Correcting Mobility Entropy from CDR Data for large-scale Comparison of Individual Movement Patterns

Maarten Vanhoof
Open Lab, Newcastle University (1)
Orange Labs, Paris, France (2)
Maarten.vanhoof@orange.com

Willem Schoors
Division of Geography; KU Leuven (3)

Anton Van Rompaey (3)
Thomas Plötz (1)
Zbigniew Smoreda (2)

ABSTRACT

In this paper we use a very large mobile phone dataset for France to create a comparable spatial pattern of human mobility on a nation scale. We focus on the construction of the so called 'mobility entropy' from CDR data as an indicator for the diversity of individual movement. In a first part we show the dependency of the standard formula for mobility entropy on the density of observation point and propose a correction to ensure comparison can be done on a nation-wide scale. In a second part we calculate the proposed Corrected Mobility Entropy (CME) measure for ~18.5 million users in France. Our results show CME to be correlated with several mobility related variables gathered over the entirety of France. In addition, we show distributions of the calculated mobility entropy values to differ significantly between several large-scale regions in France that we define by i) an unsupervised classification task of small-scale regions based on mobility related variables and ii) an official delineation of Urban Areas as proposed by the French National Statistical Institute. Our main results show diversity of mobility to be highest in (sub)-urban regions, compared to agricultural, natural, or remote areas. In addition sub-urban regions depict higher values of mobility entropy compared to urban centers; while both decrease considerably when lowering urban center sizes. Our results serve as a showcase for the potential of deploying CDR data for the description, understanding and delineation of regions with respect to individual mobility, which has clear applications in official statisctis, urban planning, mobility research and policy.

Keywords

CDR data, Mobility Studies, Regional Geography, Urban Areas.

1. Introduction

Individual mobility is influenced by a wide range of constraints, including socio-economic characteristics and the direct environment [Asgari et al. 2013]. Differences in mobility hence exist between people, but also between geographical regions as several constraints might be common for larger populations. Understanding the differences in human mobility between regions can yield information on how environmental and societal factors relate to the movement and thus behavior of local populations. Such information is valuable to planners and policy makers as it allows for a more detailed examination of the reasons for mobility differences, hereby enabling more directed policy implementations.

To study mobility differences between regions, the movement of large populations needs to be recorded. Traditional approaches, such as interviews, questionnaires or travel diaries are ill-suited for this purpose because they require considerable efforts from both researcher as participant, resulting in smaller sample sizes and limited observation periods (Chen et al. 2013).

Recent advances in information technologies allow for novel approaches to movement detection. A prime example is GPS-technology which can record participant positions at regular intervals, resulting in a consistent detection of location. Other examples are Location Based Services (LBS) and Location Based Social Networks (LBSN), for which the connection to a service requires an active sharing of location. GPS, LBS and LBSN technologies all have the potential to collect movement data at the individual level for large populations and have been widely deployed as such to study large-scale human mobility patterns (Wolf et al. 2001, Pappalardo et al. 2013, Chen et al. 2014, Bojic et al. 2015).

The biggest drawback of these technologies is that they require the active participation of users, which for many users means a burden, raises privacy concerns and as such often results in limited tracing (Chen et al. 2014). As mobility is in essence very dynamic and private, the high granularity of observations from GPS will induce users to limit information sharing to a specific time-period or to specific trajectories. The low and irregular granularity of LBS and LBSN observations on the other hand will induce incomplete observations of individual movement since these services are only relied on in specific circomstances.

Passive location recording, often as a by-product of other activities, are less prone to these disadvantages as they do not require additional participant actions. Examples of passive location recording are, for instance, the 'geotag' in Online Social Networks (OSN), or Call Detail Records (CDR) data derived from mobile phone activities on an operator network.

In this paper we use a very large mobile phone dataset for France to create a comparable spatial pattern of human mobility on a

nation scale. We focus on the derivation of the so called 'mobility entropy' from CDR data as an indicator for the diversity of individual movement. Before calculating this indicator for ~18.5 million users, we propose an adaptation of the standard formula for entropy to ensure comparison can be done on a nation-wide scale. Our results show mobility entropy to be correlated with several mobility related variables over the entirety of France. In addition, we show distributions of the calculated mobility entropy values to differ significantly between several large-scale regions in France that we define by i) an unsupervised classification task of small-scale regions based on mobility related variables and ii) an official delineation of Urban Areas as proposed by the French Statistical Office. Our main results show that diversity of mobility is highest in urban regions, compared to agricultural, natural, or remote areas. More specific, we observe sub-urban regions to depict higher values of mobility entropy compared to urban centers. A decrease of mobility entropy is observed together with a decrease of urban center size resulting in small city centers and sub-urban areas to depict similar or even lower degrees of mobility diversity compared to agricultural areas. Our results serve as a showcase for the potential of deploying CDR data for the description, understanding and delineation of regions with respect to individual mobility which has clear applications in official statisctis, urban planning, mobility research and policy.

2. Related work

Call Detailed Records (CDR) data is generated as a by-product of mobile phone activity and often captured by the provider for billing or maintenance purposes. It contains metadata of each interaction (call or text) of the subscriber with the operator's network, including location data at the level of the contacted cellular tower. Based on CDR data movement patterns of individuals can be re-constructed at cell-tower network resolution and with the temporal granularity of a user's cell phone activity.

CDR data have extensively been used to study large scale human mobility patterns (e.g. Kung 2014, Ranjan 2012 and Simini 2012), population presence (e.g. Deville and Linard 2014) and other topics[1]. The real strength of CDR data in the mobility perspective is their ability to capture movement on a very fine grained resolution for populations of, often, millions of users. In general, the gathering of CDR data is more cost-effective, less biased and available on a much larger scale in terms of users, geographical coverage and time periods compared to traditional data gathering methods (Järv, Ahas, et Witlox 2014; Liu et al. 2013).

A disadvantage of CDR data is that location detection is only done when a person initiates a call or sends a text message. This might lead to only sporadic observation for certain users (Ranjan et al., 2012; Tanahashi et al., 2012). In addition, biases may occur when selecting users based on minimal location visits requirements. Ranjan et al. (2012) for instance have found a bias towards larger (calculated) mobility entropies in such cases and advised adapted sampling schemes. A second disadvantage when using CDR data is the coarse spatial resolution since locations are only recorded at cell-tower level (Tanahashi et al., 2012). The amount of cell-towers in an area can vary, hereby rendering the accuracy of location detection to correlate strongly with tower density.

[1] For an extensive literature review on the different applications of CDR data in research, see Blondel et al. (2015)

The wide deployment of CDR data in mobility research has led to the empirical discovery of several statistical properties of large-scale human mobility, like power-law-like distributions of displacements, of mobility motifs, of visitation frequency and, accordingly, of staying time (González, Hidalgo, et Barabási 2008; Song, Qu, et al. 2010; Song, Koren, et al. 2010; Schneider et al. 2013; Szell et al. 2012; Wang, Han, et Wang 2014). At the individual level, CDR data allow for a reconstruction and quantification of the individual movement patterns. Several indicators for individual mobility as derived from CDR data exist and have been deployed in different research applications. Table 1 gives a short, non exhaustive overview.

In this work we will focus on the mobility entropy indicator. De Montjoye et al. (2013) define entropy as "a quantitative measure reflecting how many different categories there are in a given random variable, and simultaneously takes into account how evenly the basic units are distributed among those categories". We can relate this interpretation to individual movement as captured by CDR data. When moving trough space and time, a user distributes his events over a set of cell-towers. Cell-towers thus represent categories while units are observed events on these towers. The distribution of events over a set of towers can therefore be characterized by an entropy value. Calculating entropy for a movement pattern derived from CDR data renders a single value expressing the variety of visited towers: the mobility entropy. An exact definition of mobility entropy is provided for in section 4 but an intuitive interpretation can already by formulated: mobility entropy is the uncertainty of location visits or, in other terms, the diversity of visited locations by a certain person.

One of the first applications of mobility entropy was discussed in Song et al. 2010. They show that for a 3-month long CDR dataset of 45.000 users (selected on amount of activities) the distribution of mobility entropies peaks at 0.8, indicating that the uncertainty of locations visited by a person is $2^{0.8} = 1.78$ or thus fewer than two locations. Using mobility entropies to calculate the potential predictability of individual users, they arrived at a surprising 93% of potential predictability; meaning that "despite the apparent randomness of the individuals'trajectories, a historical record of the daily mobility pattern of the users hides an unexpectedly high degree of potential predictability" (Song et al. 2010, page 1020). These findings have opened the door for the use of mobility entropy in mobility models where they serve as an indicator for the difficulty of movement prediction. Users with a high mobility entropy are more difficult to predict, and model results are therefore expected to be less accurate (Baumann & Santini, 2013).

A second application of mobility entropy is in activity tracking. For such applications, entropy is calculated for short time periods in order to obtain a view on how entropy evolves throughout the day. Research has observed a decrease of entropy during nighttime when compared to daytime and in increase of weekend-compared to weekdays (Cho et al., 2011). Similarly, comparing mobility entropy between different population groups yields insights in behavioral differences as some groups have a more diverse movement patterns at different moments in the week compared to others (Bajardi et al., 2015; Cranshaw et al., 2010).

A third application of mobility entropy covers its relation with other variables, often investigated on a nation-wide scale. Pappalardo et al. (2016) for instance, show the relation between mobility entropy and the European Deprivation Index (EDI) in France.

Table 1: A non-exhaustive list of CDR-based mobility metrics and their applications in research.

Topic	Definition	Applications
Visited towers	The amount of unique towers a person visits throughout the observation period	- Basic indicator of mobility. (Williams et al., 2014) - Footprint of a user.(Sridharan & Bolot, 2013)
Radius of Gyration	The standard deviation of distances between the user's locations and the user's center of mass.	- Distance traveled by an user (González et al., 2008) - Variation in different datasets (Zhang, 2014) - Influencing factors (Yan et al., 2010)
Meaningful locations	The location where the users are observed more frequently and have a certain meaning	- Anchoring the mobility network in meaningful places (Ahas et al., 2010) - Distance between social ties (Phithakkitnukoon et al., 2012) - Database anonymity (Zang & Bolot, 2011) - Urban planning (Ahas & Mark, 2005)
Trajectory	The temporal sequence of visited locations.	- Mobility models (Calabrese et al., 2010) - Mechanisms for daily mobility (Schneider et al., 2013)
Entropy	The diversity of the mobility pattern as observed by the sequence of visited location.	-Limits to mobility predictability (Song et al., 2010) - Temporal variations (Cho et al., 2011) - Differences between populations groups (Bajardi et al., 2015; Cranshaw et al., 2010) - Relations with socio-economic variables (Pappalardo et al., 2016)

Such investigations open interesting perspectives for the use of behavioural indicators from large-scale data sources as they could provide for a more frequently updated or even real-time view on the state of nation-wide systems (Giannotti et al. 2012).

Although all applications form interesting contributions to our understanding of the drivers of individual mobility, it is remarkable that very little research has been published that discusses the spatial patterns and/or regional differences of observed human mobility on a nation-wide scale. One apparent observation, for instance, is that, to the best of the authors' knowledge, no studies have yet explored the relationship between mobility entropy and the direct environment like could be expressed in physical features or land use. In general, it is fair to state that research on mobility (entropy) indicators derived from large-scale passive location detection methods is still in its very beginnings despite several notable, already existing applications.

In the next section we explain the construction of the mobility entropy indicator for all users in a CDR dataset from France. We explore the feasibility of using mobility entropy for a nation-wide comparison and show that different environments and especially different cell-tower densities are influencing the calculation of mobility entropy. For this reason, we propose a correction for cell-tower densities in order to make mobility entropy values comparable between regions. A second section investigates the corrected mobility entropies and their relation to several socio-economic and environmental variables, like income, land use and share of public transportation. In a final section, we investigate how delineations of space, like for instance done in official statistics, results in significantl different mobility entropy distributions for the regional populations concerned. Our main results translate in:

1. A proposition to correct measures of mobility entropy for uneven distributions of observation points in space.
2. A re-appropriation of the role of the direct environnment complementary to socio-economic characteristics for understanding collective mobility behavior.
3. The observation of statistically significant differences in mobility between regions as captured by CDR data.
4. A showcase for the potential of deploying CDR data for the description, understanding and delineation of regions with respect to mobility in e.g. official statisctis, planning or policy.

3. Data

In this study we use access to a CDR dataset from France provided by Orange and based on the activities of ~19 million subscribers during a period of 154 consecutive days in 2007 (May 13 to October 14). The data gathers locational (the used cell-tower), temporal (time of action) and interactional information (who contacts whom) every time a cell or text is initiated or received by the subscribers of the network. The spatial resolution of the dataset, hence, is restricted to the spatial distribution of 18.273 cell-towers in the Orange network. The locations of all cell-towers are known but there distribution in space is not uniform. In general, higher densities of cell-towers can be found in more densely populated areas like cities or coastlines. Lower densities of cell-towers are observed in more rural areas, as well as in mountain or natural areas.

In a first step of pre-processing, the data is limited to 44 consecutive days between September 1st 2007 and October 15th 2007. This period was chosen to avoid as much as possible the inclusion of holiday undertakings in the mobility pattern which typically occur during summermonths, while still having a long enough observation period to capture daily routines. Previous studies indicate that a time period over one month is sufficient to capture habitual behavior (Schlich and Axhausen, 2003). Secondly, users from other providers or countries are omitted as their registration on the network only happens occasionally and thus no sound mobility patterns can be recorded. Finally, service numbers that do not belong to individuals as well as numbers that display machine-like activities, like highly frequent, repetitive or volumnial call patterns were omitted. The filtered, ultimately deployed dataset, stores a total of 3.926.725.446 location traces for 18.581.513 users of which 72,95% are derived from call and 27,05% from text actions.

4. Calculating Mobility Entropy

The calculation of mobility entropy in this paper bases on the entropy concept developed by Claude Shannon in its seminal work: 'A mathematical theory of communication' (Shannon 2001). In this work, Shannon laid the basis for the field of information theory by discussing the transmission of messages trough a system as a sequence of transitions in a Markov chain. Due to the different transition probabilities in such a Markov chain, an uncertainty of the next transition can be calculated. Here

is where the Shannon entropy was introduced. Based on the probabilities associated with the possible transitions of one state in a Markov chain to another, at each transition an entropy value could be calculated, representing the "rate at which new information is produced" (Shannon, 2001).

The basic formula for entropy is rather simple; certainly regarding the complexity of the system it is trying to characterize (eq. 1). The one parameter (p_i) is the probability of the occurrence of a state (i) in the system. This probability is multiplied by the logarithm of the probability. The base of the logarithm can vary, and results in different types of entropy calculated. Shannon used base two, as he was studying communication systems, which use bits (either zero or one). In this paper base ten will be used, resulting in the decimal entropy (H). The entropy is calculated for each state of the sytem after which all values are summed, resulting in the total entropy for the system under consideration. A minus is added to convert the results to a positive value. One can interpret the formulae as a counteracting between the probability of a state (p_i) and its own logarithm ($log(p_i)$) resulting in a function (H) reaching a maximum when equal probabilities are used for all states.

$$eq.1 \qquad H = -\sum_{i=1}^{n} p_i \log(p_i)$$

Shannon's entropy can be applied to several systems as a measure capturing the degree of predictability, including individual mobility networks. In this perspective, individual movement trough space, as captured in a mobility network, forms a system in which each location forms a state and the probability of being in a state equals the time a user spends in this location. The mobility entropy, hence, becomes nothing more than the Shannon entropy for a mobility network and expresses the degrees of predictability of the movement pattern of a person or, in re-verse thinking, the diversity of an individual movement pattern.

Applying Shannon's entropy to individual mobility networks derived from CDR data, Song et al. (2010) propose three different ways of calculating mobility entropy: i) random entropy, ii) temporaly-uncorrelated entropy and iii) real entropy. For all three measures, locations (or thus states of the sytems) are based on the activated cell-towers. However, the probabilities of being in a certain state (~at a certain cell-tower) are calculated differently and are given by i) whether the user visited the cell-tower before or not, ii) the amount of times the user visited a cell-tower and iii) the full spatio-temporal presence of a person. As the latter is not obtainable given the sparse temporal resolution of most users in the CDR dataset and given the idea that the first neglects to take full advantage of the available information, we will continue this work by using the temporally-uncorrelated entropy only. The full equation of the temporally-uncorrelated entropy is given in equation 2 and a graphic illustration of it's calculation for one user in the dataset is given in figure 2.

$$eq.2 \quad ME = -\sum_{i=1}^{n} p_i \log(p_i) \qquad p_i = \frac{e_i}{e}$$

$$\begin{aligned} &with\ i = the\ tower\ under\ consideration \\ &e = the\ number\ of\ events \\ &n = the\ total\ number\ of\ visited\ towers \\ &p = the\ share\ of\ events\ per\ tower \end{aligned}$$

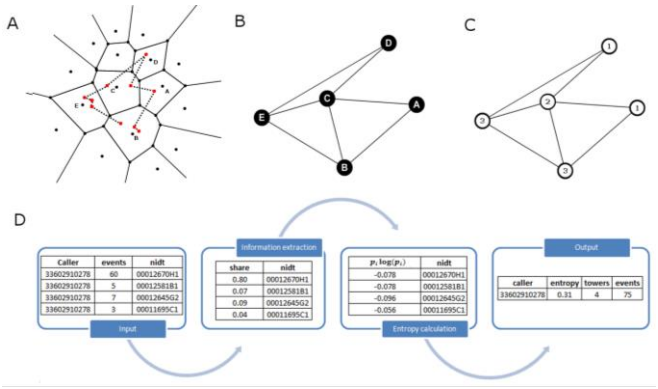

Figure 1: Graphic illustration for the calculation of mobility entropy from CDR data. A) The actual path a user follows in space (dotted lines) and the events he initiates (red dots) at cell-towers (black dots). The voronoi polygons (full lines) represent the service area borders of the different cell phone. B) The mobility network based on the visited towers. This representation accords with the random entropy in Song et al. (2010). C) The mobility network based on the amount of visits to the towers.This representation accords with the temporal-uncorrelated entropy in Song et al. (2010) and will be used troughout the rest of the work. D) Database representation of the different steps to calculate the temporal-uncorrelated mobility entropy for one user.

4.1 Correcting Mobility Entropy

Calculating mobility entropy like proposed in equation 2 means that differences of mobility entropy between regional populations can be induced by

i) Behavioral characteristics (e.g. patterns of mobile phone use resulting in different amount of calls),
ii) The direct environment (e.g. influence of physical geography like mountains and seas on movement patterns)
iii) Structural properties (e.g. the infrastructure by which we perform passive detection)

While the first two are differences we actually want to capture, the third element needs, by all means, to be reduced to a minimum as it would introduce a bias in our observations.

One clear, however often neglected, structural bias when calculating mobility entropy from detected movement is related to the density of potential observation points (which in the case of CDR data means cell-towers) in an area. Imagine for instance, as illustrated in Figure 2, a person X displaying exactly the same movement pattern in areas A, B, C, which are characterized by a decreasing density of cell-towers. The calculated entropy value based on the passively detected mobility networks from this person will differ significantly simple because the resolution of information derived from the cell-tower network is different. Logically, this will implicate a huge bias for comparison of mobility entropy values between regions. Regions with high densities of antennas, like cities or popular tourist destinations, will automatically depict higher mobility entropy values simply because the chances of observing one individual on several cell-towers are higher.

This structural bias in the calculation of mobility entropy is especially relevant when using CDR data on a nation-wide scale. As the spatial distribution of the antenna network is heavily heterogenous across different spatial regions, the density of

antennas will vary substantially inducing a structural bias with different scopes for different locations. As said before, density of antennas vary heavily for different areas of the country, resulting in the possibility to capture more, or less, specific informations on the spatial whereabouts of an individual. In terms of mobility entropy, the effect of antenna density is not to be underestimated.

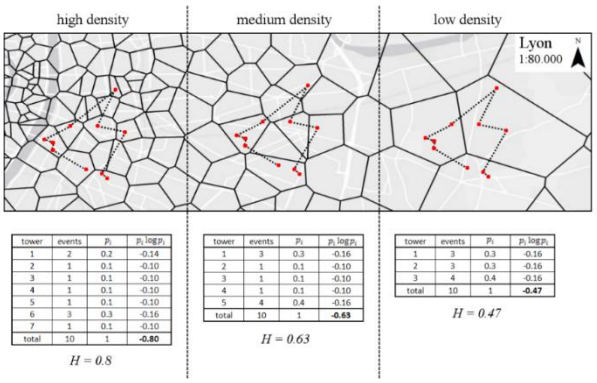

Figure 2 : Illustration of the effect of different densities of observation points on the calculation of the temporally-uncorrelated entropy (ME). Mobility entropy for the same path (dotted line) of a users in tree different density settings is calculated and shown to be different.

To balance the influence of cell-tower density on the calculated mobility entropy we propose a simple, yet effective, normalization of the entropy formula for cell-tower density.

Existing entropy normalization consists of dividing the entropy by the logarithm of either the number of visited towers or the number of events (Y. De Montjoye et al., 2013). Such approach might seem promising but in fact corrects for the amount of actions on the cell-tower, not for the distribution of cell-towers.

For this reason our proposition focusses on the correction of each cell-tower in the entropy formula based on its surrounding antenna density. The mobility entropy formula then becomes:

$$eq.2 \qquad CME = -\sum_{i=1}^{n} p_i \log(p_i)\, c_i$$

$$\begin{aligned} with\ i &= the\ tower\ under\ consideration \\ e &= the\ number\ of\ events \\ n &= the\ total\ number\ of\ visited\ towers \\ p_i &= the\ share\ of\ events\ for\ tower\ i \\ c_i &= the\ correction\ factor\ for\ tower\ i \\ CME &= the\ corrected\ mobiltiy\ entropy \end{aligned}$$

This approach is easy to understand intuitively. When cell-tower density is high, visiting a new cell-tower, and hereby increasing the mobility entropy, becomes easier and should therefore have a lower weight compared to visiting a cell-tower in a low-density area, where the registration of a user is more unique.

To estimate cell-tower density different metrics can be deployed. We opted for the most straightforward method by looking at the voronoi polygons based on the locations of the different cell-towers. Both the voronoi surface and circumference were examined, and the circumference was selected because it was less influenced by irregular voronoi shapes. The intuitive meaning of using the voronoi circumference is clear: high density areas will have smaller voronoi polygons, while the opposite happens in low density areas. A disadvantage of this metric is the limited density detection range, as the circumference is only influenced by the towers directly surrounding the tower under consideration.

Converting the voronoi circumference to correction factor c_i was done based on equation 3. Linear scaling is performed and bounded by parameters a and b. The scaling range between a and b express the the extent to which entropy values in high- and low-density areas differ. As a guideline in determining the values for a and b we looked at the scaling range of entropy modification as proposed by Montjoye et al. (2013) and approach the range of the correction factors used there. Multiple iterations where performed with different scaling ranges to investigate the impact on final entropy values. Based on visual inspection of the resulting spatial entropy pattern, and the validation discussed later, we decided to use 0.7 and 1.3 for parameters a and b respectively. The symmetrical scaling around 1 implicates in a correction which becomes larger when density deviates more from the mean density.

$$eq.3 \qquad c_i = \frac{(a-b)*(d_i - min(d))}{max(d) - \min(d)} + b$$

$$\begin{aligned} with\ d &= the\ density\ values\ for\ all\ towers \\ i &= the\ tower\ under\ consideration \\ c &= the\ scaled\ correction\ factor \end{aligned}$$

Remark that another option to address this problem could be to apply grid resampling of the cell-towers like, for instance, proposed by Bajardi et al. (2015) or Williams et al. (2014). Given the nation-wide extent of our dataset such approach is rather inapplicable. For practical reasons (18.273 cell-towers versus 269 in Williams et al. 2014) but mostly because of the high diversity of different areas under investigation (versus one single urban area in Bajardi et al. 2015) which complicate the choice of grid resolution. Low resolution grids would lead to data and precision loss in high density areas, while high resolution grids would cause a many empty grid cells in low density areas.

5. Results

5.1 Correcting mobility entropy

Mobility Entropy (ME) and Corrected Mobilty Entropy (CME) indicators were calculated for all ~18.5 million users in the French CDR dataset. Validation of both measures for human mobility is hard given that no ground truth is available, nor can possibly be collected in the future due to the large span of the dataset. Assesment of the proposed correction for mobility entropy, however, can be done by investigating the relation between ME, CME and the cell-tower density as done in figure 4. Remember that the main reason why we urged for a correction of the ME measure was because of the non-homogenous distribution of cell-towers in different areas. Given the formule of standard mobility entropy calculations, we would expect cell-tower density to influence the calculation of ME, a structural bias that we want to avoid when comparing values of mobility entropy between different (larger) regions.

To investigate the relation between ME, CME and cell-tower density we approximate the density of cell-towers by the circumference of the Voronoi polygon of that cell-tower. Cell-towers that are situated in high cell-tower density areas will have small Voronoi polygons, whereas cell-towers whose neigbours are further away will have larger Voronoi circumferences. The average values of ME and CME for each cell-tower are based on the populations of users that have a detected 'home' at the different cell-tower. Homes were defined for all users in the

dataset based on the cell-tower that that had the highest amount of activities during weekends and between 19 p.m and 9 a.m. on weekdays.

Plotting the linear regression between voronoi circumference and average ME, as done in figure 4 a), the influence of cell-tower density on the ME indicators becomes clear (R^2= 0.34). A different picture is observed when relating voronoi circumferences with average CME as visible in figure 4 b). The R^2 in this regression drops to 0.0289. Although still showing a small negative trend, most probably because of the influence of the highest tower density areas (lowest voronoi circumferences), it is fair to say the CME calculation largely diminishes the effect of antenna density, as was its initial goal. The result is promising. CME values are almost completely liberated from the structural bias of cell-tower density, rendering them comparable across France.

The relationship between ME and CME sheds light on the working of the proposed correction and is plotted in figure 3 c). Populations grouped by the 'home cell-tower' that see no change in average entropy values are situated on the 1:1 line, but are few. Rather, populations with smaller average degrees of ME tend to be corrected with small increases only whereas populations with higher average ME values were substantially corrected towards lower averages of CME as can be derived from their diverging trend towards the 1:1 line. Figure 3 d) shows the cumulative density distribution of the correction factors (c_i) as used in equation 2. The bimodal distribution clearly indicates that the latter group depicts a larger proportion of the populations having an average correction factor <1, with a peak around 0.78. Populations that had an average correction factor >1 were less numerous and had a much smaller correction, as can be observed by the peak around 1.03.

The spatial pattern of the correction factors, as showed in figure 4 clearly shows the first peak of (strongest) corrections to be located in the different city centres troughout the country. This is logical pattern as city centres typically have high densities of cell-towers. Areas in which the correction factor augments the mobility entropy values are mostly rural and mountainous areas.

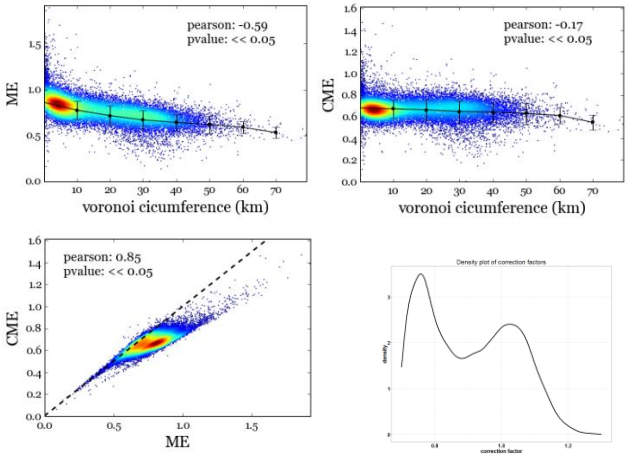

Figure 3: Correlations between antenna density (estimated by means of the voronoi circumference of all cell-towers) and (a) average ME, (b) average CME. (c) Correlation between average ME and average CME. All averages are calculated by the average value for all users having a detected 'home' at the concerned cell-tower. All parameter estimations for the linear regression are p<0.001 (d) The cumulative density function of the different correction factors (c_i) used for the calculation of CME.

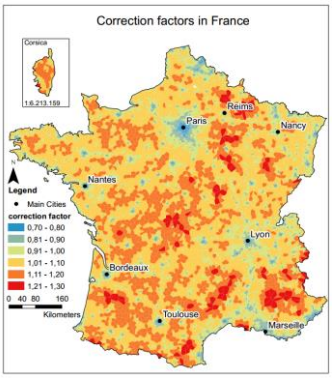

Figure 4: Spatial pattern of the applied correction factors (c_i) when calculating CME indicators. Values are depicted for all cell-towers in France (shown by their related voronoi polygons) and were calculated as the average value of all users that have a 'home' at these cell-towers.

5.2 Spatial Patterns of (Corrected) Mobility Entropy in France

The comparion of the spatial pattern of average ME and CME values per cell-towers in France offers the opportunity to evaluate the effect of the correction of ME. Figure 5 clearly shows the difference. The standard ME values (Figure 5a) show a typical, highly centered pattern with corridors that relate to, respectively, areas with high population densities and roads.

Although logical interpretable in terms of: "people living in cities and near roads move more divers", this spatial pattern clearly corresponds with the spatial pattern of the antenna density shown in Figure 5b, hence underscoring once more our argument that standard ME calculations are structurally biased. The proposed CME renders a much more homogenous spatial pattern (Figure 5c) in which regional differences in human mobility become less pronounced.

Highlighting the spatial pattern for CME values by means of the Local Moran's I statistics is done in Figure 5d. This analysis reveals statistically significant clusters of high(/low) values that are positioned in wider environments themselves depicting high(/low) values. As can be observerd several areas in France have statistically lower than expected values of CME (low-low clusters). As most of these areas are remote or mountainous areas one could interprete human mobility to be different here. However, seen that the average amount of antennas visited by users in these areas are extremely low (see Figure 5b), it is far more likely that limited observations resuls in extremely low entropy values that CME is not capable of correcting. In other words, the spatial resolution of our observations in this area is not sufficient to study the diversity of human mobility. Other areas that depict low-low clusters of CME are border regions where, plausibly, only parts of human mobility are detected by the French operator.

More interesting are the the high-high clusters of CME values that are positioned in the direct environment of medium to large-sized cities like Nantes, Rennes, Toulouse and Lyon but not in their city centres. Hypothetically, these clusters represent areas with high shares of commuters, who, besides commuting also have a rather mobile lifestyle as commuting solely would result in lower entropy values. The observation that such high-high cluster are found in the vicinity of almost all medium-sizes cities in France forms a strong suggestion towards the nation-wide comparability of the CME values.

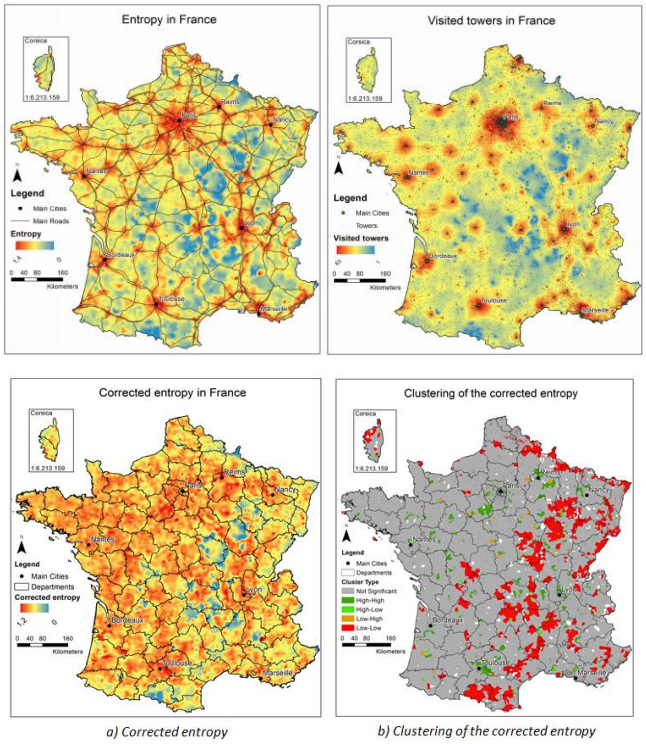

Figure 5: Spatial patterns of a) ME values, b) number of visited towers per person, c) CME values, d) Local Moran's I clustering of high-high, low-low, high-low and low-high combinations of CME values. Values are calculated for all cell-towers in France as the average value of all users that have a 'home' at these cell-towers.

6. Relations between CME and mobility related variables

Having constructed a measure for the diversity of human mobility that allows for objective comparison between different areas, the question becomes whether we can distinguish factors that influence mobility on a nation-wide scale. In this section we elaborate on this idea. We cross information obtained from national statistics and remote sensing with the distributions of corrected mobility entropy at the cell-tower level as calculated before. Deploying multiple linear regressions models we identify several significant variables, both socio-economic and related to the direct environment, that allow for the prediction of observed differences in mobility entropy for different areas.

6.1 Multivariate analysis

Since drivers of human mobility can be divided into personal factors, like socio-demographic aspects, lifestyle and accessibility of transport modes, and external factors like distances, natural environment and location of opportunities [Asgari et al.2013] we perform three different multiple regression analyses. All of them are based on a Generalized Linear Model (GLM) using a Gaussian error distribution and taking CME as dependent variable. The explaining variables used in the three models reflect:

i) Socio-economic aspects related to mobility and accessibility to transport.
ii) Environmental variables relating to the direct environment in which mobility is performed and
iii) A combination of i and ii.

A description of the deployed variables is given in Table 2.. All socio-economic variables were obtained from the Franch National Statistics office (INSEE). Indicators where retrieved for the year 2007 for all municipalities in France and information was crossed with the cell-towers on the basis of location within a municipality. Information for the environmental variables was retrieved from satellite images and GIS analysis. Land use classes stern from 2006 Corine data whereas the mean elevation is based on SRTM data, both of which are freely available to the public[2]. Here also, the spatial resolution of the variable was the administrative municipality level. Attribution to cell-tower level bases on the location of the cell-tower within a municipality. The results of the different GLM runs are showed in Table 3**.**

Table 2: Collected variables related to mobility to be used as independent variables in the GLM model.

Variable	Description
Socio-Economic	
Median Income	The median income per household member
Active population	The share of the municipality population between 15 and 65 years old
Working population	The employed share of the active population
Commuting distance	The median commuting distance for a municipality based on the CDR database
Car share	The share of trips made by car
Public transportation share	The share of trips made by public transportation
Car ownership	The average amount of cars owned per household
Employment in municipality	The share of people working in their municipality of residence
Environment	
Mean elevation	The mean elevation of the municipality, based on averaged SRTM data
Surface	The surface area of the municipality
Artificial land use	The share of artificial land use in a municipality, based on 2006 Corine data
Agricultural land use	The share of agricultural land use in a municipality, based on 2006 Corine data
Natural land use	The share of natural land use in a municipality, based on 2006 Corine data
Road distance	Average distance in the municipality to the nearest major road (regional road network)
City distance	Average distance in the municipality to the nearest city with >20,000 inhabitants
Station distance	Average distance in the municipality to the nearest train station

Table 3: Estiamted parameters from three runs of the GLM models. All results are significant at the p=0.05 level except for the ones indicated with (*). Variables that were omitted due to multi-collinearity are indicated with a '/'.

GLM	Run 1	Run 2	Run 3
Intercept (p-value)	0,6534 (***)	0,6534 (***)	0,6534 (***)
R^2	0,11	0,12	0,19
Socio-Economic			
Median Income		0,0059	0,0074
Active population		/	/
Working population		0,0159	0,015
Commuting distance		-0,0063	-0,0019
Car share		0,033	0,0019
Public transportation share		0,0057	0,0059
Car ownership		0,0019	0,0018
Employment in municipality		-0,0025	/
Environment			
Mean elevation	-0,0082		-0,0108
Surface	0,0066		0,0089
Artificial land use	/		/
Agricultural land use	0,0085		0,0086
Natural land use	-0,0049		-0,0029
Road distance	-0,0018		0,0006 (*)
City distance	-0,00142		-0,0091
Station distance	0,0024		0,0007 (*)

[2] http://www2.jpl.nasa.gov/srtm/france.htm for SRTM data

http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3 for CORINE data

7. Regional delineation and CME values

Building on three observations:

1. A rather homogenous spatial pattern of CME values for France
2. Clear and nation-wide high-high clusters of CME values in peri-urban areas for small and medium-sized cities
3. A high amount of indicators available at municipality level that are significant in representing differences between observed CME

We ask ourselves the question whether it is possible to compare differences in mobility diversity between areas based on their CME values, or not. And if so, what delineation of the French territory would depict significant, consistent interpretable differences in mobility entropy? In a first part of this section we tackle this question by creating our own delineation of space based on the significant variables from the GLM models. To our own surprise, a simple k-means unsupervised learning approach renders exactly what we want to obtain: A simple classification of territory that is easily interpreted, consistent on a nation-scale and depicts significant differences in mobility entropy distributions of inhabiting users. In a second part we recur to an official nine-fold classification of Urban Areas by the French National Statistics which offers a much more advised classification of space than our own unsupervised classification. Here too, distributions of CME differ significantly but interpretations are more in-depth and with a higher degree of certainty.

7.1 K-means clustering.

We perform an unsupervised k-means classification for all cell-towers in our dataset, using the significant variables of the third GLM run as features (see table 3). Based on a slowing decay of within group sum of squares (see appendix) we fix our number of classes to be 5. The emergent clusters are remarkable in the sense that they are easily interpretable. They show a delineation of the territory in five classes that can be observed in figure 6 and which we, based on the constitution of the clusters (see appendix) can be labeled as Urban, Suburban, Agricultural, Forest and Remote areas. Some deviations, however, are possible, as is the case along the Mediterranean coast. Here the indicators related to land use outweigh the socio-economic characteristics of living near an urban center (see appendix), classifying these regions in remote, or forest areas mostly. Additionaly remarkable is that distributions of the CME values between these clusters differ significanttly, as can be seen in the density plots of figure 6 and the results of the Wilcoxon tests in table 4. All of this indicates that the obtained clusters are relevant when discussing differences in mobility entropy at a nation-wide scale.

The main interpretations of our analysis are straightforward and can be derived from figure 7. Suburban and urban areas are inhabited by users with the highest diversity in mobility, as was already suggested in figure 5. Agricultural, forest and remote areas follow in that respective rank. The differences of mobility entropy within urban areas are the smallest (narrow distribution) indicating that mobility diversity is very similar in different French urban regions.

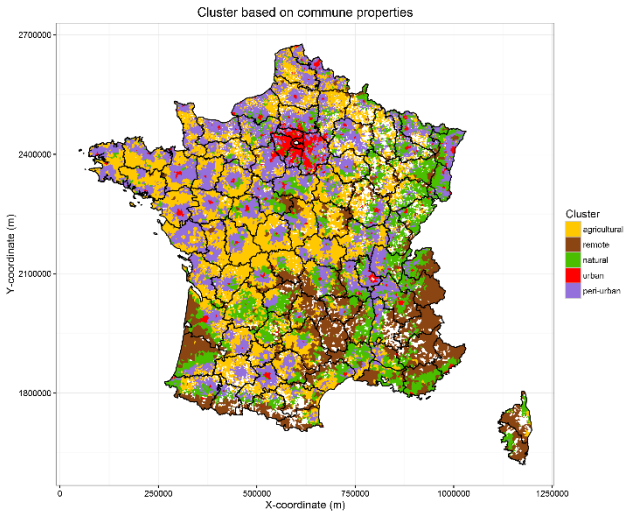

Figure 6: Spatial distribution of the clusters obtained from k-means classification. The spatial granularity is the ones of estimated service areas from French cell-towers (voronoi polygon) based on the significant variables for the full GLM model (run 3).

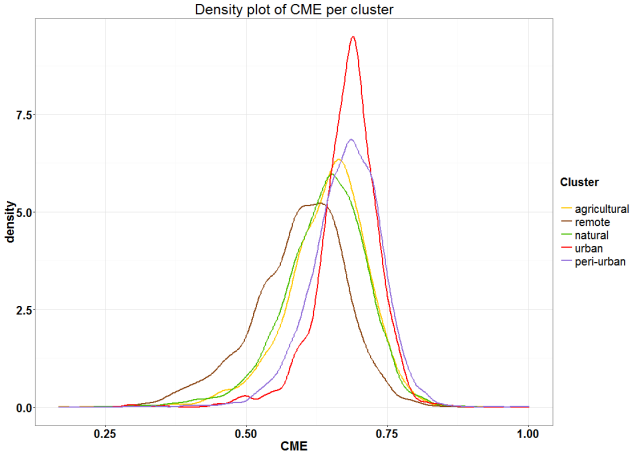

Figure 7: **Density plot of CME values per cluster in the k-means classification**. Values are calculated for all cell-towers in France as the average value of all users that have a 'home' at these cell-towers.

Table 4: P-values from Wilcoxon tests on the pairwise differences between CME distributions of classes obtained from the k-means classification.

	Agricultural	Remote	Natural	Urban
Remote	<0.01 (***)			
Natural	<0.01 (***)	<0.01 (***)		
Urban	<0.01 (***)	<0.01 (***)	<0.01 (***)	
Peri-urban	<0.01 (***)	<0.01 (***)	<0.01 (***)	0,08 (*)

7.2 Official Urban Areas

The French National Statistical Institute (INSEE) produces, about every five year, a territorial classification of Urban Areas. This classification is based on the identification of employment centres and their area of influence through commuting data[3]. Its conceptualization stretches beyond the typical physical borders defined by continuity of buildings often used in Urban Unit delineation and allows to study cities organization and development based on dynamic interactions between locations (Combes, de Bellefon et Vanhoof, 2016).

As listed in table 1, the Urban Area classification consists of 9 classes, being distinguished mainly by the size of the employment centre in the 'central urban unit'. Major, medium and small poles are centers offering respectively >10.000, between 5.000 and 10.000 and between 1.500 and 5.000 jobs. Surroundings of poles are defined by identifying municipalities of which more than 40% of the working population commutes daily to an adjacent pole. Special cases are being recognized for municipalities that have several poles to commute to (multipolarized municipalities) or municipalities that are not influenced by major, medium or small poles. A spatial distribution of the Urban Areas is shown in figure 8 and, especially for the major poles, relates strongly to the obtained clusters in section 7.1.

Table 5: Urban Areas as defined by INSEE, including the proportion of the number of municipalities and the amount of cell-towers of Orange categorized in these classes. For the spatial pattern of this classification, see appendix.

Code	Location of municipalities	Proportions	Antennas
111	Major pole (more than 10.000 employments)	9%	54%
112	Surroundings of a major pole	34%	18%
120	Multipolarized, in a large urban area	11%	5%
211	Medium pole (5.000 to 10.000 employments)	1%	3%
212	Surroundings of a medium pole	2%	1%
221	Small pole	2%	3%
222	Surroundings of a small pole	2%	0.3%
300	Other multipolarized municipality	19%	6%
400	Isolated municipality outside influence	20%	10%

The distributions of CME values within the Urban Areas classes are given in Figure 8. Results align with observations found in Figure 7, although some distributions do not longer differ statistically as can be seen from the results of the statistical test in Table 6. The most important new insights that can be gained from Figure 8 are on the relation between poles (urban centers) and surroundings of urban poles. In this analysis, peri-urban areas depict higher mobility entropies compared to their center. What is remarkable, however, is the decrease of mobility entropy with size of the urban pole, both for the urban poles themselves as for the surroundings of the poles. Bigger cities depict higher diversity of mobility both in their centers as surrounding areas, suggesting there is a scaling law between human mobility and population numbers.

[3]The 2010 Urban Area classification is based on data collected between 2006 and 2010 in the national census survey.

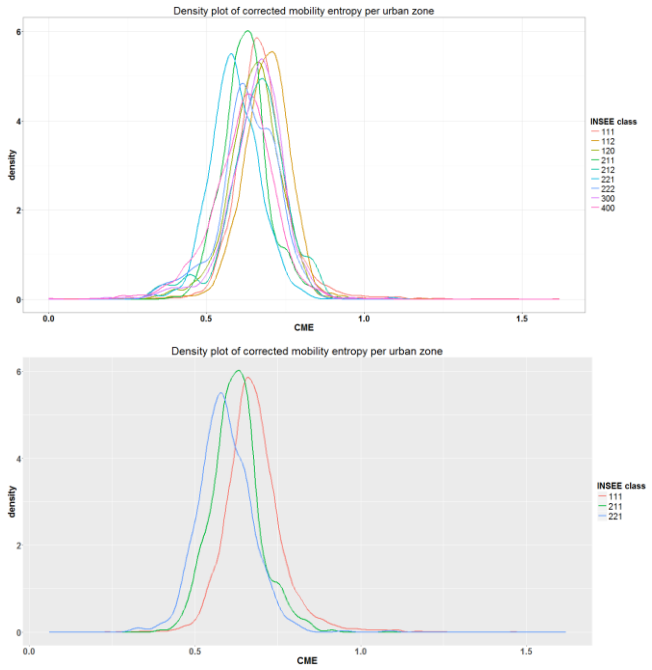

Figure 8: Density plot of CME values per class in the official Urban Area classification (top). Focus on the 3 main urban pole classes (bottom). Values are calculated for all cell-towers in France as the average value of all users that have a 'home' at these cell-towers.

Table 6: Selection of p-values from Wilcoxon tests on the pairwise differences between CME distributions of different classes in the official the Urban Area classification.

	111	112	211	212	221
112	<0.01 (***)				
211	<0.01 (***)	<0.01 (***)			
212	Non-sign.	<0.01 (***)	<0.01 (***)		
221	<0.01 (***)	<0.01 (***)	<0.01 (***)	<0.01 (***)	
222	0.02 (**)	<0.01 (***)	Non-sign.	Non-signi	<0.01 (***)

8. Discussion

[To be elaborated.]

1. *Standard mobility entropy is biased.*
2. *CME seems to work quite well, except for some outliers which we identify as border regions or regions with a very limited coverage.*
3. *Relationship between CME and variables related to mobility is clear, although predictive force of these variables is limited (link to overlap of classes in the density plots).*
4. *Exploration of delineations of space in which CME would differ significantly is rather easy to do based on the mobility-related variables and renders, remarkably, rather easily interpretable clusters that differ significantly in mobility entropy*

5. *A more fixed classification of space (INSEE – Urban Areas, confirms our data-driven action and renders similar insights, and signifancies.*
6. *Clearly, mobility entropy is higher in (sub)-urban regions, however when decreasing city size, mobility entropy decreases drastically (urban scaling laws)*
7. *Description and discussion of our main findings in terms of geography, related to the construction of clusters in the appendix, Time budgets, money for displacement, Hägerstrand. Focus on large – medium and small poles and their suburban areas.*
8. *Despite rather large overlaps of CME values for different clusters/classes differences in CME distributions are still significant. This is because of the huge numbers of observations, or thus, the force of big data that speaks, and it is questionable whether other data-collection efforts would be capable of detecting similar distances.*

9. Conclusion

In this paper we used a very large CDR dataset for France to create a comparable spatial pattern of human mobility on a nation scale. We focused on the derivation of the so called 'mobility entropy' from mobile phone records as an indicator for the diversity of individual movement. In a first part of this paper we argued that the calculation of typical mobility entropy measures needs correction for the differing density of observation points (cell-phone towers in the case of CDR data) between regions in order to be comparable for an entire country. We proposed a correction of the mobility entropy calculation to account for this structural bias and calculated a corrected mobility entropy (CME) for all ~18.5 million users in the dataset. Our results show the corrected mobility entropy (CME) to be less influenced by the density of cell-towers and to render a more homogenous spatial pattern compared to the traditional mobility entropy measure which depicts a clear bias towards high-density areas. In a second part, we investigated the relationship between CME and several socio-economic and environmental indicators that are of relevance when studying mobility. By means of a Generalized Linear Model (GLM) we show that multiple mobility-related inidcators prove to be significant in predicting, although to a limited degree, the average CME values of residents in all municipalities in France. Next, an unsupervised k-means clustering, using the significant indicators from the GLM as features resulted in a classification of the French territory in 5 classes. Remarkably, these 5 classes were easily interpreted as urban, sub-urban, agricultural, forest and remote areas and depicted statistically significant differences CME distributions. Similarly, significant differences in CME distributions where obtained when comparing differenct classes in an official classification of Urban Areas created by the French National Statistics (INSEE). All our findings suggest that the proposed correction of mobility entropy from CDR data results in a more reliable measure for comparing detected human mobility between large-scale-regions. Our main results show that diversity of mobility is highest in sub-urban regions where high proportions of the populations commute to urban centers. However, a decrease of mobility entropy is observed together with a decrease of the urban center size, both for the centers as the surrounding areas. Our results serve as a showcase for the potential of deploying CDR data for the description, understanding and delineation of regions with respect to individual mobility, which has clear applications in official statisctis, urban planning, mobility research and policy.

10. References

Ahas, R., & Mark, Ü. (2005). Location based services—new challenges for planning and public administration?. Futures, 37(6), 547-561.

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. Journal of Urban Technology, 17(1), 3-27.

Asgari, F., Gauthier, V., & Becker, M. (2013). A survey on human mobility and its applications. arXiv preprint arXiv:1307.0814.

Bajardi, P., Delfino, M., Panisson, A., Petri, G., & Tizzoni, M. (2015). Unveiling patterns of international communities in a global city using mobile phone data. EPJ Data Science, 4(1), 1.

Baumann, P., & Santini, S. (2013). On the use of instantaneous entropy to measure the momentary predictability of human mobility. In 2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC) (pp. 535–539). IEEE.

Blondel, V. D., Decuyper, A., & Krings, G. (2015). A survey of results on mobile phone datasets analysis. EPJ Data Science, 4(1), 1.

Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010, September). Human mobility prediction based on individual and collective geographical preferences. In Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on (pp. 312-317). IEEE.

Chen, C., Bian, L., & Ma, J. (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces? Transportation Research Part C: Emerging Technologies, 46, 326–337. http://doi.org/10.1016/j.trc.2014.07.001

Cho, E., Myers, S. A., & Leskovec, J. (2011, August). 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 (pp. 1082-1090). ACM.

Combes, S., De Bellefon, M-P. Vanhoof, M. (Submitted) Classifying Urban Areas in France Based on CDR data. SIGSPATIAL 2016.

Cranshaw, J., Toch, E., Hong, J., Kittur, A., & Sadeh, N. (2010, September). Bridging the gap between physical location and online social networks. In Proceedings of the 12th ACM international conference on Ubiquitous computing (pp. 119-128). ACM.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... & Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences, 111(45), 15888-15893.

de Montjoye, Y.-A., Hidalgo, C. a, Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports, 3, 1376. http://doi.org/10.1038/srep01376

Giannotti, F., Pedreschi, D., Pentland, A., Lukowicz, P., Kossmann, D., Crowley, J., & Helbing, D. (2012). A planetary nervous system for social mining and collective awareness. The European Physical Journal Special Topics, 214(1), 49–75. http://doi.org/10.1140/epjst/e2012-01688-9

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. Nature, 453(7196), 779-782.

Järv, O., Ahas, R., & Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. Transportation Research Part C: Emerging Technologies, 38, 122-135.

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. PloS one, 9(6), e96180.

Liu, F., Janssens, D., Wets, G., & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. Expert Systems with Applications, 40(8), 3299-3311.

Phithakkitnukoon, S., Smoreda, Z., & Olivier, P. (2012). Socio-geography of human mobility: A study using longitudinal mobile phone data. PloS one, 7(6), e39253.

Pappalardo, L., Rinzivillo, S., Qu, Z., Pedreschi, D., & Giannotti, F. (2013). Understanding the patterns of car travel. The European Physical Journal Special Topics, 215(1), 61–73. http://doi.org/10.1140/epjst/e2013-01715-5

Pappalardo, L., Vanhoof, M., Gabrielli, L., Smoreda, Z., Pedreschi, D., Gianotti, F. (2016/Forthcoming) An Analytical Framework to Nowcast Well-Being Using Mobile Phone Data. International Journal for Data Science and Analytics.

Ranjan, G., Zang, H., Zhang, Z. L., & Bolot, J. (2012). Are call detail records biased for sampling human mobility?. ACM SIGMOBILE Mobile Computing and Communications Review, 16(3), 33-44.

Schlich, R., & Axhausen, K. W. (2003). Habitual travel behaviour: evidence from a six-week travel diary. Transportation, 30(1), 13-36.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling daily human mobility motifs. Journal of The Royal Society Interface, 10(84), 20130246.

Shannon, C. E. (2001). A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), 3-55.

Simini, F., González, M. C., Maritan, A., & Barabási, A. L. (2012). A universal model for mobility and migration patterns. Nature, 484(7392), 96-100.

Song, C., Koren, T., Wang, P., & Barabási, A. L. (2010). Modelling the scaling properties of human mobility. Nature Physics, 6(10), 818-823.

Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in human mobility. Science, 327(5968), 1018-1021.

Sridharan, A., & Bolot, J. (2013, April). Location patterns of mobile users: A large-scale study. In INFOCOM, 2013 Proceedings IEEE (pp. 1007-1015). IEEE.

Szell, M., Sinatra, R., Petri, G., Thurner, S., & Latora, V. (2012). Understanding mobility in a social petri dish. Scientific Reports, 2. http://doi.org/10.1038/srep00457

Tanahashi, Y., Rowland, J. R., North, S., & Ma, K. L. (2012, December). Inferring human mobility patterns from anonymized mobile communication usage. In Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia (pp. 151-160). ACM.

Wang, X.-W., Han, X.-P., & Wang, B.-H. (2014). Correlations and scaling laws in human mobility. PloS One, 9(1), e84954.

Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., & Dobra, A. (2015). Measures of human mobility using mobile phone records enhanced with gis data. PloS one, 10(7), e0133630.

Wolf, J. (2000). Using GPS data loggers to replace travel diaries in the collection of travel data (Doctoral dissertation, Georgia Institute of Technology).

Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from GPS travel data. Transportation Research Record, (1768), 125–134. http://doi.org/10.3141/1768-15

Xiao-Yong, Y., Xiao-Pu, H., Tao, Z., & Bing-Hong, W. (2011). Exact solution of the gyration radius of an individual's trajectory for a simplified human regular mobility model. Chinese Physics Letters, 28(12), 120506.

Zang, H., & Bolot, J. (2011, September). Anonymization of location data does not work: A large-scale measurement study. In Proceedings of the 17th annual international conference on Mobile computing and networking (pp. 145-156). ACM.

Zhang, Y. (2014, April). User mobility from the view of cellular data networks. In IEEE INFOCOM 2014-IEEE Conference on Computer Communications (pp. 1348-1356). IEEE.

11. Appendix

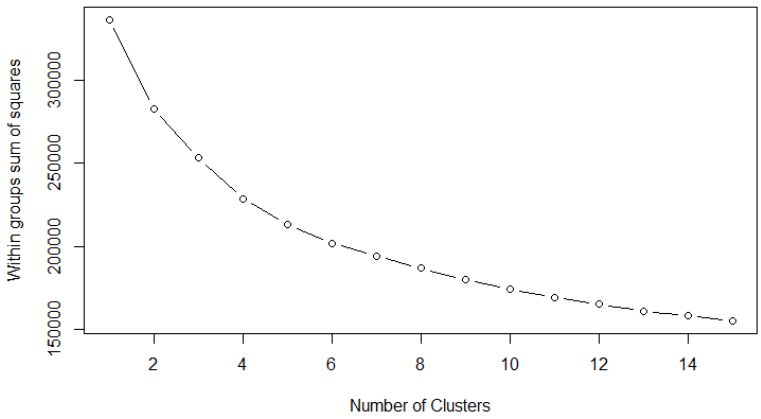

Figure 9: Within group sum of square for different number of clusters in the k-means unsupervised classification based on the all the significant variables from the full GLM model.

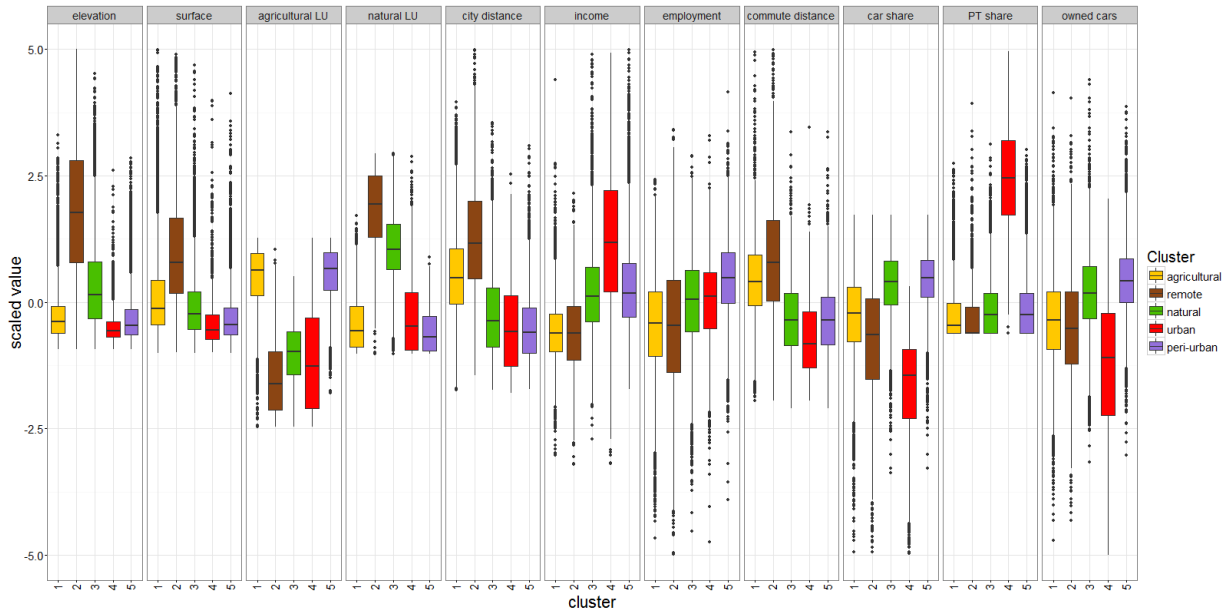

Figure 10: Construction of the different 5 clusters obtained from the the k-means unsupervised classification based on the all the significant variables from the full GLM model. Distributions are based on the observations at cell-tower level.

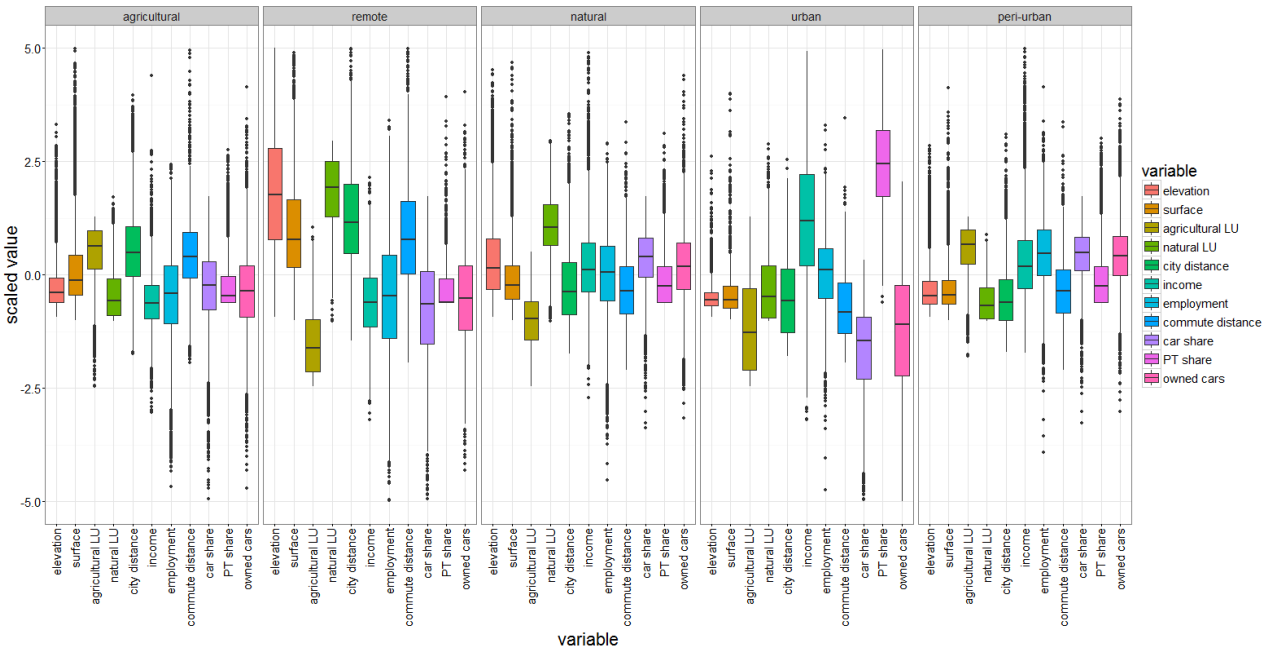

Figure 11: Construction of the different 5 clusters obtained from the the k-means unsupervised classification based on the all the significant variables from the full GLM model. Distributions are based on the observations at cell-tower level.

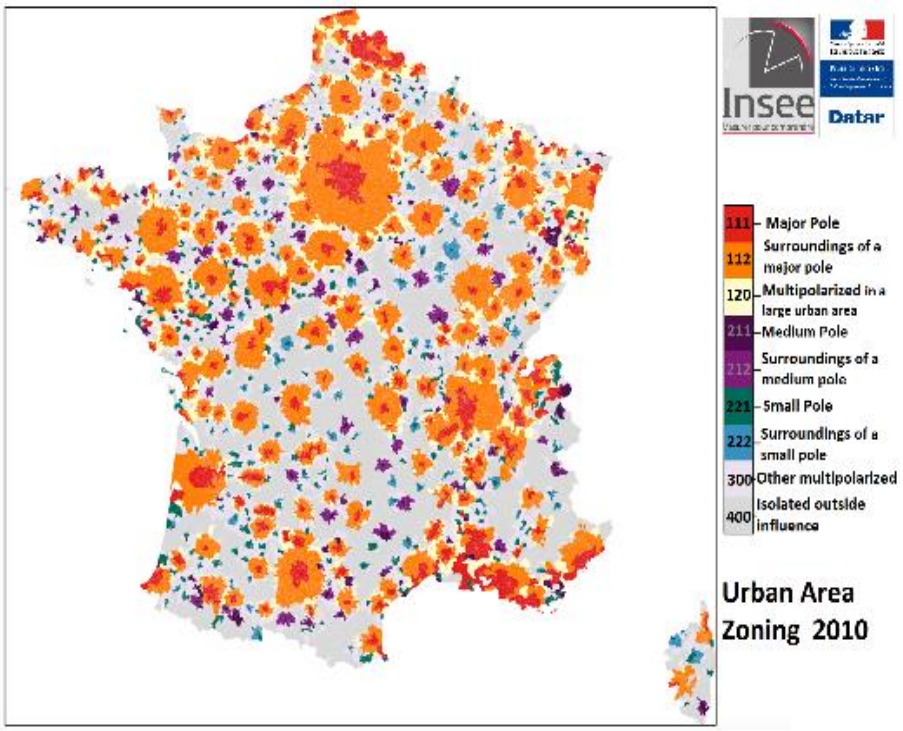

Figure 12: Spatial distribution of the Urban Area Classes as published by the French National Statistics Office in 2010.

INSEE zone	111	112	120	211	212	221	222	300
112	<2e-16	/	/	/	/	/	/	/
120	8.7e-11	<2e-16	/	/	/	/	/	/
211	<2e-16	<2e-16	1.3e-11	/	/	/	/	/
212	1	0,0013	1	6e-6	/	/	/	/
221	<2e-16	<2e-16	<2e-16	5.9e-11	7.9e-15	/	/	/
222	0,0277	9.5e-5	1	1	0.7047	0.0031	/	/
300	0,0011	<2e-16	0.1123	<2e-16	1	<2e-16	0.3930	/
400	<2e-16	<2e-16	<2e-16	1	8.8e-6	1.2e-10	1	<2e-16

Table 7. Full table of p-values from Wilcoxon tests on the pairwise differences between CME distributions of different classes in the official the Urban Area classification.