DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 263–281, http://dx.doi.org/10.1515/JOS-2015-0017

Small Area Model-Based Estimators Using Big Data Sources

Stefano Marchetti[1], *Caterina Giusti*[1], *Monica Pratesi*[1], *Nicola Salvati*[1], *Fosca Giannotti*[2], *Dino Pedreschi*[3], *Salvatore Rinzivillo*[2], *Luca Pappalardo*[2,3], *and Lorenzo Gabrielli*[2,4]

The timely, accurate monitoring of social indicators, such as poverty or inequality, on a fine-grained spatial and temporal scale is a crucial tool for understanding social phenomena and policymaking, but poses a great challenge to official statistics. This article argues that an interdisciplinary approach, combining the body of statistical research in small area estimation with the body of research in social data mining based on Big Data, can provide novel means to tackle this problem successfully. Big Data derived from the digital crumbs that humans leave behind in their daily activities are in fact providing ever more accurate proxies of social life. Social data mining from these data, coupled with advanced model-based techniques for fine-grained estimates, have the potential to provide a novel microscope through which to view and understand social complexity. This article suggests three ways to use Big Data together with small area estimation techniques, and shows how Big Data has the potential to mirror aspects of well-being and other socioeconomic phenomena.

Key words: Social mining; auxiliary information; poverty measures.

1. Introduction

The huge amounts of digital information about human activities produced by a wide range of high-throughput tools and technologies nowadays offer an objective description of human behaviour. These rich large-scale datasets, often referred to as Big Data, generally cover a large and considerable portion of the population within a territory, often reaching nationwide and even worldwide coverage. Today, statistical methods and social mining from Big Data represent a concrete opportunity to track human behaviour and understand social complexity.

[1] Department of Economics and Management – University of Pisa, Via Ridolfi 10, 56124 Pisa, Italy. Emails: stefano.marchetti@unipi.it, caterina.giusti@unipi.it, m.pratesi@ec.unipi.it and salvati@ec.unipi.it

[2] KDD Lab – ISTI – National Research Council, Via G. Moruzzi 1, 56124 Pisa, Italy. Emails: fosca.giannotti@isti.cnr.it, rinzivillo@isti.cnr.it, lpappalardo@di.unipi.it and lorenzo.gabrielli@isti.cnr.it

[3] Department of Computer Science – University of Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy. Emails: pedre@di.unipi.it and lpappalardo@di.unipi.it

[4] Department of Information Engineering – University of Pisa, Via G. Caruso 16, 56122 Pisa, Italy. Email: lorenzo.gabrielli@isti.cnr.it

Acknowledgments: The authors would like to acknowledge the valuable comments and suggestions of the Editor, the Associate Editor and three referees. These led to a considerable improvement of the article. The authors also would like to thank ISTAT – *Ufficio Territoriale per la Toscana e Umbria* for providing the direct estimates from the EU-SILC 2011 survey, in the framework of the Letter of Intent with the Department of Economics and Management of the University of Pisa. This work is financially supported by the European Project InGRID ''Inclusive Growth Research Infrastructure Development'', funded by the European Commission's 7th Framework Programme (www.inclusivegrowth.be).

This is especially true when the focus is on the spatial distribution of the phenomena under study. In fact, digital tracks such as those from GPS data, calls from mobile phones, Internet searches and networking enable the analysis of the ground truth of individual and collective behaviour at an unprecedented spatial and temporal detail, sometimes in real time (Giannotti et al. 2012). This happens when statistical agencies are not always prepared to provide statistically sound and accurate local estimates and when, at the same time, small area estimation techniques (Rao 2003) are requested by policymakers for more detailed information about the geographic distribution of poverty, inequality and life condition indicators.

There are a number of different initiatives within the official statistics community related to these two issues. Many national statistical agencies and researchers are now developing, evaluating, and implementing poverty estimation and poverty-mapping methodologies, while also promoting Big Data methodologies and best practices. For example, the European Commission has funded projects such as SAMPLE (Small Area Methods for Poverty and Living Condition Estimates) and AMELI (Advanced Methodology for European Laeken Indicators) related to this topic. More recently, the ESS Big Data Action Plan and Roadmap 1.0 (Eurostat 2014) emphasises how Big Data could be of real interest for official statistics.

The perspective is twofold. On one hand, there are many available sources of Big Data that are proxies of social behaviour along various dimensions:

- Social networks, blogs, and web search keywords can trace desires, opinions, and sentiments.
- Emails and phone contacts can trace social relationships.
- Transaction records of our purchases act as a proxy for lifestyle and shopping patterns.
- Records of our mobile phone calls and GPS trajectories can trace individuals' movements.

On the other hand, we must admit that we are just at the beginning of a data revolution and there is still a gap between the Big Data and the big picture (Giannotti et al. 2012). We can identify four main reasons behind this gap:

1. Each Big Data source has different characteristics. Sensor data are fragmented, low level and poorly correlated with key concepts; social data are highly unstructured and rarely accompanied by metadata, hence a quality evaluation of such data is subject to a preliminary metadata enrichment phase.
2. The lack of a "simple" data structure requires a preproduction treatment that is different from the usual one practised in statistical agencies on surveys and/or administrative registers. Data validation and editing may also need methodological developments due to the large volume and high frequency of Big Data.
3. There are many regulatory, business-related, and technological barriers to unleashing the potential of Big Data for social and data mining. These need to be overcome so that all individuals, businesses, and institutions can safely access the knowledge opportunities. The use of Big Data requires an adaptation of/to legislation regarding privacy issues.

4. In order to integrate these new tools within the statistical estimation process of many target parameters, there is the need for specific statistical know-how with strong Information Technology (IT) foundations for handling configuration and maintenance of Big Data repositories. In addition, accessing data and performing analysis require IT applications that are not customarily used by statistical offices, while statisticians obviously find it more convenient to handle data with traditional analytical tools.

In this article, we do not intend to bridge all the gaps outlined in the four groups of issues described above (Pentland 2012). In our opinion they represent a research roadmap and a challenge that will face statisticians and computer scientists over the next years. Here we present a contribution to the use of small area estimation methods combined with Big Data and social mining aimed at improving our ability to measure, monitor, and predict social performance, well-being, deprivation, poverty, exclusion, and inequality on a fine-grained spatial and temporal scale.

We identify three possible approaches to the use of Big Data in the small area estimation framework:

1. Use Big Data sources to create local indicators and compare them to those obtained with small area estimation methods.
2. Use Big Data sources to generate new covariates for small area models.
3. Use survey data to check and remove the self-selection bias of the values of the indicators obtained using Big Data.

In Section 2 we describe the two first approaches in depth, focusing on the study of well-being. In Sections 3 and 4 we present two applications of these approaches using EU-SILC (European Union – Statistics on Income and Living Conditions, European Commission 2015) survey data and Big Data on individuals' mobility in the region of Tuscany in Italy. In Section 5 we address the last approach mentioned above, and conclude with some final remarks on the combined use of Big Data and small area estimation in a statistical framework.

2. The Use of Big Data in Small Area Estimation

Social mining provides analytical methods for understanding human behaviour by means of the automated discovery of patterns from massive records of human activities. Although data mining and statistical learning from traditional databases are relatively mature technologies (Tan et al. 2006; Hastie et al. 2009), the emergence of Big Data on human activities, their networked format, their heterogeneity and semantic richness, their magnitude and their dynamicity pose new exciting scientific challenges (Giannotti et al. 2012).

In this context of understanding social complexity, the connection of social mining with statistical modelling and statistical data collection and analysis is fundamental. Generally, statistical data are collected by means of sample surveys or censuses. Administrative data and registers can also be exploited to produce statistical data. Censuses are complex and expensive to carry out, so sample surveys represent a common way of collecting data. In order to draw inferences on the target population, surveys should be representative of the whole population. However, to measure social complexity with a focus on the

identification and quantification of social exclusion and deprivation, there is a demand for local-level estimates of the most relevant poverty and well-being indicators. Generally the local level (local administrative area, zone of local governance) constitutes a so-called unplanned domain of estimation in sample surveys. Oversampling to increase the sample size in the domains of interest could be a feasible solution for assessing poverty and deprivation at a local level, say at Local Administrative Units levels 1 and 2 (LAU 1 and LAU 2, levels in the Nomenclature of Territorial Units for Statistics used by Eurostat), as is often required by policymakers. However, the high cost in terms of time and financial resources makes this approach impractical for obtaining accurate estimates. Big Data can represent an alternative source of data for the same areas, usually reaching a very high level of geographical detail. 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 this article we choose to follow the first perspective. A short discussion of the two alternative points of view will be given in Section 4.

The first opportunity to reconcile data from the two independent sources – Big Data and sample surveys – is to use available local measures extrapolated from Big Data to compare and benchmark measures on related aspects of the phenomenon under study (e.g., poverty and social exclusion) obtained from survey data and vice versa. Measures from Big Data sources are usually obtained very quickly; however, they can be affected by a serious self-selection bias. Conversely, small area estimates are methodologically sound, but they require timely survey and population data that can be difficult to obtain. Comparing the two alternative sets of measures referring to the same areas can provide useful insights on the potential of Big Data to benchmark small area estimates. If there is accordance between Big Data and survey data in a given small domain/area with respect to the recorded level of deprivation and poverty, then analysts and policy makers may rely on a strong evidence. Otherwise, if there is a discrepancy between the results obtained from the two sources of data, then there is a need for further investigation of those domains/areas. This is the rationale underlying the application we show in Section 3, where Big Data on individuals' mobility are compared with small area poverty estimates computed using EU-SILC data.

Alternatively, a second possibility is to use Big Data directly as a covariate in a small area model. At its heart, poverty mapping is about combining survey data that measures poverty incidence with auxiliary information, spatial or nonspatial, about the population of interest. On the one side we have survey data collected *ad hoc*, such as consumption and income, and on the other side we have auxiliary information that is obtained from other surveys, from population censuses, or from administrative registers. Variables shared by survey and auxiliary information are used together to improve the precision of the small area estimates. Auxiliary information can also consist of georeferenced data about the spatial distribution of these domains and units, obtained via geographic information systems. Attributes derived from spatial information are helpful in the analysis of socioeconomic data, as these often exhibit a spatial structure that corresponds to the definition and the characteristics of the small areas. It is here that Big Data come into play. This huge amount of data and information can be integrated with statistical modelling for small area estimation, extending the type of covariate information used in the small area model.

However, the extension of the covariates to include variables such as social media search loads or remote-sensing images (e.g., in crop-yield surveys, and also in social surveys) or tracking of human mobility opens up difficulties and challenges. Due to technical problems and legal restrictions, it is unfeasible at this stage to have unit-level data that can be linked with administrative archives, census or survey data. To overcome this problem we can use the so-called area-level models, such as the Fay-Herriot model (Fay and Herriot 1979). In this class of models, direct estimates obtained from survey data are modelled with area-level auxiliary variables, that is, a unique variable value for each area. Auxiliary variables can be known either at the population level or at the survey level; in any case, knowledge at unit level is not required, nor are the auxiliary variables required to match the survey variables in definition, as they are under the unit level approach. Thus it is relatively easy to aggregate Big Data in the domains/areas of interest and use them in the Fay-Herriot model. For example, Porter et al. (2014) use Google Trends searches as covariates in a spatial Fay-Herriot model. However, attention should be paid to the fact that under the Fay-Herriot model it is assumed that the auxiliary variables are measured without error, that is, that they are available for all the areas and they come from census or archives covering the entire population of interest. When auxiliary variables come from surveys, they suffer from sampling errors and may also suffer from nonsampling errors, and thus we consider them as measured with error. Generally, auxiliary variables coming from Big Data are not measured on all (or on a big proportion) of the units of the target population, nor are they collected using a random sample. For these reasons we consider that Big Data are subject to measurement error. In Section 4 we present an application where measures derived from Big Data are used as covariates in a Fay-Herriot model to estimate poverty indicators, accounting for the presence of measurement error in the covariates (Ybarra and Lohr 2008).

Finally, Big Data could be used directly to measure poverty and social exclusion, appropriately taking into account the self-selection problem. We envision that survey data could be used to check and remove this bias, provided that unit-level information from Big Data sources will be available. We revisit this problem in the final section of the article.

3. The Use of Big Data to Make Comparisons With Results Obtained With Small Area Estimation Methods

There has been rising interest in research on poverty mapping over the last decade, with the European Union proposing a core of statistical indicators on poverty commonly known as Laeken Indicators. These indicators can be computed for each of the EU Member States using data from sample surveys such as the EU-SILC survey. In particular, the EU-SILC provides information on the household equivalised income for each of the sampled households: this information is fundamental to compute monetary poverty indicators, such as the Head Count Ratio (HCR), for any domain or area of interest.

HCR – also known as the At-Risk-of-Poverty Rate – measures the incidence of poverty. It is a special case of the generalised measures of poverty introduced by Foster et al. (1984), hereafter FGT. Denote by t the poverty line, that is the level of welfare that

defines the state of poverty, by d the domain/area of interest, $d = 1, \ldots, D$, and by α a sensitivity parameter; the class of FGT poverty measures for a given area d is defined as:

$$F_{\alpha,d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left(\frac{t - w_{jd}}{t} \right)^{\alpha} I\left(w_{jd} \leq t\right). \quad (1)$$

Here w_{jd} is a measure of welfare (i.e., income) for unit or household j, N_d is the number of units or households in domain/area d, I is the indicator function that is equal to 1 when $w_{jd} \leq t$, 0 otherwise. When $\alpha = 0$, $F_{\alpha,d}$ is equal to the HCR indicator, which is simply the proportion of units or households in the domain/area with a measure of welfare at or below the poverty line. Since this index is easy and fast to compute and can be interpreted easily, it is widely used in poverty estimation. However, it should always be supplemented by other poverty indicators referring to the same areas, such as the mean of the household equivalised income.

The computation of these indicators using data from the EU-SILC survey results in accurate estimates for planned domains, for example for the administrative areas corresponding to regions in Italy (NUTS level 2). When the interest is in obtaining estimates at a more detailed level than the one foreseen, it can be necessary to resort to small area estimation techniques, since the sample size in many areas may be too small to obtain accurate direct estimates or it can even be equal to zero, making it impossible to compute direct estimates.

The unplanned domains of interest can correspond to administrative areas that represent local levels of government, such as provinces and municipalities, or to areas that can be used to analyse the socioeconomic structure of the territory, such as the Local Labour Systems (LLSs). Under the framework of the SAMPLE project, using unit-level small area models applied to household EU-SILC data and Population Census 2001 data, FGT poverty estimates were produced for provinces, LLSs and municipalities in Italy. The availability of unit-level data coming from both the population census and the EU-SILC survey made flexible small area estimation modelling possible.

In this article, we present the estimates of the HCR and the mean of the household equivalised income for the ten provinces of the Tuscany region, Italy. These estimates were obtained by applying the M-quantile estimators proposed by Tzavidis et al. (2010) and Marchetti et al. (2012) to data from EU-SILC 2008 and the Population Census 2001 (Pratesi et al. 2010). M-quantile models (Chambers and Tzavidis 2006) relax the parametric assumptions of random effects models traditionally used for small area estimation (Rao 2003), which can represent an advantage in many real data applications (Giusti et al. 2012b; Fabrizi et al. 2014). The household-level covariates included in the model for the mean of the household equivalised income – common to the EU-SILC survey and to the population census – are the house-ownership status, the age of the head of the household, the employment status of the head of the household, the gender of the head of the household, the years of education of the head of the household and the household size.

It is important to note that although the 2008 EU-SILC data were collected six years after the census, the 2001–2007 period was one of relatively slow growth and low inflation in Italy, so it is reasonable to assume that there was relatively little change. Moreover,

using an enlarged sample of the EU-SILC 2008 survey for the Province of Pisa, obtained as a side output of the SAMPLE project, Giusti et al. (2012b) showed that M-quantile small area estimates of the HCR and of the household equivalised income were coherent with EU-SILC 2008 direct estimates using the oversample (reliable) estimates.

The use of Big Data as a covariate in the M-quantile unit-level models described above can be considered unfeasible. There are two main reasons for this. First, as stated above, the auxiliary variables used in the small area model should be measured at the unit level and they should be the same as those available from the survey: the problem here is that Big Data with such characteristics are usually unavailable due to confidentiality reasons. Second, even if available, the Big Data cannot be considered to cover all the population of interest, due to the self-selection problem.

For these reasons, Big Data on mobility are used separately in this application to produce a measure of entropy of individuals' movements for the same areas, the provinces of the Tuscany region. Generalising the approach of Eagle et al. (2010), the aim is to study the possible agreement between the level of poverty and the diversity of its inhabitants' mobility in the areas under study, under the first approach to the joint use of small area estimators and Big Data sources mentioned in Sections 1 and 2.

In more detail, we used a large dataset of private vehicles in central Italy, tracked with a GPS device. 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. The GPS traces were collected by OCTO Telematics S.p.a., a company that provides a data collection service for insurance companies. The GPS device is automatically turned on when the car is started, and the global trajectory of a vehicle is formed by the sequence of GPS points that the device transmits every 30 seconds to the server. When the vehicle stops no points are logged or sent. We exploited these stops to split the global trajectory into several sub-trajectories, which corresponded to the single journeys undertaken by a vehicle. Vehicle traces were then mapped on the road network and their position during the stops was associated with the census sectors, provided by the Italian National Institute of Statistics (ISTAT). In this way, each car journey was described by a tuple composed of the timestamp and a pair of coordinates corresponding to the origin and destination of the journey. Table 1 shows an example of the records in the final dataset.

The mobility for a given vehicle ν is given by:

$$M_\nu = -\sum_{l_1=1}^{L}\sum_{l_2=1}^{L} p_\nu(l_1,l_2)\log\left(p_\nu(l_1,l_2)\right), \tag{2}$$

Table 1. Structure of the dataset with time, origin and destination of each trip. Source: OCTO Telematics S.p.a.

Timestamp	Origin	Destination	Car id
2011/05/12 at 08:31:20	Florence_01	Florence_423	00001
2011/05/24 at 17:53:08	Pisa_231	Prato_23	00003
. . .	. . .	. . .	. . .

a measure of entropy where (l_1,l_2) represents a pair of locations, $p_v(l_1,l_2)$ is the probability of observing a movement of vehicle v between the locations l_1 and l_2, and L is the total number of locations. The probability $p_v(l_1,l_2)$ is given by the ratio between the number of trips of v between l_1 and l_2 and the total number of trips of v. When l_1 is equal to l_2, $p_v(l_1,l_2)$ is set to 0. Then, we define the mobility of an area d as:

$$M_d = \frac{1}{V_d} \sum_{v \in d} M_v, \tag{3}$$

where V_d is the number of vehicles resident in area d. A vehicle is considered resident in the area where it most frequently stops during the night. The mobility value tends to zero when the vehicle v visits few distinct locations, showing low mobility diversity. On the other hand, when the mobility measure (2) increases, it means that the vehicle v makes journeys with several locations as destinations. We calculate the standard deviation of the mobility M_d for each area. For a given area d we measure the standard deviation of the mobility by:

$$s_{M_d} = \left\{ \frac{\sum_{v \in d} (M_v - M_d)^2}{V_d - 1} \right\}^{1/2}, \tag{4}$$

where M_v and M_d are defined by (2) and (3).

Figure 1 shows the scatterplot of the HCR values plotted against the s_{M_d} values computed for the ten provinces of the Tuscany region. Their linear correlation coefficient, used as a mere descriptive index, is equal to -0.74. This result suggests that higher levels of heterogeneity of mobility (M_d), expressed by the standard deviation s_{M_d}, are in the provinces where there are lower levels of poverty. In other words, the diversification of mobility within an area with respect to its mean value can be a proxy

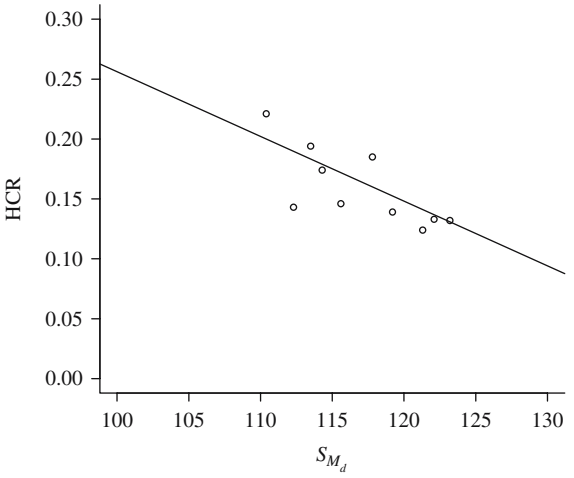

Fig. 1. Scatterplot of the standard deviation of the mobility vs. estimates of the HCR at province level in Tuscany.

of the level of poverty. Conversely, the mean level of the mobility it is not able to discriminate across areas because the values are all very similar and not correlated with the HCRs.

The result of this application is a first interesting example of how to implement the first approach presented in Section 1. However, we think that more evidence is required to confirm the relation between the two indexes. In any case, our aim is to show that Big Data may have the potential to mirror aspects of well-being and other socioeconomic phenomena, supporting the evidence emerging from survey data.

4. The Use of Big Data As Covariates in Area-Level Small Area Models

In this section we present an application of the Fay-Herriot model proposed by Ybarra and Lohr (2008) to estimate poverty indicators for the LLSs of the Tuscany region using Big Data as covariates, under the second approach to the joint use of Big Data and small area estimation mentioned in Sections 1 and 2.

As already discussed in Section 2 and 3, the limitations to the use of Big Data as covariates in unit-level small area models can be overcome using area-level small area models. Fay and Herriot (1979) proposed a model that can be used to reduce the variability of small area direct estimators based on survey data by using auxiliary information coming from other data sources. For example, the Fay-Herriot model and its spatial extension have been used to produce small area estimates of poverty indicators such as the HCR and the mean household equivalised income using EU-SILC income data and auxiliary data coming from the Population Census (Salvati et al. 2014).

It is important to underline that the properties of the small area estimators derived under the Fay-Herriot model are based on the hypothesis that the auxiliary data are available for all the areas and that they are measured without error. Thus, when (recent) Census data are not available and covariate information comes from alternative data sources – such as surveys, administrative data or Big Data – one should take into account that these kinds of data can suffer from sampling and nonsampling errors that could seriously affect the produced small area estimates.

In this section we present an application where the Fay-Herriot model is used to produce estimates of the HCR and of the mean household equivalised income for the LLSs of the Tuscany region using area-level data from the EU-SILC survey 2011 and, as covariate information, data from the EU-SILC survey itself and from Big Data on mobility. To use these data as covariate information, we propose to use the modified version of the Fay-Herriot model proposed by Ybarra and Lohr (2008) to allow for measurement error in the auxiliary variables. An alternative approach can be based on the Bayesian framework, but it is not explored here (see Ghosh et al. 2006; Torabi et al. 2009).

Ybarra and Lohr (2008) assume that for a direct estimator y_d of the target variable Y_d in area d, under the sampling design, $E[y_d] = Y_d$, and that the auxiliary data source provides an estimator $\hat{X}_d$ of a p-vector X_d of population characteristics, where the estimator $\hat{X}_d$ has mean squared error $MES(\hat{X}_d) = C_d$ under the sample design. They show that when the auxiliary variables are measured with error, the traditional Fay-Herriot estimator can be

worse than the direct estimator in terms of precision and in addition the estimated mean squared error gives a misleading notion of precision.

Suppose that X_d is the true value of the auxiliary variable in small area d available for small area estimation. Since X_d may be measured with error, we substitute an estimator $\hat{X}_d$ for X_d and use the following model:

$$y_d = \hat{X}_d^T \beta + r_d(\hat{X}_d, X_d) + e_d \tag{5}$$

where $r_d(\hat{X}_d, X_d) = u_d + (X_d - \hat{X}_d)^T \beta$, with $u_d \sim N(0, \sigma_u^2)$ and the random error $e_d \sim N(0, \psi_d^2)$, with known ψ_d. Here, u_d is independent from both e_d and $\hat{X}_d$, and random variables in different small areas are independent. They also assume that $\hat{X}_d$ and y_d are independent for each area, as when X_d and Y_d are estimated using different data sources. In our application this is the case for Big Data auxiliary variables, while for auxiliary variables from the EU-SILC survey this hypothesis is violated. However, this problem can be solved changing the model according to Ybarra (2003).

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

$$\hat{Y}_{dME} = \hat{\gamma}_d y_d + (1 - \hat{\gamma}_d) \hat{X}_d^T \hat{\beta} \tag{6}$$

where $\hat{\gamma}_d = (\hat{\sigma}_u^2 + \hat{\beta}^T C_d \hat{\beta}) / (\hat{\sigma}_u^2 + \hat{\beta}^T C_d \hat{\beta} + \psi_d^2)$ and the regression vector β and the variance component σ_u^2 are estimated according to an iterative procedure for the modified least squares as in Cheng and Van Ness (1999). Ybarra and Lohr (2008) prove the consistency of (6) and propose an analytic and a jackknife estimator of $MSE(\hat{Y}_{dME})$.

We now present the results of our application, where we are interested in producing mean estimates of the household equivalised income – equivalised according to the OECD (Organisation for Economic Co-operation and Development) modified scale (Hagenaars et al. 1994) – and estimates of the HCR for the 57 LLSs in Tuscany. Note that 24 out of the 57 LLSs are "out-of-sample areas" with a zero sample size in the EU-SILC 2011. Local Labour Systems are the areas in which most of the daily activity of the people who live and work in them takes place; their definition is similar to that of the travel-to-work-areas (TTWAs) widely used in US and UK territorial analyses (ISTAT 1997). According to the official EU nomenclature of local units they are intermediate between levels LAU 1 and LAU 2.

To compute the direct estimates of the mean household incomes and of the HCRs we used data from the 2011 census wave of EU-SILC, available at LLS level. Data from the same survey was also considered as covariate information. We did not consider the Population Census 2001 data here, since there was a structural change in the economic system after the 2008 financial crisis and the use of census 2001 information may thus lead to biased small area estimators.

As covariate information available for all 57 LLSs we also used Big Data on individuals' mobility, under the hypothesis that mobility data could be predictive of well-being measures. More specifically, we used the measure of mobility described in Section 3 and another measure of mobility based on the radius of gyration (RG), which for each vehicle measures how spread out its visited locations are from its centre of mass. The centre of mass $\mathbf{l}_{cm,\nu}$ of a vehicle ν is defined as a two-dimensional vector representing the weighted mean point of the locations visited by that vehicle. We can measure the

mass associated with a location with its visitation frequency, obtaining the following definition:

$$\mathbf{l}_{cm,\nu} = \frac{1}{\sum_{i \in L} \delta_{i,\nu}} \sum_{i \in L} \delta_{i,\nu} \mathbf{l}_i \tag{7}$$

where L is the set of all the visited locations, $\mathbf{l}_i$ is a two-dimensional vector describing the geographic coordinates of location i and $\delta_{i,\nu}$ is its visitation frequency by vehicle ν. Then, the RG of a vehicle ν is defined as:

$$RG_\nu = \left\{ \frac{1}{\sum_{i \in L} \delta_{i,\nu}} \sum_{i \in L} \delta_{i,\nu} (\mathbf{l}_i - \mathbf{l}_{cm,\nu})^T (\mathbf{l}_i - \mathbf{l}_{cm,\nu}) \right\}^{1/2}. \tag{8}$$

We can then define the radius of gyration in area d as:

$$RG_d = \frac{1}{V_d} \sum_{\nu \in d} RG_\nu. \tag{9}$$

The radius of gyration provides a measure of the volume of mobility, indicating the typical distance travelled by a vehicle and provides an estimation of its tendency to move.

We used two different models to estimate the small area income means and HCRs. To estimate the mean incomes (Y_d) using estimator $\hat{Y}_{dME}$, let $\hat{X}_d$ be the vector of the auxiliary variables of area d: it contains a constant term, the direct estimate of the proportion of male as the head of the household, the direct estimate of the mean of the squared metres of the house (both from the EU-SILC 2011 survey) and, finally, the values of the RG_d (from Big Data sources). Let C_d be the corresponding variance-covariance matrix of the auxiliary variables, with the covariances set to zero. Let y_d be the direct estimate in area d of the mean of the household equivalised income and ψ_d its standard deviation. In the model for the HCR the auxiliary variables vector $\hat{X}_d$ contains a constant term, the direct estimate of the mean of the age of the head of household (from EU-SILC 2011) and the mobility index M_d (from Big Data on mobility). Here y_d is the direct estimate of the HCR in area d.

Estimates obtained using data taken from the EU-SILC survey are design unbiased. As variance-covariance matrix C_d we used the estimated variances of the auxiliary variables' mean estimates, setting the covariances equal to zero. As regards Big Data, it can be argued that they come about according to a survey design or not. In the second case, there is no need to make any inference about unobserved population units. The first case – the one we choose here – follows a design perspective, and then there is uncertainty in the data. From this perspective, we consider our Big Data on mobility as collected on a self-selected sample of car journeys. However, as shown by Bethlehem (2002), the bias due to the self-selection process is related to the correlation between the target variable (mobility index) and the response behaviour (having or not having a GPS). Using the results shown in Pappalardo et al. (2013), we argue that this correlation coefficient can be considered very small in this application, and hence the bias due to the self-selection process could be

negligible. In fact, Pappalardo et al. (2013) show that the mobility index measured using the sample of cars with GPS is coherent with the mobility registered for all the vehicles in the municipality of Pisa (data derived from traffic sensors spread around the city). Given this evidence, it seems reasonable to use the hypothesis of independence between the mobility indexes and "having a GPS", so that we can handle these data as if they were a simple random sample from the population of vehicles. The variances in the C_d matrix are then computed using a simple random sample design variance formula (considering negligible the correction term for finite populations).

Irrespective of whether the design perspective is chosen or not, the use of Big Data as auxiliary variables in small area models is motivated by their predictive power, which results in improved efficiency of the small area estimates for sampled and out-of-sample areas.

For these reasons we also computed the small area estimates without the use of Big Data covariates in the model (results are not reported here). In about a half of the 32 sampled areas we found a gain in precision when adding the Big Data covariates. Moreover, the use of Big Data covariates allows us to obtain synthetic estimates for out-of-sample areas, as they are the only auxiliary variables available for the out-of-sample areas.

An important problem in small area estimation is the synthetic prediction for out-of-sample areas: that is, areas where there are no sampled units, even if there are population units with the characteristics of interest in those areas. The conventional approach for estimating a small area characteristic, say the mean, is the synthetic estimation (Rao 2003): $\hat{Y}_{d,OUT} = X^T_{d,OUT}\hat{\beta}$, where $X_{d,OUT}$ is the auxiliary information for the out-of-sample area d and $\hat{\beta}$ is the vector of estimated coefficients under a small area model. In the application presented here the problem is serious, since there are 24 out-of-sample areas (42 percent of the total number of small areas). Moreover, the predictor $\hat{Y}_{d,OUT} = \hat{X}^T_{d,OUT}\hat{\beta}$ according to Equation (1.6) cannot be applied because the EU-SILC auxiliary variables selected in our models are not available for the out-of-sample areas. In contrast, Big Data auxiliary variables are available instead for all the areas. One possible synthetic predictor is $\hat{Y}_{d,OUT} = \hat{\bar{X}}^T\hat{\beta} + \hat{X}_{d,BD}\hat{\beta}_{BD}$, where $\hat{\bar{X}}$ is the matrix of the direct estimators of the EU-SILC auxiliary variables at a regional level, $\hat{X}_{d,BD}$ is the value of the Big Data auxiliary information for area d and finally $\hat{\beta}$ and $\hat{\beta}_{BD}$ are the estimated regression coefficients (see Giusti et al. 2012a for an example). Accordingly, using Big Data it is possible to obtain area-specific synthetic estimates for the out-of-sample areas, taking into account the variability between areas that cannot be specified by only basing predictions on the values of $\hat{\bar{X}}$. This represents one of the major advantages in the use of Big Data sources in small area estimation.

Finally, to estimate the mean squared error of $\hat{Y}_{dME}$ for both sampled and out-of-sample areas we use a parametric bootstrap approach, since the jackknife approach described in Ybarra and Lohr (2008) was too unstable with our data, often producing negative estimates of the mean squared error.

In the parametric bootstrap we first estimated β and σ^2_u, then we parametrically generated the errors $u^*_d \sim N(0, \hat{\sigma}^2_u)$ and $e^*_d \sim N(0, \psi^2_d)$. Using these random errors and the matrix of auxiliary variables, we generated the bootstrap true values $Y^*_d = \hat{X}^T_d\hat{\beta} + u^*_d$ and the bootstrap direct estimates $y^*_d = Y^*_d + e^*_d$. In the next step, we generated a bootstrap matrix of auxiliary variables with errors $\hat{X}^*_d = \hat{X}_d + \epsilon_d$, where $\in_d \sim N_p(0, C_d)$ with N_p

a multivariate normal of dimension p. Using (6) with $\hat{X}_d^*$ and y_d^* we obtained a bootstrap estimate $\hat{Y}_{dME}^*$ of Y_d^*. Repeating this process B times to obtain B values of $\hat{Y}_{dME}^{*b}$ and Y_d^{*b}, $b = 1, \ldots, B$, the bootstrap mean squared error estimator of $\hat{Y}_{dME}$ was

$$mse(\hat{Y}_{dME}) = B^{-1} \sum_{b=1}^{B} (\hat{Y}_d^{*b} - Y^{*b})^2. \tag{10}$$

In the application of this article we have used $B = 500$.

We checked the performance of this bootstrap mean squared error estimator with a small simulation following the setting used in Ybarra and Lohr (2008) in their simulation study. The bootstrap scheme seemed to work properly, showing an expected slight underestimation of the real (i.e., Monte Carlo) mean squared error.

As an alternative to the bootstrap, for the out-of-sample areas we predicted the ψ_d values using a linear model based on the same variables used in the estimation process (Wolter 2007). This method is feasible given that data coming from Big Data sources are available for all the small areas.

Results for the means of both the equivalised income and for the HCR, obtained using (6), are mapped in Figure 2. These estimates referring to the LLSs show intraregional differences that would be lost if the scope of the analysis were to be limited to the regional level.

What is even more important is that for both the target parameters we achieved a remarkable gain in terms of precision with respect to the direct estimates. Even if this gain is marginally overestimated because the bootstrap mean squared error of $\hat{Y}_{dME}$ underestimates the real mean squared error, the gain in precision is evident. Figure 3 shows a comparison of the bootstrap mean squared error estimates of $\hat{Y}_{dME}$ and the mean

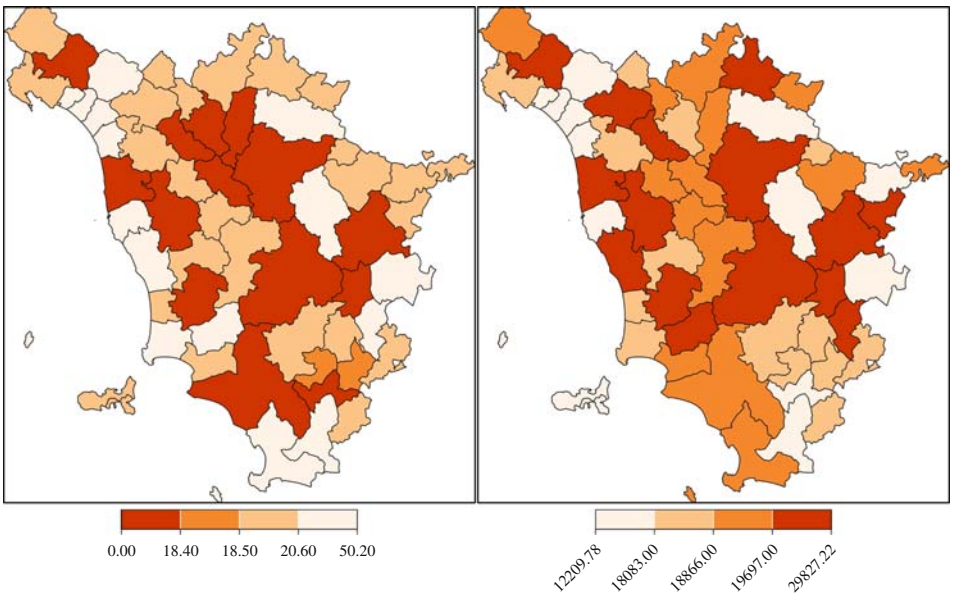

Fig. 2. Estimates of the mean equivalised income in Euros (right) and of the HCR (left) for the Local Labour Systems of Tuscany region. Small area estimates based on EU-SILC 2011 and Mobility Data 2011. Out-of-sample areas are estimated using a synthetic estimator.

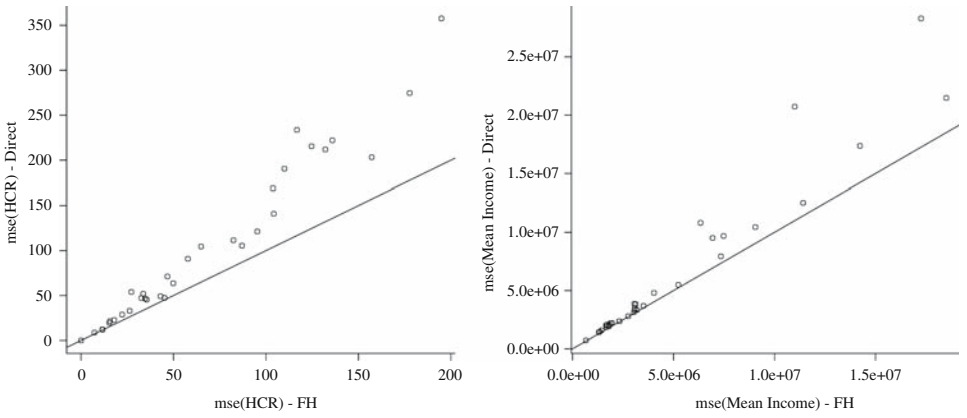

Fig. 3. The plot on the right shows the mean squared error estimates of the small areas, obtained using (10) with $B = 500$ bootstrap replications, vs. the direct estimates of the mean of the equivalised income; the plot on the left shows the same for the HCR estimates. Results are reported for the 33 Local Labour Systems (LLSs) sampled in the EU-SILC 2011 survey for Tuscany.

squared error estimates of the direct estimates y_d for both the mean equivalised income (right) and the HCR (left). Since the mean squared errors for the direct estimators are only available for the sampled areas, we only report these ratios for the 33 sampled areas.

A gain in precision is observed for all the areas. In most of the areas the gain in precision is about 5%–20% for the mean and 10%–40% for the HCR. In some areas the gain is more than 50 percent. However, the mean squared error estimator of the small area estimator $\hat{Y}_{dME}$ should be treated with caution due to its observed underestimation. Nonetheless, these first promising results encourage further research on this topic.

5. Final Remarks

Big Data on a societal scale provides a powerful microscope that can help us understand and forecast many complex socioeconomic phenomena, from the diffusion of information, innovation and crises to the unequal distribution of resources and opportunities. In particular, here we used indicators from Big Data sources in the study of poverty and living conditions and found that they have much to offer when we combine them with small area estimation methods.

Big Data can be much faster in providing auxiliary variables for small area models than official data sources. Even when individual-level data are not available due to privacy restrictions, area-level summaries can provide useful flash covariates for area-level models. Among these, the Fay and Herriot (1979) model, one of the primary tools used in small area estimation, can be used proficiently as a way to borrow strength across locations and thereby reduce the mean squared errors (MSE) of the small area estimates (Rao 2003). When Big Data covariates are used, this model can be generalised to include covariates affected by error of measurement and to obtain an indirect estimate of the small area variable of interest, rather than using a direct survey estimate.

At a time of spending reviews and budget restrictions, auxiliary information that is relatively inexpensive and readily available but is still representative of the population under consideration is of substantial interest. Covariates based on social media, GPS, or other sources (e.g., remotely sensed image data) may augment or replace traditional auxiliary information for a wide variety of poverty and living conditions indicators.

The advantage of these types of covariates is that they are often readily available and provide a considerable amount of relevant information related to a diverse set of demographic and other survey outcomes. The disadvantage is that they may require a more complex small area estimation model than the traditional Fay-Herriot model, since it is not always possible to assume that the available Big Data are known without error.

Big Data could also result in new insights and new proxies of the study variables. Our evidence of correlation between local poverty estimates obtained by surveys and independent estimates from Big Data sources is encouraging. However, the extent to which these advantages occur depends greatly on the specific case in terms of e.g., the data available, the model assumptions made and the precise small area estimation methods used.

In addition, while data quality has been widely studied and discussed and many contributions have been produced in the field of survey estimates, the quality of Big Data sources has not been thoroughly considered, as researchers and statistical agencies have only started to focus on it relatively recently.

In particular, representativeness and coverage of the populations of interest are crucial issues when using Big Data. This is because of the specific nature of the sources, which do not select units according to a sampling procedure and which generally do not cover the whole population. The problem is less relevant when the target population is actively using Information Communication Technologies (ICT) – such as Internet, mobile phones, wireless networks, and other communication mediums and is involved in economic activities traceable by electronic devices, but it is crucial when the objective is to study segments of populations close to social exclusion, such as at-risk-of-poverty individuals and their families.

Nevertheless, whatever the target of the social mining is, the use of Big Data could generate a self-selection problem in relation to the population units, which could have severe effects, leading to biased final estimates. From this point of view, the large amount of information coming from a Big Data source rather than a survey is not enough in itself. A large amount of information is very important, but is not sufficient to produce an assessment of data reliability. There are many similar well-known problems such as census undercoverage, incomplete frames, and informative nonresponse (as in web surveys). These problems have been addressed by many interesting strategies based on weighting procedures and model-based approaches (Bethlehem and Biffignandi 2012). Further methodological and experimental work remains to be done to provide a solution to the problem. We also envision that survey data could be used to check and remove the self-selection bias from estimates obtained using Big Data.

One could try to use a quality survey to check the differences between Big Data and survey data in the distribution of common variables. Alternatively, if no common variables are available, known correlated data could be used. These differences could then be

employed to compute weights that allow the reduction of bias due to the self-selection of the Big Data. The main issue here is the availability of unit-level information from Big Data sources due to confidentiality problems. However, as the use of Big Data will require an adaptation to legislation (and the adaptation of legislation itself) regarding use aspects (i.e., with respect to the access and use of data) and privacy aspects (i.e., managing public trust and acceptance of data reuse and its links to other sources, security of private data) we think that this difficulty will be overcome.

Limiting our focus to poverty studies and poverty mapping, on the one hand Big Data represents an incredible and huge source of data on social complexity and human behaviour; however, it does not ensure a representation of the population of interest's social phenomena. On the other hand, survey data are high-quality sources of data representative of the population of interest, but they are expensive in terms of money and time if they are to be collected properly. Interaction between and integration of these two sources of data is an important challenge for research in statistics and informatics. It is also extremely important that sound and effective statistical methodology be developed to accommodate this abundantly rich class of Big Data resources (Horrigan 2013).

Provided that statistics and social mining develop the ability to glean knowledge from these data, we are of the opinion that scientific research will be revolutionised by this new wave. Furthermore, policy making is going to have new evidence, because Big Data and social mining are providing statistical agencies with novel means for measuring and monitoring well-being in our societies more precisely, continuously, and ubiquitously. Poverty and inequality remain at the top of the global economic agenda, and the methodology for measuring poverty continues to be a key area of research. Measures of poverty and inequality are most useful to policy makers and researchers when they are finely disaggregated into small geographic units. Thus using Big Data together with small area estimation techniques could provide very useful insights into socioeconomic phenomena. Indeed, the Big Data source is, by its very nature, a candidate to become one of the pillars of a statistical system that produces data using social network measures, as proxies of socioeconomic indicators of poverty, well-being and progress.

We have not considered ICT-related issues here, which are also important. 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 Big Data will be a challenge for the statistical agencies of the future.

The risk is that the level of analysis is dictated by the available – appealing and continuous – information but, in any case, we must not give up on providing an accurate assessment of the reliability of the statistics derived from that information (Filippucci 2011).

6. References

Bethlehem, J.G. 2002. "Weighting Nonresponse Adjustments Based on Auxiliary Information." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little. New York: John Wiley and Sons.

Bethlehem, J. and S. Biffignandi. 2012. *Handbook of Web Surveys*. Hoboken, NJ: John Wiley and Sons.

Chambers, R.L. and N. Tzavidis. 2006. "M-Quantile Models for Small Area Estimation." *Biometrika* 93: 255–268. Doi: http://dx.doi.org/10.1093/biomet/93.2.255.

Cheng, C.L. and J.W. Van Ness. 1999. *Statistical Regression with Measurement Error*. London: Arnold.

Eagle, N., M. Macy, and R. Claxton. 2010. "Network Diversity and Economic Development." *Science* 328: 1029–1031. Doi: http://dx.doi.org/10.1126/science.1186605.

European Commission. 2015. *EU-SILC USER DATABASE DESCRIPTION Version 2007–1*. Luxembourg: EC. Available at: http://ec.europa.eu/eurostat/web/income-and-living-conditions/methodology/list-variables (accessed April 26, 2015).

Eurostat. 2014. Summary Record of 22nd Meeting of the European Statistical System Committee, Riga, September 26, 2014. Available at: https://www.google.it/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCIQFjAA&url=http%3A%2F%2Fec.europa.eu%2Ftransparency%2Fregcomitology%2Findex.cfm%3Fdo%3DSearch.getPDF%26IU%2BSnl%2FK6tfKhHQT6oxF31qBB7fI4EnisQ1BdEUO8vC5SVAw47eF02NzJJLXFBE7MymAolL%2BDBgWkUQAUSR0vEUBA1Uxa7mJl1GidS%2BHNzw%3D&ei=5OE8VYKOLozfU9nNgtAH&usg=AFQjCNFEydu1g4aGiE_rpFJfOBD4EnRW9Q&sig2=qEgQ4yw9epL7R7eVYmTmQA&bvm=bv.91665533,d.d24 (accessed April 26, 2015).

Fabrizi, E., C. Giusti, N. Salvati, and N. Tzavidis. 2014. "Mapping Average Equivalized Income Using Robust Small Area Methods." *Papers in Regional Science* 93: 685–701. Available at: http://onlinelibrary.wiley.com/doi/10.1111/pirs.12015/abstract (accessed April 2015).

Fay, R. and R. Herriot. 1979. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74: 269–277. DOI: http://dx.doi.org/10.1080/01621459.1979.10482505.

Filippucci, C. 2011. "Statistical Sources and Statistical Systems in the Information Society." *Statistica* 71: 189–211.

Foster, J., J. Greer, and E. Thorbecke. 1984. "A Class of Decomposable Poverty Measures." *Econometrica* 52: 761–766.

Ghosh, M., K. Sinha, and D. Kim. 2006. "Empirical and Hierarchical Bayesian Estimation in Finite Population Sampling Under Structural Measurement Error Models." *Scandinavian Journal of Statistics* 33: 591–608.

Giannotti, F., D. Pedreschi, A. Pentland, P. Lukowicz, D. Kossmann, J. Crowley, and D. Helbing. 2012. "A Planetary Nervous System for Social Mining and Collective Awareness." *European Physics Journal – Special Topics* 214: 49–75. Doi: http://dx.doi.org/10.1140/epjst/e2012-01688-9.

Giusti, C., S. Marchetti, M. Pratesi, and N. Salvati. 2012a. "Semiparametric Fay-Herriot Model Using Penalized Splines." *Journal of the Indian Society of Agricultural Statistics* 66: 1–14.

Giusti, C., S. Marchetti, M. Pratesi, and N. Salvati. 2012b. "Robust Small Area Estimation and Oversampling in the Estimation of Poverty Indicators." *Survey Research Methods* 6: 155–163.

Hagenaars, A.J.M., K. de Vos, and M.A. Zaidi. 1994. *Poverty Statistics in the Late 1980s: Research Based on Micro-data*. Luxembourg: Eurostat.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd ed. New York: Springer.

Horrigan, M.W. 2013. "Big Data: A Perspective From the BLS." *Amstat News January 2013*: 25–27. Available at: http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/ (accessed April 26, 2015).

ISTAT 1997. *I Sisitemi Locali del Lavoro*. Rome: ISTAT. Available at: http://www.istat.it/it/strumenti/territorio-e-cartografia/sistemi-locali-del-lavoro (accessed April 26, 2015).

Marchetti, S., N. Tzavidis, and M. Pratesi. 2012. "Non-Parametric Bootstrap Mean Squared Error Estimation for M-Quantile Estimators of Small Area Averages, Quantiles and Poverty Indicators." *Computational Statistics and Data Analysis* 56: 2889–2902. Doi: http://dx.doi.org/10.1016/j.csda.2012.01.023.

Pappalardo, L., S. Rinzivillo, Z. Qu, D. Pedreschi, and F. Giannotti. 2013. "Understanding the Patterns of Car Travel." *The European Physical Journal – Special Topics* 215: 61–73. Doi: http://dx.doi.org/10.1140/epjst/e2013-01715-5.

Pentland, A. 2012. "Society's Nervous System: Building Effective Government, Energy, and Public Health Systems." *Computer* 45: 31–38.

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. Doi: http://dx.doi.org/10.1016/j.spasta.2014.07.001.

Pratesi, M., C. Giusti, S. Marchetti, N. Salvati, N. Tzavidis, I. Molina, M. Durban, A. Grané, J.M. Marìn, M.H. Veiga, D. Morales, M.D. Esteban, A. Sanchez, L. Santamaria, Y. Marhuenda, A. Perez, M. Pagliarella, C. Ferretti, and J.N.K. Rao. 2010. *SAMPLE Project – Pilot Application*. Brussels: European Commission – Directorate General for Research and Innovation. Available at: http://www.sample-project.eu/SAMPLEwp2d17.pdf (accessed April 26, 2015).

Rao, J.N.K. 2003. *Small Area Estimation*. New York: John Wiley and Sons.

Salvati, N., C. Giusti, and M. Pratesi. 2014. "The Use of Spatial Information for the Estimation of Poverty Indicators at the Small Area Level." In *Poverty and Social Exclusion, New Methods of Analysis*, edited by G. Betti and A. Lemmi. London: Routledge.

Tan, P.N., M. Steinbach, and V. Kumar. 2006. *Introduction to Data Mining*. Boston: Addison-Wesley.

Torabi, M., G.S. Datta, and J.N.K. Rao. 2009. "Empirical Bayes Estimation of Small Area Means under a Nested Error Linear Regression Model with Measurement Errors in the Covariates." *Scandinavian Journal of Statistics* 36: 355–368. Doi: http://dx.doi.org/10.1111/j.1467-9469.2008.00623.x.

Tzavidis, N., S. Marchetti, and R. Chambers. 2010. "Robust Prediction of Small Area Means and Distributions." *Australian and New Zealand Journal of Statistics* 52: 167–186. Doi: http://dx.doi.org/10.1111/j.1467-842X.2010.00572.x.

Wolter, K.M. 2007. *Introduction to Variance Estimation*. New York: Springer.

Ybarra, L.M.R. 2003. *Small Area Estimation Using Data from Multiple Surveys*. Unpublished PhD thesis, Arizona State University.

Ybarra, L.M.R., and S.L. Lohr. 2008. "Small Area Estimation When Auxiliary Information is Measured With Error." *Biometrika* 95: 919–931. Doi: http://dx.doi.org/10.1093/biomet/asn048.

Received July 2013
Revised February 2015
Accepted February 2015