Int J Data Sci Anal (2016) 2:75–92
DOI 10.1007/s41060-016-0013-2

REGULAR PAPER

An analytical framework to nowcast well-being using mobile phone data

Luca Pappalardo[1,2] · **Maarten Vanhoof**[3,4] · **Lorenzo Gabrielli**[2] · **Zbigniew Smoreda**[3] · **Dino Pedreschi**[1] · **Fosca Giannotti**[2]

Received: 16 December 2015 / Accepted: 5 June 2016 / Published online: 27 June 2016

Abstract An intriguing open question is whether measurements derived from Big Data recording human activities can yield high-fidelity proxies of socio-economic development and well-being. Can we monitor and predict the socio-economic development of a territory just by observing the behavior of its inhabitants through the lens of Big Data? In this paper, we design a data-driven analytical framework that uses mobility measures and social measures extracted from mobile phone data to estimate indicators for socio-economic development and well-being. We discover that the diversity of mobility, defined in terms of entropy of the individual users' trajectories, exhibits (i) significant correlation with two different socio-economic indicators and (ii) the highest importance in predictive models built to predict the socio-economic indicators. Our analytical framework opens an interesting perspective to study human behavior through the lens of Big Data by means of new statistical indicators that quantify and possibly "nowcast" the well-being and the socio-economic development of a territory.

Keywords Big data · Complex systems · Human mobility · Social networks · Economic development · Economic data · Nowcasting · Forecasting

1 Introduction

Big Data, the masses of digital breadcrumbs produced by the information technologies that humans use in their daily activities, allow us to scrutinize individual and collective behavior at an unprecedented scale, detail, and speed. Building on this opportunity we have the potential capability of creating a digital nervous system of our society, enabling the measurement, monitoring and prediction of relevant aspects of the socio-economic structure in quasi real time [23]. An intriguing question is whether and how measurements made on Big Data can yield high-fidelity proxies of socio-economic development and well-being. Can we monitor and possibly predict the socio-economic development of our societies just by observing human behavior, like human movements and social relationships, through the lens of Big Data?

This fascinating question, also stimulated by the United Nations in recent reports [2,30], has attracted the interest of researchers from several disciplines, who started investigating the relations between human behavior and economic development based on large experimental datasets collected for completely different purposes [17,38]. As a first result along this line a seminal work exploited a nationwide mobile phone dataset to discover that the diversity of social contacts

✉ Luca Pappalardo
luca.pappalardo@isti.cnr.it; lpappalardo@di.unipi.it

Maarten Vanhoof
m.vanhoof1@newcastle.ac.uk

Lorenzo Gabrielli
lorenzo.gabrielli@isti.cnr.it

Zbigniew Smoreda
zbigniew.smoreda@orange.com

Dino Pedreschi
pedre@di.unipi.it

Fosca Giannotti
fosca.giannotti@isti.cnr.it

1 Department of Computer Science, University of Pisa, Pisa, Italy

2 Institute of Information Science and Technologies (ISTI), National Research Council (CNR), Pisa, Italy

3 SENSE Orange Labs, Paris, France

4 Open Lab, Newcastle University, Newcastle upon Tyne, UK

of the inhabitants of a municipality is positively associated with a socio-economic indicator of poverty, independently surveyed by the official statistics institutes [17]. This result suggests that social behavior, to some extent, is a proxy for the economic status of a given territory. However, little effort has been put in investigating how human mobility affects, and is affected by, the socio-economic development of a territory. Theoretical works suggest that human mobility is related to economic well-being, as it could nourish economy and facilitate flows of people and goods, whereas constraints in the possibilities to move freely can diminish economic opportunities [33]. So, it is reasonable to investigate the role of human mobility with respect to the socio-economic development of a given territory.

In this paper we design a data-driven analytical framework that uses Big Data to extract meaningful measures of human behavior and estimate indicators for the socio-economic development. The analytical framework we propose is repeatable for different countries and geographic scales since it is based on mobile phone data, the so-called CDRs (Call Detail Records) of calling and texting activity of users. Mobile phone data, indeed, can be retrieved in every country due to their worldwide diffusion [7]: There are 6.8 billion mobile phone subscribers today over 7 billion people on the planet, with a penetration of 128 % in the developed world and 90 % in developing countries. Moreover, CDR data have proven to be a hi-fi proxy for individuals' movements and social interactions [24,41].

We apply our analytical framework on large-scale mobile phone data—20 million users and 5.7 billions calls—and quantify the relations between human mobility, social interactions and economic development in France using municipality-level official statistics as external comparison measurements. We first define four individual measures derived from mobile phone data which describe different aspects of individual human behavior: the volume of mobility, the diversity of mobility, the volume of sociality and the diversity of sociality. Though just a subset of the many possible behavioral aspects that can be extracted from mobile phone data, the four measures we consider in this paper are widely accepted by the scientific community and have been proven to capture important aspects of both human mobility [18,24,47,59] and social relationships [3,4,17]. Each individual measure is computed for every user in our dataset based on locations and calls as recorded in the mobile phone data. In a second stage, we aggregate the four individual measures at the level of French municipalities and explore the correlations between the four measures and two external indicators of socio-economic development. We find that the average mobility diversity of individuals resident in the same municipality exhibits a superior correlation degree with the socio-economic indicators. We confirm these results against two different null models, an observation that allows us to reject the hypothesis that our discovery occurred by chance.

Next, we build regression and classification models to predict the external socio-economic indicators from the population density and the social and mobility measures aggregated at municipality scale. We show that the diversity of human mobility significantly adds a predictive power in both regression and classification models, substantially more than the diversity of social contacts and demographic measures such as population density, a factor that is known to be correlated with the intensity of human activities [42,62]. The importance of this finding is twofold. On one side, it offers a new stimulus to social research: Diversity is a key concept not only for natural ecosystems but also for the social ecosystems, and can be used to better understand the complexity of our interconnected society. On the other side, our results reveal the high potential of Big Data in providing representative, relatively inexpensive and readily available measures as proxies of socio-economic development. Our analytical framework opens an interesting perspective to engineer official statistics processes to monitor human behavior through mobile phone data. New statistical indicators can be defined to describe and possibly "nowcast" the economic status of a territory, even when such measurements would be impossible using traditional censuses and surveys [2,30].

The paper is organized as follows. Section 2 revises the scientific literature relevant to our topic, Sect. 3 describes the proposed data-driven analytical framework. In Sects. 4, 5 and 6 we apply our analytical framework on a nationwide mobile phone dataset covering several weeks of call activity in France. We introduce the mobile phone data in Sect. 4.1, the measures of individual mobility behavior and individual social behavior in Sect. 4.2, and the computations of the measures on a nation-wide mobile phone dataset in Sect. 4.3. In Sect. 5 we describe the results of the correlation analysis and validate them against two null models. In Sect. 6 we present and validate predictive models for socio-economic development. In Sect. 7 we discuss the results and finally Sect. 8 concludes the paper describing the opportunities and the challenges that arise from our research.

2 Related work

The interest around the analysis of Big Data, which provide nowadays the possibility to study human behavior at both individual and collective level, has infected all branches of human knowledge, from economy [49], to human mobility [47], social networks [4] and even sports [11,12]. In this section we briefly revise the existing work on Big Data analytics relevant to our work. We first summarize the main results on human mobility and social network analysis, and then revise the studies on Big Data analytics for the estimation of well-being.

Human mobility and complex networks Studies from different disciplines document a stunning heterogeneity of human travel patterns [24,44], and at the same time observe a high degree of predictability [18,59]. The patterns of human mobility have been used to build generative models of individual human mobility and human migration flows [32,45,47,56], to construct methods for profiling individuals according to their mobility patterns [47], to discover geographic borders according to recurrent trips of private vehicles [55], or to predict the formation of social ties [10,61], and to predict the kind of activity associated to individuals' trips on the only basis of the observed displacements [31,35,54]. There are widely accepted mobility models and measures, e.g., radius of gyration [24,47], mobility entropy [48,59], individual mobility networks [31,54] and origin-destination matrices [55], that can be used to study different aspects of both individual and collective mobility. In the context of social network analysis, the observation of social interactions data provided by emails, mobile phones, and social media allowed to reveal the complexity underlying the social structure [4]: Hubs exist in our social networks that strongly contribute to the so-called small world phenomenon [3], and social networks are found to have a tendency to partition into social communities, i.e., clusters of densely connected sets of individuals [19]. There are a large number of both individual and global network features, such as degree, social diversity and clustering coefficient to name a few [40], that can be used to study the structure and the evolution of social relationships.

Big data for official statistics The last few years have also witnessed a growing interest around the usage of Big Data to support official statistics in the measurement of individual and collective well-being [14,60]. Even the United Nations, in two recent reports, stimulate the usage of Big Data to investigate the patterns of phenomena relative to people's health and well-being [2,30]. The vast majority of works in the context of Big Data for official statistics are based on the analysis of mobile phone data, the so-called CDRs (Call Detail Records) of calling and texting activity of users. Mobile phone data, indeed, guarantee the repeatability of experiments in different countries and on different scales as they, nowadays, can be retrieved in every country given the worldwide diffusion of mobile phones [7]. A set of recent works use mobile phone data as a proxy for socio-demographic variables. Deville et al., for example, show how the ubiquity of mobile phone data can be exploited to provide accurate and detailed maps of population distribution over national scales and any time period [16]. Brea et al. study the structure of the social graph of mobile phone users in Mexico and propose an algorithm for the prediction of the age of mobile phone users [9]. Another recent work uses mobile phone data to study inter-city mobility and develop a methodology to detect the fraction of residents, commuters and visitors within each city [21].

A lot of effort has been put in recent years on the usage of mobile phone data to study the relationships between human behavior and collective socio-economic development. The seminal work by Eagle et al. analyzes landline calls and a nationwide mobile phone dataset to shows that, in the UK, regional communication diversity is positively associated with a socio-economic ranking [17]. Gutierrez et al. address the issue of mapping poverty with mobile phone data through the analysis of airtime credit purchases in Ivory Coast [26]. Blumenstock shows a preliminary evidence of a relationship between individual wealth and the history of mobile phone transactions [8]. Decuyper et al. use mobile phone data to study food security indicators finding a strong correlation between the consumption of vegetables rich in vitamins and airtime purchase [15]. Frias-Martinez et al. analyze the relationship between human mobility and the socio-economic status of urban zones, presenting which mobility indicators correlate best with socio-economic levels and building a model to predict the socio-economic level from mobile phone traces [20]. Pappalardo et al. analyze mobile phone data and extract meaningful mobility measures for cities, discovering interesting correlation between human mobility aspects and socio-economic indicators [48]. Lotero et al. analyze the architecture of urban mobility networks in two Latin-American cities from the multiplex perspective. They discover that the socio-economic characteristics of the population have an extraordinary impact in the layer organization of these multiplex systems [37]. Amini et al. use mobile phone data to compare human mobility patterns of a developing country (Ivory Coast) and a developed country (Portugal). They show that cultural diversity in developing regions can present challenges to mobility models defined in less culturally diverse regions [1]. Smith-Clarke at al. analyze the aggregated mobile phone data of two developing countries and extract features that are strongly correlated with poverty indexes derived from official statistics census data [57]. Other recent works use different types of mobility data, e.g., GPS tracks and market retail data, to show that Big Data on human movements can be used to support official statistics and understand people's purchase needs. Pennacchioli et al. for example provide an empirical evidence of the influence of purchase needs on human mobility, analyzing the purchases of an Italian supermarket chain to show a range effect of products: The more sophisticated the needs they satisfy, the more the customers are willing to travel [50]. Marchetti et al. perform a study on a regional level analyzing GPS tracks from cars in Tuscany to extract measures of human mobility at province and municipality level, finding a strong correlation between the mobility measures and a poverty index independently surveyed by the Italian official statistics institute [38].

Position of our work Despite an increasing interest in this field of research, a review of the state-of-the-art cannot avoid to notice that there is no unified methodology to exploit Big Data for official statistics. It is also surprising that widely accepted measures of human mobility (e.g., radius of gyration [24] and mobility entropy [59]) have not been used so far. We overcome these issues by providing an analytical framework as support for official statistics, which allows for a systematic evaluation of the relations between relevant aspects of human behavior and the development of a territory. Moreover, our paper shows how standard mobility measures, not exploited so far, are powerful tools for official statistics purposes.

3 The analytical framework

Our analytical framework is a knowledge and analytical infrastructure that uses Big Data to provide reliable measurements of socio-economic development, aiming at satisfying the increasing demand by policy makers for continuous and up-to-date information on the geographic distribution of poverty, inequality or life conditions. Figure 1 describes the structure of the framework, highlighting the complexity of the nowcasting process: between data and predictions, many complex steps are required to transform Big Data into reliable estimates for a municipality's socio-economic development. The analytical framework is based on mobile phone data, which guarantee the repeatability of the process for deployment in different countries and geographical scales. In particular the CDRs, generally collected by mobile phone operators for billing and operational purposes, contain an enormous amount of information on how, when, and with whom people communicate. This wealth of information allows to capture different aspects of human behavior and stimulated the creativity of scientists from different disciplines, who demonstrated that CDRs are a high-quality proxy for studying individual mobility and social ties [24,41].

Starting from the collected mobile phone data (Fig. 1a) a set of measures are computed which grasp the salient aspects of individuals' mobility and social behavior (Fig. 1b). This step is computationally expensive when the analytical framework is applied on massive data such as the CDRs of an entire country for a long period. To parallelize the computations and speed up the execution a distributed processing platform can be used such as Hadoop (see Sect. 4.3). A wide set of mobility and social measures can be computed during this phase, and the set can be enlarged with new measures as soon as they are proven to be correlated with socio-economic development aspects of interest. In Sect. 4.2 we propose, as an example, a set of standard measures of individual mobility and sociality and show how they can be computed on mobile phone data.

As generally required by policy makers, official statistics about socio-economic development are available at the level of geographic units, e.g., regions, provinces, municipalities, districts or census cells. Therefore, the individuals in the dataset have to be mapped to the corresponding territory of residence, in order to perform an aggregation of the individual measures into a territorial measure (Fig. 1c, d). When the city of residence or the address of the users is available in the data, this information can be easily used to assign each

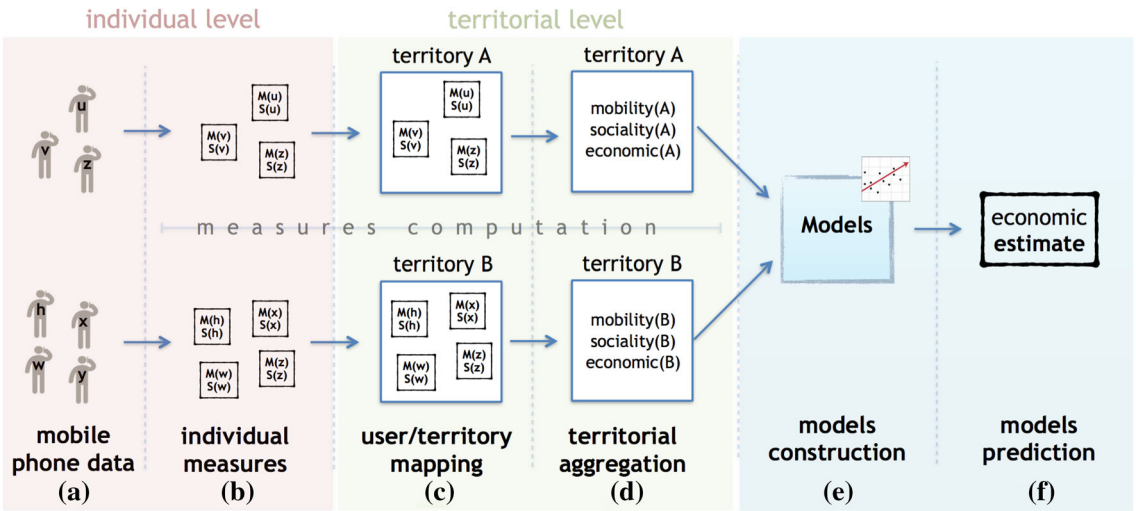

Fig. 1 The data-driven analytical framework. Starting from mobile phone data (**a**) mobility and social measures are computed for each individual in the dataset (**b**). Each individual is then assigned to the territory where she resides (**c**) and the individual measures are aggregated at territorial level (**d**). Starting from the aggregated measures predictive models are constructed (**e**) in order to estimate and predict the socio-economic development of the territories (**f**)

Table 1 Example of call detail records (CDRs)

(a)				(b)		
Timestamp	Tower	Caller	Callee	Tower	Latitude	Longitude
2007/09/10 23:34	36	4F80460	4F80331	36	49.54	3.64
2007/10/10 01:12	36	2B01359	9H80125	37	48.28	1.258
2007/10/10 01:43	38	2B19935	6W1199	38	48.22	−1.52
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Every time a user makes a call, a record is created with timestamp, the phone tower serving the call, the caller identifier and the callee identifier (a). For each tower, the latitude and longitude coordinates are available to map the tower on the territory (b)

individual to corresponding city of residence. Unfortunately these socio-demographic data are generally not available in mobile phone data for privacy and proprietary reasons. This issue can be solved, with a certain degree of approximation, by inferring the information from the data source. In literature the phone tower where a user makes the highest number of calls during nighttime is usually considered her home phone tower [51]. Then with standard Geographic Information System techniques it is possible to associate the phone tower to its territory (see Sect. 4.3).

The obtained aggregated measures are compared with the external socio-economic indicators to perform correlation analysis and learn predictive models (Fig. 1e). The predictive models can be aimed at predicting the actual value of socio-economic development of the territory, e.g., by regression models (Sect. 6.1), or to predict the class of socio-economic development, i.e., the level of development of a given geographic unit as done by classification models (Sect. 6.2). Finally, the predictions produced by the models are the output of the analytical framework (Fig. 1f). The measures, the territorial aggregation and the predictive models can be updated every time new mobile phone data become available. This provides policy makers with up-to-date estimates of the socio-economic situation of a given territory, in contrast to indicators produced by official statistics institutes which are generally released after several months or even once a year.

In the following sections we apply the proposed analytical framework on a large-scale nation-wide mobile phone dataset and describe its implementation step by step: from the definition of measures on the data (Sects. 4.1 and 4.2), to their computation and territorial aggregation (Sect. 4.3), and the construction of predictive models (Sects. 5.1, 6.1 and 6.2).

4 Measuring human behavior

We now discuss steps (a), (b) and (c) in Fig. 1, presenting the experimental setting consisting in the computation of the individual measures on the data and their aggregation at territorial level. First, we describe the mobile phone data we use as proxy for individual behavior, together with details about data preprocessing (Sect. 4.1). Then we define the individual measures capturing diverse aspects of individual mobility and social behavior (Sect. 4.2). Finally we show how we compute the individual measures and aggregate them at municipality level (Sect. 4.3).

4.1 Mobile phone data

We have access to a set of CDRs gathered for billing and operational purposes by mobile phone operator Orange. The dataset records the geographic location of 87,000 phone towers and 5.7 billion calls made during 45 days by 20 million anonymized mobile phone users, resulting in a total size of 900 GB of information. CDRs collect geographical, temporal and interaction information on mobile phone use and show an enormous potential to empirically investigate human dynamics on a society wide scale [28]. Each time an individual makes a call the mobile phone operator registers the connection between the caller and the callee, the duration of the call and the coordinates of the phone tower communicating with the served phone, allowing to reconstruct the user's time-resolved trajectory. Table 1 illustrates an example of the structure of CDRs.

CDR data have been extensively used to study human mobility due to the following advantages: They provide a means of sampling user locations at large population scales; they can be retrieved for different countries and geographic scales given their worldwide diffusion; they provide an objective concept of location, i.e., the phone tower. Nevertheless, it is worth noting that CDRs suffer different types of bias [29,53], such as: (i) the position of an individual is known at the granularity level of phone towers; (ii) the position of an individual is known only when she makes a phone call; and (iii) phone calls are sparse in time, i.e., the time between consecutive calls follows a heavy tail distribution [5,24]. In other words, since individuals are inactive most of their time, CDRs allow to reconstruct only a subset of the mobility of an individual. Several works in literature study the bias in CDR

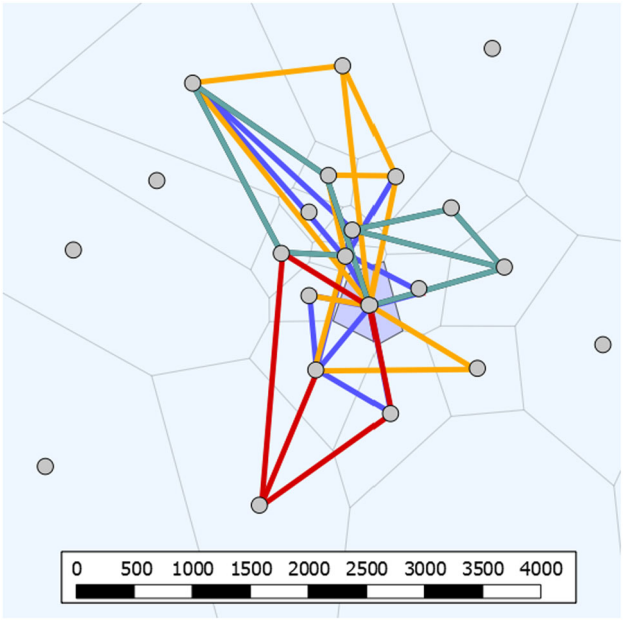

Fig. 2 The detailed trajectory of a single user. The phone towers are shown as *gray dots*, and the Voronoi lattice in *gray marks* approximate reception area of each tower. CDRs records the identity of the closest tower to a mobile user; thus, we can not identify the position of a user within a Voronoi cell. The trajectory describes the user's movements during 4 days (each day in a *different color*). The tower where the user made the highest number of calls during nighttime is depicted in bolder gray (Color figure online)

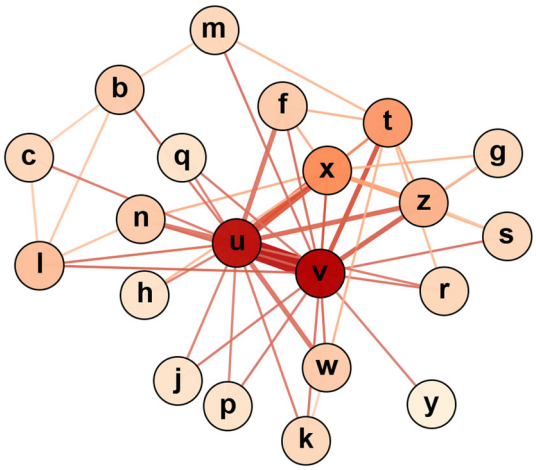

Fig. 3 A fraction of the call graph centered on a single user u. *Nodes* represent users, *edges* indicate reciprocated calls between the users, the size of the edges is proportional to the total number of calls between the users during the 45 days

Table 2 Measures and indicators used in our study

Individual measures		
Sociality	Social volume	*SV*
	Social diversity	*SD*
Mobility	Mobility volume	*MV*
	Mobility diversity	*MD*
Socio-economic indicators		
Demographic	Population density	*PD*
Development	Deprivation index	*DI*
	Per capita income	*PCI*

Social volume, social diversity, mobility volume and mobility diversity are individual measures derived from mobile phone data. Population density, deprivation index, and per capita income are external socio-economic indicators provided by INSEE

data by comparing the mobility patterns observed on CDR data to the same patterns observed on GPS data [43,44,46,47] or handover data (data capturing the location of mobile phone users recorded every hour or so) [24]. The studies agree that the bias in CDR data does not affect significantly the study of human mobility patterns.

In order to cope with sparsity in time of CDR data and focus on individuals with reliable statistics we carry out some preprocessing steps. First, we select only users with a call frequency higher than the threshold $f = N/D > 0.5$, where N is the number of calls made by the user and $D = 45$ days is the length of our period of observation. In practice, we delete all the users with less than one call every two days (in average over the observation period).

Second, we reconstruct the mobility trajectories and the social network of the filtered users. We reconstruct the trajectory of a user based on the time-ordered list of cell phone towers from which she made her calls during the period of observation (see Fig. 2). We then translate the CDR data into a social network representation by linking two users if at least one reciprocated pair of calls exists between them during the period of observation (i.e., A called B and B called A). This procedure eliminates a large number of one-way calls, most of which correspond to isolated events and do not represent meaningful communications [41]. Figure 3 shows a fraction of the social network centered on a single user. The resulting dataset contains the mobility trajectories of 6 million users and a call graph of 33 million edges.

4.2 Measure definition

We introduce two measures of individual mobility behavior and two measures of individual social behavior, studying them in two aspects: volume and diversity (see Table 2).

We define two measures that capture aspects of individual social interactions: *social volume* (*SV*), the number of social contacts of an individual; and *social diversity* (*SD*), the diversification of an individual's calls over the social contacts. Within a social network, we can express the volume of social interactions by counting the amount of links an individual possesses with others. This simple measure of connectivity is widely used in network science and is called the *degree* of an individual [40]. In a call graph the degree of an individual is the number of different individuals who are in contact by

mobile phone calls with her. We can therefore see the degree as a proxy for the volume of sociality for each individual:

$$SV(u) = degree(u) \tag{1}$$

The degree distribution is well approximated by a power law function denoting a high heterogeneity in social networks with respect to the number of friendships [34,41].

The social diversity of an individual u quantifies the topological diversity in a social network as the Shannon entropy associated with her communication behavior [17]:

$$SD(u) = -\frac{\sum_{v=1}^{k} p_{u,v} \log(p_{u,v})}{\log(k)} \tag{2}$$

where k is the degree of individual u, $p_{u,v} = \frac{V_{u,v}}{\sum_{v=1}^{k} V_{u,v}}$ and $V_{u,v}$ is the number of calls between individual u and individual v during the period of observation. SD is a measure for the social diversification of each individual according to its own interaction pattern. In a more general way, individuals who always call the same few contacts reveal a low social diversification resulting in lower values for SD, whereas individuals who distribute their calls among many different contacts show high social diversification, i.e., higher SD. The distribution of SD across the population is peaked, as measured in CDRs and landlines data [17].

Starting from the mobility trajectories of an individual we define two measures to describe individual mobility: *mobility volume* (MV), the characteristic traveled distance of an individual, and *mobility diversity* (MD), the diversification of an individual's movements over her locations. The radius of gyration [24] provides with a measure of mobility volume, indicating the characteristic distance traveled by an individual (see Fig. 4). In detail, it characterizes the spatial spread of the phone towers visited by an individual u from the trajectories' center of mass (i.e., the weighted mean point of the phone towers visited by an individual), defined as:

$$MV(u) = \sqrt{\frac{1}{N} \sum_{i \in L} n_i (r_i - r_{cm})^2} \tag{3}$$

where L is the set of phone towers visited by the individual u, n_i is the individual's visitation frequency of phone tower i, $N = \sum_{i \in L} n_i$ is the sum of all the single frequencies, r_i and r_{cm} are the vectors of coordinates of phone tower i and center of mass, respectively. It is known that the distribution of the radius of gyration reveals heterogeneity across the population: most individuals travel within a short radius of gyration but others cover long distances on a regular basis, as measured on CDR data and GPS data [24,44].

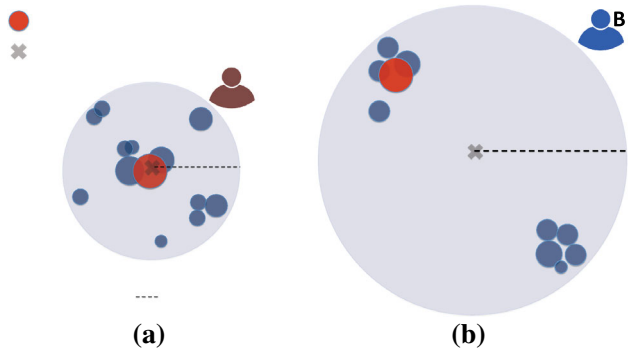

Fig. 4 The radius of gyration of two users in our dataset. The figure shows the spatial distribution of phone towers (*circles*). The size of circles is proportional to their visitation frequency, and the *red location* indicates the most frequent location L_1 (the location where the user makes the highest number of calls during nighttime). The *cross* indicates the position of the center of mass, and the *black dashed line* indicates the radius of gyration. User *A* has a small radius of gyration because she travels between locations that are close to each other. User *B* has high radius of gyration because the locations she visits are far apart from each other (Color figure online)

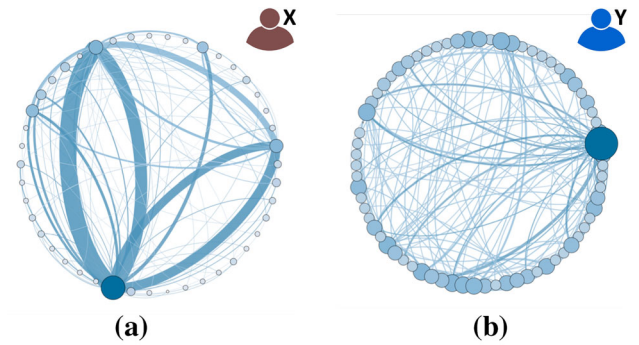

Fig. 5 The mobility entropy of two users in our dataset. *Nodes* represent phone towers, edges represent trips between two phone towers, the *size of nodes* indicates the number of calls of the user managed by the phone tower, and the *size of edges* indicates the number of trips performed by the user on the edge. User *X* has low mobility entropy because she distributes the trips on a few large preferred edges. User *Y* has high mobility entropy because she distributes the trips across many equal-sized edges

Besides the volume of individual mobility we define the diversity of individual mobility by means of the Shannon entropy of individual's trips:

$$MD(u) = -\frac{\sum_{e \in E} p(e) \log p(e)}{\log(N)} \tag{4}$$

where $e = (a, b)$ represents a trip between an origin phone tower and a destination phone tower, E is the set of all the possible origin-destination pairs, $p(e)$ is the probability of observing a movement between phone towers a and b, and N is the total number of trajectories of individual u (Fig. 5). Analogously to SD, MD is high when a user performs many different trips from a variety of origins and destinations; MD

is low when a user performs a small number of recurring trips. Seen from another perspective, the mobility diversity of an individual also quantifies the possibility to predict individual's future whereabouts. Individuals having a very regular movement pattern possess a mobility diversity close to zero and their whereabouts are rather predictable. Conversely, individuals with a high mobility diversity are less predictable. It is known that the distribution of the mobility diversity is peaked across the population and very stable across different social groups (e.g., age and gender) [59].

Clearly, there are many other behavioral aspects that can be derived from mobile phone data and studied in relation to socio-economic development. As a first sample of mobility measures, we propose radius of gyration and mobility entropy for two reasons: (i) They have been extensively studied in literature [18,24,44,47,59]; (ii) it is not clear which of the two mobility measures better correlates with individual and collective well-being. As a first sample of sociality measures, we choose the network degree since it provides a clear and simple measurement of an individual's social engagement. Moreover, we choose social diversity because it is known to correlate with indicators of socio-economic development at territorial level [17]. This provides us ground for comparison between the predictive power of mobility and behavior and the predictive power of social behavior. The set of measures can be extended by including novel measures as soon as they are proposed to describe different aspects of individual human behavior.

4.3 Measure computation

We implement step (b) in Fig. 1 by computing the four behavioral measures for each individual on the filtered CDR data. Due to the size of the dataset (900 GB), for the computations we use a Hadoop cluster of five Linux servers, each one possessing a quadri-core processor and 2 Terabyte of hard drive. We use the Apache Ambari distribution to manage and monitor the Hadoop cluster and to perform distributed storage and processing on the mobile phone dataset.[1]

We find no relationship between the mobility and the social measures at individual level: The correlation between *SV* and *MV*, as well as the correlation between the *SD* and the *MD*, is close to zero. This suggests that the mobility measures and the sociality measures capture different aspects of individual behavior.

We apply step (c) in Fig. 1 by aggregating the individual measures at the municipality level through a two-step process: (i) We assign to each user a home location, i.e., the phone tower where the user performs the highest number of calls during nighttime (from 10 p.m. to 7 a.m.) [51]; (ii) based on these home locations, we assign each user to the corresponding municipality with standard Geographic Information Systems techniques. Figure 6 shows the spatial distribution of Orange users in French municipalities. We aggregate the *SV*, *SD*, *MV* and *MD* at municipality level by taking the mean values across the population of users assigned to that municipality. We obtain 5, 100 municipalities each one with the associated four aggregated measures.

[1] We developed our own programs to compute the individual behavioral measures by using the Hadoop cluster.

5 Correlation analysis

Here we realize step (d) in Fig. 1 and study the interplay between human mobility, social interactions and socio-economic development at municipality level. First, in Sect. 5.1 we introduce the external socio-economic indicators and investigate their correlation between the behavioral measures aggregated at municipality level. Then in Sect. 5.2 we compare the results with two null models to reject the hypothesis that the correlations appear by chance.

5.1 Human behavior versus socio-economic development

As external socio-economic indicators, we use a dataset provided by the French National Institute of Statistics and Economic Studies (INSEE) about socio-economic indicators for all the French municipalities with more than 1,000 official residents. We collect data on population density (PD) and two socio-economic indicators: per capita income (PCI) and deprivation index (DI) (Fig.14).

PCI measures the average wealth of individuals living in a municipality, i.e., the sum of the incomes gained by the residents of a given municipality divided by the number of residents. Due to its simplicity, PCI does not capture important aspects of well-being such as income inequality, costs, environmental impacts, sustainability, and quality of life in general [13]. For this reason, we also collect for every municipality its European Deprivation Index (DI), which is constructed by combining several variables reflecting individual experience of deprivation such as overcrowding, no access to a car or electric heating, unemployment, and low education level. The different variables are combined into a single score by a linear combination with specific choices for coefficients (see "Appendix 1") [52]. Therefore deprivation index is a composite index: The higher its value, the lower is the well-being of the municipality. Preliminary validation showed a high association between the French deprivation index and both income values and education level in French municipalities, partly supporting its ability to measure socio-economic development and well-being [52].

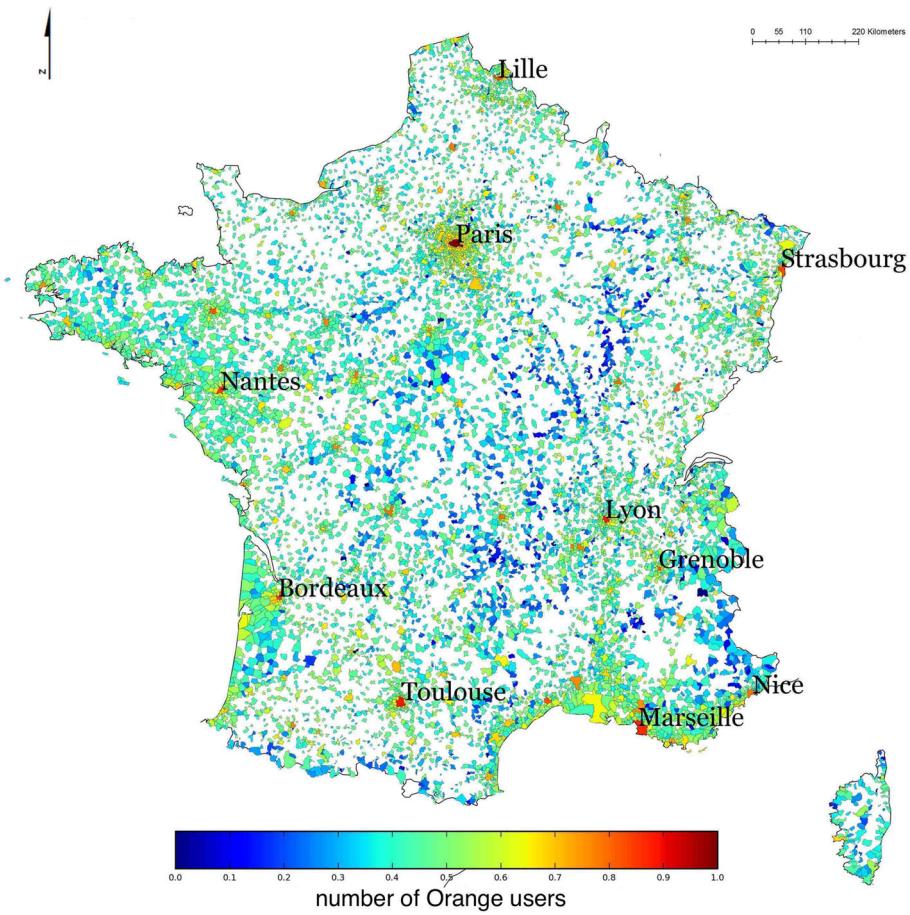

Fig. 6 The spatial distribution of users over French municipalities with more than 1000 official residents. Each user is assigned to a municipality according to the geographic position of her home location. The *color* of municipalities, in a gradient from *blue* to *red*, indicates the number of Orange users assigned to that municipality (normalized in the interval [0, 1]) (Color figure online)

We investigate the correlations between the aggregated measures and the external socio-economic indicators finding two main results. First, the social volume is not correlated with the two socio-economic indicators (Fig. 7c, d), while mobility volume is correlated with per capita income (Fig. 7b). Second, we find that mobility diversity is a better predictor for socio-economic development than social diversity. Figure 7e–h shows the relations between diversity measures and socio-economic indicators. For mobility diversity clear tendencies appear: As the mean mobility diversity of municipalities increases, deprivation index decreases, while per capita income increases (Fig. 7e, f). Social diversity, in contrast, exhibits a weaker correlation with the deprivation index than mobility diversity and no correlation with per capita income (Fig. 7g, h).

Figure 8 provides another way to observe the relations between the diversity measures and socio-economic development. We split the municipalities in ten deciles according to the values of deprivation index. For each decile we compute the distributions of mean mobility diversity and mean social diversity across the municipalities in that decile. For mobility diversity the deciles of the economic values increase, while the mean decreases and the variance increases, highlighting a change of the distribution in the different groups. This is consistent with the observation made in the plots of Fig. 7e. Conversely, for social diversity distribution we do not observe a significant change in the mean and the variance. The observed variation of the mobility diversity distribution in the different deciles is an interesting finding when compared to previous works such as Song et al. [58] which states that mobile predictability is very stable across different subpopulations delineated by personal characteristics like gender or age group. Figures 7 and 8 suggest us that the diversity of human mobility aggregated at municipality level is better associated with the socio-economic indicators than with socio-demographic characteristics.

The relation between mobility diversity and deprivation index is stronger and more evenly distributed over the different levels of deprivation index for municipalities.

5.2 Validation against null models

In order to test the significance of the correlations observed on the empirical data we compare our findings with the results produced by two null models.

In null model NM1 we randomly distribute the users over the French municipalities. We first extract uniformly N users

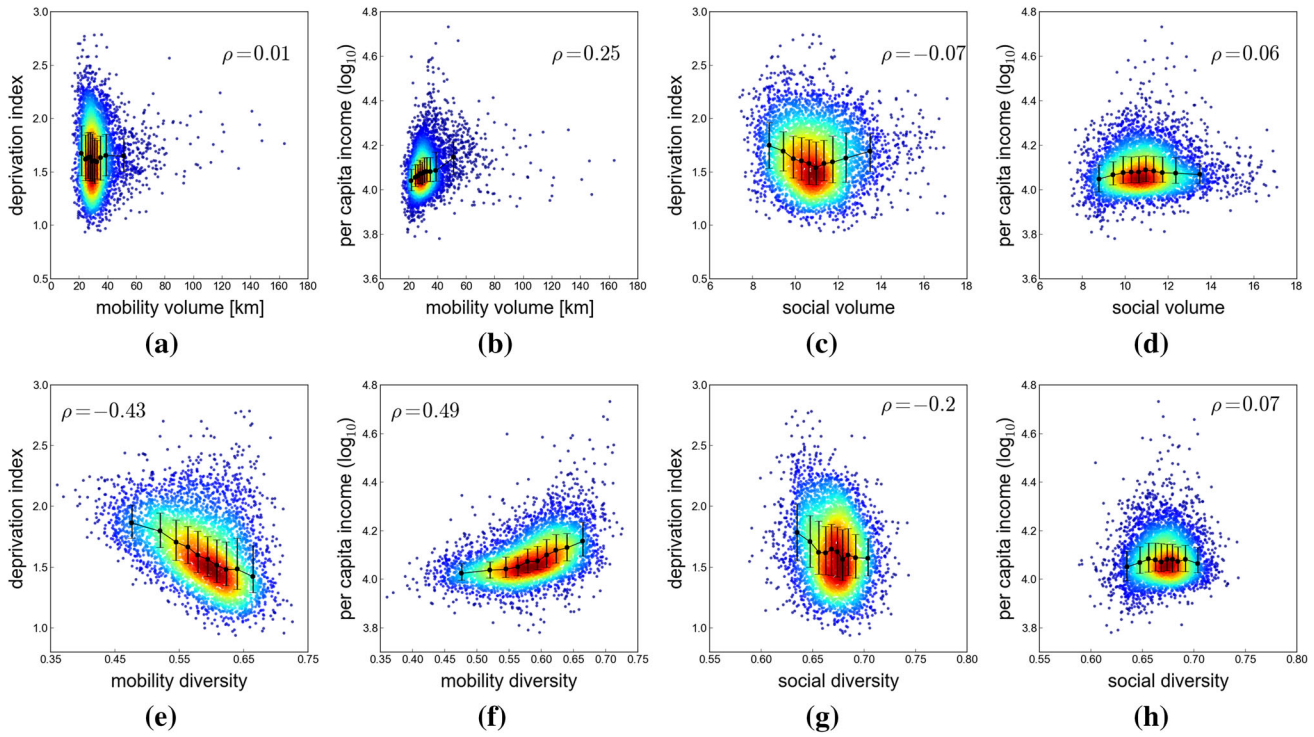

Fig. 7 The relation between the aggregated diversity measures and the socio-economic indicators: **a** mobility volume vs deprivation index; **b** mobility volume vs per capita income; **b** social volume vs deprivation index; **d** social volume vs per capita income; **e** mobility diversity vs deprivation index; **f** mobility diversity vs per capita income; **g** social diversity vs deprivation index; **h** social diversity vs per capita income. The *color of a point* indicates, in a gradient from *blue* to *red*, the density of points around it. We split the municipalities into ten equal-sized groups according to the deciles of the measures on the x axis. For each group, we compute the mean and the standard deviation of the measures on the y axis and plot them through the *black error bars*. ρ indicates the Pearson correlation coefficient between the two measures. In all the cases the p value of the correlations is <0.001 (Color figure online)

from the dataset and assign them to a random municipality with a population of N users. We then aggregate the individual diversity measures of the users assigned to the same municipality. We repeat the process 100 times and take the mean of the aggregated values of each municipality produced in the 100 experiments. In null model NM2 we randomly shuffle the values of the socio-economic indicators over the municipalities. We perform this procedure 100 times and take, for each municipality, the mean value of the socio-economic indicators computed over the 100 produced values.

In contrast with empirical data we find no correlation in the null models between the diversity measures and the socio-economic indicators, neither for mobility diversity nor for social diversity (Fig. 9). We obtain the same zero correlation when performing the null models on the two volume measures. Such a clear difference between the correlations observed over empirical data and the absence of correlations in observations on randomized data allows us to reject the hypothesis that our findings are obtained by chance.

6 Predictive models

In this section, we instantiate step (e) in Fig. 1 by learning and validating both regression models (Sect. 6.1) and classification models (Sect. 6.2) to predict the external socio-economic indicators from the aggregated measures.[2]

6.1 Regression models

To learn more about the relationship between the aggregated measures and the socio-economic indicators we implement two multiple regression models M1 and M2. We use deprivation index as dependent variable in model M1 and per capita income as dependent variable for model M2. The four aggregated measures and population density serve as regressors for both models. We determine the regression line using the least squared method. The model M1 for deprivation index

[2] The aggregated data and the code (in R) to replicate the experiments can be requested by sending an email to zbigniew.smoreda@orange.com and lpappalardo@di.unipi.it.

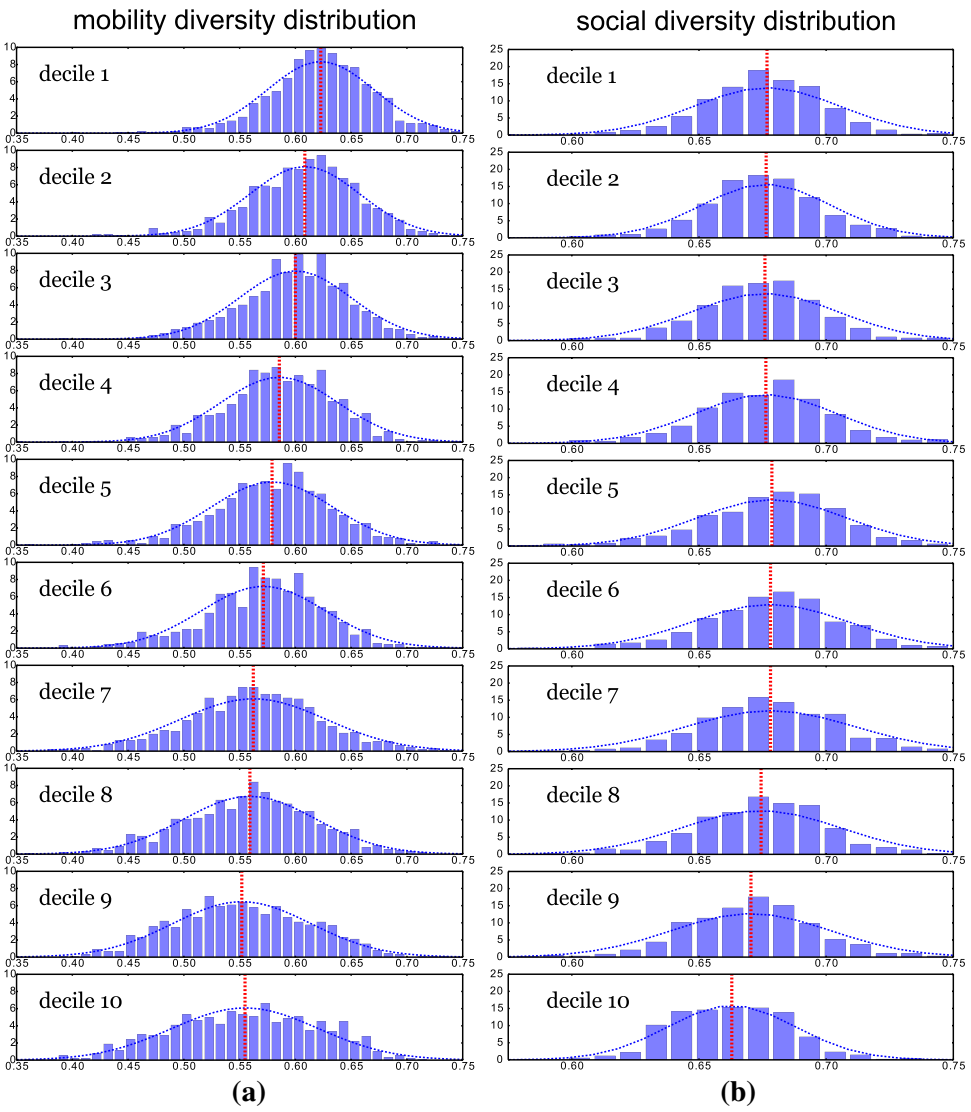

Fig. 8 The distribution of mobility diversity (**a**) and social diversity (**b**) in the deciles of deprivation index. We split the municipalities into ten equal-sized groups computed according to the deciles of deprivation index. For each group, we plot the distributions of mean mobility diversity and mean social diversity. The *blue dashed curve* is a fit of the distribution, and the *red dashed line* is the mean of the distribution (Color figure online)

produces a coefficient of determination $R^2 = 0.43$, meaning that regressors explain 43 % of the variation in the deprivation index. The model M2 for per capita income explains 25 % of the variation in the per capita income producing a coefficient of determination $R^2 = 0.25$. Tables 3 and 4 show the coefficients of the regression equations, the standard error of the coefficients and the p values of the regressors for model M1 and model M2, respectively. For both model M1 and M2 we have verified the absence of multicollinearity between the regressors, the normality and the homoscedasticity of regression residuals.

We quantify the contribution of each regressor to the multiple regression model by computing a relative importance metric [25]. Figure 10 shows the relative importance of regressors produced by the LMG method [36] for both model M1 and model M2. We observe that mobility diversity holds the highest contribution to the regressions, accounting for 54 and 65 % of the importance for M1 and M2, respectively, while social diversity provides only a small contribution (0.7 % for M1 and 0.3 % for M2). Population density provides an important contribution in both models, and mobility volume is an important variable to model M2 only (20 % of the variance).

To validate the models we implement a cross-validation procedure by performing 1,000 experiments. In each experiment we randomly divide the dataset of municipalities into a training set (60 %) and a test set (40 %), compute model M1 and model M2 on the training set, and apply the obtained models on the test set. We evaluate the performance of the models on the test set using the root mean square error $RMSE = \sqrt{\sum_i^n (\hat{y}_i - y_i)^2)/n}$, where $\hat{y}_i$ is the value predicted by the model and y_i the actual value in the test set, and computing the CV(RMSE), i.e., the RMSE normalized to the mean of the observed values. Figure 11 shows the variation of adjusted R^2 and $CV(RMSE)$ across the 1000 experiments. We observe that the adjusted R^2 of the models

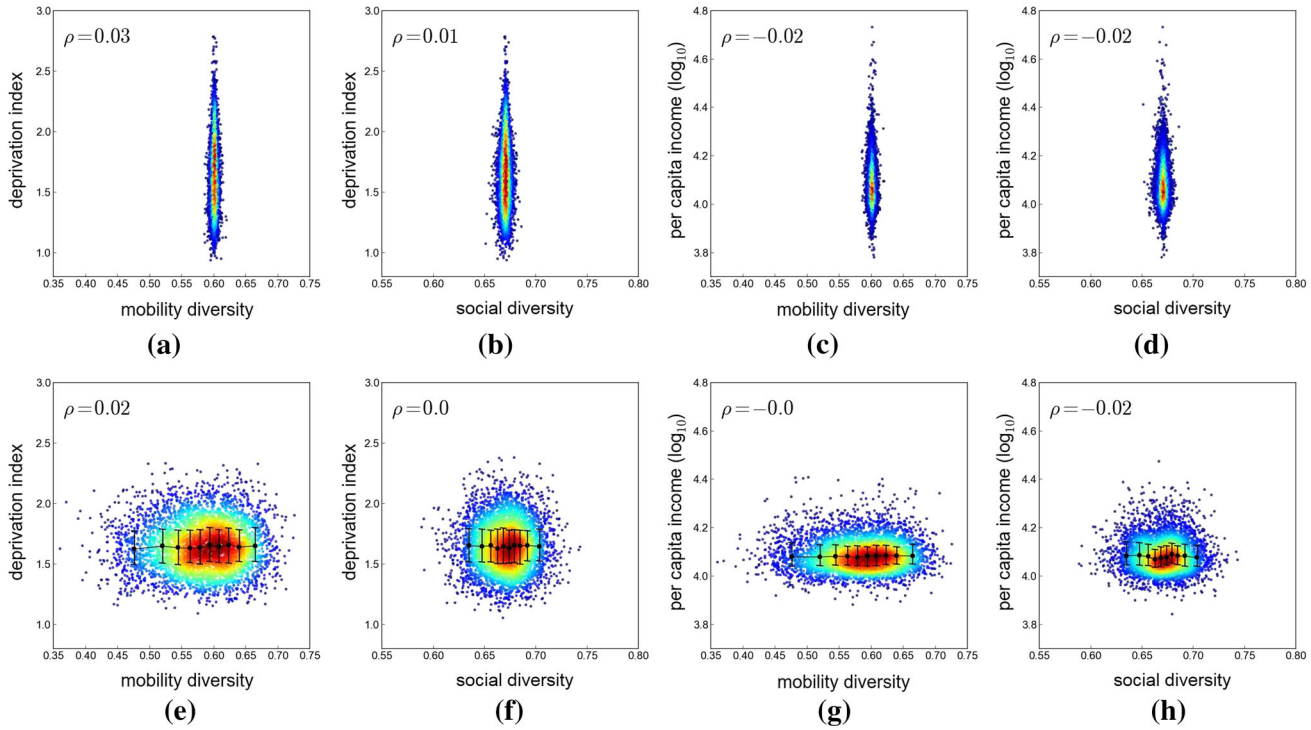

Fig. 9 The relation between the socio-economic indicators and the diversity measures computed on null model NM1 (**a**–**d**) and null model NM2 (**e**–**h**). The *color of a point* indicates, in a gradient from *blue* to *red*, the density of points around it. We split the municipalities into ten equal-sized groups according to the deciles of the measures on the x axis. For each group, we compute the mean and the standard deviation of the measure on the y axis (the *black error bars*) (Color figure online)

Table 3 The linear regression model M1 for deprivation index

	Coefficients	Std. error	p value
Model M1 (deprivation index), $R^2 = 0.4267$			
PD	0.247	0.005	$<2\times10^{-16}$
MD	−2.980	0.0575	$<2\times10^{-16}$
SD	−2.153	0.2027	$<2\times10^{-16}$
MV	0.002	0.0002	5.35×10^{-16}
SV	0.006	0.0027	0.013
Intercept	4.078	0.1281	$<2\times10^{-16}$

The *coefficients* column specifies the value of slope calculated by the regression. The *std. error* column measures the variability in the estimate for the coefficients. The p value column shows the probability the variable is not relevant

Table 4 The linear regression model M2 for per capita income

	Coefficients	Std. error	p value
Model M2 (per capita income), $R^2 = 0.25$			
PD	781.94	74.84	$<2\times10^{-16}$
MD	22,773.47	729.05	$<2\times10^{-16}$
SD	18,451.79	2569.05	7.82×10^{-13}
MV	63.116	3.64	$<2\times10^{-16}$
SV	191.16	34.62	3.56×10^{-8}
Intercept	−18,933.66	1624.36	$<2\times10^{-16}$

The *coefficients* column specifies the value of slope calculated by the regression. The *std. error* column measures the variability in the estimate for the coefficient. The p value column shows the probability the variable is not relevant

is stable across the experiments (Fig. 11a, c) and that the error of prediction is low for both models and lower for model M1 (deprivation index).

Finally, we compare the actual values of socio-economic indicators with the values predicted by the models by computing the relative error, i.e., for each municipality i we compute $(\hat{y}_i - y_i)/y_i$. We observe that the mean relative error computed across the municipalities is close to zero for both model M1 and model M2 (Fig. 12).

We also implement generalized nonlinear models (GNM) and nonlinear least squares (NLS) to predict DI and PCI, observing no significant difference in terms of residual standard error with respect to model M1 and model M2 (see "Appendix 2").

6.2 Classification models

Here, instead of predicting the value of deprivation or per capita income of municipalities we want to *classify* the level

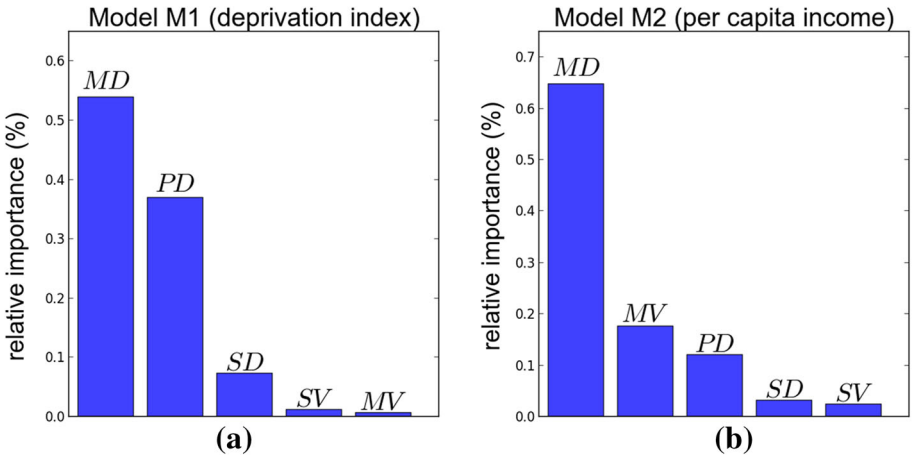

Fig. 10 The relative importance of the aggregated measures in the multiple regression models M1 (**a**) and M2 (**b**). We use the Lindeman, Merenda and Gold (LMG) method [36] to quantify an individual regressor's contribution to the model. We observe that mobility diversity is the most important variable with a contribution of about 54 and 65 % for model M1 and M2, respectively

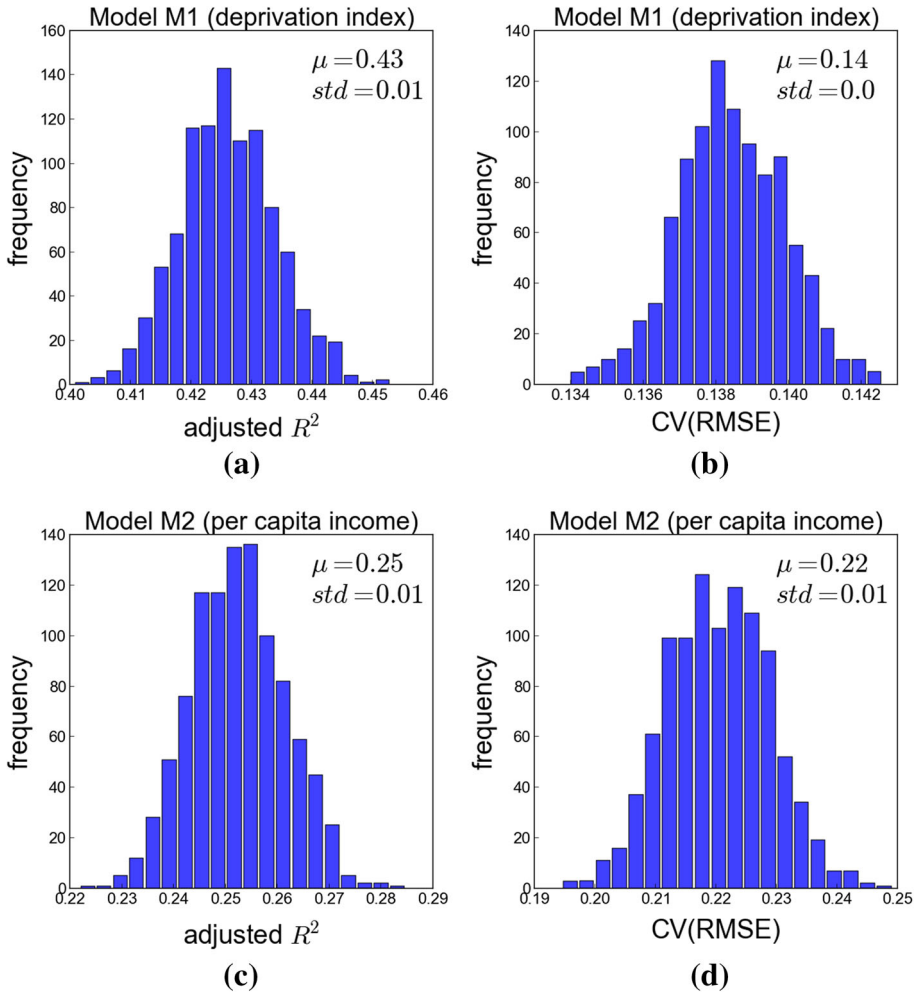

Fig. 11 Validation of regression models. We perform 1000 experiments in each of which we first divide randomly the dataset into a training set (60 %) and a test set (40 %), then we learn the model on the training set and evaluate it on the test set. **a** The distribution of the adjusted coefficient of determination R^2 across the experiments for model M1. **b** The distribution of the root mean square error (RMSE) across the experiments for model M1. **c** The distribution of the adjusted R^2 across the experiments for model M2. **d** The distribution of RMSE for model M2

of socio-economic development of municipalities. To this purpose we build two supervised classifiers C1 and C2 that assign each municipality to one of three possible categories: low level, medium level or high level of deprivation index (classifier C1) or per capita income (classifier C2). To transform the two continuous measures deprivation index and per capita income into discrete variables we partition the range of values using the 33th percentile of the distribution. This produced, for each variable to predict, three equal-populated classes. We perform the classification using Random Forest classifiers on a training set (60 % of the dataset) and validate the results on a test set (40 % of the dataset). Classifier C1 for deprivation index reaches an overall accuracy of 0.61, while the overall accuracy of classifier C2 for

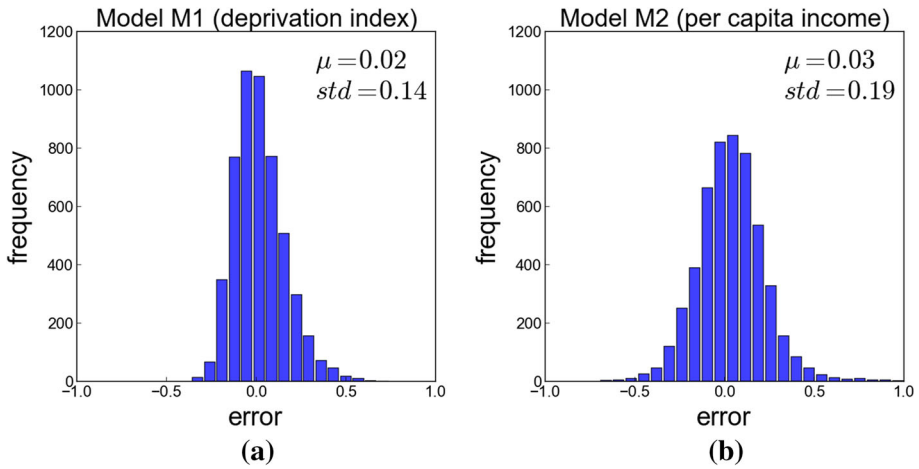

Fig. 12 Distribution of the relative error $(\hat{y} - y)/y$ across the municipalities for regression models M1 (**a**) and M2 (**b**)

Table 5 Statistics by class for classifier C1 (deprivation index) and classifier C2 (per capita income)

	Recall	Precision
Model C1: accuracy = 0.61		
Low deprivation	0.6230	0.6657
Medium deprivation	0.4970	0.4918
High deprivation	**0.7089**	**0.6721**
Model C2, accuracy = 0.54		
Low income	0.6098	0.5700
Medium income	0.3590	0.3993
High income	**0.6552**	**0.6376**

Recall is the number municipalities for which the classifier predicts the correct class divided by the number of municipalities in that class. Precision is the number of municipalities for which the classifier predicts the correct class divided by the number of municipalities the classifier predicts to be in that class. We observe that the class 'high' is the best predicted class (the corresponding results are in bold in the table)

per capita income is 0.54, against a random case accuracy of 0.33. Table 5 shows precision, recall and overall accuracy reached by classifier C1 and classifier C2 on the three classes of socio-economic development. For both models, we achieve the highest recall and precision for the 'high' class (Table 5).

We also evaluate the importance of every aggregated measure in classifying the level of socio-economic development of municipalities, using the Mean Decrease Gini measure. Similarly to the Relative Importance metrics for the regression models, in both classifier C1 and classifier C2 the mobility diversity has the highest importance, followed by population density (Fig. 13).

7 Discussion of results

The implementation of the analytical framework on mobile phone data produces three remarkable results.

First, the use of the measures of mobility and social behavior together with the standard and commonly available socio-demographic information actually *adds predictive power* with respect to the external indicators. Indeed, while a univariate regression that predicts deprivation index from

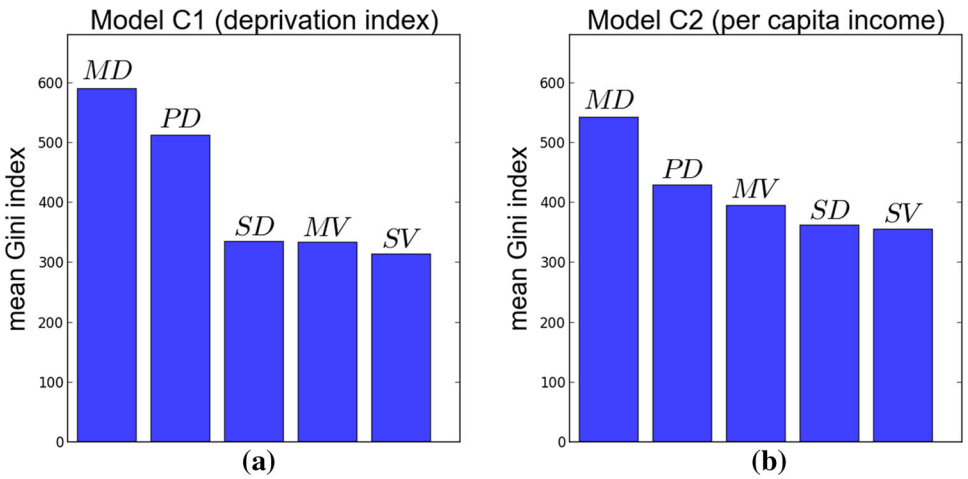

Fig. 13 The mean decrease in Gini coefficient of the variables used to learn the classifiers, for deprivation index (**a**) and per capita income (**b**). The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of nodes and leaves in the resulting random forest classifier

population density is able to explains only 11 % of the variance, we can explain 42 % of the variance by adding the four behavioral measures extracted from mobile phone data (see Table 3). This outcome suggests that mobile phone data are able to provide precise and realistic measurements of the behavior of individuals in their complex social environment, which can be used within a knowledge infrastructure like our analytical framework to monitor socio-economic development.

Second, the *diversification of human movements* is the most important aspect for explaining the socio-economic status of a given territory, far larger than the diversification of social interactions and demographic features like population density. This result, which is evident from both the correlations analysis and the contribution of mobility diversity in the models (Figs. 7, 10, 13), is also important for practical reasons. Mobile phone providers do not generally release, for privacy reasons, information about the call interactions between users, i.e., the social dimension. Our result shows that this is a marginal problem since the social dimension has a lower impact to the quality of the models than the mobility dimension (Figs. 10, 13). Hence, the implementation of our analytical framework guarantees reliable results even when the social dimension is not available in the data.

The interpretation of the observed relation between mobility diversity and socio-economic indicators is, without a doubt, two-directed. It might be that a well-developed territory provides for a wide range of activities, an advanced network of public transportation, a higher availability and diversification of jobs, and other elements that foster mobility diversity. As well as it might be that a higher mobility diversification of individuals lead to a higher socio-economic development as it could nourish economy, establish economic opportunities and facilitate flows of people and goods. In any case this information is useful for policy makers, because a difference in the diversification of individual movements is linked to a difference into the socio-economic status of a territory.

Third remarkable result is that our regression and classification models exhibit good performance when used to predict the socio-economic development of other municipalities, whose data where not used in the learning process (Fig. 11; Table 5). This result is evident from the cross-validation procedure: The accuracy and the prediction errors of the models are not dependent on the training and test set selected. The models hence give a real possibility to continuously monitor the socio-economic development of territories and provide policy makers with an important tool for decision making.

8 Conclusions and future works

In this paper we design an analytical framework that uses mobile phone to extract meaningful measures of human behavior and estimate indicators for socio-economic development. Many steps compose this complex and computationally intensive nowcasting process: data cleaning, behavioral measures computation, user/territory mapping, territorial aggregation of measures, model construction and model evaluation. We apply the analytical framework on a nationwide mobile phone dataset in France covering several weeks and find that the diversification of human movements is the best proxy for indicators of socio-economic development. We know that bio-diversity is crucial to the health of natural ecosystems, that the diversity of opinion in a crowd is essential to answer difficult questions [22] and that the diversity of social contacts is associated with socio-economic indicators of well-being [17]. The story narrated in this paper suggests that diversity is a relevant concept also in mobility ecosystems: The diversity of human mobility may be a reliable indicator of the variety of human activities, and a mirror of some aspects of socio-economic development and well-being.

We are aware that the computation of individual measures on CDR data (step (a) and (b) in Fig. 1) present privacy issues. An important next step will be to incorporate a privacy-by-design approach. We intend to use a method to assess the privacy risk of users in order to detect risk cases where the privacy of users is violated and apply privacy enhancing techniques for data anonymization [39].

In our experiments we compare the measures of mobility and sociality with two external socio-economic indicators: per capita income and deprivation index. Per capita income is a simple indicator indicating the mean income of individuals resident in a given municipality, without any information about the distribution of the wealth and the inequality. In contrast deprivation index is a composite indicator obtained as linear combination of several different variables regarding economic and ecological aspects (see “Appendix 1”). It would be interesting, as future work, to investigate the relation between the behavioral measures and the socio-economic development in a multidimensional perspective, using the single variables composing the deprivation index to understand which are the aspects of socio-economic development that best correlate with the measures of human behavior. This multidimensional approach is fostered by recent academic research and a number of concrete initiatives developed around the world [6,27] which state that the measurement of well-being should be based on many different aspects besides the material living standards (income, consumption

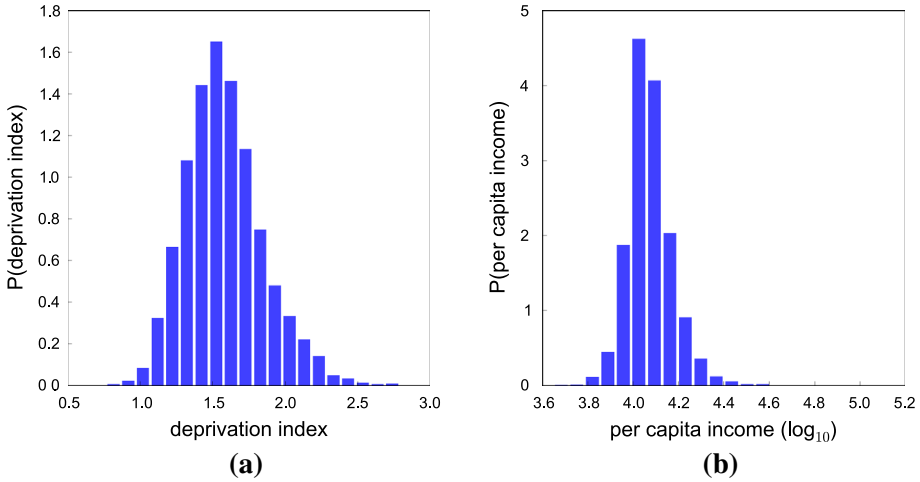

Fig. 14 Distribution of deprivation index (**a**) and logarithm of per capita income (**b**) across French municipalities

and wealth): health, education, personal activities, governance, social relationships, environment, and security. All these dimensions shape people's well-being, and yet many of them are missed by conventional income measures. Official statistics institutions are incorporating questions to capture people's life evaluations, hedonic experiences and priorities in their own surveys (see for example the Italian BES project developed by Italian National Statistics Bureau [6]). When these measures will become available, they will allow us to refine our study on the relation between measures extracted from Big Data and the socio-economic development of territories.

In the meanwhile, experiences like ours may contribute to shape the discussion on how to measure some of the aspects of socio-economic development with Big Data, such as mobile phone call records, that are massively available everywhere on earth. If we learn how to use such a resource, we have the potential of creating a digital nervous system in support of a generalized and sustainable development of our societies. This is crucial because the decisions of citizens and policy makers depend on what we measure, how good our measurements are and how well our measures are understood.

Acknowledgments The authors would like to thank Orange for providing the CDR data, Giovanni Lima and Pierpaolo Paolini for the contribution developed during their master theses. We also thank Cezary Ziemlicki for his invaluable support. We are grateful to Carole Pornet and colleagues for providing the socio-economic indicators and for computing the deprivation index for the French municipalities. This work has been partially funded by the following European projects: Cimplex (Grant Agreement 641191), PETRA (Grant Agreement 609042), SoBigData RI (Grant Agreement 654024).

Appendix 1

As described by Pornet et al. [52], the value of deprivation index for French municipalities is calculated in the following way:

$$\begin{aligned}\text{deprivation} = {} & 0.11 \times \text{Overcrowding} \\ & + 0.34 \times \text{No access to electric heating} \\ & + 0.55 \times \text{Non-owner} \\ & + 0.47 \times \text{Unemployment} \\ & + 0.23 \times \text{Foreign nationality} \\ & + 0.52 \times \text{No access to a car} \\ & + 0.37 \times \text{Unskilled worker-farm worker} \\ & + 0.45 \times \text{Household with 6 + persons} \\ & + 0.19 \times \text{Low level of education} \\ & + 0.41 \times \text{Single-parent household.}\end{aligned}$$

Appendix 2

We compare linear regression models M1 and M2 to generalized nonlinear models (GNM) and nonlinear least squares (NLS) to predict both DI and PCI. Table 6 shows the

Table 6 The residual standard error of linear and nonlinear models to predict DI and PCI

	LM	GNM	NLS
DI	0.219	0.219	0.231
PCI	2779	2779	2811

LM indicates model M1 (for DI) and model M2 (for PCI) described in Sect. 6.1. GNM indicates generalized nonlinear models. NLS indicates nonlinear least squares. The models are implemented using the R packages `lm`, `gnm` and `nls`

residual standard errors of linear and nonlinear models. We observe that the residual standard errors do not differ significantly and that, in general, nonlinear models perform equal or worse than linear models.

References

1. Amini, A., Kung, K., Kang, C., Sobolevsky, S., Ratti, C.: The impact of social segregation on human mobility in developing and urbanized regions. EPJ Data Sci. **3** (2014)
2. A world that counts: mobilizing the data revolution for sustainable development. Technical report, United Nations (2014)
3. Backstrom, L., Boldi, P., Rosa, M., Ugander, J., Vigna, S.: Four degrees of separation. In: Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12, pp. 33–42. ACM, New York, NY, USA (2012)
4. Barabasi, A.-L.: Linked: The New Science of Networks. Perseus Publishing, New York (2002)
5. Barabási, A.-L.: The origin of bursts and heavy tails in human dynamics. Nature **435**, 207 (2005)
6. Bes: il benessere equo e sostenibile in italia. Technical report, ISTAT (2014)
7. Blondel, V.D., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis (2015). arXiv:1502.03406
8. Blumenstock, J.: Calling for better measurement: Estimating an individual's wealth and well-being. In: ACM KDD (Data Mining for Social Good) (2014)
9. Brea, J., Burroni, J., Minnoni, M., Sarraute, C.: Harnessing mobile phone social network topology to infer users demographic attributes. In: Proceedings of the 8th Workshop on Social Network Mining and Analysis, SNAKDD'14. ACM (2014)
10. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'11, pp. 1082–1090. ACM (2011)
11. Cintia, P., Pappalardo, L., Pedreschi, D.: Engine matters: A first large scale data driven study on cyclists' performance. In: Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on, pp. 147–153. IEEE (2013)
12. Cintia, P., Pappalardo, L., Pedreschi, D., Giannotti, F., Malvaldi, M.: The harsh rule of the goals: data-driven performance indicators for football teams. In: Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA'15. IEEE (2015)
13. Costanza, R., Kubiszewski, I., Giovannini, E., Lovins, H., McGlade, J., Pickett, K.E., Ragnarsdóttir, K.V., Roberts, D., De Vogli, R., Wilkinson, R.: Development: time to leave GDP behind. Nature **505**(7483), 283–285 (2014)
14. Daas, P.J.H., Puts, M.J., Buelens, B.: Big data and official statistics. In: The 2013 New Techniques and Technologies for Statistics Conference (2013)
15. Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J., Krings, G., Gutierrez,T., Blondel, V.D., Luengo-Oroz, M.A.: Estimating food consumption and poverty indices with mobile phone data. CoRR (2014). arXiv:1412.2595
16. Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tandem, A.J.: Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. (PNAS) **111**(45), 15888–15893 (2014)
17. Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. Science **328**(5981), 1029–1031 (2010)
18. Eagle, N., Pentland, A.S.: Eigenbehaviors: identifying structure in routine. Behav. Ecol. Sociobiol. **63**(7), 1057–1066 (2009)
19. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
20. Frias-martinez, V., Soto, V., Virseda, J., Frias-martinez, E.: Can cell phone traces measure social development? In: Third Conference on the Analysis of Mobile Phone Datasets, NetMob (2013)
21. Furletti, B., Gabrielli, L., Giannotti, F., Milli, L., Nanni, M., Pedreschi, D., Vivio, R., Garofalo,G.: Use of mobile phone data to estimate mobility flows. measuring urban population and inter-city mobility using big data in an integrated approach. In: 47th SIS Scientific Meeting of the Italian Statistica Society, Cagliari, 06/2014 (2014)
22. Galton, F.: Vox populi. Nature **75**(7), 450–451 (1907)
23. Giannotti, F., Pedreschi, D., Pentland, A., Lukowicz, P., Kossmann, D., Crowley, J.L., Helbing, D.: A planetary nervous system for social mining and collective awareness. EPJ Spec. Top. **214**, 49–75 (2012)
24. González, M.C., Hidalgo, C.A., Barabási, A.-L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008)
25. Groemping, U.: Relative importance for linear regression in r: the package relaimpo. J. Stat. Softw. **17**(1), 1–27 (2006)
26. Gutierrez, T., Krings, G., Blondel, V.D.: Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. CoRR (2013). arXiv:1309.4496
27. Helbing, D., Balietti, S.: How to create an innovation accelerator. EPJ Spec. Top. **195**(1), 101–136 (2011)
28. Hidalgo, C.A., Rodriguez-Sickert, C.: The dynamics of a mobile phone network. Phys. A **387**(12), 3017–3024 (2008)
29. Iovan, C., Olteanu-Raimond, A.-M., Couronn, T., Smoreda, Z.: Moving and calling: Mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In: Springer, editor, 16th International Conference on Geographic Information Science (AGILE'13), May (2013)
30. Indicators and a monitoring framework for the sustainable development goals. Technical report, United Nations (2015)
31. Jiang, S., Jr, J.F., González, M.: Clustering daily patterns of human activities in the city. Data Min. Knowl. Discov. **25**, 478–510 (2012)
32. Karamshuk, D., Boldrini, C., Conti, M., Passarella, A.: Human mobility models for opportunistic networks. IEEE Commun. Mag. **49**(12), 157–165 (2011)
33. Kwan, M.-P.: Gender, the home-work link, and space–time patterns of nonemployment activities. Econ. Geogr. **75**(4), 370–394 (1999)
34. Leskovec, J., Horvitz, E.: Planetary-scale views on a large instant-messaging network. In: WWW, pp. 915–924. ACM (2008)
35. Liao, L., Patterson, D.J., Fox, D., Kautz, H.: Learning and inferring transportation routines. Artif. Intell. **171**(5–6), 311–331 (2007)
36. Lindeman, R., Merenda, P., Gold, R.: Introduction to Bivariate and Multivariate Analysis. Scott Foresman, Glenview (1980)
37. Lotero, L., Cardillo, A., Hurtado, R., Gomez-Gardenes, J.: Several multiplexes in the same city: the role of socioeconomic differences in urban mobility. SSRN 2507816 (2014)
38. Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., Gabrielli, L.: Small area model-based estimators using big data sources. J. Off. Stat. **31**(2), 263–281 (2015)
39. Monreale, A., Rinzivillo, S., Pratesi, F., Giannotti, F., Pedreschi, D.: Privacy-by-design in big data analytics and social mining. EPJ Data Sci. **10** (2014). doi:10.1140/epjdss13688-014-0010-4
40. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**(2), 167–256 (2003)
41. Onnela, J., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., Kertesz, J., Barabasi, A.L.: Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. USA **104**(18), 7332–7336 (2007)

42. Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., Pentland, A.: Urban characteristics attributable to density-driven tie formation. Nat. Commun. **4**, 1961 (2013). doi:10.1038/ncomms2961
43. Pappalardo, L., Rinzivillo, S., Pedreschi, D., Giannotti, F.: Validating general human mobility patterns on gps data. In: Proceedings of the 21th Italian Symposium on Advanced Database Systems, (SEBD2013) (2013)
44. Pappalardo, L., Rinzivillo, S., Qu, Z., Pedreschi, D., Giannotti, F.: Understanding the patterns of car travel. EPJ Spec. Top. **215**(1), 61–73 (2013)
45. Pappalardo, L., Rinzivillo, S., Simini, F.: Human mobility modelling: exploration and preferential return meet the gravity model. Procedia Comput. Sci. **83**, 934–939 (2016). The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016)/The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016)/Affiliated Workshops
46. Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F.: Comparing general mobility and mobility by car. In: Proceedings of the 2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence, BRICS-CCI-CBIC '13, pp. 665–668. IEEE Computer Society, Washington, DC, USA, (2013)
47. Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F. Barabási,A.-L.: Returners and explorers dichotomy in human mobility. Nat. Commun. **6**, (8166) (2015). doi:10.1038/ncomms9166
48. Pappalardo, L., Smoreda, Z., Pedreschi, D., Giannotti, F.: Using big data to study the link between human mobility and socio-economic development. In: Proceedings of the IEEE International Conference on Big Data (2015)
49. Pennacchioli, D., Coscia, M., Rinzivillo, S., Giannotti, F., Pedreschi, D.: The retail market as a complex system. EPJ Data Sci. **3**(1), 33 (2014)
50. Pennacchioli, D., Coscia, M., Rinzivillo, S., Pedreschi, D., Giannotti, F.: Explaining the product range effect in purchase data. In: Proceedings of the IEEE International Conference on Big Data, IEEE Big Data 2015, pp. 648–656 (2013)
51. Phithakkitnukoon, S., Smoreda, Z., Olivier, P.: Socio-geography of human mobility: a study using longitudinal mobile phone data. PLoS One **7**(6), e39253,06 (2012)
52. Pornet, C., Delpierre, C., Dejardin, O., Grosclaude, P., Launay, L., Guittet, L., Lang, T., Launoy, G.: Construction of an adaptable european transnational ecological deprivation index: the french version. J. Epidemiol Community Health **66**(11), 982–989 (2012)
53. Ranjan, G., Zang, H., Zhang, Z.-L., Bolot, J.: Are call detail records biased for sampling human mobility? SIGMOBILE Mob. Comput. Commun. Rev. **16**(3), 33–44 (2012)
54. Rinzivillo, S., Gabrielli, L., Nanni, M., Pappalardo, L., Pedreschi, D., Giannotti, F.: The purpose of motion: Learning activities from individual mobility networks. In: Proceedings of the 2014 International Conference on Data Science and Advanced Analytics, DSAA'14 (2014)
55. Rinzivillo, S., Mainardi, S., Pezzoni, F., Coscia, M., Pedreschi, D., Giannotti, F.: Discovering the geographical borders of human mobility. Künstliche Intell. **26**(3), 253–260 (2012)
56. Simini, F., González, M.C., Maritan, A., Barabási, A.-L.: A universal model for mobility and migration patterns. Nature **484**(7392), 96–100 (2012)
57. Smith-Clarke, C., Mashhadi, A., Capra, L.: Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 511–520. ACM (2014)
58. Song, C., Koren, T., Wang, P., Barabási, A.-L.: Modelling the scaling properties of human mobility. Nat. Phys. **6**(10), 818–823 (2010)
59. Song, C., Qu, Z., Blumm, N., Barabási, A.-L.: Limits of predictability in human mobility. Science **327**(5968), 1018–1021 (2010)
60. Struijs, P., Daas, P.J.H.: Quality approaches to big data in official statistics. In: European Conference on Quality in Official Statistics (2014)
61. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabási, A.-L.: Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, pp. 1100–1108. ACM, New York, NY, USA (2011)
62. Yan, X.-Y., Zhao, C., Fan, Y., Di, Z., Wang, W.-X.: Universal predictability of mobility patterns in cities. J. R. Soc. Interface **11**(100) (2014). http://dx.doi.org/10.1098/rsif.2014.0834