On The Relationship Between Socio-Economic Factors and Cell Phone Usage

Vanessa Frias-Martinez
Telefonica Research
Madrid, Spain
34-914832878
vanessa@tid.es

Jesus Virseda
Telefonica Research
Madrid, Spain
34-914832883
jvjerez@tid.es

ABSTRACT

The ubiquitous presence of cell phones in emerging economies has brought about a wide range of cell phone-based services for low-income groups. Often times, the success of such technologies highly depends on its adaptation to the needs and habits of each social group. In an attempt to understand how cell phones are being used by citizens in an emerging economy, we present a large-scale study to analyze the relationship between specific socio-economic factors and the way people use cell phones in an emerging economy in Latin America. We propose a novel analytical approach that combines large-scale datasets of cell phone records with country-wide census data to reveal findings at a national level. Our main results show correlations between socio-economic levels and social network or mobility patterns among others. We also provide analytical models to accurately approximate census variables from cell phone records with $R^2 \approx 0.82$.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human Factors; K.4.2 [**Social Issues**]: Miscellaneous

General Terms

Human Factors, Measurement

Keywords

Call detail records, socio-economic factors, census maps, sensing human behavior.

1. INTRODUCTION

The recent adoption of ubiquitous technologies by large portions of the population in emerging economies has given rise to a variety of cell-phone based services for low-income populations in areas like health, education or banking [9, 6]. Although some services have proven successful over the years, others have not survived the first months of deployment [11]. Multiple technical and human reasons lie behind these failures, the lack of service personalization being an important one. Service personalization focuses on adapting services to user needs and behavioral traits, which is specially important in emerging economies where technologies and services from developed countries are often times deployed without sensitivity to local cultures and social behaviors. To overcome this practice, service personalization seeks to reveal groups of technology users that share behavioral patterns. The identification of these behavioral niches in the population allows to better adapt the services to the needs of each group.

In order to personalize cell-phone based services for emerging economies, we focus our study on understanding the role that demographic and socio-economic factors play on the way cell phones are used in an emerging economy. Our aim is to reveal whether specific gender, age or socio-economic groups use cell phones differently from others. These discriminant features will provide critical information for the personalization and adaptation of mobile-based services to the behavioral segments identified. Furthermore, the relationship between socio-economic or demographic factors and cell phone usage is also important from a policy perspective, given that such analyses can provide an understanding of the success (or failure) of specific technology-based programs across different social groups.

Analyzing the use of cell phones and its relationship with specific human factors has typically been carried out through questionnaires and personal interviews [2]. However, the widespread presence of cell phones in emerging economies is generating millions of digital footprints from cell phone usage. These large-scale datasets contain call records that provide thorough information of user interactions with their cell phones. As such, these records can be useful to model the use of cell phones through variables like consumption levels, social networks or mobility patterns. Recently, researchers have studied the relationship between cell phone usage patterns from subscribers in Rwanda and their demographic or socio-economic characteristics [1]. To carry out such analyses, the researchers computed usage patterns from a large-scale dataset of cell phone calls collected by a Rwandan telecommunications company. On the other hand, the authors carried out personal interviews over the phone with the subscribers, who self-reported their own socio-economic and demographic information. Unfortunately, such mixed methods approach limits the amount of cell phone users that can be modelled to the number of interviews that can be carried out, thus losing the large-scale component of the analysis provided by the calls' dataset.

ICTD'12 March 12–15, 2012, Atlanta, GA, USA
Copyright 2012 ACM 978-1-4503-1045-1/12/03 ...$10.00.

In an attempt to overcome these issues, we propose a new analytical approach that combines large-scale datasets of cell phone records with country-wide census data gathered by National Statistical Institutes (NSIs). On one hand, we use cell phone records collected by a telecommunication company to reveal subscribers' phone usage patterns for millions of users at an emerging economy. On the other hand, we utilize the census data gathered by the country's National Statistical Institute to obtain a set of social, economic and demographic variables by geographic area within the emerging economy. The combination of both sources of information reveals relationships between cell phone usage and census data at a large scale without the need to carry out personal interviews.

In addition, we also provide analytical models to formalize the relationship between cell phone usage and demographic or socio-economic variables. These models might be used to *approximate* the unknown census variables of an individual or a geographic region based only on their cell phone usage records. Given that the computation of census maps is typically highly expensive and time-consuming, such predictive models might prove useful, specially for low-resource emerging economies. In fact, the analytical models could be used as a complement or *soft substitute* of the expensive national campaigns that NSIs carry out to compute the census maps. To sum up, the contributions of our paper are twofold:

- **Statistical evaluation of the relationship between cell phone usage and demographic or socio economic factors.** We provide national evaluations of an emerging economy by combining large-scale datasets of call records with country-wide census data by geographic region.

- **Analytical models to approximate census variables from cell phone records.** We infer mathematical models that could be used as inexpensive *soft substitutes* of national census campaigns.

The rest of the paper is organized as follows: Section 2 describes recent literature related to the study of human factors and cell phone usage; Section 3 presents the call detail records and census datasets used for our study; and Sections 4 and 5 describe the statistical evaluation and the analytical models. Finally Section 6 details conclusions and future work.

2. RELATED WORK

There exists a large body of work on qualitative studies analyzing the relationship between socio-economic factors and cell phone usage. Donner *et al.* presented a survey of 277 microentrepreneurs and mobile phone users in Kigali, Rwanda, to understand the types of relationships with family, friends and clients, and its evolution over time [2]. Among other findings, the author discovered an inverse correlation between the age of the user and the probability of adding new contacts to its mobile-based social network. The author also claims that users with higher educational levels were also more prone to add new contacts to their social networks. Similar qualitative studies were carried out by Kwon *et al.* [7]. The authors conducted a study to understand the impact of demographics and socio-economic factors on the technology acceptance of mobile phones. For that purpose, they circulated a four-page survey with 33 questions to 500 cell phone subscribers and found that older subscribers felt more pressure to accept the use of mobile phones than their younger counterpart. In fact, cell phones were generally given as presents by family members for security purposes. The analyses we present in this paper offer the ability to expand these studies to millions of users by correlating their cell phone behavioral usage to public datasets with socio-economic information.

The literature covering large-scale quantitative analyses of the relationship between cell phone usage and human factors is very limited, given the recent availability of large datasets with cell phone call records. Eagle *et al.* studied the correlation between communication diversity and its index of deprivation in the UK [3]. The communication diversity was derived from the number of different contacts that users of a UK cell phone network had with other users. Eagle combined two datasets: (i) a behavioral dataset with over 250 million cell phone users whose geographical location within a region in the UK was known, and (ii) a dataset with socio-economic metrics for each region in the UK as compiled by the UK Civil Service. The author found that regions with higher communication diversity were correlated with lower deprivation indexes. Although this result represents an important first step towards understanding the impact of socio-economic parameters on mobile use at a regional level, we seek to elaborate more fine-grained impact analyses that can draw correlations between human factors and cell phone usage at even smaller scales like cities, neighborhoods or blocks.

Blumenstock *et al.* analyzed the impact that factors like gender or socio-economic status have on cell phone use in Rwanda [1]. Similarly to Eagle *et al.*, the authors combined two datasets, one containing call detail records from a telco company in Rwanda and the other one containing socio-economic variables computed from personal interviews with the company's subscribers. Their main findings revealed gender-based differences in the use of cell phones and large statistically significant differences across socio-economic levels with higher levels showing larger social networks and larger number of calls among other factors. This approach succeeds to reveal findings at an individual level, however, it limits the scalability of the results to the availability of the subscribers and to the amount of time and money available to carry out personal phone interviews to hundreds of users. To overcome these problems, we propose an approach that combines two large-scale datasets to understand the relationship between cell phone use and specific human factors. Specifically, we use the call detail records collected by telecommunications companies and the census data collected by local National Institutes of Statistics.

3. DESCRIPTION OF DATASETS

In this section, we describe the datasets and variables used in our analysis and discuss how to combine them to analyze the relationship between cell phone usage and demographic or socio-economic factors.

Call Detail Records

Cell phone networks are built using a set of base transceiver stations (BTS) that are responsible for communicating cell phone devices within the network. Each BTS or cellular tower is identified by the latitude and longitude of its geographical location. The area of coverage of a BTS can be approximated with Voronoi diagrams [12]. Call Detail Records (CDRs) are generated whenever a cell phone connected to the network makes or receives a phone call or uses a service (e.g., SMS, MMS). In the process, the BTS details are logged, which gives an indication of the geographical position of the user at the time of the call. It is important to clarify that the maximum geolocation granularity that we can achieve is that of the area of coverage of a BTS *i.e.,* we do not know the whereabouts of a subscriber within the coverage area. From all the information contained in a CDR, our study only considers the encrypted originating number, the encrypted destination number, the time and date of the call, the duration of the call, and the BTS that the cell phone was connected to when the call was placed.

Our CDR dataset contains 5 months of cell phone calls and SMSs from over $10,000,000$ pre-paid and contract subscribers across twelve large- and middle-sized cities in a Latin American country. From these call detail records, we compute three sets of variables per subscriber so as to model cell phone usage: (1) consumption variables; (2) social network variables and (3) mobility variables. The *consumption variables* characterize the general cell phone use statistics of a person, measuring, among others, the number of input or output calls, the duration of the calls or the expenses. The *social network variables* compute measurements relative to the social network that subscribers build when communicating with others. These variables approximate parameters like the number of people a person typically calls to or receives calls from (*i.e.,* input and output degree of the social network), the social distance between contacts (*diameter of the social network*) or the strength of the communication ties (strong/frequent contacts versus weak ones). Finally, the *mobility variables* characterize the geographic areas where a person typically spends most of his/her work and leisure time as well as the spatio-temporal mobility patterns of individuals with the granularity of the area of coverage of a BTS.

Census Data

In order to gather country-wide demographic and socio-economic information from the Latin American country under study, we use census data collected by the local National Statistical Institute (NSI). The local NSI carries out individual and household surveys at a national level every five years. These surveys employ a large staff of enumerators (*census takers*) that are responsible for interviewing every household head within their assigned geographical area. The enumerators have been specially trained to be able to gather all the required information in a proper manner. Although in some cities the census information is collected with laptops, in general, paper survey forms are still very common, which makes the collection process even more expensive and time consuming. Given the private nature of the individual census information, the NSI only makes public average values per geographical units (*GUs*), which divide a city into blocks of a few square kilometers.

For analytical purposes, we make use of three groups of variables from the census data: *education variables*, *demographic variables* and *goods' ownership variables* to characterize each *GU*. Table 1 shows a list with all the variables under study. Education variables measure the level of education of the citizens determining whether they are illiterate or have finished up to a certain educational grade. The demographic variables measure gender and age variables as well as the presence of indigenous population. Finally, the goods' ownership variables might be used as a proxy of the purchasing power of a person, measuring parameters like the existence of electricity, water or a computer in the household. We also use another variable provided by the local NSI: the socio-economic level (SEL). This is a unique value computed as a weighted average of all the census variables and represents the average socio-economic level of a geographical unit *GU*. The SEL is expressed as a letter that ranges from A/B (very high socio-economic level) to E (very low socio-economic level) with intermediate values C+, C, D+ and D.

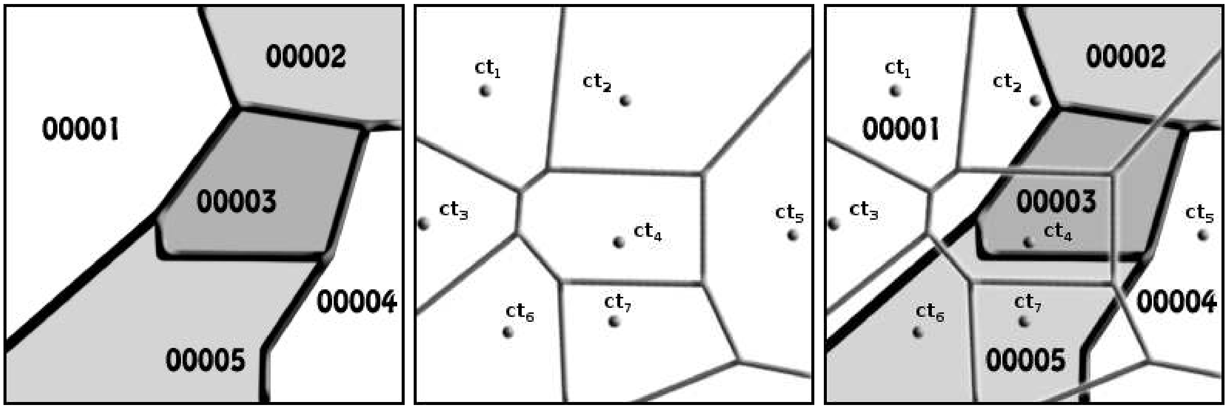
Census Variables	
Variable Type	Description
Education	% of Population with Primary School
	% of Female Population with Primary School
	% of Male Population with Primary School
	% of Population with Secondary School
	% of Female Population with Secondary School
	% of Male Population with Secondary School
	% of Illiterate Population
	% of Female Illiterate Population
	% of Male Illiterate Population
Demographics	% of Female Population
	% of Male Population
	% of Young Population (< 16)
	% of Middle-Age Population $(16 - 60)$
	% of Senior Population (> 60)
Goods	% of Houses with Cement Floor
	% of Houses with 1 room
	% of Houses with 3+ rooms
	% of Houses with Electricity
	% of Houses with Water
	% of Houses with TV
	% of Houses with PC
	% of Houses with All
SEL	Socio-Economic Level

Table 1: List of Census variables computed per household. The National Statistical Institute defines three groups: education variables, demographic variables and goods.

Combining Call Records with Census Data

In order to understand the relationship between cell phone usage and census information, we first need to map cell phone usage variables to the census information of different geographical units. The mapping is carried out through a three-step process: (1) associate a BTS residential location to each subscriber; (2) compute average cell phone usage variables per BTS region; and (3) associate census information to each BTS region. Step one focuses on approximating the geographical location of the residence of an individual. These locations allow us to associate cell phone subscribers to geographical units and thus to specific census data. However, the residential location of cell phone subscribers is only known for clients that have a contract with the carrier, which in the emerging economy under study accounts for less than 10% of the total population.

In order to approximate the residential location of the



(a) (b) (c)

Figure 1: Merging Cell Phone Use Maps With Census Maps: (a) Geographical Units (GUs) with census data for an urban area; (b) Distribution of BTSs in the same urban area (Voronoi Diagrams); (c) Overlapping of the Voronoi Diagrams with the Geographic Units (GUs).

subscribers with the pre-paid option, we use the residential detection algorithm described in [4]. The algorithm assigns the home location of an individual to a region covered by a BTS, based on general calling patterns detected in cell phone records. We apply the algorithm to all the pre-paid subscribers in our sample ($\approx 9 millions$) and subsequently assign to each of them a BTS representing his/her residential location. In the second step, we compute -for each BTS area- the average of each cell phone usage variable across all users whose residential location is that same BTS. These averages represent the aggregate cell phone usage of the subscribers that live in the geographical area covered by a BTS.

The last step focuses on the mapping between geographical units (GU) and BTS areas of coverage. Figure 1 shows an example of a census map and a cell phone usage map for an urban area. Figure 1(a) represents the geographical units (GU) that a city is divided into to carry out the census surveys. Each GU is associated to a set of census variables (educational, demographic, goods and SELs) that represent the average values for the population that lives in that area. On the other hand, Figure 1(b) shows the BTS coverage map for the same city (approximated with Voronoi diagrams), where each BTS is associated to a set of cell phone usage variables (consumption, social and mobility variables) averaged across all subscribers whose residential location lies within that BTS's coverage area. Finally, Figure 1(c) represents the merge between the two maps clearly showing that geographical units and BTS coverage areas within a city do not necessarily match.

To carry out this last step, we use a scan line algorithm [8] that associates to each BTS area the set of geographical units (GUs) whose areas are partially (or totally) included in the geographical area enclosed by each Voronoi polygon [4]. With this approach, we can represent each BTS BTS_i as $BTS_i = s{*}GU_a + v{*}GU_b + ... + w{*}GU_d$ where $s, v, ...w$ represent the fractions of the geographical units GU_a, GU_b,...,GU_d that cover BTS_i. These mappings leverage the weight of each GU, and its census variables, on the BTS area and allow us to compute for each census variable an average value per BTS. For example, if we were to compute the *percentage of indigenous citizens* for BTS_4 in Figure 1, we would first represent BTS_4 as the list of geographical units that cover its area *i.e.,* $BTS_4 = 0.6GU_3 + 0.21GU_1 + 0.19GU_5$ and next, apply the formula using the values of the census variable for each geographical unit. Thus, if the percentage of illiterate population in GU_3 is 30%, in GU_1 is 40% and in GU_5 is 30%, the final value for the percentage of illiterate population in BTS 4 would be 32.1%. Repeating this process for each census variable and BTS across all cities in our study, results in a final map that associates to each BTS a set of cell phone usage and census variables representing average values over the population that lives within the BTS's coverage area.

4. SOCIO-ECONOMIC FACTORS AND CELL PHONE USAGE

To carry out the analyses, we run statistical tests that determine whether there exist significant differences or correlations between cell phone use variables and census variables. Nevertheless, the relationships that such analyses reveal are only valid, in principle, for the subscribers in our sample cities. Although the sample contains millions of pre-paid and contract subscribers, that does not guarantee they represent the whole population. In order to be able to extend our findings beyond the subscribers in our sample and to all the population in the emerging economy under study, we need to guarantee that they represent a distribution of citizens similar to the general distribution of the country under study.

Given that in the emerging economy under study the law requires that both contract and pre-paid subscribers provide personal information regarding gender, age and residential location, we can compare the subscribers in our study with the overall population. Figures 2 and 3 show the distribution of socio-economic levels, age and gender per cell phone subscriber in our sample, and the distribution of the same variables across all the population obtained from the NSI census dataset. We can observe that the distributions are very similar except for a few exceptions. In the case of gender and age groups, we observe that our subscriber distribution is a little bit skewed towards male cell owners. A similar bias is observed in the age groups. This is probably due to the fact that sometimes the cell phone number is

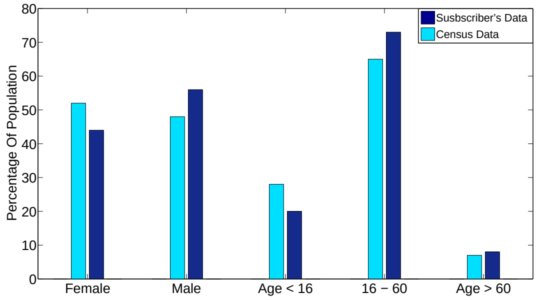


Figure 2: Histogram representing the percentage of subscribers divided by gender and age groups in our sample (dark) and for the whole emerging economy (light). We observe that both follow similar distributions.

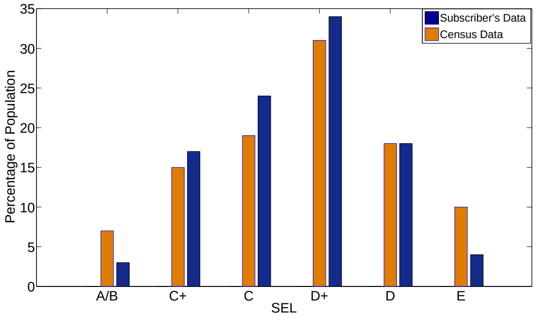


Figure 3: Histogram representing the percentage of subscribers in our sample from each socio-economic level (dark) and the percentage of citizens in the country under study per socio-economic level (light).

associated to the buyer of the cell phone (typically the father) and not to the real user of the cell phone (who might be the wife or the children). In the case of the SELs, we observe a similar distribution except for the extreme values (A and E). This is an artifact caused by our mapping methodology: since we average the SEL values of all the GUs contained within a BTS coverage area, extreme values tend to be *dispersed* across the central ones. Additionally, given that our subscriber sample does not take into account citizens from rural areas, who are typically associated to the lowest SEL, we might also loose some of the granularity for SEL E. Although not shown here, individual census variables in our sample also showed similar distributions to the country's population. These similarities allow us to extend the findings in this section to the whole country and not just to the subscribers in our sample.

Statistical Methodology

To understand the relationship between cell phone usage and census variables, we run ANOVA tests and Pearson's correlations on the census and cell phone data distributions. In our context, ANOVA tests are used to understand whether there exist statistically significant differences in cell phone usage across different social groups. It is important to highlight that this test only reveals that there exist significant differences in the mean between the distributions of one or more groups. Additionally, we compute Pearson's correlations to be able to quantify the general linear relationships between cell phone usage and census variables.

To carry out the ANOVA tests, we first divide the range of each of the census variables specified in Table 1 into four quartiles ($q1, ...q4$) that are used to represent the *social groups* in the statistical test. The quartiles for each census variable are computed by dividing the range between the minimum and the maximum percentage for that variable into four different subsets *e.g.,* $q1 = (min, min + \frac{max-min}{4})$. Each q_i represents a social group in our population associated to low, medium-low, medium-high or high values for a specific census variable. In the case of the SEL census variable, we use the ranges defined by the National Statistical Institute that differentiate six social groups: A/B, $C+$, C, D, $D+$ and E. The ANOVA test is then used to determine whether there exist statistically significant differences in the mean of cell phone usage variables across the four social groups with lower or higher values for a specific census variable.

The procedure laid out in figure 4 shows the steps to compute the statistical tests for all the cell phone usage and census variables. Assuming that our initial dataset contains a list of triads (BTS, cell phone usage variables, census variables), we compute for each cell phone usage variable $pvar$ and each census variable $cvar$ the BTSs that have a $cvar$ value within each quartile q_i. Next, for each quartile q_i, we build a distribution with the $pvar$ values of all the BTSs that have a $cvar$ value within q_i. This process builds four distributions $D[q_i]$ representing the cell phone usage variable values $pvar$ for each social group with a low, medium-low, medium-high or high value for the census variable $cvar$. By running an ANOVA on these four distributions, we can determine whether there exists a statistically significant difference between the cell phone usage variable $pvar$ across the four social groups.

For example, if we were to study the relationship between the average number of input calls in a geographical area and the percentage of illiterate people within that area, we would first compute four quartiles representing low, medium-low, medium-high or high percentages of illiteracy. Then, for each quartile, we would compute the number of input calls for the BTSs that have a percentage of illiterate people within that quartile; and finally, we would compute the ANOVA test between the four quartile distributions and report statistical significance.

Additionally, in order to quantify the differences between each pair of cell phone usage and census variables, we compute the Pearson's correlation between the distribution containing the values of the $pvar$ cell phone usage variables across all BTSs ($set(pvar)$) and the distribution $set(cvar)$ which contains the $cvar$ census variable values for the same BTSs, as shown in Figure 4.

Next, we report statistical results for each group of variables characterizing cell phone usage (consumption, social and mobility) and the census variables. Given the large number of (consumption variable,census variable) pairs that have to be evaluated, for clarity purposes we only report variables that gave significant statistical results. For the

Procedure 1 Process to compute statistical tests for all cell phone usage and census variables.

```
for each cell phone usage variable $pvar$ do
  for each census var $cvar$ do
    $q_1 = (min, min + \frac{max-min}{4})$
    $q_2 = (min + \frac{max-min}{4}, min + \frac{max-min}{2})$
    $q_3 = (min + \frac{max-min}{2}, max - \frac{max-min}{4})$
    $q_4 = (max - \frac{max-min}{4}, max)$
    for each $q_i$ do
      $list[q_i]$ = select {BTS} with $cvar$ value in $q_i$
      for each $BTS$ in list$[q_i]$ do
        $D[q_i] = D[q_i] \cup pvar$
      end for
    end for
    ANOVA($D[q_1]$,$D[q_2]$,$D[q_3]$,$D[q_4]$)
    $set(pvar) = pvar\ all\ BTS$
    $set(cvar) = cvar\ all\ BTS$
    correlation({$set(pvar)$},{$set(cvar)$})
  end for
end for
```

Figure 4: Process to compute statistical tests for all cell phone usage and census variables.

ANOVA tests, we will report results with $p < 0.05$ and for Pearson's correlations we will discuss moderate ($0.4 <| r |< 0.7$) and strong correlations ($| r |> 0.7$). Additionally, we only report voice-based mobility information since we do not have geolocation information for SMS traffic.

Consumption Variables

The *consumption variables* that were found to have statistical significance are the total number of calls made or received (Total), the total number of output calls (Output), the duration of the calls (Duration) and its expenses (Expenses). In terms of SMS behavior, the total number of SMS sent or received (Total); the output SMSs (Output) and the SMS expenses (Expenses). All variables refer to weekly averages computed for the subscribers that live within the same BTS coverage area.

Table 2 shows the statistical results for the ANOVA tests and the Pearsons' correlation. The numerical values in each cell represent ANOVA p-values, the (*) represents a moderate correlation and (**) a strong correlation; and (+) and (-) signs refer to a positive or a negative correlation between each two distributions, respectively. Our main findings show that there exists a statistically significant difference between the socio-economic level and the average number of calls made by the subscribers. The average number of output calls is also significantly different depending on the socio-economic level, with even more significance ($p < 0.01$). We also see that, as expected, the average expenses on calls are significantly different across socio-economic levels. However, no significant difference is observed for the duration of the calls. In terms of SMSs, we observe statistically significant differences between the socio-economic level and the average total number of SMSs consumed (sent and received) as well as the average number of output SMSs and expenses. As for the educational level achieved by the population, we observe that there exist significant differences between the percentage of the population in a geographical area with secondary school finished and the number of calls, input calls and expenses, but no significant differences are found in SMS usage.

The correlation results show moderate positive correlations between the SEL and the number of calls or the expenses *i.e.,* the higher the socio-economic level, the more calls a person tends to make and the larger the amount of money they spend on calls. We also observe a positive correlation between the percentage of citizens with the secondary school finished and the number of calls and expenses, meaning that the larger the number of citizens with a secondary degree in a geographical area, the larger the number of calls and number of SMS these citizens consume.

The demographic variables also show interesting results. The percentage of middle-age population (16 to 60) in a region, seems to determine a significant difference in the number of calls made as well as their associated expenses. We have also find statistically significant differences in the average number of SMS sent and received as well as in their expenses. However, age does not seem to have any moderate or strong correlation with the consumption variables. Gender seems to determine the existence of statistically significant differences for the number of output calls, the duration of the calls as well as for the average number of output SMS. In fact, we observe a moderate negative correlation between the percentage of male population and the number of output calls *i.e.,* the larger the male population, the smaller the number of output calls that are made. The symmetric correlation is found for the female gender whereby the higher the female population in a region, the larger the number of output calls that are made.

Finally, the goods variables that revealed statistically significant differences where the presence of a house with a cement floor, the presence of a PC at the household, and the presence of all goods explained in Table 1 at a household. We observe that the three variables show significant differences in the number of calls, the expenses associated to the calls, the average number of SMS sent and received and their expenses. In terms of correlations, the presence of a PC at a household appears to be moderately positively correlated to the number of cell phone calls and SMSs used by the subscribers. To sum up, we observe that higher socio-economic levels, including education and access to goods, tends to be correlated to higher consumption levels of calls, SMS and expenses. On the other hand, gender might have an impact on social aspects of the communication (duration of the calls) but not on general consumption levels.

Social Variables

The *social network variables* that revealed statistical significance with respect to socio-economic information are the reciprocity of the communications and the physical distance between contacts. By reciprocity (R), we refer to the number of reciprocal voice or SMS communications between a person and her contacts. We evaluate three possible values: at least one, at least two and at least five reciprocal communications. On the other hand, physical distance ($Phys.Dist.$) refers to the average distance between a person's residential area and the residential area of her cell phone contacts. All variables are computed as weekly averages.

Table 3 shows the results for the ANOVA and Pearson's

Table 2: Census Variables and Consumption Variables. Numbers represent ANOVA p-values; (*) represents a moderate correlation and (**) a strong correlation; (+) and (-) signs refer to a positive or a negative correlation between each two distributions, respectively.

	CALLS				SMS		
CENSUS VARIABLE	Total	Output	Duration	Expenses	Total	Output	Expenses
SEL	0.018 +*	0.004	–	0.028 +*	0.015	0.009	0.023
S.School	0.003 +*	–	–	0.002 +*	–	–	–
Middle-Age	0.004	–	–	0.021	0.002	–	0.003
Male	–	0.003 -*	0.002	–	–	0.002	–
Female	–	0.002 +*	0.004	–	–	0.001	–
Cement Floor	0.020	–	–	0.012	0.001	–	0.021
PC	0.002 +*	–	–	0.004	0.001 +*	–	0.040
All	0.031	–	–	0.012	0.022	–	0.030

Table 3: Census Variables and Social Network Variables.

	CALLS				SMS			
CENSUS VARIABLE	R(5)	R(2)	R(1)	Phys.Dist.	R(5)	R(2)	R(1)	Phys.Dist.
SEL	0.002 +*	0.008 +*	0.010	0.030 +*	0.003 +*	0.006	–	0.012
S.School	0.008 +*	0.010	–	0.003 +*	0.013	–	–	–
Middle-Age	0.001 -*	0.002	–	0.010 +*	–	–	–	–
Male	0.010 -*	0.002 -*	–	0.002 +*	–	–	–	–
Female	0.002 +*	0.001 +*	–	0.010 -*	–	–	–	–
PC	0.010 +*	0.008	–	0.002 +*	0.015 +*	–	–	0.009 +*
All	0.009 +*	0.003	–	0.004 +*	0.023 +*	–	–	0.010 +*

correlation tests. We observe that there exists a significant difference between the socio-economic level of a person and the number of reciprocal calls or SMSs sent and received. In fact, the statistical significance is observed when there exist one or more reciprocal calls between a person and her contacts. In the case of SMSs, the reciprocity needs to be two or higher in order to determine a statistical significance. Additionally, we also observe that the physical distance is also statistically different across diverse socio-economic levels, for both calls and SMSs. In terms of correlations, we observe that the socio-economic level SEL is moderately positively correlated with the number of reciprocal calls whenever the reciprocity is, on average, two or higher for calls, and five or higher for SMSs. In terms of physical distance between a person and her contacts, we see a positive correlation with the SEL in the case of cell phone calls, but not for SMS.

The percentage of population with secondary school finished also seems to determine statistically significant relationships with the number of reciprocal calls and physical distance. In fact, reciprocity of two calls/SMS or higher as well as the physical distance between contacts with at least one reciprocal call show statistical differences across the four quartiles of percentage of population with secondary education finished. The percentage of population with the secondary school finished is found to be positively correlated to the number of reciprocal calls between contacts that call each other at least five times per week on average *i.e.,* the larger the percentage of population with higher education, the more reciprocal calls these subscribers make. Similarly, we also observe that the larger the percentage of population with secondary studies finished, the larger the physical distance between a person and her contacts *i.e.,* a person with higher studies tends to have a social network more geographically extended that those with lesser studies.

Age and gender also reveal significant differences although only for voice: the age group $(16 - 60)$ as well as the gender seem to have an impact in the number of reciprocal calls as well as in the physical distance. We observe that the middle-age group (ages between 16 and 60) show negative correlations between the number of reciprocal calls and the age *i.e.,* the older, the less reciprocal calls subscribers appear to make. On the other hand, we can report that the older the person is, the farther away her voice contacts tend to be. We also see that the higher the percentage of male population in a geographical area, the fewer reciprocal calls between subscribers and the larger the physical distance between voice contacts. Inversely, the higher the percentage of female population, we observe a higher number of reciprocal calls as well as smaller physical distances between women and their contacts.

Finally, the *goods' variables* that refer to the percentage of population with a PC at the household as well as the percentage of population with all amenities at home (electricity, water, PC, fridge, TV and washing machine) reveal statistical significant differences both in terms of high reciprocal calls as well as reciprocal SMSs. Analogously, the physical distance between a person and her voice contacts is also associated to significant differences. Specifically, we observe a positive correlation between the percentage of population with a PC at home and the number of reciprocal calls made by such population. In fact, the higher the presence of a PC or of all amenities, the higher the number of high reciprocal calls made by the subscribers as well as the higher the physical distance between these subscribers and their voice/SMS

Table 4: Census Variables and Mobility Variables.

	CALLS			
CENSUS VARIABLE	N.BTS	Dist.Travelled	Radius	Diameter
SEL	0.009 +**	0.020 +*	0.010 +**	0.023 +*
P.School	0.010 +**	0.020 +*	0.012 +*	0.021 +*
Middle-Age	0.010 +*	0.020 +*	0.031 +*	0.010 +*
Male	0.010 +**	0.020 +*	0.023 +*	0.030 +*
Female	0.012 -**	0.009 -*	0.012 -*	0.020 -*
PC	0.012	0.0023	0.031	0.038
All	0.023	0.012	0.004	0.014

contacts.

Mobility Variables

The *mobility variables* that show statistically significant differences with the socio-economic factors are: (1) number of different BTSs visited by a person; (2) distance travelled by a person, computed as the distance between each pair of consecutively visited BTSs; (3) the radius of gyration, computed as the weighted average of the number of visits to each BTSs for a given person; and (4) the diameter, obtained as the maximum distance between the BTSs typically visited by a person. These last two measures reveal very important findings since they can be interpreted as an approximation of the distance between *home and work* (radius of gyration [5]), and of the area where a person typically spends most of her time (diameter [10]).

Table 4 shows that there exist significant differences between the socio-economic level of a person and the number of BTSs visited by the person, the radius of gyration and the diameter. In fact, we observe a strong positive correlation between the number of BTSs and the SEL such that the larger the number of BTSs, the higher the SEL of the citizens at a specific geographical area. A similar strong correlation is found for the radius of gyration whereas distance travelled and diameter show moderate correlations. In general, the SEL seems to be highly related to the mobility of a subscriber seeing more mobility across higher socio-economic classes. The percentage of citizens with primary studies finished also seems to be statistically differentiating across all mobility variables. We observe that areas with larger numbers of educated people seem to be correlated to larger distances travelled as well as larger home-work distances.

In terms of age and gender, we observe similar results for the statistical differences and correlations. In fact, we see that the younger the person, the larger the distances tend to be *i.e.,* young people tend to visit more BTSs (different geographic regions) as well as to show larger radiuses and diameters meaning that the areas where they typically move about are larger than the ones experienced by their older counterparts. As for gender, we observe opposite results between male and female. The higher the percentage of male population present in a region, the higher the number of BTSs visited, distances travelled, and the larger the radius and diameters observed. On the contrary, the higher the percentage of female population within a geographical region, the smaller the average number of BTSs visited by the population as well as the more geographically constrained the mobility patterns are. Finally, in terms of goods, although there exist statistically significant differences ($p < 0.05$), we do not observe any moderate or strong correlation among these and the mobility variables.

5. PREDICTIVE MODELS

In this last section we explore whether cell phone usage variables can be used to build predictive models for the census variables. Such models could provide *cheap tools to approximate* the variables of the census maps, which are expensive to compute specially for resource-constrained emerging economies. Formally, we seek to find a model (Formula 1) that predicts the value of a census variable C_i based on sets of behavioral $\vec{B}$, social $\vec{S}$ and/or mobility $\vec{M}$ cell phone usage variables computed from the call records:

$$C_i = \vec{x} * \vec{B} + \vec{y} * \vec{S} + \vec{z} * \vec{M} \tag{1}$$

In order to carry out this analysis, we make use of multivariate linear regression (with ordinary least squares) over the cell phone usage variables that showed a statistically significant difference with the census variables (as presented in the previous section). In our analyses, $\vec{B}$ refers to total number of calls, total number of output calls, duration and expenses; $\vec{S}$ refers to reciprocal calls and physical distance and $\vec{M}$ to number of BTSs, distance travelled, radius of gyration and diameter. To evaluate the *goodness of the fit*, we report the adjusted R^2.

Figure 5 shows the R^2 values for the prediction models built using the cell phone behavioral variables $\vec{B}$, the cell phone social variables $\vec{S}$ and the mobility variables $\vec{M}$. We show results for the prediction of the census variables using only one group of variables (either $\vec{B}$ or $\vec{S}$ or $\vec{M}$) or using any combination of the three groups. As for the census variables, we only report regression results for those whose adjusted R^2 values where higher than 0.50. We assume that smaller values do not guarantee the existence of *good* predictive models for a specific census variable. Our main findings show that the combination of all three variable groups $(\vec{B}, \vec{S}, \vec{M})$ builds a highly predictive model specially for the SEL (with $R^2 = 0.83$), for the percentage of population with primary school ($R^2 = 0.82$) and for the percentage of population with *All* basic goods at home (with $R^2 = 0.86$). We also observe that the group of behavioral and mobility variables $(\vec{B}, \vec{M})$ outperform the predictive accuracy of the social and mobility variables $(\vec{S}, \vec{M})$. In fact adjusted R^2 values are considerably larger for $(\vec{B}, \vec{M})$ than for $(\vec{S}, \vec{M})$. However, the group of behavioral and social variables $(\vec{B}, \vec{S})$ seem to be the worst predictors than any other combined

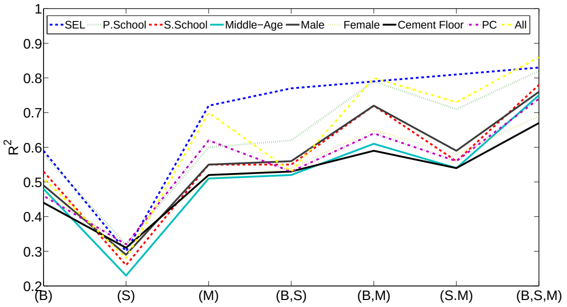


Figure 5: Multivariate Regression Results. Prediction of census variables from cell phone usage behavioral (B), social (S) and mobility variables (M). R^2 values represent the goodness of the fit.

pair. Finally, in terms of individual groups of variables, the mobility variables $\vec{M}$ are, by far, the best predictive group across all census variables.

6. CONCLUSIONS AND FUTURE WORK

We have presented a large-scale study to understand the relationship between various socio-economic variables and the way people use cell phones in an emerging economy in Latin America. Our novel approach combines large-scale datasets of cell phone records with country-wide census data gathered by National Statistical Institutes (NSIs) to reveal large-scale findings without the need to carry out expensive personal interviews or questionnaires. The main findings reveal that there exist moderate and strong correlations between the socio-economic level of a person and the expenses, the reciprocity of her communications, the physical distance with her contacts or the geographical areas where the person typically moves about. Additionally, we have provided a predictive model that allows to predict a variety of socio-economic variables exclusively from cell phone records. Such predictive models can be used as cheap approximators of the census maps which are highly expensive to compute, specially for emerging economies. Future work will focus on computing similar analyses for other emerging economies to understand cross-culture differences in the use of cell phones and studying whether there exist country-based differences across urban and rural areas.

7. REFERENCES

[1] J. Blumenstock and N. Eagle. Mobile divides: Gender, socioeconomic status, and mobile phone use in rwanda. In *Proceedings of the 4th International Conference on Information and Communication Technologies and Development*, 2010.

[2] J. Donner. The use of mobile phones by microentrepreneurs in kigali, rwanda: Changes to social and business network. *Information Technologies and International Development*, 3(2), 2007.

[3] N. Eagle. Behavioral inference across cultures: Using telephones as a cultural lens. *IEEE Intelligent Systems*, 23:4:62–64, 2008.

[4] V. Frias-Martinez, J. Virseda, A.Rubio, and E. Frias. Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. *Int. Conf. on Information & Communication Technologies and Development (ICTD), London, UK)*, 2010.

[5] M. Gonzalez, C. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, pages 453:479–482, 2008.

[6] N. Hughes and S. Lonie. M-pesa: mobile money for the âĂIJunbankedâĂİ turning cellphones into 24-hour tellers in kenya. *Innovations, MIT Press Journals*, 2(1-2), 2007.

[7] H. Kwon and L. Chidambaram. A test of the technology acceptance model: The case of cellular telephone adoption. In *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[8] J. M. Lane, L. C. Carpenter, T. Whitted, and J. F. Blinn. Scan line methods for displaying parametrically defined surfaces. *Communications ACM*, 23(1):23–34, 1980.

[9] A. Molnar and V. Frias-Martinez. Educamovil: Mobile educational games made easy. In *Proceedings of the World Conference Ed-Media*, 2011.

[10] A. Rubio, V. Frias-Martinez, E. Frias-Martinez, and N. Oliver. Human mobility in advanced and developing economies: A comparative study. In *AAAI Spring Symposium on Artificial Intelligence for Development (AI-D)*, 2010.

[11] K. Verclas. Scaling mobile services for development: What will it take?, mobileactive.org, 2010. `http://mobileactive.org/scaling-mobile-services-development-what-will-it-take`.

[12] G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal fur die Reine und Angewandte Mathematik*, 133:97–178, 1907.