# MOBILE PHONE DATA FOR POPULATION ESTIMATES AND FOR MOBILITY AND COMMUTING PATTERN ANALYSES

Fabrizio De Fausti, Roberta Radini, Tiziana Tuoto, Luca Valentino

**Abstract.** The use of mobile phone data (MPD) for statistical production has been widely explored in the past decade. In official statistics, MPD can be used to supplement and enrich the information available through administrative data and social surveys, taking advantage of the richness of MPD both in terms of timeliness and greater spatial availability. The National Statistical Institute (Istat) has been undertaking this investigation some years ago, particularly regarding population density estimation and small-scale mobility analysis.

In this contribution, we analyze the usability and potential of MPD, highlighting the stages at which MPDs can augment the information already available through administrative data and social surveys. First, we assess the reliability of MPDs through comparison with official estimates. In addition, to analyze and understand people's spatio-temporal behavior, it is important to better understand the location of MPDs, i.e., information about the geographic reference of cell phones during their activity. Finally, we highlight the assumptions underlying our elaborations, as well as additional potentials and limitations of the available data.

## 1. Objective and data description

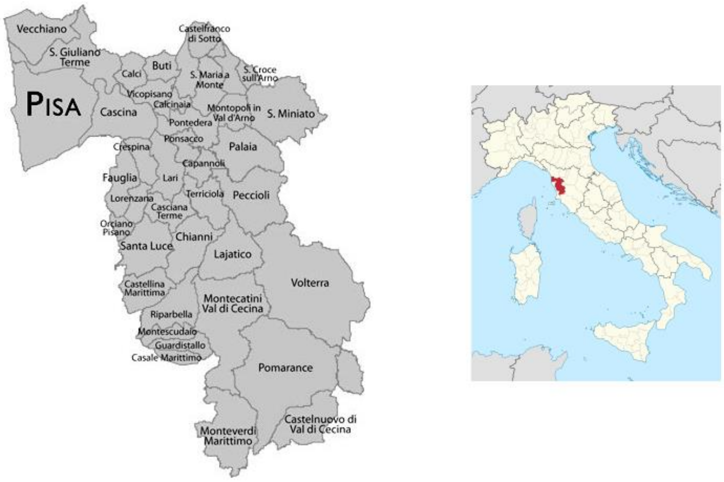
New sources of data, including those held by private entities, are attractive for reuse for statistical purposes. Mobile phone data (MPD) are promising, since today almost everyone carries (at least) a mobile device with them during their daily activities and travels. Over the past decade, many National Statistical Institutes around the world have studied the potential and limitations of these data sources, in several fields: dynamic and present population, tourism, commuting, etc. The MPD can be used to complement and enrich the traditional surveys and the administrative data, thanks to their timeliness and greatest spatial availability in representing human behaviors. The Italian National Statistical Institute (Istat) is interested in exploring the usability and potential of Mobile Phone Data (MPD) in the production of official statistics, first assessing its reliability through comparison with official estimates.

In this contribution we describe a particular type of MPD, so-called Call Detail Records (CDRs), and the data processing steps to be undertaken to estimate the location of the MPD, i.e., information about the geographic reference of cell phones during their activity. Finally, we show some potential applications of the MPD, for estimating population density and analysing mobility patterns.

### *1.1. Data*

ISTAT received from a mobile network operator (MNO, the data provider) Call Detail Records (CDRs) of its subscribers' calls over a 6-week period in an Italian province. This provision was managed as part of a Persons & Places research project[1]. Specifically, the data are for the weeks between January 1 and February 12, 2017, and for the province of Pisa, a medium-sized province in central Italy (Tuscany). The province is organized into 37 municipalities (as shown in Figure 1) among which the most populous are Pisa, Cascina, San Giuliano Terme, Pontedera and San Miniato.

**Figure 1 –** *The province of Pisa and all its municipalities.*



[1] This project complies with the privacy regulation: the data collection complies with the regulations of DL 6.9.1989 n.322; anonymity complies with DL 30.06.2003 n 196 art. 4 paragraph 1 lett. b) and n) and Opinion No. 9802796 of 09.06.2022 as reported in PSN document IST-03434. 4 paragraph 1 lett. b) and n) and Opinion No. 9802796 of 09.06.2022 as reported in PSN document IST-02834.

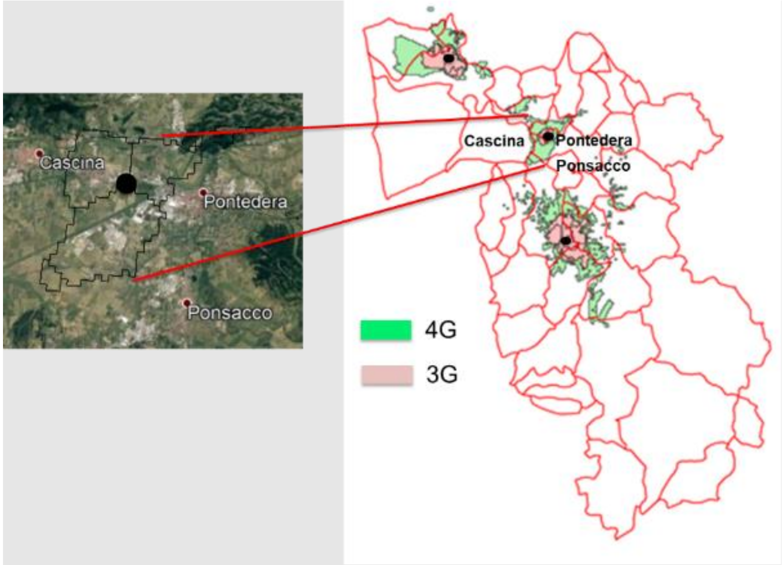
The CDR data available to ISTAT are composed as follows: Caller ID, which is a numeric code associated with each Subscriber Identity Module (SIM) by an algorithm that guarantees its anonymity; the network cell from which the call originated; the date and time the call originated; the duration of the call; and the network cell where the call ended. For text messages, the data report the date and time of the text message and the network cell from which the text message was sent.

The supply consists of about eighteen million CDRs, divided into: less than eleven million calls and seven million text messages. The total number of calling SIMs is just over four hundred thousand. Descriptive analyses are shown in Section 3. CDRs are processed to ensure anonymity.

To analyze and understand people's spatiotemporal behavior, it is important to evaluate MPD localization, i.e., information related to the geographic referencing of cell phones during their activity. In the case of CDR, we have passive localization that corresponds to the code of the antenna/sector to which the calling device has been connected. In fact, the cellular signal is picked up by an antenna and enters the network. The antennas are the cellular phone installations that receive and retransmit signals from cellular phones that are distributed throughout the territory in a capillary manner, based on population density. Each antenna is designed to serve a limited portion of territory, called a "cell". Cells are divided into different sectors. Each sector is a service characterized by a technology, a direction, and an antenna coverage area (this area is named Service Area, as shown in Figure 2).

Call location information can be obtained by different techniques. In the simplest, localization is based on the position of the antenna. This localization concentrates all calls and text messages in the municipality where the antenna is located, although antenna coverage is very wide and often covers areas that belong to more than one municipality. For example, in Figure 2 on the left, the antenna mast is in the municipality of Cascina, but the coverage covers the municipalities of Cascina, Pontedera, and Ponsacco. The methodology we have applied in this work, in collaboration with the MNO, divides the proportion of the territory according to the BSA, and assigns each call and text message as a percentage to all the municipalities served by the specific BSA.

**Figure 2. –** *Left: An example of best service areas (BSAs) of an antenna in 4G technology. The antenna tower is located at the point represented by the black circle, contains three sectors, and the different areas represent the corresponding three BSAs. Right: an example of BSAs for different technologies in the same antenna tower. Three antennas are shown, BSAs for 3G technology are represented in pink and those for 4G technology in green.*



Specifically, knowing the percentage of the coverage area of each sector for each municipality and considering the uniform distribution of calls for each BSA, the call rates of each BSA for each municipality were calculated based on the percentage of coverage. The following analyses are based on these calculations. Potentially, additional information can be introduced, e.g., land use of the area covered by the BSA, population counts from other sources, the presence of specific points of interest (universities, large local units and businesses, large shopping centers, large hospitals, ...), and as a result, different assumptions can be applied to distribute the calls from each BSA to each municipality. In this application, we chose the simplest hypothesis, with the intention of testing its robustness in further applications based on more sophisticated methods and hypotheses. In addition, it is worth noting that many of the additional data sources that can be exploited to enrich the current hypothesis are often not available at the level of the BSA, which is an irregular polygon not linked to any administrative territorial unit. For competitive exploitation of all additional sources, a further step of mapping the BSA on a regular grid is necessary and recommended.

## 2. First analyses on Pisa CDR

Figure 3 shows the number of SIM and the number of calls per day. In the period between two holidays (New Year's Day and Epiphany), the trend is irregular with respect to the following weeks. During weekdays, the trend is regular, with a sharp drop on weekends. This trend is also documented in phone traffic in other countries (de Jonge *et al.* 2012, Douglas *et al.* 2015, Furletti *et al.* 2017).

**Figure 3 –** *SIMs (blue) and calls (red) per day in the period between 1st January and 12th February.*

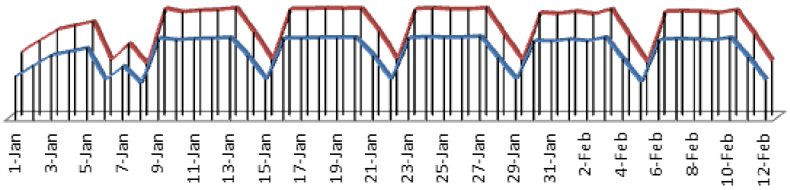


Figure 4 shows that the daily pattern of voice call events and SMS events considered together is the same as SMS events alone. For this reason, the data were used without any distinction between voice and text data.

**Figure 4 -** *Voice calls and text messages per day (pink), SMSs only (red).*

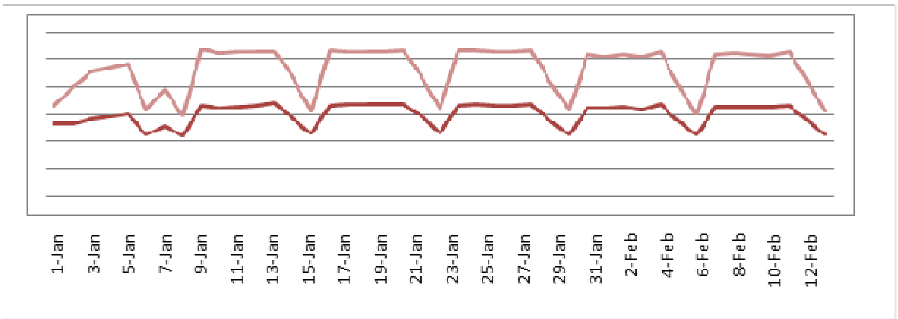
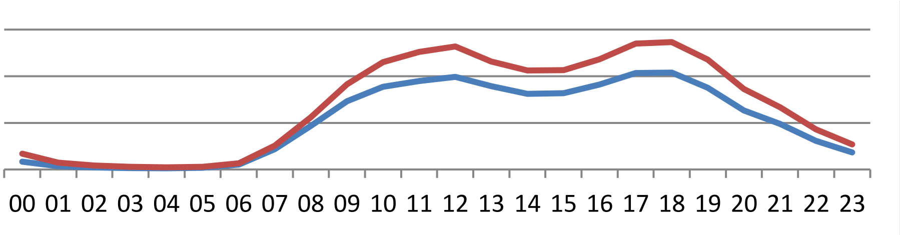


Figure 5 shows that the hourly number of voice calls is similar on weekdays and weekends. Only the volumes are different. Two peaks of voice calls are noted, around 12:00 noon and 6:00 p.m. on each day.

**Figure 5 –** *Voice calls during the working days and the weekend, red and blue, respectively.*



## 3. Results

### 3.1. Mobile phone data for population estimates

To properly use MPD for population estimates, we first studied the correlation between the MPD and official population counts, i.e., the count of people residing in the reference area at the reference time.

CDRs provide information on the activity of MP (mobile phone) users at a given date (with detailed time) and at a very small spatial scale. Calls and text messages can be used to produce population count estimates given certain basic assumptions:

- High level of MP penetration, which is the number of active MP users per 100 people within a specific population;
- High level of mobile network coverage in the territory, i.e., the portion of the geographical area, throughout the country, in which people can make calls and send messages from their cell phones;
- knowledge of the MNO market share, that is, the percentage of total subscribers belonging to a particular MNO.

High values for the previous indicators allow us to use the MPD to derive some estimates of population counts under reasonable assumptions; see Deville *et al.* 2014 and Douglas *et al.* 2015 for a discussion of this topic.

Italy has one of the highest MP penetration rates in developed countries; in fact, the percentage of MP per 100 citizens is about 154% in 2016[2], which means that, on average, each person owned 1.54 cell phones in 2016. Italy also shows a high coverage of MP networks in the territory, for example, 4G technology coverage is

[2] This information is published on AGCOM web site: https://www.agcom.it/servizi-di-rete-mobile (visited 22-01-2024)

97% and 3G technology coverage is 99% of the Italian territory[3]. In addition, the close cooperation with MNO ensures that the market share can be assessed on a small spatial scale, under the confidentiality constraint of business secrecy.

To investigate the correlation of MPD with official population data, we first focused the analysis on the nighttime population. The approximation of residential population with nighttime mobile phone users has been already exploited in several works (Ma and Wu 2012, Deville *et al.* 2014, Douglas *et al.* 2015). In this paper, we identify mobile phone users with SIM cards.

An initial study was conducted by considering SIM localization using antenna tower location. This work highlighted the limitations of this type of localization and suggested the need to implement a finer geolocation methodology. For this purpose, in collaboration with the MNO, we adopted the BSA-based approach mentioned in the previous section. The municipality of residence is then assigned to each SIM according to the following procedure: a SIM's home is located in the BSA most frequently recorded in the CDR records during the nighttime hours, from 8 p.m. to 7 a.m. If this BSA covers several municipalities, the SIM is counted with a percentage for each of them, and the percentage assigned to each municipality is proportional to the area covered by the BSA.

Figure 6 shows a scatter plot of the count of SIMs active at night against the January 2017 residential population estimates for the province of Pisa at the municipal level. There is a reasonably good relationship, approximately linear as indicated by the LOESS regression interpolation, in blue in the graph. In the linear regression model, the correlation coefficient is 0.94, reflecting the adequacy of the model in predicting the residential population through nighttime mobile phone users. Similar results in terms of high correlation are also obtained when the logarithmic transformation is considered, and the extreme value represented by the city of Pisa is excluded from the analysis.

The high correlation between phone users and residential population is also confirmed when the analysis focuses on SIMs active during the daytime. Specifically, we analyzed SIMs calling between 5 p.m. and 6 p.m.: the most frequent peak hour during the observed period, Monday through Friday. We used the population of phone users identified by these SIMs as a predictor of the residential population, and the SIMs are assigned to the municipality resulting from the nighttime location.

[3] Data processed from open data published by AGCOM referring to 2018, https://maps.agcom.it/

**Figure 6 -** *Scatter plot of nighttime mobile phone users versus residential population (municipalities in Pisa province).*

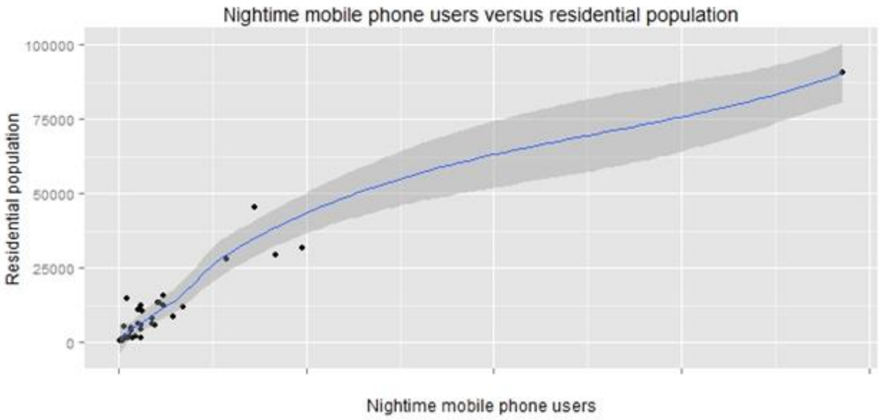
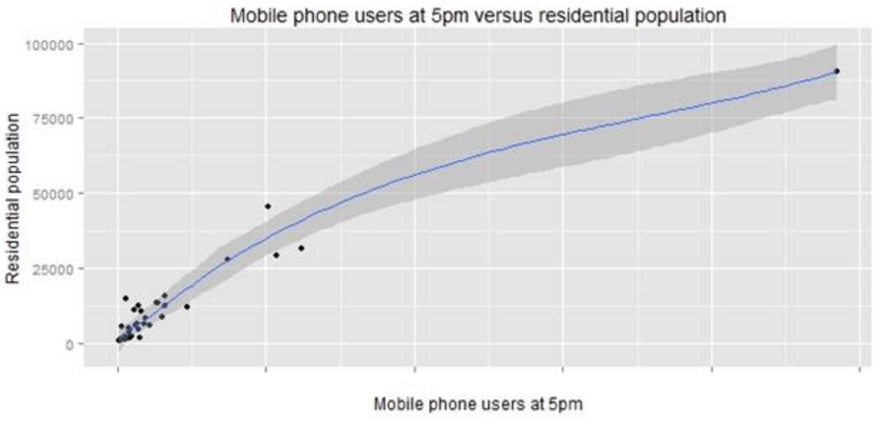


Figure 7 shows the scatter plot of the count of active SIMs from 5 a.m. to 6 p.m. against the January 2017 residential population estimates for the province of Pisa at the municipal level. Again, a reasonably good approximation of the linear relationship is observed, as indicated by the LOESS regression interpolation (in blue in the plot). In the linear regression model, the correlation coefficient improves to 0.95.

**Figure 7 -** *Scatter plot of mobile phone users at 5 pm versus residential population (municipalities in Pisa's province).*

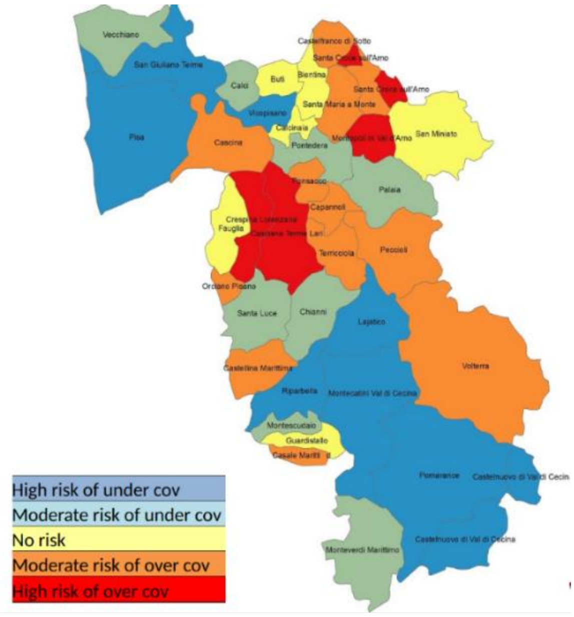


These results are considered satisfactory and encourage us to use CDRs to estimate something that cannot currently be observed with sample surveys and administrative data, but which could be a new output within official statistics.

One example of these new opportunities offered by MPDs is related to their use in combination with the coverage surveys included in the new population census framework. In the Census Transformation Program, MPDs allow us to identify areas that might be problematic for census counts; for example, areas at risk of over- or under-coverage can be identified by comparing population estimates from MPDs with counts of persons 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 sampling stage, when designing the coverage sample survey, and at the estimation stage, when small area population estimates are to be provided. Figure 8 shows the areas at risk of under/over-coverage for the province of Pisa based on a comparison of MPD population estimates with the January 2017 residential population.

**Figure 8 -** *Municipalities at risk of over/under-coverage, Pisa province.*



The MPD population estimates are based on the nighttime MP users by applying a multiplier estimator (EMCDDA, 1999, Jandl 2009) that considers the market share of the MP operators in the Pisa province. The municipalities where the MP population estimates are similar to the official estimates, with differences lower that 10% are in yellow. The municipalities with lower MP population estimates are in the light red area (up to 50% lower than the official estimates) thus highlighting a moderate risk of over-coverage. At the same time the higher ones are in the light blue area and show a moderate risk of under-coverage. The municipalities with the highest risk of over-coverage are in dark red, since the MP population estimates are lower up to 1.5 times than the counts enrolled in the registers; instead, the data referring to under coverage, as the estimates are higher, are reported in blue.

### *3.2. Mobility pattern analysis: the Origin-Destination Matrix*

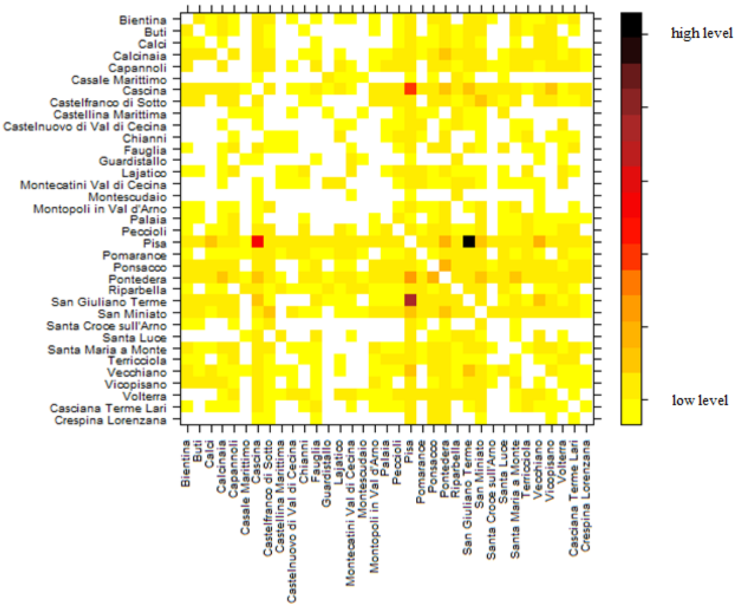
Another opportunity offered by MPD is related to understanding how and where people move, so-called mobility pattern analysis. There are at least two ways to study mobility through MPD: the first is based on the relative densities of CDRs, both across areas and over time, as shown in the previous section; the second is based on anonymized data at the individual level. Some results of the first type of analysis were shown in the previous section, while in this section we show some results using the second type as input. In this case, meaningful placement for MPD such as "home" and "work/study" is determined as follows:

- "home" is the municipality where a MP user is more frequently found during the nighttime, as in the previous section for the residential population estimates;
- "work/study" is the municipality where the MP user is repeatedly observed during the daytime hours.

By aggregating individual-level data for which home and work/study were previously derived, the origin-destination flows of home and work/study can be produced. In Figure 8 we propose an origin-destination matrix for the Pisa province at municipal level, where only movements within the province are considered. The main diagonal represents people who live ("home") and "work/study" in the same municipality. These intra-municipal movements are not of interest in this analysis, and are not shown in the matrix, although they account for 70% of the data analysed. The intensity of the movement is represented by the intensity of the colors on the matrix. Again, the results are in line with the information coming from other sources: the most used routes are those involving Pisa and its neighboring municipalities (Cascina and San Giuliano Terme), as well as the municipalities where the province's largest establishments are located (e.g., Pontedera). This result is in line with expectations since the city of Pisa is a national attraction for tourism and is home to an important university. It also has a much larger population than other centers in the province.

One disadvantage of this analysis, compared with results derived from administrative data, is that it is not possible to identify the reason for mobility; on the other hand, MPD allow us to assess the frequencies of mobility, which cannot be derived from administrative data.

**Figure 9 -** *This Origin-Destination (OD) Matrix describes people movement in the Pisa Province.*



## 4. Lesson learned and next steps

The results of the CDR analyses described in this report are definitively encouraging regarding the potential of MPD for both population estimates and the study of mobility patterns.

A key issue for effective exploitation of CDRs is small-scale localization of MP users' activities (calls and text messages). Currently, localization based on antenna position (i.e., all calls and text messages are assigned to the municipality where the antenna is located) has weaknesses that compromise the reliability of population estimates and all other statistics related to this concept. We overcame these drawbacks through collaboration with the MNO, which provided us with the percentage of territory served by the BSA. On this basis, we were able to develop a procedure that assigns each MP activity as a percentage to all municipalities served by the specific BSA, and in this way, we greatly improved the location of CDRs and obtained reliable population estimates at the municipal level.

In the future, in agreement with the MNO, we will be able to produce statistics at a smaller scale than the individual municipality, i.e., census sections, to take full advantage of the enormous amount of information that MPDs provide us on "urban rhythms" to design and optimize mobility in dense urban centers.

Finally, the analyses proposed in this report are still a local observation, limited to one province. The availability of new data will enable us to examine these results on a larger scale, such as at the regional level and at the national level.

## References

DE JONGE E., VAN PELT M., ROOS M. 2012. Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. *Discussion paper 201214,* Statistics Netherlands.

DEVILLE P., LINARD C., MARTIN S., TATEM A.J. 2014. Dynamic population mapping using mobile phone data, *PNAS*, Vol. 111, No 45, pp. 15888-15893.

DOUGLASS R.W., MEYER D.A., RAM M., REDEOUT D., SOND D. 2015. High resolution population estimates from telecommunications data, *EPJ Data Science*, Vol. 4, No 1.

EUROPEAN MONITORING CENTRE FOR DRUGS AND DRUG ADDICTION (EMCDDA) 1999 Scientific Review of the Literature on Estimating the Prevalence of Drug Misuse on the Local Level. Lisbon: EMCDDA, July 1999.

FURLETTI B., TRASARTI R., CINTIA P., GABRIELLI L. 2017. Discovering and Understanding City Events with Big Data: The Case of Rome, *Information*, Vol. 8, No 3, p.74.

JANDL M. 2009. A multiplier estimate of the illegally resident third-country national population in Austria based on crime suspect data. *Working Papers 2*, Hamburg Institute of International Economics. Database on Irregular Migration.

MA X., WU L. 2012. Towards Estimating Urban Population Distributions from Mobile Call Data, *Journal of Urban Technology*, Vol. 19, No 4, pp. 3-21.

Fabrizio DE FAUSTI, ISTAT, defausti@istat.it
Roberta RADINI, ISTAT, radini@istat.it
Tiziana TUOTO, ISTAT, tuoto@istat.it
Luca VALENTINO, ISTAT, valentino@istat.it