Using Big Data to study the link between human mobility and socio-economic development

Luca Pappalardo
Department of Computer Science
University of Pisa, Italy
Email: lpappalardo@di.unipi.it

Zbigniew Smoreda
SENSE
Orange Lab, France
Email: zbigniew.smoreda@orange.com

Dino Pedreschi
Department of Computer Science
University of Pisa, Italy
Email: pedre@di.unipi.it

Fosca Giannotti
Institute of Information Science and Technologies
National Research Council (CNR), Italy
Email: fosca.giannotti@isti.cnr.it

***Abstract*—Big Data offer nowadays the potential capability of creating a digital nervous system of our society, enabling the measurement, monitoring and prediction of relevant aspects of socio-economic phenomena in quasi real time. This potential has fueled, in the last few years, a growing interest around the usage of Big Data to support official statistics in the measurement of individual and collective economic well-being. In this work we study the relations between human mobility patterns and socio-economic development. Starting from nation-wide mobile phone data we extract a measure of mobility volume and a measure of mobility diversity for each individual. We then aggregate the mobility measures at municipality level and investigate the correlations with external socio-economic indicators independently surveyed by an official statistics institute. We find three main results. First, aggregated human mobility patterns are correlated with these socio-economic indicators. Second, the *diversity of mobility*, defined in terms of entropy of the individual users' trajectories, exhibits the strongest correlation with the external socio-economic indicators. Third, the volume of mobility and the diversity of mobility show opposite correlations with the socio-economic indicators. Our results, validated against a null model, open an interesting perspective to study human behavior through Big Data by means of new statistical indicators that quantify and possibly "nowcast" the socio-economic development of our society.**

I. Introduction

The Big Data originating from the digital breadcrumbs of human activities, sensed as a by-product of the ICT systems we use everyday, allow us to scrutinize the ground truth of individual and collective behavior at an unprecedented detail [39]. Multiple dimensions of our social life have Big Data proxies nowadays. Our social relationships leave traces in the network of our phone or email contacts, in the friendship links of our favorite social networking site. Our shopping patterns leave traces in the transaction records of our purchases. Our movements leave traces in the records of our mobile phone calls, in the GPS tracks of our on-board navigation systems. Sensing Big Data at a societal scale has the potential of providing a powerful social microscope, which can help us understand many complex and hidden socio-economic phenomena. Such challenge clearly requires high-level analytics, modeling and reasoning across all the social dimensions above, an activity that it is often referred to as "social mining": the task of making sense of Big Data by extracting meaningful information from large, messy and noisy data [17]. In recent years, also stimulated by national official statistics institutes and the United Nations [41], researchers from different disciplines have started to use Big Data and social mining to support official statistics in the measurement of individual and collective well-being [8][37]. The majority of works in literature focus on the analysis of mobile phone data to study the relations between communications patterns and well-being [11][4]. In this paper we analyze Big Data from mobile phones to study the link between individuals' mobility patterns and the socio-economic development of cities. We try to answer the following intriguing question: Can we monitor and possibly predict the socio-economic development of cities just by observing human movements of their residents through the lens of Big Data? The answer to this fascinating question, as we show in this paper, is related to the concepts of *mobility volume* and *mobility diversity*.

We know that bio-diversity is crucial to the health of natural ecosystems and for the balance, or well-being, of plant and animal species that inhabit them. Diversity is a key concept also for the social ecosystems: from Francis Galton who showed that the diversity of opinion in a crowd is essential to answer difficult questions [15] to more recent works showing that the diversity of social contacts is associated to socio-economic indicators of well-being [11][19][4], social diversity has proven to be essential in many contexts [38][23]. In this paper we argue that diversity is a key concept also for the mobility ecosystem, and that the volume and the diversity of mobility patterns have high predictive power with respect to the socio-economic development of cities.

Starting from large-scale mobile phone data we quantify the relations between human mobility and socio-economic development in France using municipality-level official statistics as external comparison measurements. We first define two individual measures over mobile phone data which describe two aspects of individual mobility behavior: the volume of mobility, i.e. the characteristic traveled distance of an individual [18], and the diversity of mobility, i.e. the diversification of movements of an individual over her locations [36]. Each

individual measure is computed for each of the several million users in our dataset based on their locations and calls as recorded in the mobile phone data. We then aggregate the two individual measures at the level of French municipalities and explore the correlations between the aggregated measures and external indicators covering different aspects of socio-economic development: wealth, employment, education and deprivation. We find that both mobility measures correlate with the external socio-economic indicators, and in particular the measure of mobility diversity shows much stronger correlations. We validate our results against a null model which produces zero correlations allowing us to reject the hypothesis that our results occurred by chance. Finally, we observe that at municipality level mobility volume and mobility diversity show negative correlations and opposite correlations with the socio-economic indicators, suggesting that they play different roles in the socio-economic development of cities.

The importance of our findings is twofold. On one side, we show that mobility diversity and mobility volume are key concepts for the well-being of our cities that can be used to understand deeply 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 and well-being. New statistical indicators can be defined to describe the well-being of a territory, in order to support official statistics when such measurements are not possible using traditional censuses and surveys [41][27].

The paper is organized as follows. Section II revises the main works in the study of Big Data for measuring development and well-being. Section III and Section IV present respectively the mobile phone data and the socio-economic indicators we use in our study. Section V introduces the measures of mobility and explains how to compute them on the mobile phone data. In Section VI we show the main results of our study and discuss them in Section VII. Finally, Section VIII concludes the paper discussing open lines of new research.

II. Related Work

Big Data offer nowadays the potential capability of creating a digital nervous system of our society, enabling the measurement, monitoring and prediction of relevant aspects of human behavior [17]. For example the availability of massive digital traces of human whereabouts, such as GPS traces from private vehicles and mobile phone data, has offered novel insights on the quantitative patterns characterizing human mobility [6][18][16]. Studies from different disciplines document a stunning heterogeneity of human travel patterns as measured by the so-called radius of gyration [18][28], and at the same time observe a high degree of predictability as measured by the mobility entropy [36][12]. The patterns of human mobility have been used to build generative models of individual human mobility [21][29], generative models to describe human migration flows [34], methods for profiling individuals according to their recurrent and total mobility patterns [29], methods to discover geographic borders according to recurrent trips of private vehicles [33], methods to predict the formation of social ties [7][40], and classification models to predict the kind of activity associated to individuals' trips on the only basis of the observed displacements [22][20][32].

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 [8][37]. 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 [41][27]. 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 CDR (Call Detail Records) of calling and texting activity of users. Mobile phone data, indeed, guarantee the repeatability of experiments on different countries and geographical scales since they can be retrieved nowadays in every country due to their worldwide diffusion [3]. 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 [10]. Brea et al. study the structure of the social graph of mobile phone users of Mexico and propose an algorithm for the prediction of the age of mobile phone users [5]. Another recent work use mobile phone data to study inter-city mobility and develop a methodology to detect the fraction of residents, commuters and visitors within each city [14].

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 a nationwide mobile phone dataset and shows that, in the UK, regional communication diversity is positively associated to a socio-economic ranking [11]. Gutierrez et al. address the issue of mapping poverty with mobile phone data through the analysis of airtime credit purchases in Ivory Coast [19]. Blumenstock shows a preliminary evidence of a relationship between individual wealth and the history of mobile phone transactions [4]. 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 [9]. 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 [13]. 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 [24]. 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 census data [35].

Other recent works use different types of mobility 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 [30]. 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 [25].

III. Mobile Phone Data

Mobile phones are nowadays very common technological devices carried out by individuals in their daily routine, offering a good proxy to study the patterns of human mobility. In our study, we exploit the access to a dataset of Call Detail Records (CDR) gathered by Orange mobile phone operator, recording 200 million calls made during 45 days (from 2007/09/01 to 2007/10/15) by 20 million anonymized users in France. CDRs collect geographical, temporal and interaction information on mobile phone use and show a great potential to empirically investigate human dynamics on a society wide scale [18][2][26]. Each time an individual makes a call or sent a text message the mobile phone operator registers the connection between the caller and the callee, the time of the phone activity and the phone tower communicating with the served phone, allowing to reconstruct the user's time-resolved trajectory [18]. Table I shows the format of CDR data and tower location data in our dataset. To make sure that users' private information are protected, all the users are anonymized by translating their identifiers into hash formats. The item "timestamp" records the exact time of the phone activity, while "tower" is the identifier of wireless tower that is serving the caller's call or text message. The item "mode" is simply used to distinguish between calls and text messages.

(a)

caller	callee	timestamp	tower	mode
4F80460	4F80331	2007/09/10 23:34	36	call
2B01359	9H80125	2007/10/10 01:12	38	SMS
2B19935	6W1199	2007/10/10 01:43	38	call
⋮	⋮	⋮	⋮	⋮

(b)

tower	latitude	longitude
36	49.54	3.64
37	48.28	1.258
38	48.22	-1.52
⋮	⋮	⋮

TABLE I. The formats of CDR data (a) and tower data (b).

To focus on individuals with reliable statistics we carry out some preprocessing steps. First, we select only users with a call frequency higher than a threshold $f = N/45 > 0.5$, where N is the number of calls made by the user and 45 days is the length of our period of observation, we then delete all the users with less than one call every two days (in average over the observation period). The resulting dataset contains the mobility trajectories of 6 million active users.

IV. Socio-Economic Data

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 in 2007 for all the French municipalities with more than 1,000 official residents. We collect data about four aspects of socio-economic development: (i) *per capita income*, the mean income in a given municipality; (ii) *education rate*, the fraction of residents of a municipality with primary education only; (iii) *unemployment rate*, the ratio between unemployment individuals and all the residents of a municipality; (iv) *deprivation index*, constructed by selecting among variables reflecting individual experience of deprivation and combining them into a single score by a linear combination with specific choices for coefficients [31][1]:

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

Preliminary validation showed a high association between the French deprivation index and both income and education values in French municipalities, partly supporting its ability to measure socio-economic status [31]. Figure 1 shows the distribution of the four socio-economic indicators across the French municipalities.

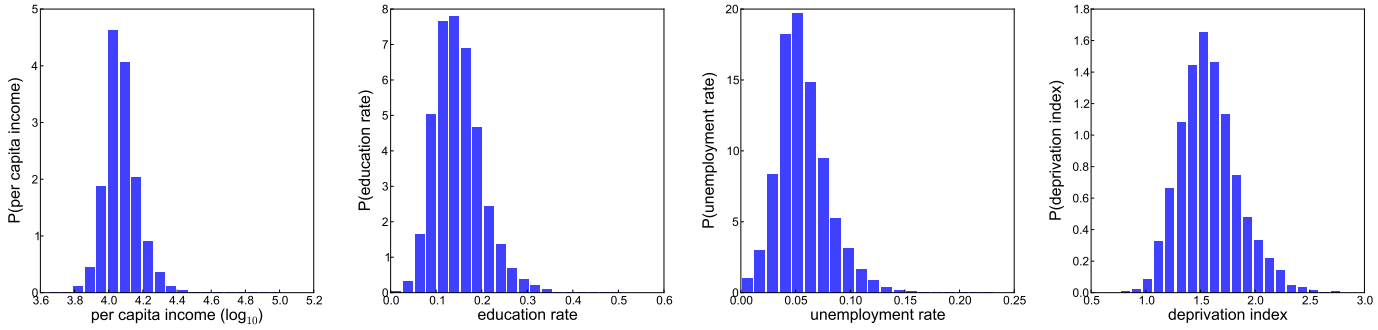
V. Measuring Human Mobility

Starting from the trajectories of an individual we consider two aspects of individual mobility: the volume of mobility, i.e. how large the typical distance traveled by an individual is, and the diversity of mobility, i.e. how the trips of an individual are distributed over the locations visited. The radius of gyration r_g is a measure of mobility volume and indicates the characteristic distance traveled by an individual [18][28][29]. It characterizes the spatial spread of the phone towers visited by an individual u from her center of mass (i.e. the weighted mean point of the phone towers visited by an individual), defined as:

$$r_g(u) = \sqrt{\frac{1}{N}\sum_{i \in L} n_i (\mathbf{r}_i - \mathbf{r}_{cm})^2} \qquad (1)$$

where L is the set of phone towers visited by the individual, 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, $\mathbf{r}_i$ and $\mathbf{r}_{cm}$ are the vectors of coordinates of phone tower i and center of mass respectively. To clarify the concept, let us consider Figure 2 which displays the radius of gyration of two individuals in our dataset. User A travels between

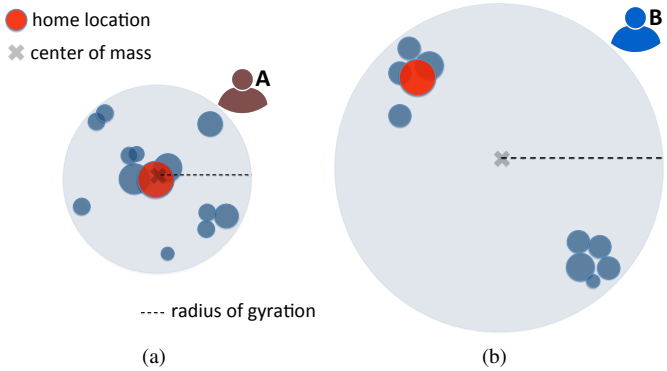
[1]The variables used to compute the deprivation index (Overcrowding, No access to electric heating, etc.) refer to socio-economic status in 2007. We used the procedure in [31] to compute the deprivation index on these variables.



(a) (b) (c) (d)

Fig. 1. **The distribution of socio-economic variables across the French municipalities.** (a) Distribution of logarithm of per capita income; (b) distribution of education rate; (c) distribution of unemployment rate; (d) distribution of deprivation index. We observe that all the distributions show clear peaks highlighting the presence of typical socio-economic values across the French municipalities.

locations that are close to each other, resulting in a low radius of gyration $r_g(A)$. In contrast, user B has a large radius of gyration since the locations she visits are far apart from each other. Figure 3a shows the distribution of radius of gyration across the individuals in our dataset. The distribution is well approximated by a heavy tail distribution indicating a large variability of the radii, a confirmation of previous results on both GSM data [18] and GPS data [28].



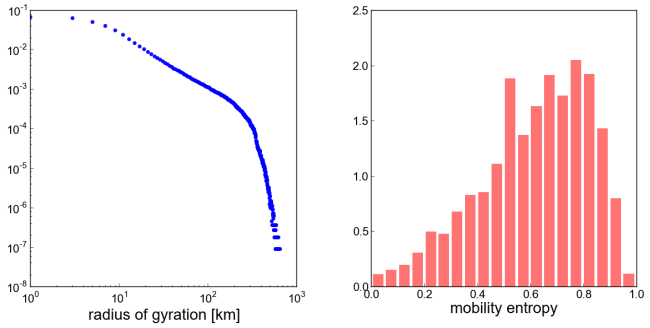
(a) (b)

Fig. 2. **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, 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, 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.

We measure the mobility diversity of an individual u by using the Shannon entropy [36]:

$$S(u) = -\frac{\sum_{e \in E} p(e) \log p(e)}{\log N} \qquad (2)$$

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. Mobility entropy is high when an individual performs many different trips from a variety of origins and destinations; it is low when

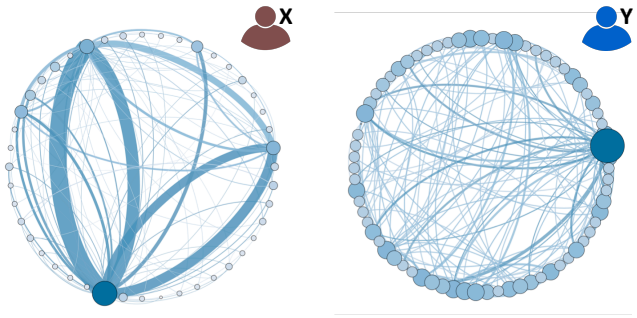


(a) (b)

Fig. 3. **The distributions of radius of gyration and mobility entropy of individuals in our dataset.** (a) Distribution of radius of gyration. We observe a heavy-tail distribution indicating a large variability of radius of gyration across the population. (b) Distribution of mobility entropy, denoting a high mean degree of unpredictability of human mobility patterns.

she performs a small number of recurring trips. To clarify the concept let us consider Figure 4 which shows a network visualization of the mobility entropy of two individuals in our dataset. In the figure nodes represent phone towers, edges represent trips between two phone towers, and the size of edges is proportional to the number of trips performed on the edge. User X has low mobility entropy since she distributes her trips on a few preferred edges. Conversely user Y has high mobility entropy because she distributes her trips across many equal-sized edges. Mobility entropy also quantifies the possibility to predict individual's future whereabouts. Individuals having a very regular movement pattern possess a mobility entropy close to zero and their whereabouts are rather predictable (the case of user X). Conversely, individuals with a high mobility entropy are less predictable (the case of user Y). Figure 3b shows the distribution of mobility entropy across the users in our dataset, and indicates a high mean degree of predictability of individual human mobility patterns [36].

The most frequented location $L_1(u)$ is the place where an individual u is found with the highest probability when stationary, most likely her home. In Figure 2 the red circles indicate $L_1(A)$ and $L_1(B)$, i.e. the phone towers where users A and B make the highest number of calls during the period of observation.



(a) (b)

Fig. 4. **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, 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.

VI. Correlation Analysis

We compute the two mobility measures for each individual on the CDR data. Due the size of the dataset, we use the MapReduce paradigm implemented by Hadoop to distribute the computation across a cluster of coordinated nodes and reduce the time of computation. We then aggregate the individual measures at the municipality level through a two-step process: (i) we assign to each user u a home location $L_1(u)$, i.e. the phone tower where the user performs the highest number of calls during nighttime (from 10 pm to 7 am) [31]; (ii) based on these home locations, we assign each user to the corresponding municipality with standard Geographic Information Systems techniques. We aggregate radius of gyration and mobility entropy at municipality level by taking the mean, median and standard deviation values across the population of users assigned to that municipality. We obtain a set of 5,100 municipalities each one with the associated two aggregated indicators.

We investigate the correlations between the aggregated mobility measures and the four external socio-economic indicators presented in Section IV. Table II summarizes the correlation between the aggregated mobility measures and the socio-economic indicators. Four main results emerge. First, mobility diversity is a better predictor for socio-economic development than mobility volume (Figure 5 and Table II). Mobility diversity indeed has much stronger correlations than mobility volume regardless the type of aggregation (Table II). Secondly per capita income, primary education rate and deprivation index show stronger correlations with the mobility measures than the unemployment rate. Third, mobility diversity and mobility volume show opposite correlations with the socio-economic indicators: where the correlation is positive for mobility diversity, the same correlation is negative for mobility volume, and vice versa. Figure 6 provides another way to observe the relations between mobility diversity and socio-economic development. We split the municipalities in deciles based on the values of deprivation index, and for each decile we compute the distributions of mobility entropy at municipality level. We observe that as the deciles of the economic values increase both the mean and the variance of the distribution change, consistently with the plots of Figure 5a.

In order to test the significance of the correlations observed on the empirical data, we compare our findings with the results produced by a null model where we randomly distribute the users over the French municipalities. We first extract uniformly N users 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. The outcomes of the null model have zero correlations with all the socio-economic indicators, allowing us to reject the hypothesis that our results occurred by chance.

measure	DI	PCI	PER	UR
mean S	**-0.43**	**0.49**	**-0.49**	**-0.17**
mean r_g	0.01	-0.25	0.01	-0.04
median S	-0.43	0.48	-0.47	-0.17
median r_g	0.16	-0.21	0.47	-0.1
std S	0.20	-0.26	0.27	0.11
std r_g	0.01	0.28	-0.21	0.13

TABLE II. Correlations between aggregated mobility measures and socio-economic indicators.

VII. Discussion of the Results

The most remarkable result in our study is the observation that human mobility, and mobility diversity in particular, is associated with socio-economic indicators on a municipality scale. To be specific, on a municipality level mobility entropy is positively correlated with per capita income and negatively correlated with deprivation index, primary education rate and unemployment rate (Figure 5). Generalizing our empirical findings, we state that a greater diversification of human mobility is linked to a higher overall wealth, to a more educated territory and to a lower level of deprivation. Remarkable is that a systematic variation of the mobility entropy distribution exists across geographical units defined on socio-economic indicators (Figure 6), delineating subpopulations where a different distribution of entropy emerges based on the occurrence of socio-economic indicators. This is an important finding when compared to Song et al. [36], a seminal work on the predictability of human mobility, which states that mobility entropy is very stable across different subpopulations delineated by personal characteristics like gender or age group. The contrast between our findings and the result of Song et al. suggests that socio-economic situations on a city scale are more related to individual mobility than individual demographic characteristics. The observed variation also suggests a relation between socio-economic development and predictability: people resident in more developed and richer territories show a higher mobility entropy and hence more unpredictable mobility patterns.

Although the relations between mobility diversity and socio-economic indicators appear clearly, it is difficult to formulate a hypothesis to explain their connections. Without a doubt, the relation between socio-economic indicators and

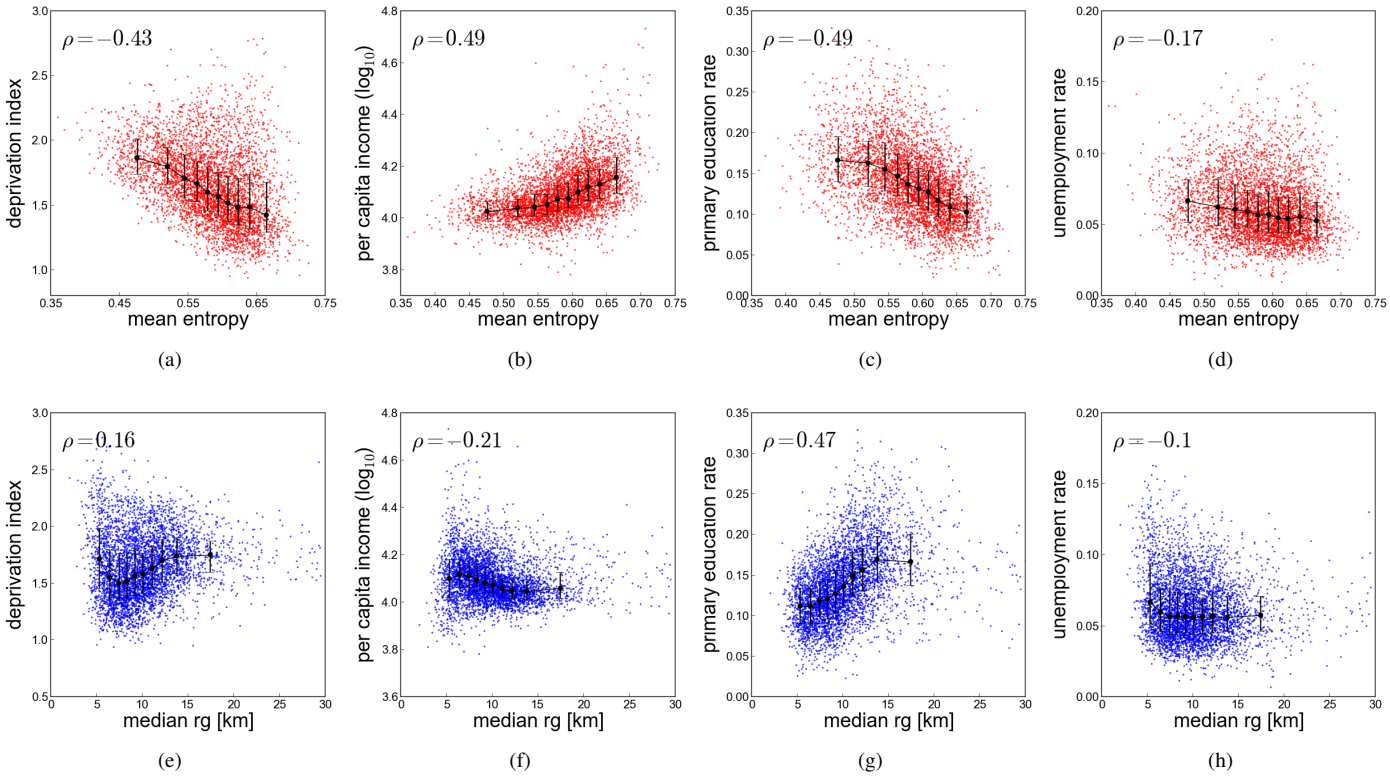


Fig. 5. **The correlations between human mobility measures and socio-economic indicators:** (a) mobility entropy vs deprivation index; (b) mobility entropy vs logarithm of per capita income; (c) mobility entropy vs education rate; (d) mobility entropy vs unemployment rate; (e) radius of gyration vs deprivation index; (f) radius of gyration vs logarithm of per capita income; (g) radius of gyration vs education rate; (h) radius of gyration vs unemployment rate. 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 < 0.001). We observe that mobility entropy has stronger correlations with socio-economic indicators than radius of gyration.

mobility diversity is 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 economic well-being as it could nourish economy, establishes economic opportunities and facilitate flows of people and goods. Interpretations of the relation between mobility diversity and socio-economic development are not directly derivable from the empirical results and should therefore be combined with more thorough theoretical insights.

Another interesting result is that mobility volume and mobility diversity show opposite correlations, i.e. high values of aggregated mobility volume correspond to low socio-economic development, while high values of aggregated mobility diversity correspond to high socio-economic development (Figure 5). Assuming that human mobility is driven by people's daily activities, a possible explanation is that people living in well developed municipalities have a wide availability of activities, resulting in high mobility diversity. In contrast, people living in less development municipalities, like municipalities in the countryside, are forced to travel in search of activities that cannot be found in their municipality, resulting in a wide mobility volume. To investigate this hypothesis we compute the correlation between the aggregated mobility diversity and the aggregated mobility volume. We find a negative correlation ($\rho = -0.38$) confirming our insight: at municipality scale high mobility diversity is linked to low mobility volume (Figure 7). We plan to investigate deeply this aspect in order to understand the reason of this interesting correlation.

VIII. Conclusion

In this paper we investigate the relationships between human mobility patterns and socio-economic development in French municipalities. Starting from nation-wide mobile phone data we extract for each individual two mobility measures: radius of gyration, the characteristic distance traveled by an individual, and mobility entropy, the diversification of movements over her locations. We then aggregate the individual mobility measures at municipality level by taking the mean, the median and the variance across the population of users assigned to each municipality. Finally, we compare the aggregated mobility measures with external socio-economic indicators measuring education level, unemployment rate, income and deprivation. We find that both mobility measures show correlations with the socio-economic indicators, and mobility entropy shows the strongest correlations. We confirm our results against a null model which produces zero correlations, allowing us to reject the hypothesis that our discovery occurred by chance. Starting from our interesting results, we plan to extend our study in three directions.

First, since mobile phone data also provide information about social interactions, it would interesting to extract mea-

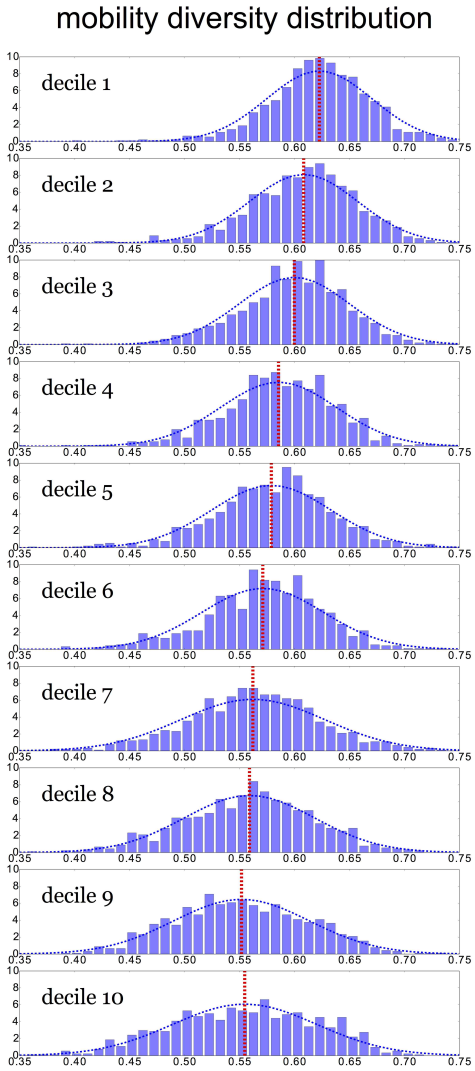


Fig. 6. **The distributions of mobility entropy in the different deciles of deprivation index.** We split the municipalities into ten equal-sized groups computed according to the deciles of deprivation index. For each group, we plot the distributions of mobility entropy. The blue dashed curve represents a fit of the distribution, the red dashed line represents the mean of the distribution. We observe a systematic variation of both mean and variance of the distribution of mobility entropy across the deciles defined by deprivation index.

sures capturing the social behavior of individuals. The seminal work by Eagle et al. showed that social diversity is a good proxy for socio-economic development of territories [11]. It would be interesting to compare the correlations produced by social diversity and mobility diversity in order to understand and quantify the different roles they play in the socio-economic development of a territory. Is mobility diversity a better proxy for socio-economic development than social diversity?

Second, to learn more about the relationship between the aggregated mobility measures and the socio-economic indicators it would be useful to implement and validate predictive models. The predictive models can be aimed at predicting the actual value of socio-economic development of the territory, e.g. by regression models, or to predict the class of socio-economic development, i.e. the level of socio-economic development of a given geographic unit as done by classification

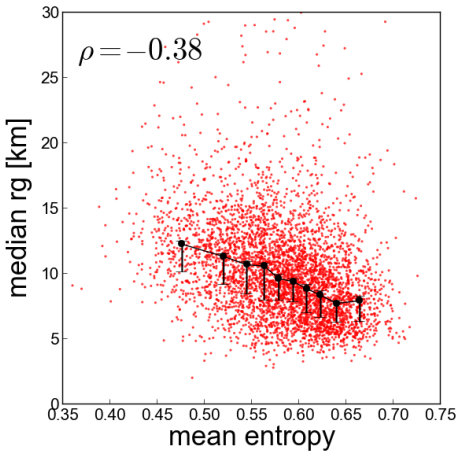


Fig. 7. **The correlation between aggregated mobility diversity and aggregated mobility volume.** We split the municipalities into ten equal-sized groups according to the deciles of the measures on the x axis. For each group, we compute the mean and the standard deviation of the measures on the y axis and plot them through the black error bars. ρ indicates the Pearson correlation coefficient between the two measures (p-value < 0.001). We observe a negative correlation suggesting that high mobility entropy is linked to low mobility volume, and vice versa.

models. If we find that the accuracy and the prediction errors of the models are not dependent on the training and test set selected, we would have a further confirmation that mobility measures extracted from Big Data give a real possibility to continuously monitor the socio-economic development of territories and provide policy makers with an important tool for decision making.

Third, we plan to investigate the relation between human mobility patterns and socio-economic development in a multidimensional perspective by including many other indicators to understand which are the aspects of socio-economic development that best correlate with the proposed mobility measures. The new indicators will allow us to refine our study on the relation between mobility 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 well-being with Big Data that are 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, sustainable development of our societies.

ACKNOWLEDGMENT

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 are grateful to Carole Pornet and colleagues for providing the socio-economic indicators and for computing the deprivation index for the French municipalities. We also thank Maarten Vanhoof and Lorenzo Gabrielli for the insightful discussions.

This work has been partially funded by projects: Cimplex (grant agreement 641191), PETRA (grant agreement 609042), SoBigData RI (grant agreement 654024).

REFERENCES

[1] A. Amini, K. Kung, C. Kang, S. Sobolevsky, and C. Ratti. The impact of social segregation on human mobility in developing and urbanized regions. *EPJ Data Science*, 3, 2014.

[2] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[3] V. D. Blondel, A. Decuyper, and G. Krings. A survey of results on mobile phone datasets analysis, 2015. cite arxiv:1502.03406.

[4] J. Blumenstock. Calling for better measurement: Estimating an individual's wealth and well-being. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'14. ACM, 2014.

[5] J. Brea, J. Burroni, M. Minnoni, and C. Sarraute. 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.

[6] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439:462, 2006.

[7] E. Cho, S. A. Myers, and J. Leskovec. 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, pages 1082–1090. ACM, 2011.

[8] P. J. H. Daas, M. J. Puts, and B. Buelens. Big data and official statistics. In *The 2013 New Techniques and Technologies for Statistics conference*, 2013.

[9] A. Decuyper, A. Rutherford, A. Wadhwa, J. Bauer, G. Krings, T. Gutierrez, V. D. Blondel, and M. A. Luengo-Oroz. Estimating food consumption and poverty indices with mobile phone data. *CoRR*, abs/1412.2595, 2014.

[10] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tandem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 111(45):15888–15893, 2014.

[11] N. Eagle, M. Macy, and R. Claxton. Network Diversity and Economic Development. *Science*, 328(5981):1029–1031, May 2010.

[12] N. Eagle and A. S. Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[13] V. Frias-martinez, V. Soto, J. Virseda, and E. Frias-martinez. Can cell phone traces measure social development? In *Third Conference on the Analysis of Mobile Phone Datasets, NetMob*, 2013.

[14] B. Furletti, L. Gabrielli, F. Giannotti, L. Milli, M. Nanni, D. Pedreschi, R. Vivio, and G. Garofalo. 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 Statistical Society*, Cagliari, June 2014.

[15] F. Galton. Vox populi. *Nature*, 75(7), 1907.

[16] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20(5):695–719, 2011.

[17] F. Giannotti, D. Pedreschi, A. Pentland, P. Lukowicz, D. Kossmann, J. L. Crowley, and D. Helbing. A planetary nervous system for social mining and collective awareness. *EPJ Special Topics*, 214:49–75, 2014.

[18] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[19] T. Gutierrez, G. Krings, and V. D. Blondel. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *CoRR*, abs/1309.4496, 2013.

[20] S. Jiang, J. F. Jr, and M. González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25:478–510, 2012.

[21] D. Karamshuk, C. Boldrini, M. Conti, and A. Passarella. Human mobility models for opportunistic networks. *IEEE Communications Magazine*, 49(12):157–165, 2011.

[22] L. Liao, D. J. Patterson, D. Fox, and H. Kautz. Learning and inferring transportation routines. *Artif. Intell.*, 171(5-6):311–331, Apr. 2007.

[23] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences (PNAS)*, 108(22), 2011.

[24] L. Lotero, A. Cardillo, R. Hurtado, and J. Gomez-Gardenes. Several multiplexes in the same city: The role of socioeconomic differences in urban mobility. *Available at SSRN 2507816*, 2014.

[25] S. Marchetti, C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli. Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31(2), 2015.

[26] J. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks. *Proceeding of the National Academy of Sciences (PNAS)*, 104(18):7332–7336, 2007.

[27] Indicators and a monitoring framework for the sustainable development goals: Launching a data revolution for the sdgs. A report by the Leadership Council of the Sustainable Development Solutions Network, 20 march 2015.

[28] L. Pappalardo, S. Rinzivillo, Z. Qu, D. Pedreschi, and F. Giannotti. Understanding the patterns of car travel. *EPJ Special Topics*, 215(1):61–73, 2013.

[29] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási. Returners and explorers dichotomy in human mobility. *Nature Communications*, 6(8166), 2015.

[30] D. Pennacchioli, M. Coscia, S. Rinzivillo, D. Pedreschi, and F. Giannotti. Explaining the product range effect in purchase data. In *IEEE International Conference on Big Data*, pages 648–656, 2013.

[31] C. Pornet, C. Delpierre, O. Dejardin, P. Grosclaude, L. Launay, L. Guittet, T. Lang, and G. Launoy. Construction of an adaptable european transnational ecological deprivation index: the french version. *Journal of Epidemiol Community Health*, 66(11):982–9, 2012.

[32] S. Rinzivillo, L. Gabrielli, M. Nanni, L. Pappalardo, D. Pedreschi, and F. Giannotti. The purpose of motion: Learning activities from individual mobility networks. In *Proceedings of International Conference on Data Science and Advanced Analytics*, DSAA'14, 2014.

[33] S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti. Discovering the geographical borders of human mobility. *Künstliche Intelligenz*, 26(3):253–260, 2012.

[34] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[35] C. Smith-Clarke, A. Mashhadi, and L. Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 511–520. ACM, 2014.

[36] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[37] P. Struijs and P. J. H. Daas. Quality approaches to big data in official statistics. In *European conference on Quality in Official Statistics*, 2014.

[38] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday Books, New York, 2004.

[39] Data, data, everywhere. The Economist, 25 February 2010.

[40] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabási. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM.

[41] A world that counts: mobilizing the data revolution for sustainable development. A report by the United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG), November 2014.