Socio-Economic Levels and Human Mobility

V. Frias-Martinez, J. Virseda, E. Frias-Martinez

***Abstract*— Socio-economic levels provide 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 mobility has been researched mostly in areas like epidemic spreading and public transportation [2]. However, such studies have been limited in scope mostly due to the lack of large scale human mobility information. Anyhow, the recent adoption of cell phones by large social groups in both emerging and emergent economies has made it possible to capture large scale data about human mobility, which combined with regional socio-economic levels allow to study the impact that socio-economic indices might have in human mobility. In this paper we study the relationship between human mobility variables and socio-economic levels using cell phone traces. Our results indicate that population with higher socio-economic levels is strongly linked to larger mobility ranges than population from lower socio-economic status.**

***Index Terms*— human mobility, socio-economic indicators, ICTs for development.**

I. INTRODUCTION

THE relation between human mobility and socio-economic levels 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 $1 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 –for the first time in human history– of large scale quantitative data about 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. Given that high penetration rates are also common across emerging economies, CDRs can be used to model human mobility worldwide.

In this paper we analyze whether there is a relation between different 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 have relevant implications for a variety of fields ranging from transport planning to virus spreading containment.

The rest of the paper is organized as follows: first we describe 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, indicating a strong correlation between the socio-economic indicators and the range of mobility.

II. 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 [3]; rural-urban differences regarding cancer [2] or travel behaviour of inter-city bus passengers [4]. However, to the best of our knowledge there are no studies that measure the impact of socio-economic levels (SEL) on individual human mobility at the large scale we propose.

Related research has also compared mobility variables between a developed and a developing economy [5]. The difference with our study is that while in [5] 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.* study the relation between socio-economic levels and social network diversity using also cell phone records and social development indicators in the UK [6]. Their findings indicate that social network diversity seems to be a very strong indicator of the development of large online social communities.

V. Frias-Martinez, J. Virseda and E. Frias-Martinez are with the Data Mining and User Modeling Group, Telefonica Research, Madrid, Spain (email: vanessa@tid.es; jvjerez@tid.es; efm@tid.es).

III. DATA ACQUISITION

A. 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 1 km² in dense urban areas to more than 3 km² in rural areas. For simplicity, we assume that the cell of each BTS tower is a 2-dimensional non-overlapping region and we use Voronoi diagrams to define the areas 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.

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.

For our study we collected the anonymized and encrypted CDR traces of a main city in a Latin-American country for a period of 6 months, from February 2010 to July 2010. 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 were considered in order to filter those individuals with insufficient information to characterize their mobility. The total number of users considered after filtering was close to 500,000. The city selected was covered by 1,000 BTS towers.

B. Aggregated Mobility Variables

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 second call started. This distance approximates the route that the user has followed.

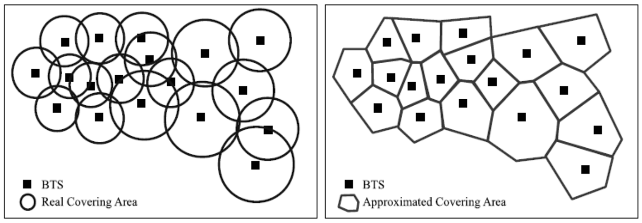


Fig. 1 (left) Original coverage areas of BTSs and (right) approximation of coverage areas by Voronoi diagram.

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 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).

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 computing the centre of masses across all the weighted BTSs. Gonzalez *et al.* [11] and Song *et al.* [12] 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 radious 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 the $i=1,\ldots,n(t)$ is the position recorded as longitude and latitude of a BTS and r_{cm} the center of mass of the trajectory, defined as:

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

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.

Residence Location: 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 analysis, we need to approximate the residential location of the clients that use 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 [7]. The algorithm computes the residential location based on the calling patterns detected in his/her calling detail records.

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 behaviour of the users that live in the geographical area that approximates the coverage of each BTS (Voronoi polygon).

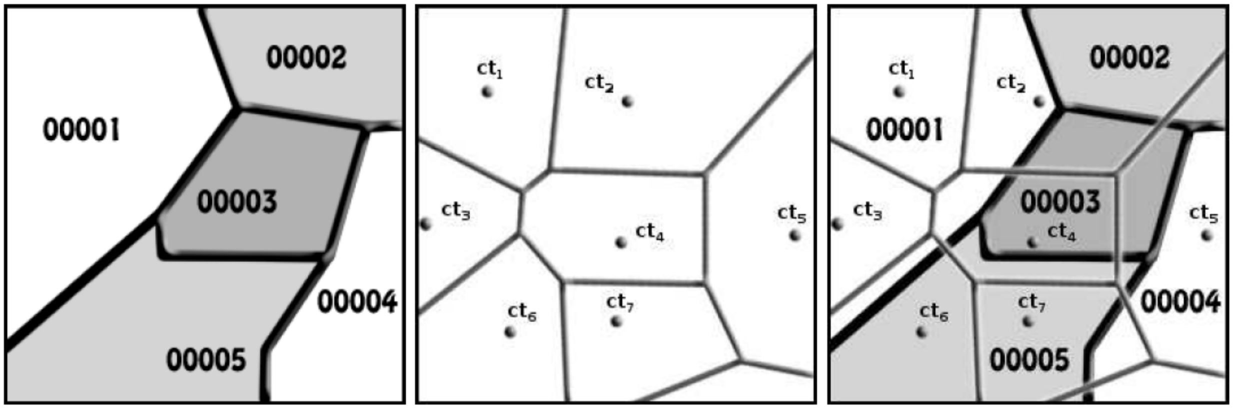
C. *Socio Economic Levels (SEL)*

The distribution of the socio-economic levels 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 SEL. 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 defined by the National Institute. Each GR has between one and four km^2. Our city under study is composed of 1,200 geographical regions (GR) 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%.

IV. MATCHING MOBILITY DATA WITH SEL

The data obtained in the previous section provides aggregated human mobility characteristics at a BTS level and SELs that characterize each geographical region (GR). In order to study the relationship between SELs and human mobility we first need to geographically map BTS coverage areas (cells) with GRs by computing a SEL value for each geographical area covered by each BTS.

Formally, we seek to associate to the area of coverage of each BTS a set of GRs that are totally or partially included in it. Each GR within the BTS area of coverage will have a weight associated to it. The weight represents the percentage of the BTS cell covered by the GR. A graphical example is shown in Figure 2. Figure 2(left) presents the set of GRs (00001 through 0005) defined by the National Statistics Institute. Each GR has an associated SEL value (A...D). Figure 2(center) represents, for the same geographical area, the BTS towers (ct1 though ct7) and their coverage simulated with Voronoi tessellation. Finally Figure 2 (right) shows the overlap



(left) (center) (right)

Figure 2. (left) Example of Geographical Regions (GR) that have a SEL associated; (center) the same geographical areas with the BTS towers (coverage approximated with Voronoi tessellation) and (right) the correspondence between GRs and BTS towers used by a scanning algorithm to assign a SEL to a BTS tower area.

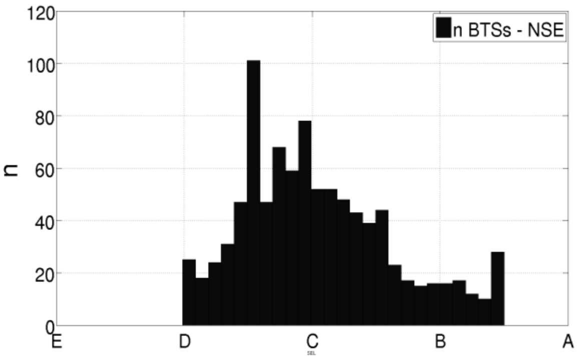


Fig.3. Number of BTS towers for each SEL after applying the transformation using bins of size 2.5.

between both representations. This mapping allows to express the area of coverage of each BTS cell (ct) as a function of the GRs as follows:

$$\mathrm{ct}_{\mathrm{i}} = \mathrm{w}_1 GR_1 + \ldots + w_n GR_n \qquad (3)$$

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 2, 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 compute 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 Formula (3) with the numerical SEL values for each GR (instead of its letter). 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 3 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 an A SEL, these have disappeared in the BTSs as a consequence of the mapping and weighting introduced by Formula (3). Also because in the city under study there were no GRs with an E SEL, no BTS has a SEL smaller that 25 (D).

TABLE I
CORRELATION VALUES FOR EACH PAIR OF MOBILITY VARIABLE AND SEL ORDERED BY THEIR RELEVANCE

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

V. 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 [13].

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.

VI. 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 of *[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 indicate 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 SEL is correlated to an increase in the radius, diameter and number of different BTSs used by an individual, this increase is not correlated to an increment in the total

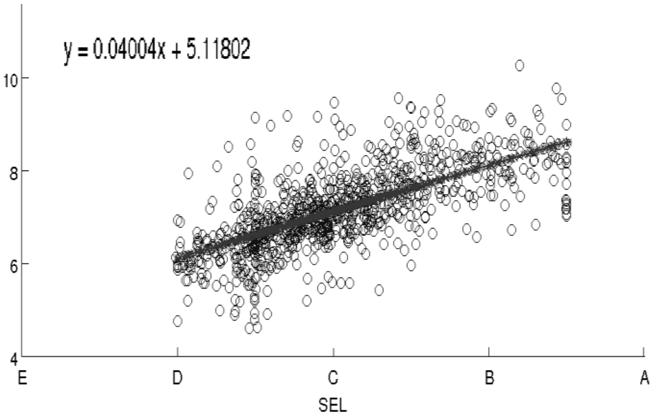


Fig. 4. Socio-economic level (SEL), in the X axis, vs. the weekly average of different number of BTSs used (Y axis) and the fit obtained from applying linear least squares regression.

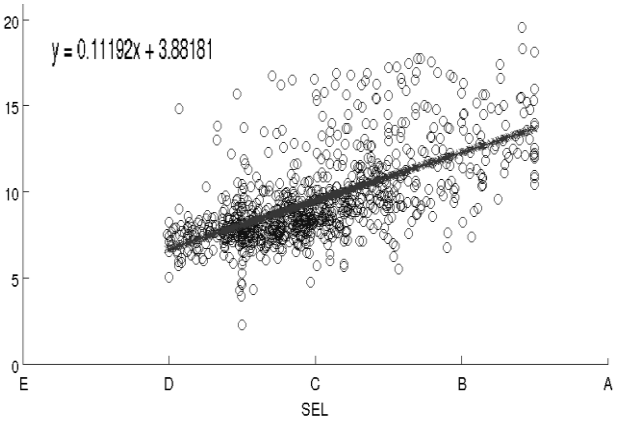


Fig. 5. Socio-economic level (SEL), in the X axis, vs. the weekly average of the radius of gyration measured in km (Y axis) and the fit obtained from applying linear least squares regression.

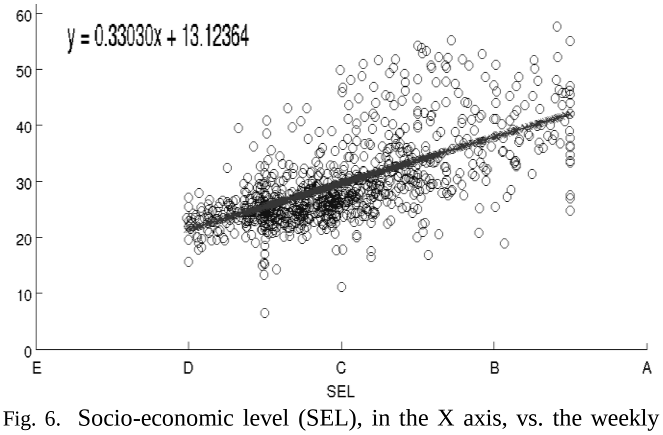


Fig. 6. Socio-economic level (SEL), in the X axis, vs. the weekly average of the diameter of the area of influence measured in km (Y axis) and the fit obtained from applying linear least squares regression.

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 to travel distances being longer.

Figures 4, 5 and 6 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 kms (in Figures 5 and 6) or absolute number of towers in Figure 4. Each Figure also shows the regression equation where X stands for the socio-economic level (SEL) and Y is the corresponding mobility variable. The average quadratic errors of the regressions are 0.66, 6.42 and 58.13 respectively. 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 3.

As explained earlier, researchers have shown that the Radius of Gyration could be used as an indication of the distance between home and work [11]. Thus, Figure 5 shows that individuals with higher SEL tend to live further way from their jobs than individuals with lower SELs. In particular, while socio-economic levels C and D have a radius of gyration of approximately 5-10km, individuals with socio-economic levels B and above have a radius of 13km. 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 15km.

As opposed to the radius of gyration, the area of influence considers all BTS towers equally important and as such, is a good indicator of the area where the most and the least frequent daily activities take place. Such variable has been used, for example, to measure the risk and the speed of the spread of a virus: the larger the diameter, the higher the risk of virus spreading [5]. Figure 6 shows that while individuals of 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.). We could thus 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.

When comparing the diameter of the area of influence with the radius of gyration (consider 2*radius=diameter), it can be seen that the radius of the area of influence is approximately twice as large as the radius of gyration across all socio-economic levels. This indicates that the relation between the two variables is the same independently of the SEL.

The variable *number of different BTSs used* (presented in Figure 4) has the strongest correlation among all mobility variables. Although it is not a very intuitive variable, it has a lot of implications for the design of cellular wireless networks and for characterizing mobility.

Figure 4 shows that while lower socio-economic levels use around 6 different BTSs on average every week (an individual with a SEL of E will approximately use 5), higher socio-economic levels use a number of BTSs higher than 8, with an estimate that the number of BTSs used by an individual of SEL A is around 9. These results are aligned with both the radius of gyration and the diameter of the area on influence: higher SELs use more BTSs than lower SELs, which is expected as an increase in the SEL is related to an increase in the radius and/or the diameter of the area of influence.

Given the high correlation between the different number of BTSs used on average each week and the socio-economic level, and assuming that the same correlation holds true for individuals (and not aggregated models) this variable could be used as an estimator of the SEL of a given individual. Just by reversing the linear model presented in Figure 4, the SEL can be estimated as:

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

The equation will be valid for:

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

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. The other two mobility variables could also be used to predict the SEL although because the correlation is not as strong, the prediction would not be as good.

VII. CONCLUSIONS AND FUTURE WORK

The relation between socio-economic levels and a variety of health or behavioural human variables has been widely studied and presented in the literature in fields such as medicine, psychology, ecology and public policy. Although there are studies that measure the relation between SEL and human mobility in some scenarios [1]-[4], none of these studies have been able to directly measure human mobility and in general their conclusions are based on a limited number of individuals.

By using cell phone records we have matched SEL with a group of six mobility variables at a BTS (aggregated) level and 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. In the three cases there is a linear behaviour in which the increase in SEL is related to an increase in the mobility variable. Such analyses can be useful for policy makers in the areas of transport planning or epidemiology.

It is important to clarify that the results presented in this paper are not necessarily the same for other regions of the same country and it remains to be seen to which extent the linear models obtained are valid to other areas. Future work will study the development of prediction models given specific mobility variables.

In principle, the same technique here presented can be used to study the relation between SEL and a large variety of variables modelling how technology and specially cell phones are used at a scale and detail never done before.

References

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

[2] R. Maheswaran, T. Pearson, H. Jordan, D. Black, “Socioeconomic deprivation, travel distance, location of service, and uptake of breast cancer screening in North Derbyshire, UK”, in *J Epidemiol Community Health* 2006;60:208-212

[3] A. Carlsson-Kanyama, A. Liden, “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, 1999

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

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

[6] N. Eagle, M. Macy, R. Claxton, “Network Diversity and Economic Development” in Science 328 2029-1030, 2010.

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

[8] M. Lane, L. C. Carpenter, T. Whitted, and J. F. Blinn, “Scan line methods for displaying parametrically defined surfaces,” *Communications ACM*, vol. 23, no. 1, 1980.

[9] D. Brockmann and F. Theis, “Money Circulation, Trackable Items and the Emergence of Universal Human Mobility Patterns”, Pervasive Computing 7, Nr. 4, 28, 2008.

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

[11] Gonzalez M., Hidalgo C. A., and Barabási A.-L., “Understanding individual human *mobility* patterns”,. Nature, 453, 779 – 782, 2008.

[12] Song C. et al., “Limits of Predictability in Human Mobility”, Science 2010, Vol. 327. no. 5968, pp. 1018 – 1021.

[13] Cohen J., “Statistical Power Analysis for the Behavioral Sciences”, Lawrence Erlbaum Associates, 2nd Edition 1988.