On the Relation between Socio-Economic Status and Physical Mobility

In emerging economies, the socio-economic status is a key element to evaluate social improvement as it provides an understanding of the population's access to housing, education, health or basic services like water and electricity. The relationship between such indicators and human physical mobility has been researched mostly in areas like access to medical infrastructures and public transportation. However, such studies have been limited in scope mostly due to the lack of large scale human mobility information. Nevertheless, the recent adoption of cell phones by large social groups in emerging economies has made it possible to capture large scale data about human physical mobility, which combined with regional socio-economic levels, allows to study the relationship between socio-economic indices and human mobility. In this paper we study the relationship between mobility variables and socio-economic levels using cell phone traces. Our results indicate that populations with higher socio-economic levels are strongly linked to larger mobility ranges than populations from lower socio-economic status. Finally, we also present a model that formalizes our findings on the relation between socio-economic levels and human mobility.

Keywords: human physical mobility, socio-economic indicators, policy guidelines, ICTs and mobiles for development.

1. INTRODUCTION

The socio-economic status is an index used in the social sciences to characterize an individual or a household economic and social position relative to the rest of the society, and is typically computed as a combination of variables that act as proxies to the underlying social and economic realities. The relation between human mobility and socio-economic status has been studied in a variety of scenarios, mainly related to access to health services and public transportation. In fact, research has shown that socio-economic levels might be correlated to travel distance, access to health clinics or energy consumption [1, 2]. The vast majority of these studies suffer from two important limitations: (1) they approximate human mobility through the use of proxy data such as public transport routes [10] or by tracking the travels of dollar bills [9]; and (2) the majority of these studies are based on qualitative and quantitative interviews with individuals, which highly limits the scope of the study and might bias the data. As a result, to date, the relation between human mobility and socio-economic levels has not been clearly measured, mainly because of the difficulty to obtain direct human mobility data from a sufficiently large number of individuals.

Nevertheless, the recent adoption of ubiquitous computing technologies by a very large portion of the population has enabled the capture large scale quantitative data about individual human mobility. In this context, mobile phones play a key role as sensors of human behavior as these are typically owned by individuals that carry them at –almost– all times. As a result, most of the recent large scale quantitative data about human mobility has been gathered via Call Detail Records (CDRs hereafter) from cell phone networks. This fact is also true for emerging economies where, in the last 10 years, the penetration rate of cell phones has experienced a steady growth, even surpassing landline infrastructures. For example, recent studies carried out by the International Telecommunication Union (ITU), show penetration rates of 96% in Venezuela, 42% in Kenya and 30% in India, as well as ratios of mobile cellular subscriptions to fixed telephone lines of 4.3:1, 25.2:1 or 9.2:1 respectively [3]. Just in Africa, in 2008, the number of mobile phone subscribers surpassed the number in North America. The pervasiveness of cell phones in emerging countries across Asia, Africa and Latin America has promoted the creation of cell phone-based services specifically designed to tackle emerging problems in areas like health, education or agriculture. In fact, there are many examples of successful cell phone-based services for emerging economies such as the Village Phone initiative, which allows to generate profit from cell phone rentals [15]; txtEagle that is based on crowd-sourcing techniques and generates revenues through the execution of small tasks [16]; EducaMovil, that provides educational contents through games for children in low-resource and isolated schools [17] or mobile health solutions like E-IMCI to improve treatment adherence in low-income areas [18]. For these reasons, it is fair to say that in many countries cell phones constitute an important part of the citizen's livelihoods.

In this paper we analyze the relation between socio-economic levels and human mobility by characterizing human mobility with a set of variables measured from the information contained in cell phone call detail records. The CDRs used for our analysis have been managed at an aggregated level and have also been encrypted to preserve privacy. Our findings are relevant for a variety of areas in policy design for emerging economies, ranging from transport planning to virus spreading containment.

Additionally, we use these results to predict the socio-economic level of a geographical area using the mobility information of the individuals that live within that domain. In fact, although the availability of socio-economic data is common in developed nations, that is not the case for emerging economies where the information is not necessarily as available, and/or economic activities might happen informally. As a result, the identification of socio-economic levels in emerging economies is more complex and less reliable than in developed economies. Although this limitation can be overcome with national surveys that capture income, education and consumer goods to identify socio-economic status, these surveys tend to be time and resource intensive as well as expensive. Being able to analyze the distribution of socio-economic status is a key element for policy design, outcome evaluation and impact assessment. As a result it is very relevant for developing economies to have means to effectively capture data on the incomes of their populations, and its evolution over time, in a cost-effective fashion. In this paper, we present a mathematical model that approximately computes socio-economic levels based on human mobility variables. This approach should be viewed as a complement to traditional approaches of evaluating socio-economic status of populations in a low-cost and efficient manner.

The rest of the paper is organized as follows: first we present the related work. The following section describes the main characteristics of the datasets used in our analysis, both the call detail records and the socio-economic information available. After that, we describe the aggregation and matching techniques to combine CDRs and socio-economic levels and explain the methodology used to evaluate the impact that socio-economic levels might have on human mobility. The results section details our findings on the relation between socio-economic levels and mobility variables and presents a mathematical model that formalizes these findings.

2. RELATED WORK

A few studies have measured strong relationships between socio-economic levels and human mobility at specific scenarios such as access to hospitals [1]; travel patterns [4]; rural-urban differences regarding cancer [2] or travel behavior of inter-city bus passengers [5]. These studies tend to use proxies of mobility and are based on small-scale surveys or focus groups. However, to the best of our knowledge there are no studies that measure the impact of socio-economic levels (SEL) on individual human physical mobility at the large scale we propose.

Related research has also compared mobility variables between a developed and an emerging economy [6]. The difference with our study is that while in [6] the socio-economic levels are implicitly derived from the country where the data originates; in this paper we make use of country-based household survey data to determine socio-economic levels. Eagle *et al.* [7] studied the relation between socio-economic levels and social network diversity using also cell phone records and regional social development indicators in the UK. Their findings indicate that social network diversity seems to be a very strong indicator of the development of large online social communities. Compared to this work, our paper provides a more granular study and focuses on mobility variables instead of online social networks.

3. EXPERIMENTAL SETTING

In this section, we first describe how cell phone networks work and how the call records are captured within the network. Next, we introduce the mobility variables that we use to characterize physical mobility in our work, followed by a description of the dataset with the cell phone records and a brief analysis of the mobility variables across the population under study. Finally, we describe the dataset that contains the socio-economic levels and its main characteristics.

3.1 CELL PHONE TRACES

Cell phone networks are built using a set of base transceiver stations (BTS) that are in charge of communicating cell phone devices with the network. Each BTS has a geographical location typically expressed by its latitude and longitude. The area covered by a BTS tower is called a cell. At any given moment, a cell phone can be covered by one or more BTSs. Whenever an individual makes a phone call, the call is routed through a BTS in the area of coverage. The BTS is assigned depending on the network traffic and on the geographic position of the individual. The geographical area covered by a BTS ranges from less than one km² in dense urban areas to more than three km² in rural areas. For simplicity, we assume that the cell of each BTS tower is a two-dimensional non-overlapping region and we use Voronoi diagrams to define the area of coverage of each individual BTS. (Figure 1) presents on the left a set of BTSs with the original coverage of each cell, and on the right the approximated coverage computed using Voronoi.

[Figure 1]

CDR (Call Detail Record) databases are generated when a mobile phone connected to the network makes or receives a phone call or uses a service (e.g., SMS, MMS, etc.). In the process, and for invoice purposes, the information regarding the time and the BTS tower where the user was located when the call was initiated is logged, which gives an *indication* of the geographical position of a user at a given moment in time. Note that no information about the exact position of a user in a cell is known. From all the information contained in a CDR, our study considered the encrypted originating number, the encrypted destination number, the time and date of the call, the duration of the call, the BTS that the cell phone was connected to when the call was placed and the BTS tower that the cell phone was connected to when the call finished.

3.2. MOBILITY VARIABLES

In order to characterize the average mobility of the individuals living in the area of coverage of a BTS, we first compute a set of mobility variables for each individual and after that we aggregate these variables at a BTS level. The CDRs were used to compute the following variables that characterize individual mobility:

● *Distance travelled between phone calls (weekly average):* is the distance travelled by a user *between* consecutive calls. For a pair of calls, it is computed as the distance between the coordinates *(latitude,longitude)* of the tower where the first call ended and the coordinates *(latitude,longitude)* of the tower where the following call started. This distance approximates the route that the user has followed.

● *Distance travelled during a phone call (weekly average):* is obtained as the distance between the BTS *(latitude,longitude)* where the cell phone call started and the BTS *(latitude,longitude)* where the cell phone call ended, and is an indication of the mobility of an individual. Note that in 60% of the cases there is no mobility during a phone call i.e., the phone call starts and ends in the same geographic area (BTS).

● *Total Distance travelled (weekly average):* the total distance travelled by an individual is obtained by adding *the distance travelled between phone calls* for each pair of two consecutive calls/SMS/MMS with the *distance travelled during each phone call* as explained in the previous definitions.

● *Diameter of the Area of Influence (weekly average):* The area of influence of an individual is defined as the geographical area where a user spends his/her time doing his/her daily activities. It is computed as the maximum distance (in kilometers) between the set of BTSs used to make/receive all calls during the temporal period under study (in our case each week). Formally, being $\{BTS_i\}$ (i=1:n) the set of BTSs used in the period of time considered, the diameter is defined as $\max(dist(BTS_i, BTS_j)) \forall (i, j = 1:n)$. Note that a user can have a diameter of 0 if all the calls in the period of time considered are placed though the same BTS.

● *Radius of gyration (weekly average):* while in the previous mobility variable all BTSs are considered to be equally important, in the radius of gyration each BTS is weighted by the number of phone calls placed or received at it, and the radius is obtained by computing the centre of masses across all the weighted BTSs. Gonzalez *et al.* [12] and Song *et al.* [13] use the radius of gyration to describe the typical range of a user trajectory in their studies. The authors showed that individuals tend to typically move between two BTS and thus determined that this variable could be a good approximation of the distance between home and work. For each user, the radius of gyration is defined as:

$$r_g(t) = \sqrt{\frac{1}{n(t)} \sum_{i=1}^{n} (r_i - r_{cm})^2} \qquad (1)$$

where r_i with $i=1,...,n(t)$ is the position recorded as longitude and latitude of a BTS and r_{cm} the center of mass of the trajectory, defined as:

$$r_{cm} = \frac{1}{n(t)} \sum_{i=0}^{n} r_i \qquad (2)$$

● *Number of different BTS towers used* (weekly average): this variable is complementary to the diameter of the area of influence and the radius of gyration. In

fact, it is possible to have small values for diameter and radius and a large number of BTSs, which would indicate that although the area where the user moves is not large, the user moves frequently within it. And vice versa, a user can show large areas of influence and a reduced number of BTSs, which would indicate that user activities concentrate on a limited number of distant geographical regions.

Finally, it is important to highlight that in order to map the socio-economic levels of different regions with human mobility we need to have an approximation of the geographical location of the residence of an individual. This residential location will allow us to correlate human mobility specific to certain geographical areas with their socio-economic levels. The residential location is only known for clients that have a contract with the carrier, which in the case of emerging economies accounts for less than 10% of the total population. Thus, in order to carry out large scale analyses, we need to approximate the residential location of the clients that use also the pre-paid option. For that purpose, we used a residential detection algorithm that assigns the residential location of an individual to the region covered by a specific BTS tower. The algorithm computes the residential location based on general calling patterns detected in cell phone records. For our particular case, the pattern that identifies the home location is given by the BTS with the highest number of handled calls between 8pm and 12pm on Mondays, Tuesdays and Wednesdays. It is important to note that this pattern is highly influenced by the cultural factors of each country and cannot be generalized to other geographical areas. Details of the algorithm can be found in [8].

Once each user is assigned a BTS as a residential location, we compute -for each BTS- the average of each mobility variable for all users whose residential location is at that same BTS. These averages represent the aggregate mobility behavior of the users that live in the geographical area covered by each BTS.

3.3. MOBILITY DATASET

For our study we collected the anonymized and encrypted CDR traces from a main city in an emerging economy[1] over a period of six months, from February 2010 to July 2010 (company policy does not allow us to reveal the geographic origin of the data). The city was specifically selected due to its diversity in socio-economic levels. From all the individuals, only users with an average of two daily calls (made or received) were considered in order to filter those individuals with insufficient information to characterize their mobility.

The total number of users obtained after filtering was close to 500,000 which represents around 7% of the inhabitants of the city. Given that the CDRs are provided by a telecommunications company that is the main carrier in the region, it is fair to say that the 7% of the subscribers probably represent a large part of the population. The city considered has an area close to 1,500 km^2 and is covered by 1,000 BTS towers. Each geographical region has an average of 320 users (with a standard deviation of 165) and each BTS gives coverage to an average of 1620 interactions

[1] Based on the 2009 International Monetary Fund (IMF) World Economic Outlook Country Classification www.imf.org/external/pubs/ft/weo/2009/02/weodata/groups.htm

(voice/SMS/MMS) with a standard deviation of 604 over the period under study. In order to get a good understanding of the call records dataset, we next present a brief description of the mobility variables for the population under study.

[Figure 2]

(Figure 2) shows the CDF (Cumulative Distribution Function) for the average number of calls per day i.e., the percentage of subscribers (y axis) that have a minimum number of calls per day (x axis). We can observe that around 70% of cell phones make or receive less than two calls (SMS, MMS or voice) per day, while only the top 1% has more than 10 daily interactions on average. This behavior is typical of emerging economies, where cell phones are scarcely used when compared with developed nations. In fact, only around 20% of users in developed nations have less than two interactions per day (on average). As mentioned earlier, we only consider for analysis individuals with an average of at least 2 interactions per day. Although it might seem that such filter introduces a bias in the sample favoring wealthier individuals, our analysis will show that the proportion of socio-economic levels in our sample is maintained even after the filter is applied.

[Figure 3]
[Figure 4]
[Figure 5]
[Figure 6]

Next, we show some statistics related to the mobility variables computed for the users with at least an average of two calls per day. (Figure 3), (Figure 4), (Figure 5) and (Figure 6) present the CDF for the total distance travelled (averaged per week), average diameter of the area of influence (averaged weekly), average radius of gyration (per week) and the number of different BTSs used (weekly average), respectively. We do not show the CDFs for variables *distance travelled between phone calls* and *distance travelled within a phone call* given that these are indirectly shown in Figure 3.

(Figure 3) shows that the best part of individuals (70%) travel less than 100km on average per week, with just the top 3% travelling more than 500km. (Figure 4) reveals that the total distance travelled typically covers an area of influence with a diameter smaller than 50km i.e., 80% of individuals have a diameter of 50km or less. If we compare the diameter of the area of influence with the radius of gyration shown in (Figure 5), and considering that *2*radius=diameter*, we observe that the radius of the area of influence is approximately twice as large as the radius of gyration across all socio-economic levels. In fact, approximately 80% of the individuals have a radius of gyration smaller than 10km. This indicates that the relation between the two variables is the same independently of the SEL i.e., across all SELs we observe the radius of the area where users carry out their daily activities is twice as large as the radius of the area covered when users go from home to work. Finally, (Figure 6) presents the CDF

for the number of BTSs used weekly, which is an indirect indicator of mobility. In general, we see that the average individual uses less than 20 unique BTSs per week. Note that a small number of BTSs does not necessarily imply a small total distance travelled or a small area of influence or radius of gyration. BTSs are only an indication of the points of interest or areas visited by a user.

3.4. SOCIO-ECONOMIC LEVELS

The socio-economic levels (SELs) for the city under study were obtained from the corresponding National Institute of Statistics. These levels, gathered through national household surveys, give an indication of the social status of an individual relative to the rest of the individuals in the country.

In our particular case, the National Institute defines five SELs (A, B, C, D and E), with A being the highest. The SEL value is obtained from the combination of 134 indicators such as the level of studies of the household members, the number of rooms in the house, the number of cell phones and land lines, computers, combined income, occupation of the members of the household, etc. The SELs are computed for each geographical region (GR) defined by the National Institute. Each GR has between one and three km². Our city under study is composed of 1,200 geographical regions as determined by the National Institute. It is important to highlight that the city does not have GRs with a socio-economic level E. The rest of SEL levels are as follows: A levels represent 8% of the GRs, B 22%, C 38% and D 32%.

4. MATCHING MOBILITY DATA WITH SEL

The datasets and variables described in the previous section provide aggregated human mobility characteristics at a BTS level and SELs that characterize geographical regions (GRs). In order to study the relationship between SELs and human physical mobility we first need to geographically map BTS coverage areas with GRs, in order to compute a SEL value for each area of coverage of each BTS.

[Figure 7]

Formally, we seek to associate to the area of coverage (cell) of each BTS the set of GRs that are totally or partially included in it. Each GR within the cell of a BTS will have a weight associated to it. The weight represents the percentage of the BTS cell covered by each particular GR. A graphical example is shown in (Figure 7). (Figure 7a) presents the set of GRs (00001 through 0005) defined by the National Statistics Institute. Each GR has an associated SEL value (A...D). (Figure 7b) represents, for the same geographical area, the BTS towers (ct1 though ct7) and their coverage approximated with Voronoi tessellation. Finally, (Figure 7c) shows the overlap between both representations. This mapping allows expressing the area of coverage of each BTS cell tower (ct) as a function of the GRs as follows:

$$ct_i = w_1 GR_1 + ... + w_n GR_n \qquad (3)$$

where w_1 represents the fraction of the GR_i that partially covers a certain coverage area of BTS tower ct_i. Following the example in (Figure 7), ct_1 is completely included in GR 0001 and as such n=1 and w_1=1. The same reasoning applies to ct_3. A more common scenario is ct_4, which is partially covered by GR 00003, 00001 and 00005 with $n=3$ and weights $w_1=0.68$, $w_2=0.17$ and $w_3=0.15$. The process to obtain the mapping between the BTS coverage areas (cts) and the GRs uses a *scan line* algorithm to compute the numerical representations of each GR and BTS map [8]. These representations are then used to compute the fractions of the BTS coverage areas covered by each GR. A more detailed description of the algorithm can be found in [7].

Once each BTS tower is represented by a set of GRs and weights, we can assign a SEL value to each BTS in the city under study. To do so, we first assign numerical values to each SEL level as follows: A=100, B=75, C=50, D=25 and E=0, thus transforming the discrete SEL values into a [0-100] range. The final SEL value associated to a BTS can be obtained by computing Equation (3) with the numerical SEL values for each GR. Following the previous examples and assuming that the SEL of GRs 00001, 00005 and 00003 are respectively B, B and C, the SEL associated with BTS ct_1 and ct_3 will be 75, while the SEL associated with BTS ct_4 will be 0.68*75+0.17*50+0.15*75=70.75.

[Figure 8]

(Figure 8) shows the number of BTS towers with a specific socio-economic level in the city under study. The SELs are represented as a continuous range (0-100) and divided into bins of size 2.5. We observe that although the original data had GRs with SEL A, these are not present in the BTSs as a consequence of the mapping and weighting introduced by Formula (3). To obtain a BTS with a SEL A, its area of coverage would have to be completely defined by GRs with an A level, which does not happen in the city under study given the limited number of A level GRs present in our sample. Also because in the city under study there were no GRs with an E SEL, no BTS has a SEL value smaller that 25 (D).

5. METHODOLOGY

Once the mapping between SELs and BTSs has been done, each BTS can be characterized by the aggregated values of the human mobility variables and by a SEL value. In order to understand the relationship between SELs and human mobility, we compute the Pearson's correlation coefficient (noted as r) for each pair of SEL and mobility variable (six as defined in Section III) gathered from the 1,000 BTS towers in the city under study. The Pearson's correlation coefficient is a measure, in the range [-1,+1], of the linear dependence between two variables, where a value of 1 (or -1) implies that a linear equation describes the relationship between the two variables perfectly, and a value of 0 implies that there is no linear correlation between the variables. For our study only values of r in the range [1, 0.5) and (-0.5, -1] are considered relevant [14].

In order to validate the statistical significance of the correlation values, a p-value was computed by creating a t-statistic with *n-2* degrees of freedom, where *n* is the number of BTS towers. Only those values of correlation with a significance of $p<0.01$ are considered valid. For the pairs of SEL and mobility variables that have a $p<0.01$ and an *r* in the range *[1, 0.5)* or *(-0.5, -1]* a linear least square method is applied to estimate the parameters of the linear regression model that explains the variance of the mobility variable with the SEL. The identification of the mobility variables that have a high correlation with SEL can be used as indicators of the socio-economic wealth of the population under study.

6. RESULT ANALYSIS

(Table 1) presents the Pearson's correlation coefficients, ordered by its relevance, for each one of the mobility variables defined in Section III. All of the correlations had a $p<0.01$ probably due to the large amount of pairs (1,000) used to compute *r*. As can be seen, only three of the six mobility variables have an *r* value in the range *[1, 0.5)* or *(-0.5, -1]*: *Number of different BTSs used, Radius of gyration and Diameter of the Area of Influence*. The correlation coefficients for the other three variables show that there is no linear relation between the SEL and the distance travelled: total, during or between phone calls. This indicates that while an increase in the SEL is correlated to an increase in the radius of gyration, diameter of the area of influence and number of different BTSs used by an individual, such increase is not correlated to an increment in the total distance traveled *i.e.,* individuals with higher socio-economic levels carry out their daily activities in larger geographical areas when compared with individuals with lower socio-economic levels, but this fact is not correlated with total travelled distances being larger. Additionally, it is important to highlight that these high correlations do not imply a causality relation between the SEL and each one of the variables i.e., having a large area of influence does not cause a high SEL or *viceversa*.

[Table 1]

(Figure 9), (Figure 10) and (Figure 11) depict the data points used for the correlation analysis (pairs of mobility variable vs. SEL obtained from each BTS) and the result of the linear regression for the variables: *total number of BTSs* used, the *Radius of Gyration* and the *Diameter of the area of influence,* respectively. The X axis represents the SEL expressed in the range [0,100] and the Y axis indicates kilometers (in (Figure 10) and (Figure 11)) or absolute number of towers in (Figure 9). Each Figure also shows the regression equation where X stands for the socio-economic level (SEL) and Y for the corresponding mobility variable. (Table 2) shows the average errors and average quadratic errors for each of the regressions. The three plots show that the majority of the data points are clustered around socio-economic levels C and D in accordance with the results presented in (Figure 8).

[Table 2]

[Figure 9]
[Figure 10]
[Figure 11]

The variable *number of different BTSs used* (depicted in (Figure 9)) shows that while lower socio-economic levels use around six different BTSs on average every week (an individual with a SEL of E will approximately use five), higher socio-economic levels use a number of BTSs larger than eight, with an estimate that the number of BTSs used by an individual of SEL A is around nine. This result shows that higher SELs tend to visit more points of interest in the city than lower SELs, although this does not necessarily translate into traveling longer distances.

Regarding the *radius of gyration*, (Figure 10) shows that individuals with higher SEL tend to have a higher radius than individuals with lower SELs. In particular, while socio-economic levels C and D have a radius of gyration of approximately five to ten kilometers, individuals with socio-economic levels B and above have a radius of 13 km. or more. Using the regression equation, we can determine that individuals with an E SEL will have a radius of 3.8 km, while individuals with SEL A will have an approximate radius of 15 km. In any case, the values of radius of gyration are in accordance with the values presented in [12]. As for the *diameter of influence,* (Figure 11) shows that while individuals from low socio-economic levels have a diameter between 20 and 30 km. (individuals with an E SEL have a diameter of 13km. using the regression equation), individuals with socio-economic levels B and above have a diameter larger than 35 km (individuals with a SEL of A would have a diameter of 46 km.). The approximation given for A and E SELs is made under the assumption that for those cases the linear relation applies.

6.1. IMPLICATIONS FOR PUBLIC POLICY DESIGN

The previous results provide insights that might be useful for a variety of public policy fields. In this section, we present an analysis of these findings and its implications noting that they cannot be directly generalized to all emerging economies. In fact, our findings only hold for the particular city under study and cannot be directly extended to other cases. However, the methodology presented in this paper could potentially be used to replicate the analysis for other region given the necessary information is available.

Researchers have shown that the radius of gyration can be used as an indication of the distance between home and work [12]. Based on these findings and on the results shown in (Figure 10), we could hypothesize that individuals with higher SEL tend to live further away from their jobs than individuals with lower SELs since the geographical areas with high SELs have a radius three times larger than the areas with lower SELs. This fact could be caused by a variety of reasons, among others a limited public transportation infrastructure and the cost of having a car (1 car for every 8 citizens for the city under study compared to 1 car for every 1.5 citizens in an average developed economy). In any case, this could be interpreted as an indication that neighborhoods with lower SELs are typically more isolated than their higher SEL

counterparts which would imply that their inhabitants might have fewer opportunities to improve their quality of live since they are more geographically isolated. Thus, a strong public transportation infrastructure would highly important for improving quality of live as it opens opportunities to go beyond the original neighborhood where their daily activities take place and have access to jobs that are geographically located at larger distances. Our research findings can be used to help in the design of public transportation infrastructure using the radius of gyration provided for each BTS. By correlating such information with, for example, the existing bus routes, we can identify the cells with lower radiuses that are not covered or are poorly covered by the public transportation system. These areas will have to be prioritized to provide citizens with access to a larger variety of professional opportunities in order to enhance their quality of life.

As opposed to the radius of gyration, the diameter of the area of influence considers all BTS towers equally important and as such, is a good indicator of the geographical area where all daily activities take place. We have observed that there is a difference of a factor of three between the diameter of influence in geographical areas with high SEL and areas with low SEL. The same considerations mentioned for the radius of gyration apply here, but can be generalized to all activities such as leisure or commercial activities i.e., areas with low SEL have smaller diameters of influence and thus a limited access to leisure or commercial activities located at larger distances that their high SEL counterpart. Thus, it is important to take into account the SEL distribution when designing public transportation infrastructures that give access to work and leisure opportunities, since these will probably increase the possibility of citizens moving up the economic ladder.

Finally, the diameter of the area of influence is a relevant indicator not only for public transport policies, but also for policies in the area of epidemic spreading control. In fact, the diameter can be used to measure the risk and the speed of the spread of a virus in a geographical area i.e., the larger the diameter, the larger the area where people move and interact with others, and thus the higher the risk of virus spreading [6]. Although the epidemic spreading might depend on a variety of variables such as incubation periods, transmission channels, education of the citizens or social relationships, the mobility patterns always influence the spreading process. Thus, from our analysis, we could determine that in areas with lower socio-economic levels it would be easier to contain the epidemic than in areas with higher SEL, where individuals carry out their daily activities in larger geographical areas. It is important to highlight that although these results provide important insights to model and analyze the spreading of specific epidemics (like HIV or flu) many other variables need to be taken into account to build a realistic spread simulation as presented in [19].

6.2. SOCIO-ECONOMIC MODEL BASED ON PHYSICAL MOBILITY

In this section we present a formal model that approximates socio-economic levels using only physical mobility information. In Section 5, we have shown that there exists a high correlation between the aggregated mobility variables at a BTS level and the SELs. This finding makes it possible to design a model to approximate SEL using

physical mobility variables. Although a multivariate model using the variables with higher correlation to the SELs (different number of BTS, radius of gyration and diameter of the area of mobility) would seem like the best option, the high correlation among these three variables does not justify such approach. For that reason, we only use the mobility variable that has the highest correlation with the SELs (number of BTSs) to design the formal model. By reversing the linear regression model presented in (Figure 9), the SEL at a BTS level can be estimated as:

$$SEL = 25 * Number\_Different\_BTSs - 127,75 \quad (4)$$

with an adjusted R square value of 0.72. The equation will be valid for:

$$5,11 < Number\_Different\_BTSs < 9,11 \quad (5)$$

Nevertheless, for values smaller than 5.11 the SEL can be considered E and for values larger than 9.11 it can be considered A. As a result, Equation (4) presents a formal model or mechanism to estimate the SEL of the individuals that live in the area of coverage of a particular BTS by aggregating their mobility behavior and considering only one variable. This estimation is of great value for emerging economies where, as discussed in the Introduction, estimating the actual SEL of specific geographic regions can be difficult and/or expensive.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have studied the relation between socio-economic levels and a variety of physical mobility indicators. Although there are studies that measure the relation between SEL and human mobility in specific scenarios, such as access to hospital, cancer prevalence or access to public transport [1, 2, 4, 5], none of these studies have been able to directly measure human physical mobility and hence, their conclusions tend to be based on a limited number of individuals and on data obtained through interviews or small scale surveys.

By using cell phone records, we have at our disposal an almost unlimited number of individuals whose mobility data can be retrieved *objectively* without the need of interviews. After matching SEL with a group of six mobility variables at a BTS (aggregated) level we have identified three relevant correlations between SEL and the number of different BTS towers used, the radius of gyration and the diameter of the area of influence. All three cases share a pattern by which the increase in SEL follows a linear relation with the increase of the mobility variable. We have also discussed the potential of such analyses for public policy decision making in areas like transport planning or epidemiology. Additionally, we have presented a formal model that approximates the SEL of a geographic area from a unique mobility variable that characterizes the average cell phone usage of the citizens living in that area. This formal model might be used as a proxy to approximate regional SELs which in emerging economies can be difficult and costly to compute.

It is important to clarify that the implications and formal models presented in this

paper are not necessarily the same for other countries or even for other regions within the same country. Also, it would be very relevant to analyze the impact that changes in weekdays versus weekend physical mobility behaviors have in the analyses that we have presented in the paper. In principle, the same technique discussed here can be used to study the relation between SEL and a large variety of variables modeling how technology and specially cell phones are used at a scale and detail never done before (ranging from mobile internet access to characteristics of social networks).

References

[1] Propper, C., Diamiani, M., Leckie & G., Dixon, J. (2007). "Impact of patients' socioeconomic status on the distance travelled for hospital admission in the English National Health Service" in Journal Health Serv. Res. Policy;*12*:153-159.

[2] Maheswaran, R., Pearson, T., Jordan & H., Black, D. (2006). "Socioeconomic deprivation, travel distance, location of service, and uptake of breast cancer screening in North Derbyshire, UK", in *Journal of Epidemiology Community Health*;60:208-212.

[3] ITU ICT-Eye Free Statistics, http://www.itu.int/ITU-D/ict/statistics/

[4] Carlsson-Kanyama, A. & Liden, A. (1999). "Travel patterns and environmental effects now and in the future: implications of differences in energy consumption among socio-economic groups" in *Ecological Economics 30(3)*, pp 405-417.

[5] Mohammed Ahsan, H., Mizanur Rahman, Md. & Habib, K. (2002). "Socio Economic Status and Travel Behavior of inter-city bus passengers: Bangladesh Perspective", in *Journal of Civil Engineering* 30(2).

[6] Rubio, A., Frias-Martinez, V., Frias-Martinez E. & Oliver, N. (2010, March). "Human Mobility in Advanced and Developing Economies: A Comparative Analysis", in *AAAI Spring Symposia Artificial Intelligence for Development, AI-D*, Stanford, USA.

[7] Eagle, N., Macy, M. & Claxton, R. (2010). "Network Diversity and Economic Development", Science 328 2029-1030.

[8] Frias-Martinez, V., Virseda, J., Rubio, A. & Frias-Martinez, E. (2010, March). "Towards Large Scale Technology Impact Analyses: Automatic Residential Localization from Mobile Phone-Call Data", in *International Conference on Information & Communication Technologies and Development ICTD, London,UK.*

[9] Lane,M., Carpenter, L.C., Whitted, T. & Blinn, J.F. (1980). "Scan line methods for displaying parametrically defined surfaces," *Communications ACM*, vol. 23, no. 1.

[10] Brockmann, D. and Theis, F. (2008). "Money Circulation, Trackable Items and the Emergence of Universal Human Mobility Patterns", Pervasive Computing 7, Nr. 4, 28.

[11] Ratti, C., Liu, L., Hou, A., Biderman, A. & Chen, J. (2008). "Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen", Institute for Electrical and Electronics Engineers.

[12] Gonzalez, M., Hidalgo, C. A. & Barabasi, A.-L. (2008). "Understanding individual human *mobility* patterns", Nature, 453, 779 – 782.

[13] Song, C., Qu, Z., Blum, N. & Barabasi, A.-L. (2010). "Limits of Predictability in Human Mobility", Science, Vol. 327. no. 5968, pp. 1018 – 1021.

[14] Cohen J. (1988). "Statistical Power Analysis for the Behavioral Sciences", Lawrence Erlbaum Associates, 2nd Edition.

[15] Village Phone, Grameen Bank, http://www.villagephonedirect.org/contents/

[16] Eagle N., txtEagle, www.txteagle.com

[17] Molnar A. and Frias-Martinez V. (2011). "EducaMovil: Mobile Educational Games Made Easy", Ed-Media World Conference on Educational Multimedia, Hypermedia and Telecommunications.

[18] DeRenzi B., Lesh N., Parikh T., Sims C., Mitchell M., Maokola W., Chemba M., Hamisi Y., Schellenberg D. and Borriello G. (2008). "e-IMCI: Improving Pediatric Health Care in Low-Income Countries", Proceedings of Conference on Computer-Human Interaction.

[19] Frias-Martinez E., Williamson G. And Frias-Martinez V. (2008). "An Agent-Based Model of Epidemic Spread using Human Mobility and Social Network Information", SocialCom.

TABLE 1

Correlation	Value
(SEL, *Number of different BTS*)	0.58
(SEL, *Radius of gyration*)	0.54
(SEL, *Diameter of the Area of Influence*)	0.53
(SEL, *Distance travelled between phone calls)*	0.39
(SEL, *Total Distance travelled*)	0.37
(SEL, *Distance travelled during a phone call*)	-0.11

Table 1. CORRELATION VALUES FOR EACH PAIR OF MOBILITY VARIABLE AND SEL ORDERED BY THEIR RELEVANCE

Data Points in the Fitting	Average Quadratic Error	Average Error
(SEL, *Number of different BTS*)	0.66	0.52
(SEL, *Radius of gyration*)	6.42	1.66
(SEL, *Diameter of the Area of Influence*)	58.14	4.95

Table 2. AVERAGE ERROR AND QUADRATIC ERROR OF THE LINEAR REGRESSION FITTINGS FOR THE FOLLOWING VARIABLES: NUMBER OF DIFFERENT BTSs USED, AVERAGE DIAMETER OF THE AREA OF INFLUENCE AND AVERAGE RADIUS OF GYRATION.

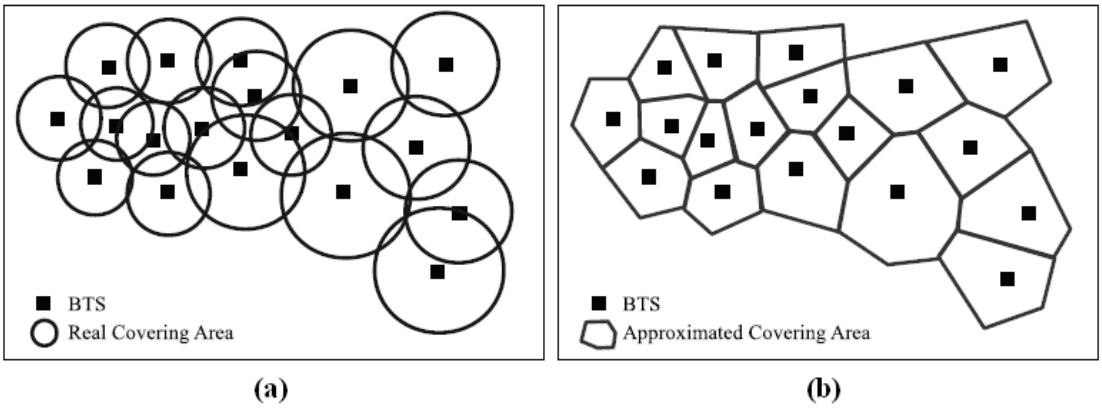

Figure 1. (1a) Original coverage areas of BTSs and (1b) approximation of coverage areas by Voronoi diagram.

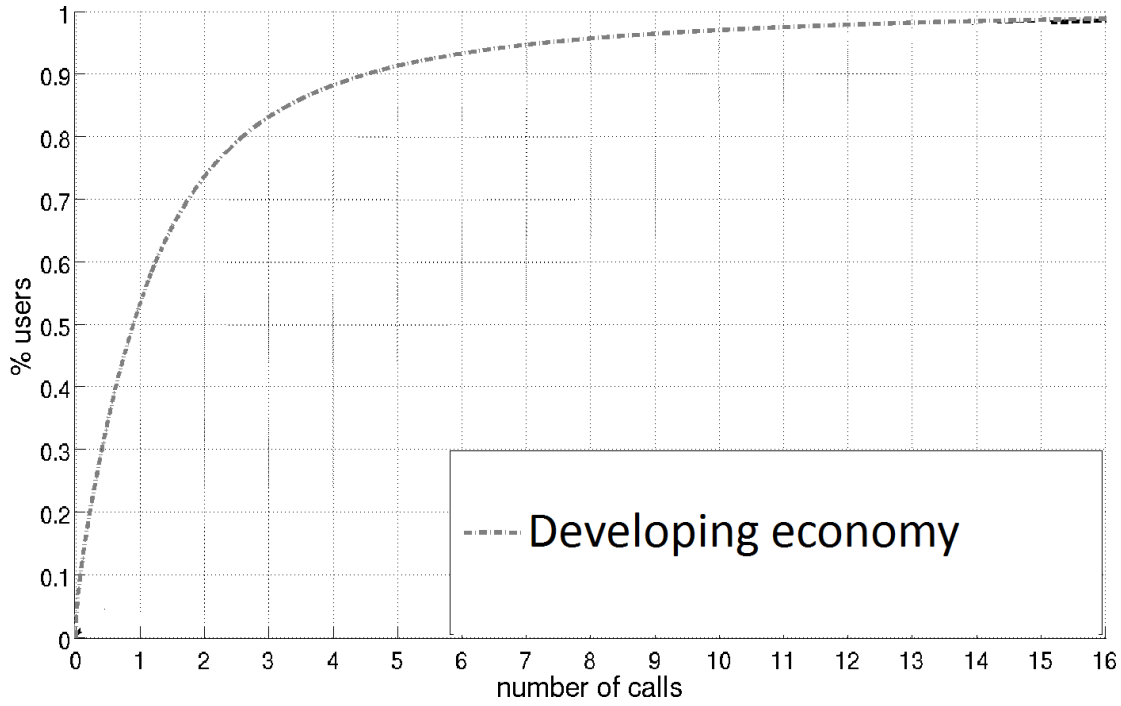

Figure 2. CDF of the average number of calls per day.

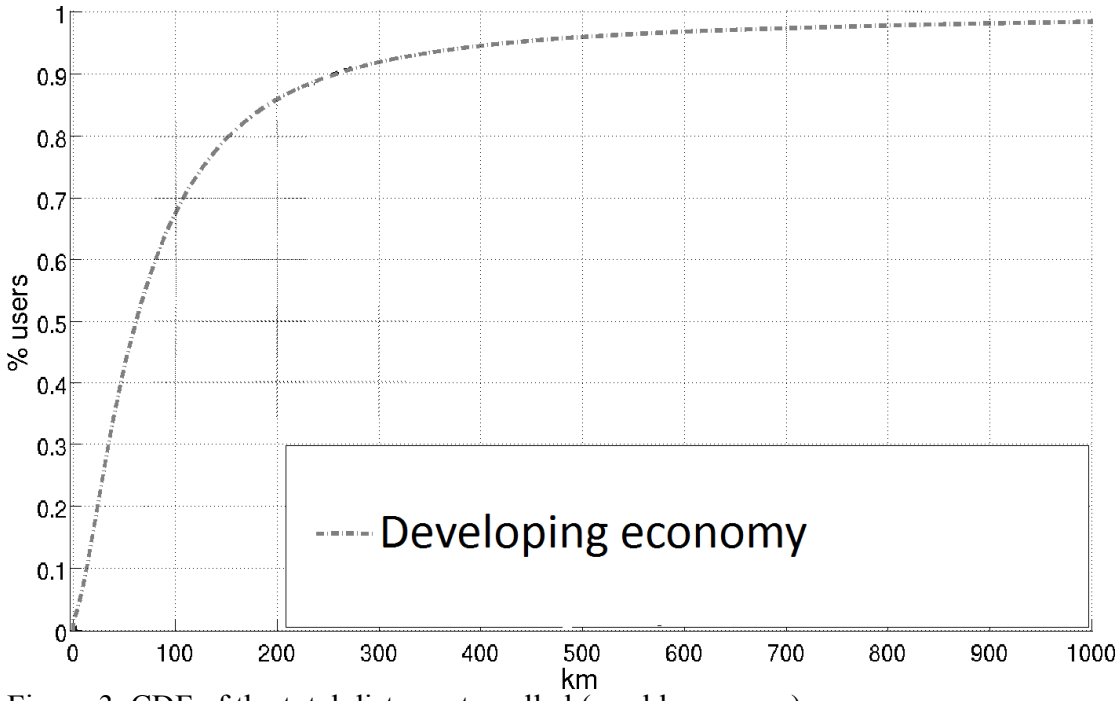

Figure 3. CDF of the total distance travelled (weekly average).

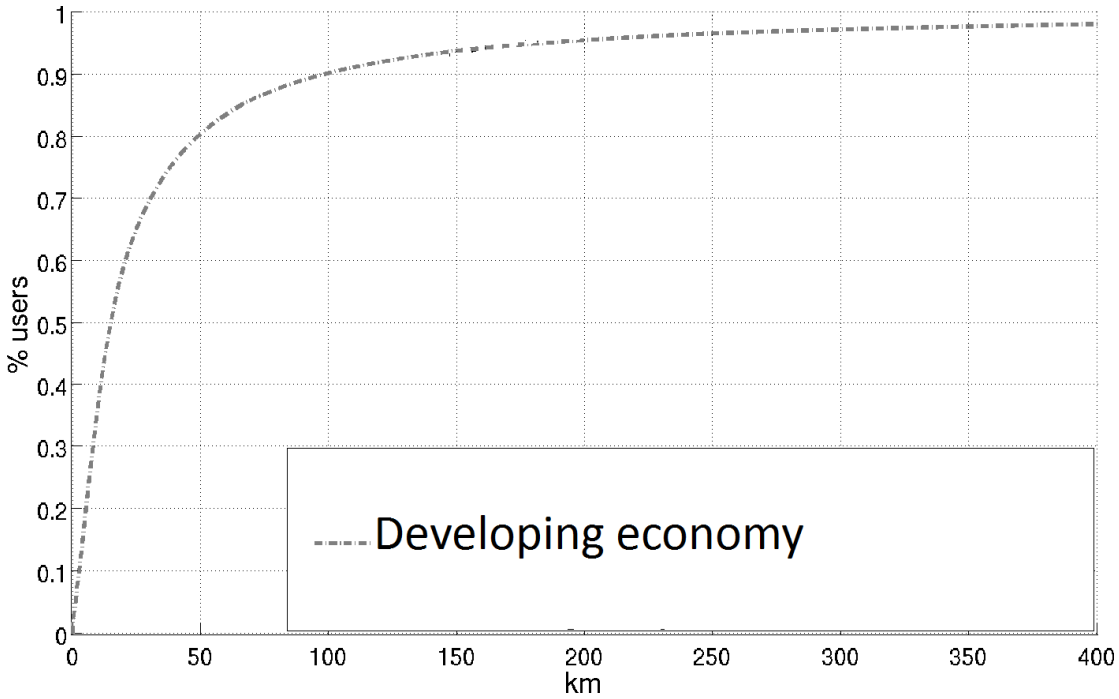

Figure 4. CDF of the average diameter of the area of influence (weekly average).

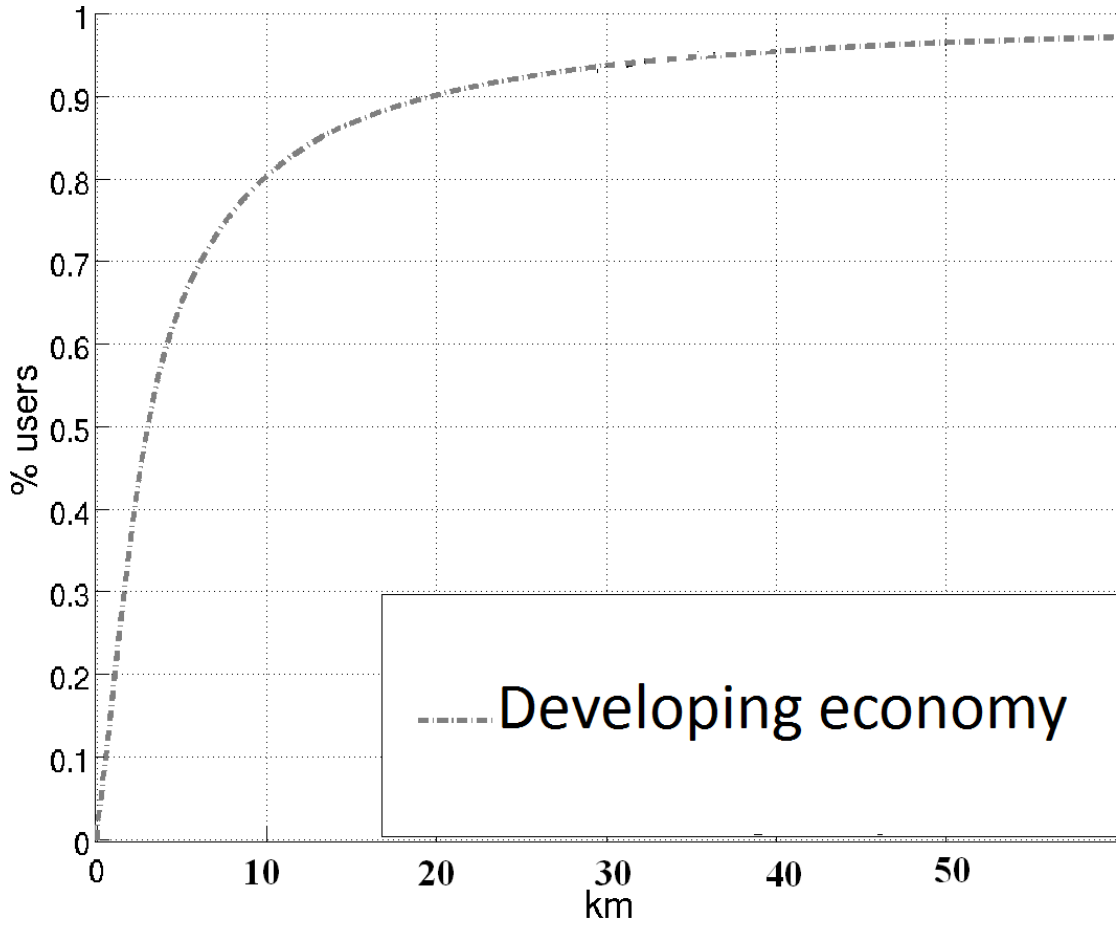

Figure 5. CDF of the average radius of gyration (weekly average).

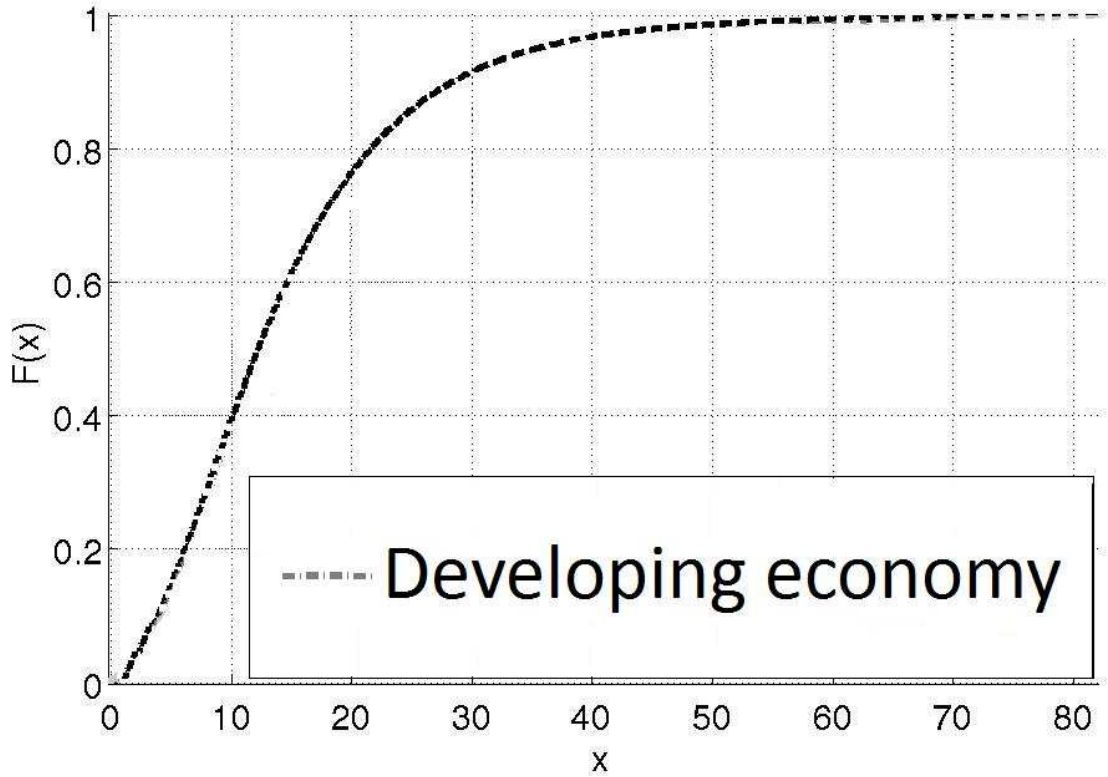

Figure 6. CDF of the number of BTSs used (weekly average).

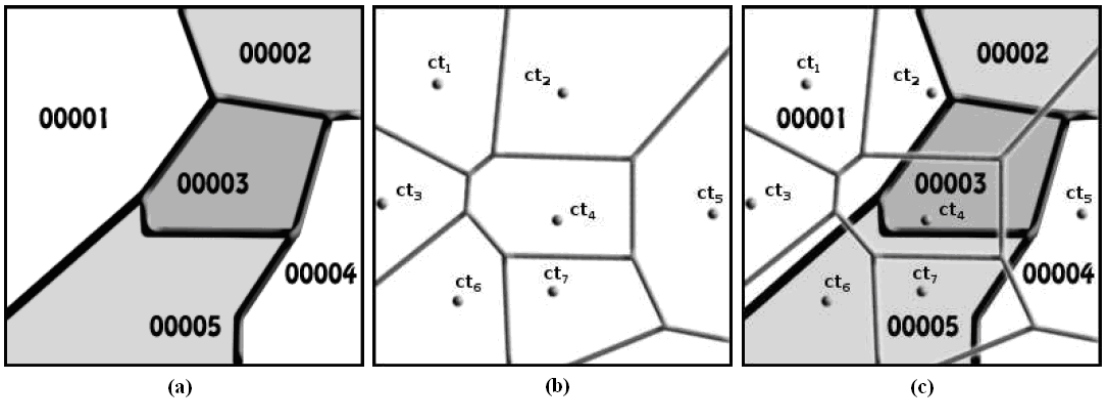

Figure 7. (7a) Example of Geographical Regions (GR) that have a SEL associated; (7b) the same geographical areas with the BTS towers (coverage approximated with Voronoi tessellation) and (7c) the correspondence between GRs and BTS towers used by a scanning algorithm to assign a SEL to a BTS tower area.

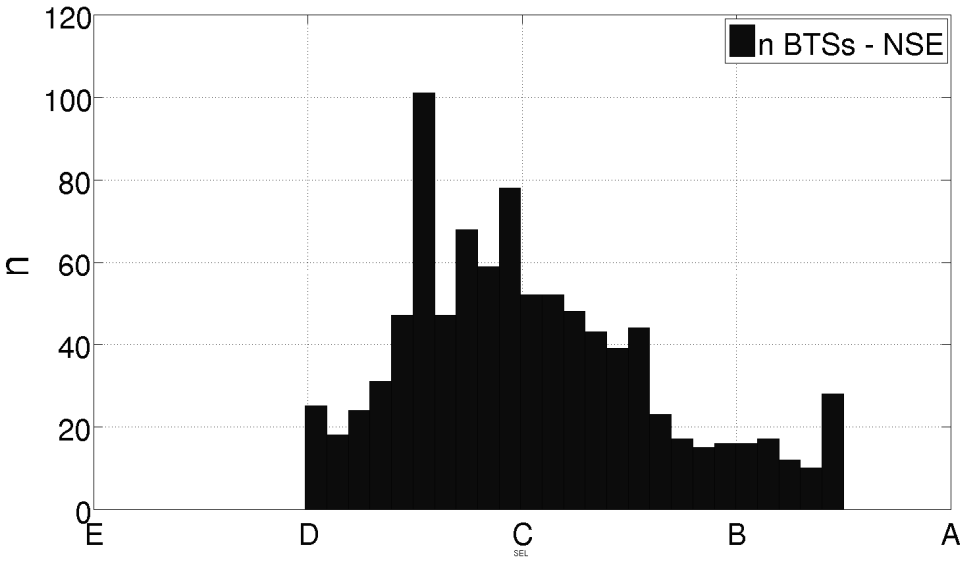

Figure 8. Number of BTS towers for each SEL after applying the discrete to continuous transformation and divided into bins of size 2.5.

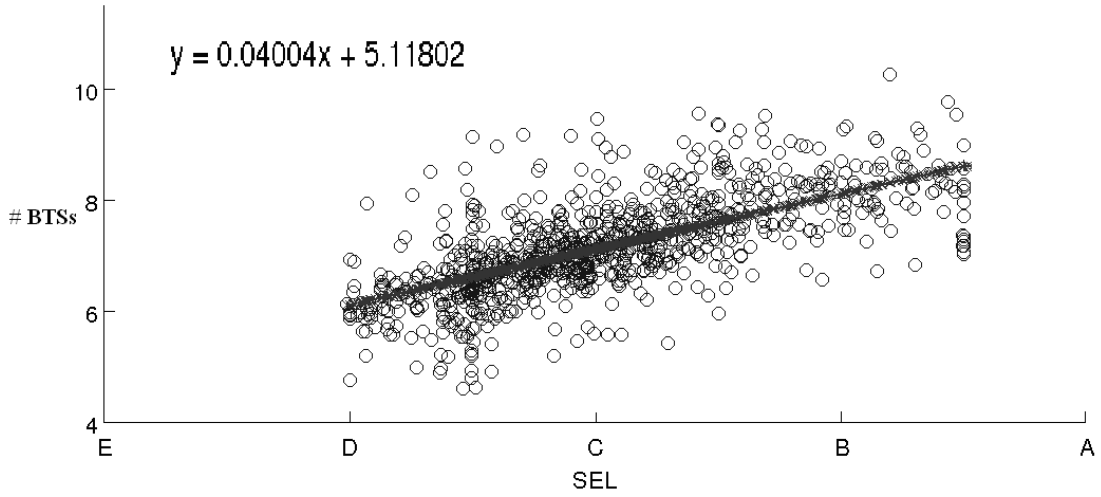

Figure 9. Socio-economic level (SEL) in the X axis versus the weekly average of different number of BTSs used (Y axis) and the fit obtained from applying linear least squares regression.

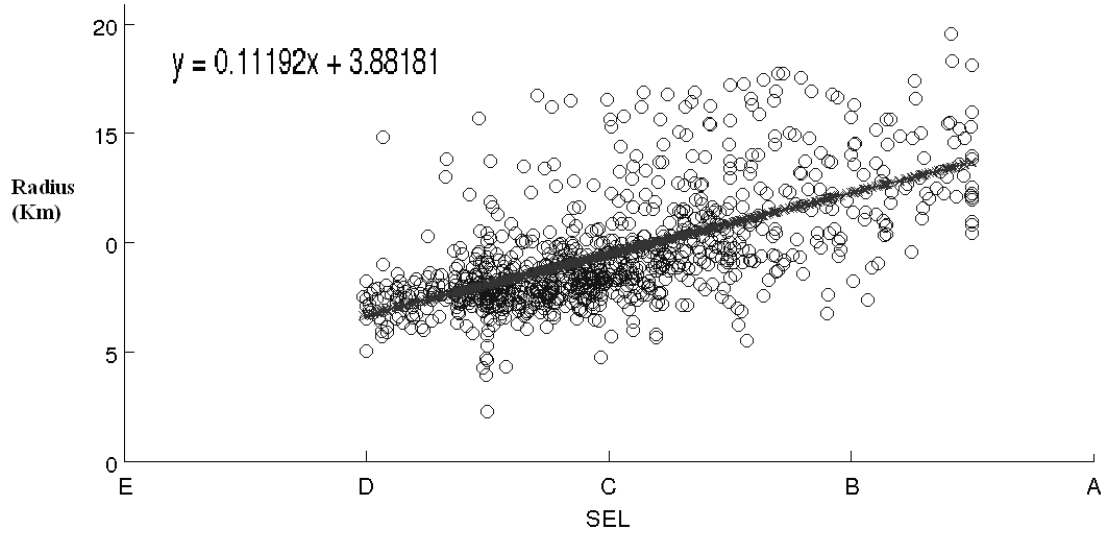

Figure 10. Socio-economic level (SEL) in the X axis versus the weekly average of the radius of gyration measured in km (Y axis) and the fit obtained from applying linear least squares regression.

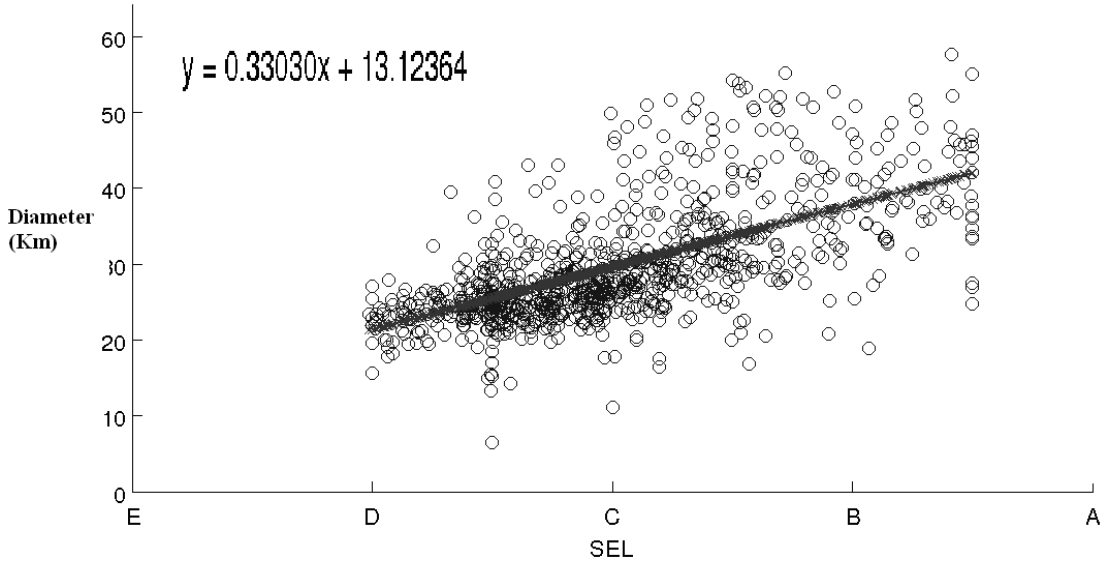

Figure 11. Socio-economic level (SEL) in the X axis versus the weekly average of the diameter of the area of influence measured in km (Y axis) and the fit obtained from applying linear least squares regression.