat{\theta}_i^{dir} + (1-\hat{\gamma}_i)\mathbf{X}_i^T\hat{\beta}, \quad \hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_i^2}. \tag{3}$$

The terms $\hat{\gamma}_i$ are commonly known as *shrinkage* factors.

When all the parameters (σ_u^2, β) are known the mean squared error (MSE) of the estimator (2) is

$$MSE(\tilde{\theta}_i^{FH}) = E[(\tilde{\theta}_i^{FH} - \theta_i)^2] = \gamma_i\psi_i^2 = g_{1i}. \tag{4}$$

When the parameters in (2) are estimated we obtain the estimator (3) that has the following MSE

$$\begin{aligned} MSE(\hat{\theta}_i^{FH}) &= \gamma_i\psi_i^2 + (1-\gamma_i)^2\mathbf{X}_i^T V(\hat{\beta})\mathbf{X}_i + \psi_i^4(\psi_i^2+\sigma_u^2)^{-3}V(\hat{\sigma_u^2}) \\ &= g_{1i} + g_{2i} + g_{3i}, \end{aligned} \tag{5}$$

where g_{2i} is the contribution to the MSE from estimating β and g_{3i} is the contribution to the MSE from estimating σ_u^2. In equation (5) $V(\hat{\beta})$ and $V(\hat{\sigma_u^2})$ are the asymptotic variances of an estimator $\hat{\beta}$ of β and an estimator $\hat{\sigma_u^2}$ of σ_u^2, respectively. An estimator of (5) is as follows

$$mse(\hat{\theta}_i^{FH}) = \hat{g}_{1i} + \hat{g}_{2i} + 2\hat{g}_{3i}, \tag{6}$$

where $\hat{g}_{1i} = \hat{\gamma}_i\psi_i^2$, $\hat{g}_{2i} = (1-\hat{\gamma}_i)^2\mathbf{X}_i^T[\sum_{i=1}^m \mathbf{X}_i\mathbf{X}_i^T/(\psi_i^2+\hat{\sigma}_u^2)]^{-1}\mathbf{X}_i$, $\hat{g}_{3i} = \psi_i^4(\psi_i^2 + \hat{\sigma}_u^2)^{-3}2[\sum_{i=1}^m 1/(\hat{\sigma}_i^2+\psi_i^2)^2]^{-1}$. More details concerning analytic MSE estimation for area level model can be found in Rao and Molina (2015); Datta and Lahiri (2000); Prasad and Rao (1990).

Usually, for $M-m$ out of sample areas, estimates are obtained using a regression-synthetic estimator

$$\hat{\theta}_l^{syn} = \mathbf{X}_l^T\hat{\beta}, \tag{7}$$

where $\mathbf{X}_l$, $l = m+1,\ldots,M$, is the vector of auxiliary variables for the out of sample area l. The MSE of $\hat{\theta}_l^{syn}$ is given by

$$MSE(\hat{\theta}_l^{syn}) = \sigma_u^2 + \mathbf{X}_l^T\Big[\sum_{i=1}^m \mathbf{X}_i\mathbf{X}_i^T/(\psi_i^2+\sigma_u^2)\Big]^{-1}\mathbf{X}_l + o(m^{-1}).$$

A second order unbiased MSE estimator under the REML estimation of σ_u^2 is given by

$$mse(\hat{\theta}_l^{syn}) = \hat{\sigma}_u^2 + \mathbf{X}_l^T\Big[\sum_{i=1}^m \mathbf{X}_i\mathbf{X}_i^T/(\psi_i^2+\hat{\sigma}_u^2)\Big]^{-1}\mathbf{X}_l. \tag{8}$$

4 Estimates of the share of food consumption in the Italian provinces *with* and *without* Twitter data

In this section we show that the use of Twitter data can improve the precision of the Share of Food Consumption Expenditure (SFCE) estimates in the Italian provinces, obtained using small area methods. As discussed in section 2, the HBS is designed to obtain reliable estimates at a regional level in Italy. Direct estimates at a finer geographical level, such as the province level, can have a too large coefficient of variation and can be considered unreliable. Small area estimation methods have been recognized as a cost effective solution to overcome to this problem. As we have already noticed, the availability of auxiliary information at provincial level and not at unit-level (household-level) restricts our choice to the area-level approach. Moreover, the low spatial autocorrelation of the SFCE at provincial level and the absence of time-series data lead us to the use of the FH estimator (3), described in section 3.

First, we estimated the SFCE at provincial level using the FH model (1) selecting the more predictive variables among the data described in section 2 without considering the iHappy variable, the one computed using Twitter 2012 data. In this way we obtained a reduction in MSE in all the provinces. Second, we added the iHappy variable to the other auxiliary variables and we estimated the SFCE again. If the iHappy variable is linearly correlated with the SFCE and this relation is not yet explained by the other auxiliary variables, then we expect a better performance in terms of MSE when using iHappy. We will show that the results obtained support this expectation.

Under both models - with and without the iHappy indicator - the target variable, the SFCE, was obtained from the HBS 2012 survey as the ratio between the consumption expenditure for food (including beverages) and the total consumption expenditure. Its direct estimate at provincial level was obtained using the Horvitz and Thompson (1952) expansion estimator,

$$\hat{\theta}_i^{dir} = \frac{\sum_{j=1}^{n_i} y_{ij} w_{ij}}{\sum_{j=1}^{n_i} w_{ij}}, \quad i = 1, \ldots, m, \tag{9}$$

where the survey weights w_{ij}, which are computed by ISTAT, are the inverse of the inclusion probability of household j in area i, calibrated according to known totals in the population and adjusted for the non-response. In (9) y_{ij} is the SFCE for household j in area i. In 2012 the Italian provinces were 110 in total. However, in 2012 no HBS sample data were available for the province of Enna (Sicily) therefore it was not possible to obtain a direct estimate for this province, so we computed a synthetic estimator given that we know the auxiliary data for this province.

The selected auxiliary variables for the model without the iHappy variable are the share of owners of the house x_1, the share of households lead by a female x_2, the per-household local government expenses to support several categories of citizens, households with children (x_3), old-aged persons (x_4), immigrants (x_5), at risk of poverty persons (x_6), services to families[3] (x_7). So let $\mathbf{X}_i = [1, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}]^T$

[3] Intended as households consisting of two or more individuals who are related by birth, marriage or adoption, although they also may include other unrelated people.

be the design p-vector for model (1) for the area i, where x_{ki}, $k = 0,\ldots,p = 7$, $i = 1,\ldots,m$, is the value of the kth auxiliary variable in area i (with $x_{0i} = 1$).

In some recent works (Jiang et al, 2001; Cordero et al, 2016) the authors do not model the raw proportions but an arcsin square-root transformation of the proportions. This transformation is used to stabilize the variance and to guarantee that the predictions fall in the space $[0,1]$. However, there are works where the raw proportions are modeled, for example see Rao and Molina (2015, Example 6.1.4), Hidiroglou et al (2007) and Salvati et al (2014). In our work we chose to model the raw proportions given that the area-level random errors can be considered normally distributed and the estimates are all in the range $0-1$ (and similar to the point estimates).

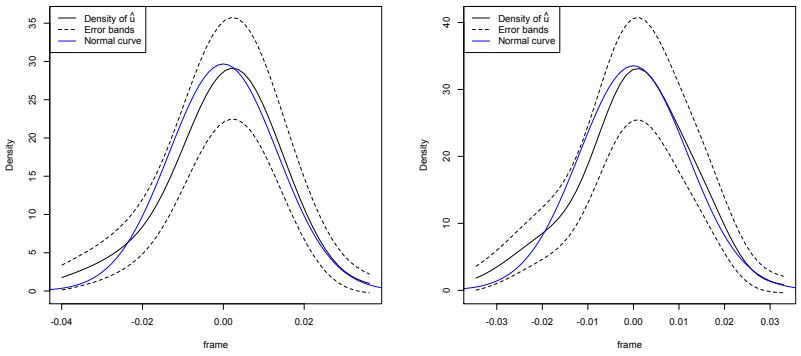
The FH model without the iHappy variable is then $\hat{\theta}_i^{dir} = \mathbf{X}_i^T\beta + u_i + e_i$. Estimates of β and σ_u^2 were obtained under the Normality assumptions made in section 3 using the restricted maximum likelihood (REML), while $\psi_i^2 = \left(\sum_{j=1}^{n_i}(w_{ij}^2 - w_{ij})y_{ij}^2\right)/\left(\sum_{j=1}^{n_i} w_{ij}\right)^2$. From the analysis of $\hat{u}_i = \hat{\gamma}(\hat{\theta}_i^{Dir} - \mathbf{X}^T\hat{\beta})$, the Normality assumption seems reasonable. Indeed, the Shapiro and Wilk (1965) Normality test is equal to 0.978 with a p-value of 0.063. In figure 1a it is represented a non-parametric density estimate with 95% confidence interval bands (obtained according to Bowman et al (1998)) of $\hat{u}_i$ with a superimposed Normal density obtained form the data. The Normal curve falls inside the confidence bands, giving us another evidence that the Normality assumption is reasonable.

To check the hypothesis that big data can help to increase the precision of the small area estimates - if used as auxiliary variables - we added the iHappy variable (x_8), obtained from the analysis of Twitter data as explained in section 2, to the set of the selected auxiliary variables ($x_1, x_2, \ldots, x_7$). Let $\mathbf{Z}_i = [\mathbf{X}_i, x_{8i}]^T$, where x_{8i} is the iHappy value for area i. The FH model is $\hat{\theta}_i^{dir} = \mathbf{Z}_i^T\beta^{BD} + u_i^{BD} + e_i^{BD}$, where the superscript ^{BD} refers to parameters under the model that makes use of big data (the Twitter data). Point and *mse* estimates are then obtained according to the methodology described in section 3 (replacing $\mathbf{X}_i$ by $\mathbf{Z}_i$).

In both the models - with and without iHappy variable - we selected the auxiliary variables using a step-wise procedure based on AIC (Hastie and Pregibon, 1992). The selected variables show a negative linear correlation with the target that range from -0.130 to -0.509 (table 1). The negative correlations were expected for all the variable, but the share of households lead by a female. In general, in Italy, households lead by a female are positively correlated with poverty indexes and deprivation variables. However, we can suppose that the households lead by a female are associated with a reduction of the household size, so the expenses in food and beverages decreases so that to increase the SFCE. This hypothesis is supported by a linear correlation between the share of the households lead by a female and the household size equal to -0.857. As done for the model without iHappy variable, we estimated β^{BD} and σ_u^{BD} under the Normality assumptions made in section 3 using the REML (ψ_is remain unchanged). The Shapiro and Wilk (1965) Normality test for $\hat{u}_i^{BD}$s is equal to 0.980 with a p-value of 0.107. Figure 1b represents a non-parametric density estimate with 95% confidence interval bands of $\hat{u}_i^{BD}$ with a superimposed Normal density estimated from the data. The Normal curve falls inside the confidence bands giving us another evidence that the Normality assumption is reasonably.

Table 1: 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



(a) $\hat{u}_i$ (b) $\hat{u}_i^{BD}$

Fig. 1: Non-parametric density estimate with 95% confidence interval bands of $\hat{u}_i$ (1a) and $\hat{u}_i^{BD}$ (1b) with a superimposed Normal density.

The regression parameters estimated for both the models - with and without iHappy - are showed in table 2. The βs obtained under the two models are similar, the introduction of the iHappy variable in the FH model does not change significantly the model, it just add predictive power to it. The parameter σ_u is estimated equal to 0.020 for the model without iHappy and to 0.019 for the model with iHappy. To verify the null hypothesis that $\sigma_u^2 = 0$, we used the test proposed by Datta et al (2011), with

$$\sum_{i=1}^{m} \psi_i^{-2}(\hat{\theta}_i^{dir} - \mathbf{Z}_i^T \hat{\beta}_{wls})^2 = T \sim \chi^2_{m-p} \text{ under } \mathrm{H}_0\,,$$

where $\hat{\beta}_{wls} = (\sum_{i=1}^m \psi_i^{-2}\mathbf{Z}_i\mathbf{Z}_i^T)^{-1}\sum_{i=1}^m \psi_i^{-2}\mathbf{Z}_i\hat{\theta}_i^{dir}$ is the weighted least square estimator of β under the null hypothesis. If $T \geq \chi^2_{m-p,\alpha}$ - where $\chi^2_{m-p,\alpha}$ is the upper α-point of χ^2_{m-p} - we reject the null hypothesis $\sigma_u^2 = 0$, as is in our application.

It is important to underline that the iHappy indicator is based on self-selected data, the Twitter data. However, in this application we are not able to treat the self-

Table 2: Regression parameters of the FH model with and without the iHappy variable.

	$\hat{\beta}^{BD}$	p-value^{BD}	$\hat{\beta}$	p-value
Intercept	0.7165	0.0000	0.6446	0.0000
iHappy2012	−0.0019	0.0067	–	–
Share of owners of the house	−0.0038	0.0000	−0.0039	0.0000
Share of household lead by female	−0.3164	0.0009	−0.3222	0.0012
Expenses for household with children	−0.0001	0.2121	−0.0002	0.0513
Expenses for old-aged persons	−0.0001	0.0123	−0.0001	0.0280
Expenses for immigrants	−0.0013	0.0003	−0.0013	0.0009
Expenses for at risk of poverty persons	0.0006	0.0009	0.0007	0.0006
Expenses for services to families	−0.0005	0.0460	−0.0006	0.0246

selection bias due to lack of information. Thus, we assume that the self-selection is negligible[4]. Moreover, the iHappy indicator can be affected by measurement error, since not any happy tweet corresponds to a happy person. As it concerns the measurement error, one could apply the area level model proposed by Ybarra and Lohr (2008), which modify the FH model by taking into account the random error in the auxiliary variables - i.e. when the auxiliary variables come from a survey. However, in our application the MSE of the iHappy is small, due to the very large sample size (43 millions of tweets), so that the model proposed by Ybarra and Lohr (2008) approximately corresponds to the traditional FH model.

Results on the SFCE estimates are summarized in table 3. From this table we can see that point estimates are very similar to each other, this is a desired result. The unique exception is for the FH estimates (with and without iHappy variable) in the highest quantiles where there is a known shrinkage effect of the FH estimator. For what concerns the gain in terms of reduction of *rmse* (estimated root mean squared error) the results are as desired. Using the FH estimator (3) with the set of auxiliary variables $\mathbf{X}_i$s the *rmse* is reduced in all the provinces. The reduction of the *rmse* is on average about 30% with a 25% of provinces where the reduction is at least about 40%, as shown in table 3. Moreover, using also the iHappy variable the reduction of the *rmse* goes from about 30% to about 32% with an average gain of 2%. A clearer picture of the gain in precision due to the introduction of the iHappy variable in the FH model can be see in the last line of table 3, which shows the ratio between the *rmse* of the FH estimator that use the iHappy variable ($\hat{\theta}_i^{FH.BD}$) and the FH estimator that does not use the iHappy variable ($\hat{\theta}_i^{FH}$). There is a gain in all the areas, but one where we observe a loss of 0.5%. The gain goes from about 2% up to about 7%. Given that the small area estimates obtained without the use of the iHappy variable show a remarkable gain in terms of reduction of *mse*, the further reduction of the *mse* due to the introduction of the iHappy variable in the model is a very good result. This is particularly important also because the iHappy variable can be computed every year, while updated census information on the population is not always available.

[4]In other social studies content analysis of the texts posted using social network seems to provide acceptable predictions of the behavior of the whole population, not only of those using Twitter. Particularly, the debate is lively on the electoral predictions (Ceron et al, 2015).

Table 3: Summary of point estimates of SFCE for 109 Italian provinces obtained using direct and small area estimators (without iHappy $\hat{\theta}_i^{FH}$ and with iHappy $\hat{\theta}_i^{FH.BD}$), and summary of the ratios between *rmse*s of direct estimates and *rmse*s of small area estimates with and without iHappy auxiliary variable.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{\theta}_i^{Dir}(\%)$	15.38	19.44	21.34	22.45	25.56	35.42
$\hat{\theta}_i^{FH}(\%)$	15.37	19.70	21.60	22.19	24.47	29.91
$\hat{\theta}_i^{FH.BD}(\%)$	15.44	19.68	21.64	22.17	24.65	29.55
$rmse(\hat{\theta}_i^{FH})/rmse(\hat{\theta}_i^{Dir})(\%)$	19.79	60.66	74.90	70.38	82.16	99.39
$rmse(\hat{\theta}_i^{FH.BD})/rmse(\hat{\theta}_i^{Dir})(\%)$	18.44	58.29	72.37	68.49	80.35	99.43
$rmse(\hat{\theta}_i^{FH.BD})/rmse(\hat{\theta}_i^{FH})(\%)$	93.18	95.73	97.33	97.02	98.22	100.50

The reduction of the *mse* obtained introducing the iHappy variable is graphically represented by the plot in figure 2. Here, we contrast the *mse* of the FH estimates obtained with and without the iHappy variable. The gain is present in all the areas, but one.

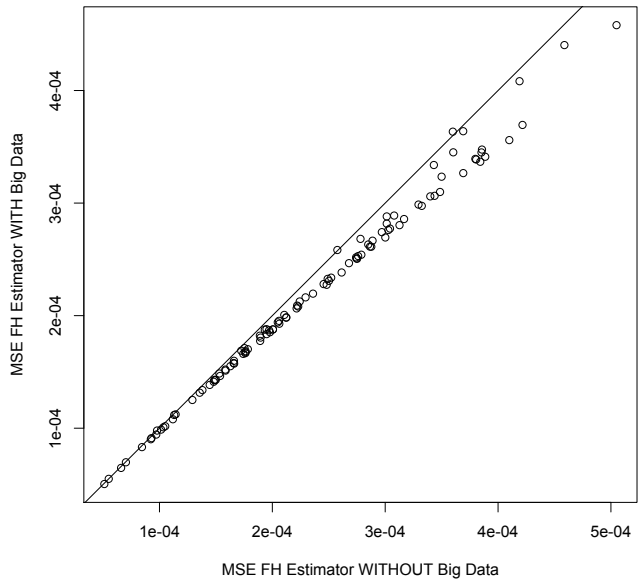
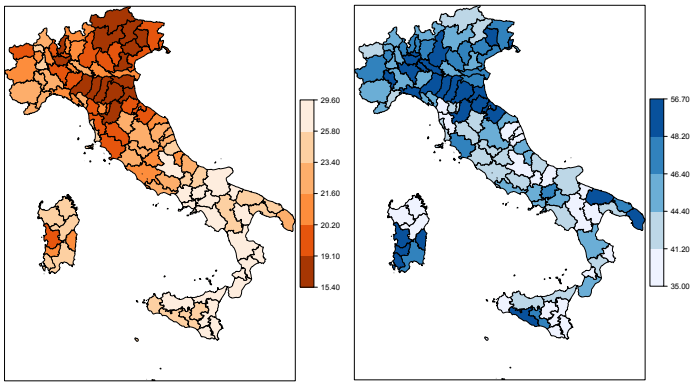


Fig. 2: *mse* of the FH estimates obtained with (*x* axis) and without (*y* axis) the iHappy variable.

In order to obtain a clearer picture of the estimates across the country, we mapped them out in figure 3. In the same figure we contrast our estimates with the map of the iHappy variable to show the relationship between the two variables. The SFCE point estimate for the out of sample province of Enna has been computed using the regression synthetic estimator (7), while its MSE has been obtained using (8). In particular, the estimated SFCE for the province of Enna is 25.29% with an rmse of 1.98%. These results seem plausible according to the estimates obtained for the neighbors provinces.



(a) SFCE estimates (b) iHappy

Fig. 3: Map of the FH estimates of the SFCE (3a) and map of the iHappy variable for 110 provinces in Italy (3b). In both the maps a darker color corresponds to a better situation.

As already discussed in section 2, the SFCE can act as a proxy to measure the living conditions in most of developed and developing countries. In Italy the SFCE is 22.2% at national level, showing that in average the consumption of food does not represent a large amount on total expenses for consumption. At provincial level, as shown in table 3, the SFCE varies between 15.44% (Ravenna, central Italy) and 29.55% (Caserta, southern Italy), so there is evidence of spatial heterogeneity of the indicator. About a quarter of the provinces has an SFCE greater than or equal to 25%. All these provinces are in the southern part of Italy. Nine provinces have a estimated SFCE that is below 18%, five of these provinces are in the central part and the other four are in the northern part of Italy. All the provinces in the lowest quartile are in the central or northern part of Italy. These results confirm the well known Italian north-south divide concerning the socio-economic indicators.

5 Conclusions

In the era of data deluge there are many new sources of data tracking human behavior. In this paper we focused on the iHappy indicator obtained from the analysis of Twitter data. The data consist of all the geo-referenced tweets posted in 2012 in the Italian provinces, classified by Curini et al (2015) as the percentage of happy tweets to the total of tweets at provincial level.

In our analysis the iHappy indicator resulted a good additional covariate to predict households' SFCE, given the net influence of other covariates characterizing the provinces, such as the tenure status of the house, the gender of the head of the households, the level of the expenses of the local government to support vulnerable groups.

In Italy the SFCE shows a territorial variability that mimics that of many socio-economic indicators: in 2014 the north-eastern and north-western part of Italy had the lowest level of SFCE (respectively 15.7% and 15.5%) while the southern part (islands included) had the highest (21%) (see ISTAT (2015)). This north-south divide is evident also from the territorial distribution of the iHappy indicator, with few exceptions (some provinces of Sardinia, Puglia and Sicily). Given that the iHappy indicator can be considered as a "thermometer of emotional mood" of Italian households, it would be interest to test its predictive power to estimate the trend and the changes in households' SFCE: however, at the moment this task is not achievable due to the cross-sectional availability of the HBS data we used to measure the SFCE.

Concluding, the iHappy indicator on happiness can provide useful covariates on yearly bases, free of charge and broken by provinces. It comes affected by self-selection bias and measurement error. In this application we assumed that the self-selection is negligible and that the measurement error appears to be a minor issue.

In the paper we have not considered issues related to Information Communication Technology (ICT) since we used the iHappy indicator as computed by Curini et al (2015). However, the heterogeneity, lack of structure (requiring important work to prepare the data for statistical production), and volume (which hampers the use of standard statistical tools) of the Twitter data are a challenge to exploit all the potentialities they have.

The above mentioned issues can be a limitation for the purpose of adopting Twitter data as current and not episodic sources of auxiliary information in small area methods. Here we limit to remark that: i) some issues share statistical methodological and IT aspects, as those linked to the application of the content analysis to the huge amount of considered tweets; ii) technological issues are important but probably easier to be solved than statistical and IT methodological issues, as some solutions are already on the market.

In few words, what is considered *big* today is going to be considered normal tomorrow. What it is necessary is the development of joint skills to treat the data and to envision their usage in statistical models, as we did in the small area estimation models. Hence, the skills for dealing with statistical models should come from the two worlds, by the data scientist profile.

Acknowledgements The research presented in this paper was developed in the framework of the European Commission FP7 project InGRID (Inclusive GRowth Research Infrastructure Diffusion, www.inclusivegrowth.eu)

References

Barigozzi M, Alessi L, Capasso M, Fagiolo G (2009) The distribution of households consumption-expenditure budget shares. Tech. rep., European Central Bank - Working papers series

Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. Science 350:1073–1076

Bowman A, Hall P, Prvan T (1998) Bandwidth selection for the smoothing of distribution functions. Biometrika 85:799–808

Ceron A, Curini L, Iacus S (2015) Using social media to fore- cast electoral results: A review of state-of-the-art. Italian Journal of Applied Statistics 25(3):237–259

Chambers R, Tzavidis N (2006) M-quantile models for small area estimation. Biometrika 93(2):255–68

Cordero C, Encina J, Lahiri P (2016) Analysis of Poverty Data by Small Area Estimation, Wiley, chap Poverty Mapping for the Chilean Comunas

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

Datta G, Lahiri P (2000) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica 10:613–627

Datta G, Hall P, Mandal A (2011) Model selection and testing for the presence of small area effects, and application to area-level data. Journal of the American Statistical Association 106:362–374

Datta GS, Rao JNK, Smith DD (2005) On measuring the variability of small area estimators under a basic area level model. Biometrika 92(1):183–196, DOI 10.1093/biomet/92.1.183, URL http://biomet.oxfordjournals.org/content/92/1/183.abstract

Deaton A (2003) Household surveys, consumption, and the measurement of poverty. Economic Systems Research 15(2):135–159

Decuyper A, Rutherford A, Wadhwa A, Bauer J, Krings G, Gutierrez T, Blondel V, Luengo-Oroz M (2014) Estimating food consumption and poverty indices with mobile phone data. Tech. rep., UNITED NATIONS GLOBAL PULSE

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

Fay R, Herriot R (1979) Estimation of icome from small places: An application of james-stein procedures to census data. Journal of the American Statistical Association 74:269–77

Giusti C, Marchetti S, Pratesi M, Salvati N (2012) Robust small area estimation and oversampling in the estimation of poverty indicators. Survey Research Methods 6(3):155–163

Giusti C, Masserini L, Pratesi M (2016) Local comparisons of small area estimates of poverty: an application within the tuscany region in italy. Social Indicators Research

Hastie T, Pregibon D (1992) Generalized linear models, Wadsworth and Brooks/Cole, chap 6

Hidiroglou M, Singh A, Hamel M (2007) Some thoughts on small area estimation for the canadian community health survey. Internal statistics canada document, Statistics Canada

Horvitz D, Thompson D (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47:663–85

ISTAT (2015) Statistiche report, la spesa per consumi delle famiglie, anno 2014. roma, 8 luglio 2015. Tech. rep., ISTAT

Jiang J, Lahiri P, Wan S, Wu C (2001) Jackknifing in the fay-herriot model with an example, unpublished manuscript

Lechene V (2000) National Food Survey: 2000, Department for Environment Food and Rural Affairs, London, chap Income and price elasticities of demand for foods consumed in the home

Marchetti S, Secondi L (2016) Estimates of household consumption expenditure at provincial level in italy by using small area estimation methods: Real comparisons using purchasing power parities. Social Indicators Research

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 31:263–281

Meyer B, Sullivan J (2003) Measuring the well-being of the poor using income and consumption. The National Bureau of Economic Research Working Paper

Pfeffermann D (2013) New important developments in small area estimation. Statist Sci 28(1):40–68, DOI 10.1214/12-STS395, URL http://dx.doi.org/10.1214/12-STS395

Porter A, Holan S, Wikle C, Cressie N (2014) Spatial fay-herriot models for small area estimation with functional covariates. Spatial Statistics 10:27–42

Prasad N, Rao J (1990) The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association 85:163–171

Pratesi M (2016) Analysis of Poverty Data by Small Area Estimation. Wiley Series in Survey Methodology, Wiley, URL https://books.google.it/books?id=TS29BgAAQBAJ

Pratesi M, Giusti C, Marchetti S (2012) Small area estimation of poverty indicators. In: Davino C, Fabbris L (eds) Survey data collection and integration, Springer

Rao J (2003) Small Area Estimation. New York:Wiley

Rao J, Molina I (2015) Small Area Estimation. Wiley Series in Survey Methodology, Wiley, URL https://books.google.it/books?id=i1B_BwAAQBAJ

Regmi A, Deepak M, Seale J, Bernstein J (2001) Changing Structure of Global Food Consumption and Trade, USDA (United States Department of Agriculture) - Economic Research Service, chap Cross-Country Analysis of Food Consumption Patterns

Salvati N, Giusti C, Pratesi M (2014) The use of spatial information for the estimation of poverty indicators at the small area level, Routledge

Shapiro S, Wilk M (1965) An analysis of variance test for normality (complete samples). Biometrika 67:215–216

Tzavidis N, Marchetti S, Chambers R (2010) Robust estimation of small area means and quantiles. Australian and New Zeland Journal of Statistics 52(2):167–186

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