Challenges and Opportunities with Mobile Phone Data in Official Statistics

Fabrizio De Fausti[1], Marcello Savarese[2], Francesco Fabbri[2], MariaRita Spada[2], Roberta Radini[1], Tiziana Tuoto[1] and Luca Valentino[1]

Keywords: mobile phone data, CDR, mobility, population estimates

1. INTRODUCTION

In recent years, the use of new data source for statistical purposes represents a big challenge for official statistics. Great potentialities are recognized to mobile phone data for supplementing most of the statistical output on population estimates and mobility at a very fine scale. The great interest for the use of these data in Official Statistics is proved also from the several Eurostat founded projects on the topics. However, statisticians from national statistical institutes have rarely the opportunity to directly handle raw data from telecommunication providers, in order to directly investigate how raw data can be treated to extract all their potentialities for official statistics production. This work describes the first results of cooperation between Istat, the Italian National Statistical Institute, and Wind Tre, Mobile Phone Operator MPO, discussing opportunities and challenges. Some uses and the applied methodologies are introduced. The mobile phone data seem promising for further developments and innovative solutions for describing complex behavior, not completely caught by other data sources, i.e. administrative data and to explore new digital solutions taking into account always the privacy constrains. This work is also important to encourage the different stakeholders to cooperate defining a new ecosystem useful for different contexts.The usability and potentialities of Mobile Phone Data (MPD) are analyzed with respect to the new census framework, underlining the steps in which MPD may increase the information already available via administrative data and social surveys. In effect, currently, Istat and other National Statistical Institutes (NSIs) is implementing a census transformation program, the new framework provides for leaving the traditional door-to-door decennial census in favor of the combined use of statistical registers based on administrative data and social surveys. Specific aspects, like coverage of sub-population and other information that cannot be derived by administrative data and ongoing social surveys, will be investigated by yearly ad-hoc sample surveys. To this aim, MPD can be used in different ways, both as complementary data source and primary data source, as well as to validate population estimates. In this report, the abovementioned aspects are investigated, even if, firstly, the reliability of MPD is assessed through the comparison with the official estimates.

2. DATA

The cellular radio-communication is constituted by a bi-directional communication. It is essentially composed of the Mobile Station (MS), also called Mobile Phone (MP), and a Mobile Phone network that includes the antennas that capture and transmit the signals, moreover there are other intermediate systems that allow managing the communication. In this work, the MNO and the NSI share some anonymized data collected for billing the phone traffic, the so-called Call Detail Records - CDRs. The shared CDRs report information on the anonymized user ID identifying the device, also called SIM (Subscriber Identity Module), on the type of the event (text message and call) and its

[1] Istat – Italian National Institute of Statistics {defausti, radini, tuoto, luvalent}@istat.it

[2] Wind-Tre S.p.A. Italy {,marcello.savarese, fra.fabbri mariarita.spada}@windtre.it

starting and ending time. In addition, the localization of the antenna capturing the call when it starts and when it ends are provided. It is worth noting that other signals exchanged by the MP devices and the MP network, for instance when the device is not active, can be exploited for statistical purposes.

A crucial point in analysing the spatio-temporal behavior of people with MPD is related to the MP localization, this is essential to define the accuracy of the population estimates and mobility at the finest local area. The collaboration between the provider and the statistical institute has allowed to focus the problems of localization and to implement solutions that combine the official statistics quality requirements and the MNO experts technical–scientific knowledge. In this light, MNO has provided some auxiliary data, such as the Best Service Areas (BSAs), which help in localizing the activities of the MP registered in the CDR.

The BSAs provide a partition of the territory defined by the MPO in order to plan and manage the radio base stations (BSs) of the MP network in the most efficient way (i.e. guaranteeing a suitable quality of service). Actually, the BSAs are defined via models able to predict the coverage of 2G, 3G and 4G technology networks with computationally-efficient optimizers in order to automatically configure large networks and achieve optimal performances in terms of throughput, served users, land use maps, road network maps, and bandwidth re-use. Generally speaking, a BSA represents the area where the signal measurement of a certain antenna sector has the best coverage.

3. METHODS AND MAIN RESULTS

To properly use MPD for official statistical purposes, we firstly investigated the correlation between MPD and official population estimates. CDRs provide information on the activity of MP users at a given date (with detailed time) and a very small spatial scale. However, calls-in and text messages can be used to produce population estimates given some basic assumptions, such as: high level of MP penetration rate, high level of MP coverage over the field; the knowledge of the MP operator market share. One of the highest MP penetration rates in developed countries can be observed in Italy; indeed, the percentage of MP connection per 100 citizens is about 154% in 2016, as well as a high coverage of MP networks over the territory. Moreover, working in strong cooperation with the MP operator ensures to be able to assess the market share at small spatial scale.

To investigate the correlation of MPD with official population figures, we firstly concentrated the analyses on the nighttime population. The approximation of residential population with nighttime mobile phone users has been stated in several works (Ma and Wu 2012, Deville et al. 2014, Douglas et al. 2015). In this work, the residential municipality is assigned to each SIM according to the following procedure: a percentage is assigned to each municipality on the basis of the coverage of the BSA that most frequently registers calls-in and text messages during the nighttime. The nighttime is from 8pm to 7am.

Figure 1 shows a scatter plot of the count of nighttime active SIMs versus the January 2017 residential population estimates for the province of Pisa at municipality level. It shows that there is a reasonable good relationship, approximately linear as depicted by the LOESS regression interpolation, in blue in the graph. In the linear regression model, the correlation coefficient is 0.94, proving the adequacy of the model in predicting residential population via the nighttime mobile phone users. Similar results in terms of high correlation are also obtained when considering logarithmic transformation and whether the extreme value represented by the city of Pisa is excluded from the analysis.

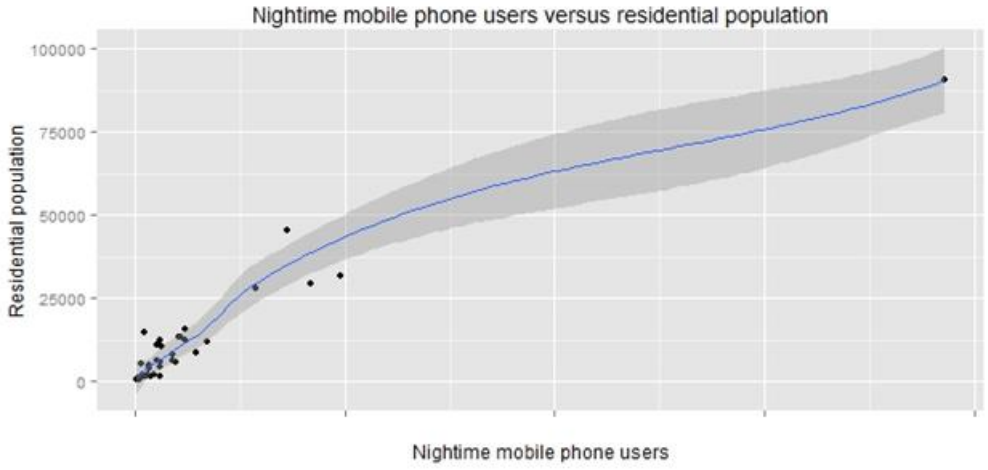

Figure 1 - Night time MP users versus residential population (Pisa province)

The CDRs can be also used to estimate mobility patterns, at anonymized individual-level. In this case, a meaningful positioning for MPD, such as "home" and "work/study", are determined as follows: the "home" is the municipality where a MP user is more frequently located during the nighttime, as abovementioned for the residential population estimates; the "work/study" is the municipality where the MP user is repeatedly observed during the daytime.

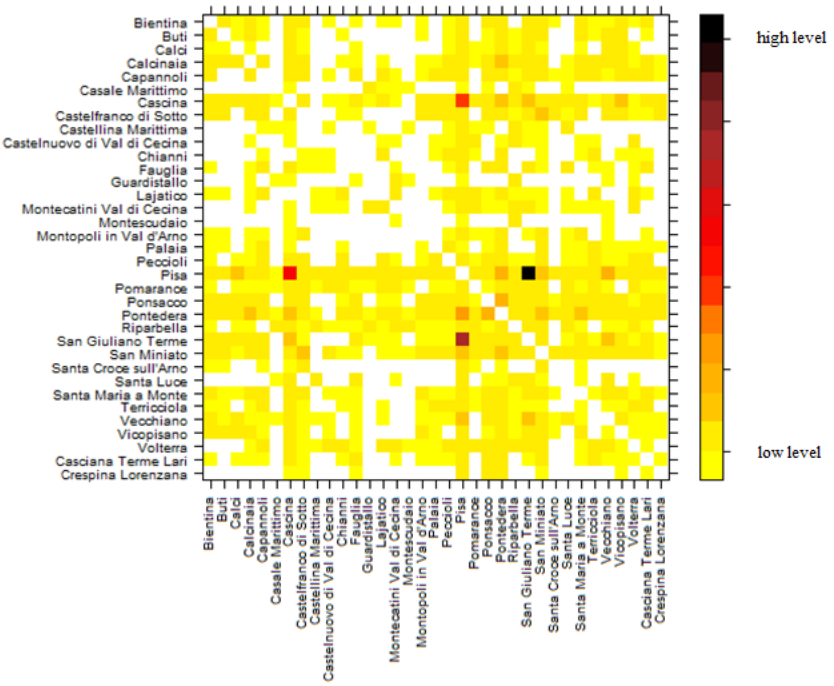

Figure 2 - The Origin-Destination matrix for Pisa province using CDR

By aggregating individual-level data for which home and work/study have previously been derived, it is possible to produce home and work/study origin-destination flows. In Figure 2, we propose an origin-destination matrix for the Pisa province at municipal level, where only movements within the province are taken into account. The main diagonal should represent people who live ("home") and "work/study" in the same municipality. To make the graph analysis clearer, this kind of people are not taken into

account, even if they represent 70% of the analyzed data, the intensity of the movement is represented by the intensity of the colors on the matrix.

4. CONCLUSIONS

The results of the analyses of the CDRs described in this report are definitively encouraging, highlighting the potentialities of MPD both for population estimates and mobility pattern study. In the Census transformation program, the MPD allow us to identify areas that might be problematic for census counts, for instance, areas at risk of over or under coverage can be identified by comparing population estimates from MPD with the counts of people enrolled in registers. The risk of over/under-coverage can be defined at a very small scale and this information can be used both at the sample stage, when designing the coverage sample survey, and at the estimation stage, when small area population estimates have to be provided. Moreover, CDRs provide information with considerable timeliness and they also allow official statisticians to capture new phenomena that currently cannot be observed with sample survey and administrative data, e.g. the presence of migrants or other hidden populations.

A key point for a successful exploitation of CDRs is the small scale localization of the MP users' activities, thanks to the availability of BSAs, in cooperation with the MPO. In the future, we may suppose to produce statistics at a smaller scale than the single municipality, i.e. at census sections, so to fully exploit the huge amount of information that MPD supply us with "urban rhythms" for designing and optimizing mobility in dense urban centers.

(A VERY SHORT LIST OF) REFERENCES

Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel D and Tatem AJ (2014) 'Dynamic population mapping using mobile phone data', PNAS, 111, 45

Douglass RW, Meyer DA, Ram M, Rideout D and Song D (2015) 'High resolution population estimates from telecommunications data' EPJ Data Science

Furletti B, Trasarti R, C Paolo and Gabrielli L (2017) Discovering and Understanding City Events with Big Data: The Case of Rome, Information, 8, 74

Jonge, E. de, Pelt, M. van, Roos, M. (2012) Time patterns, geospatial clustering and mobility statistics based on mobile phone network data, Discussion paper, Statistics Netherlands

Ma X and Wu L (2012) 'Towards Estimating Urban Population Distributions from Mobile Call Data'