Research Article

Cell Phone Analytics: Scaling Human Behavior Studies into the Millions

Vanessa Frias-Martinez
vanessa@tid.es
Researcher
Telefonica Research
Ronda de la Comunicación s/n
Edificio Oeste 1
Madrid, 28050
Spain

Jesus Virseda
jesus.virseda.jerez@gmail.com
Researcher
Planning and Learning
Research Group
Universidad Carlos III
Avenida de la Universidad, 30
Leganes, 28911
Spain

Vanessa Frias-Martinez
Jesus Virseda

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, the success of such technologies depends highly on their 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 novel methodology to analyze large-scale relationships between specific socioeconomic factors and the ways people use cell phones. Our approach combines large-scale datasets of cell phone records with countrywide census data to reveal findings at a national level. We evaluate the proposed methodology in an emerging economy in Latin America and show relevant correlations between socioeconomic levels and social network structure or mobility patterns, among others. Finally, we provide an analytical model to formalize the relationship between cell phone use and demographic or socioeconomic variables.

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 (see Hughes & Lonie, 2007; Soto, Frias-Martinez, Virseda, & Frias-Martinez, 2011). Although some services have proven successful over the years, others have not survived the first months of deployment (Verclas, 2010). Multiple technical and human reasons lie behind these failures, with the lack of service personalization being an important one. Service personalization focuses on adapting services to user needs and behavioral traits, which is especially important in emerging economies where technologies and services from developed countries are often deployed with no sensitivity to local culture or social behavior. To overcome this practice, service personalization seeks to identify groups of technology users that share behavioral patterns. The identification of these behavioral niches in the population allows vendors to better adapt their services to the needs of each group.

To personalize cell phone-based services for emerging economies, we focus our study on understanding the role that demographic and socioeconomic factors play in determining how cell phones are used in an emerging economy. Our aim is to discover whether specific gender, age, or socioeconomic 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 socioeconomic or demographic factors and cell phone use 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.

Analysis of the uses of cell phones and their relationships with specific human factors has typically been carried out through questionnaires and personal interviews (Donner, 2007). 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 and their environment. As such, these records can be useful for modeling the use of cell phones through variables such as consumption levels, social network structure, or mobility patterns. Recently, Blumenstock and Eagle (2010) studied the relationship between cell phone usage patterns from subscribers in Rwanda and their demographic or socioeconomic characteristics. 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 socioeconomic and demographic information. Unfortunately, such a mixed-methods approach limits the number of cell phone users who can be included in the model to the number of interviews that can be carried out, thus losing the large-scale component of the analysis provided by the calls dataset.

In an attempt to overcome these issues, we propose a new methodology that combines large-scale datasets of cell phone records with countrywide census data gathered by various countries' national statistical institutes (NSIs). On one hand, the methodology uses cell phone records collected by telecommunication companies to reveal subscribers' phone usage patterns for millions of users in an emerging economy. On the other hand, the methodology uses the census data gathered by countries' NSIs to obtain a set of social, economic, and demographic variables by geographic area within the country under study. The combination of both sources of information reveals relationships between cell phone use and census data on a large scale without needing to carry out personal interviews. To demonstrate the methodology, we present an evaluation using calling records from various cities in an emerging Latin American economy. The evaluation offers a wide range of quantitative results, which we proceed to analyze and compare against previous qualitative and quantitative findings in the literature.

Finally, we also provide an analytical model to formalize the relationship between cell phone use and demographic or socioeconomic variables. Such a model might be used to *approximate* the unknown census variables of a geographic region based only on its cell phone usage records. Given that the computation of census maps is typically very expensive and time-consuming, such predictive models might prove useful, especially 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 summarize, the contributions of our paper are threefold:

1. **A novel methodology to compute large-scale statistical analysis of the relationship between cell phone use and demographic or socioeconomic factors.** We describe the datasets required to apply the methodology, its main steps, and the algorithm used to compute the statistical analysis.
2. **Evaluation of the methodology using real call records and socioeconomic information from an emerging economy in Latin America.** We evaluate the methodology proposed and describe important insights regarding urban regions in Latin America. For completeness, we compare these results against qualitative and quantitative analyses in the related literature. Although some results might seem obvious or familiar, it is important to clarify that our main contribution is the methodology to reveal large-scale insights. As a result, the approach might be used to confirm or reject a plethora of cell phone-related behavioral assumptions.

3. **An analytical model to approximate census variables from cell phone records.** We infer a mathematical model that could be used as an inexpensive *soft substitute* for national census campaigns and evaluate it for Latin American urban environments.

The rest of the paper is organized as follows: Section 2 presents the novel methodology, as well as the datasets and statistical analysis it uses. Section 3 goes on to describe the evaluation of the methodology using real call records and census data from large and mid-sized cities in an emerging Latin American economy. Section 4 presents the predictive analytical model and its evaluation, and then section 5 highlights the most important findings of our evaluation and frames our results within the larger related literature. Finally, section 6 details conclusions and proposes future work.

2. Description of the Methodology

In this section, we describe the novel methodology proposed to carry out large-scale analysis of the relationship between cell phone use and socioeconomic variables. For that purpose, we discuss the datasets required: call detail records and census data, the algorithm to merge both sources of information, and the statistical analysis to carry out the large-scale evaluation.

2.1 Call Detail Records

Cell phone networks are built using a set of base transceiver stations (BTS) that are responsible for communicating with cell phone devices within the network. Each BTS or cellular tower is identified by the latitude and longitude of its geographic location. A BTS area of coverage can be approximated with Voronoi diagrams (Voronoi, 1907). 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 give an indication of the geographic position of the user at the time of the call. It is important to clarify that the maximum geolocation granularity we can achieve is that of the BTS area of coverage; 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 methodology uses these CDRs to compute three sets of variables per subscriber: 1) consumption variables, 2) social network variables, and 3) mobility variables for voice and SMS records. The *consumption variables* characterize the general cell phone usage statistics of a person, measuring the number of a) input, b) output, and c) total calls, as well as d) the duration of the call and e) the expenses related to the call. The *social network variables* compute measurements relative to the social network that subscribers build when communicating with others. These variables compute a) the number of people a person typically calls or receives calls from (i.e., input and output degree of the social network), b) the physical distance between a person and her or his contacts (the diameter of the social network), and c) the strength of the communication ties or reciprocity that determines the number of calls reciprocated by the call recipient at least one time (R[1]) or more (R[n]). Finally, the *mobility variables* characterize how citizens move about in their environment. Although, in principle, these variables do not model specific cell phone usage patterns (like consumption or social variables), we compute them because they provide insights into human behavior that can be useful for the personalization and adaptation of cell phone services to the needs and habits of different social groups. For that purpose, we measure a) the average number of BTSs visited, which gives an approximation of the volume of the mobility; b) the average distance traveled, computed from the distances between the visited BTSs; c) the radius of gyration, which is computed as a weighted average between the BTSs used by an individual and can be considered an approximation of the distance between home and work (Gonzalez, Hidalgo, & Barabasi, 2008); and d) the diameter, which represents the geographic areas where a person spends both her or his work and leisure time, and which is computed as the distance between the BTSs used by an individual.

2.2 Census Data

We gathered countrywide demographic and socioeconomic information from census data collected by local NSIs. The NSIs of each country carry out individual and household surveys at a national level, typically every five years. These surveys employ a large staff of enumerators (census takers) who are responsible for interviewing every household head within their assigned geographic area. The enumerators are specially trained to be able to gather all the required information in the proper manner. Although, in some cities, the census information is collected with laptops, in general, paper survey forms are still common, which makes the collection process even more expensive and time-consuming. Given the private nature of the individual census information, the NSIs only make public average values per geographic units (GUs). The size of the GUs varies from country to country and can represent a few city blocks or larger urban or rural regions. Our methodology can be applied to any granularity; however, the more granular the GUs, the more data are available in the statistical analysis, thus increasing the likelihood of accurate findings.

Our methodology uses three groups of variables typically present in census data, education variables, demographic variables, and goods' ownership variables, to characterize each GU (Table 1 shows an example). Education variables measure the citizens' level of education to determine whether they are illiterate or have attained a certain grade-level of education. The demographic variables measure gender and age variables, as well as the presence of indigenous populations. Finally, the goods' ownership variables might be used as a proxy for the purchasing power of a person, measuring parameters such as the availability of electricity, running water, or a computer in the household. We also use another variable typically provided by NSIs: the socioeconomic level (SEL). This is a unique value computed as a weighted average of all the census variables, and it represents the average socioeconomic level of a GU. The SEL is usually expressed as a letter that ranges from A/B (very high socioeconomic level) to E (very low) with intermediate values C+, C, D+, and D. More or less SEL granularity is also possible, though this is highly dependent on the techniques used by the NSIs.

2.3 Combining Call Records with Census Data

To understand the relationship between cell phone use and census information, we first need to map cell phone usage variables to the census information of different GUs. The mapping is carried out by a three-step process presented in Soto et al. (2011). First, the process associates a BTS residential location to each subscriber; second, it computes average cell phone usage variables per BTS region; and finally, it associates census information to each BTS region. Next, we provide an overview of the mapping process. We refer the reader to Soto et al. (2011) for further details.

Step 1 focuses on approximating the geographic location of an individual's residence. These locations allow us to associate cell phone subscribers to GUs, and thus to specific census data. Although in some emerging regions prepaid customers are legally required to provide their residential location, this is not the norm. In fact, the residential location of cell phone subscribers in emerging regions is typically only known for clients who have a contract with the carrier, which accounts for less than 10% of the total population. Thus, to approximate the subscribers' residential locations with the prepaid option, the mapping process uses the residential detection algorithm described in Frias-Martinez, Virseda, Rubio, and Frias-Martinez (2010). 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. The mapping process applies such an algorithm to all the prepaid subscribers and subsequently assigns to each of them a BTS representing her or his residential location. For the users with a contract, the mapping process uses the address determined in the contract and associates their homes with the closest BTS.

In the second step, the process computes—for each BTS area—the weekly average of each cell phone usage variable across all users whose residential location is within that same BTS. These averages represent the aggregate cell phone use of the subscribers who live in the geographic area covered by a BTS.

The last step focuses on the mapping between GUs and BTS areas of coverage. Each GU is associated with a set of census variables (educational, demographic, goods, and SEL) that represent the average values for the population living in that area. On the other hand, 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 coverage area. The mapping process uses a scan line algorithm (Lane, Carpenter, Whitter, & Blinn, 1980)

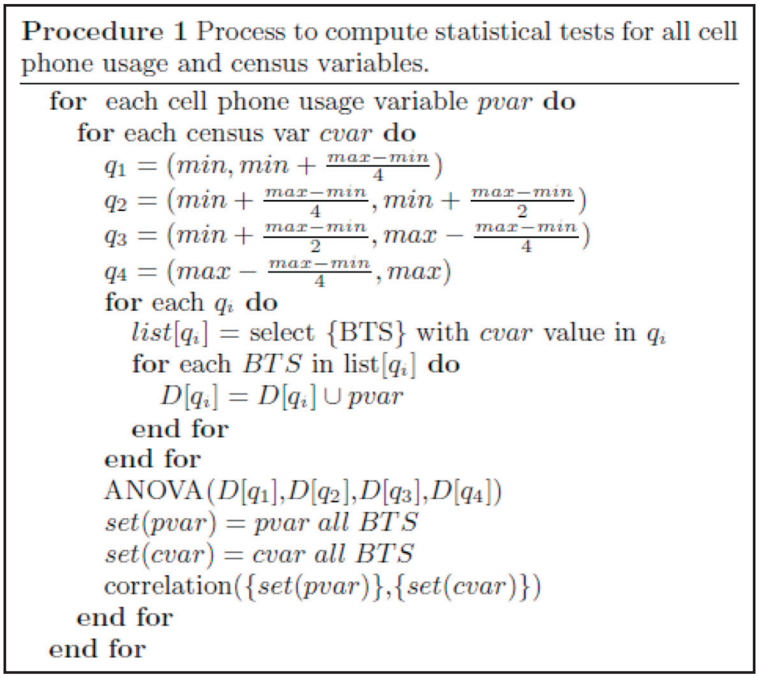
Table 1. Example List of Census Variables Computed by an NSI.

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	Socioeconomic Level

that associates to each BTS area the set of GUs whose areas are partially or totally included in the geographic area enclosed by each Voronoi polygon. With this approach, each BTS_i can be represented as $BTS_i = m * GU_a + n * GU_b + \ldots + x * GU_d$ where $\{m,n, \ldots, x\}$ are the fractions of the geographic 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 for computing an average value per BTS and census variable. Repeating this process for each census variable and BTS present in our datasets results in a final map that associates to each BTS a set of cell phone usage and census variables representing average values for the population that lives within the BTS's coverage area. As discussed earlier, if the GUs are not very granular, it might be the case that a GU geographically covers more than one BTS. The methodology presented is still valid in such a case; however, lower granularities do provide less information for the statistical analysis, and the results might not be as conclusive or accurate as they would with higher granularities.

2.4 Statistical Analysis

Once the mapping process has been computed, a list of pairs (*calling variable$_{i,}$*, *census variable$_j$*) is available for each geographic area covered by a BTS. To understand the relationship between cell phone use and census variables, the methodology runs ANOVA tests and Pearson's correlations on each set of pairs across all BTSs. In our context, ANOVA tests are used to understand whether there exist statistically significant differences in cell phone use across different census variables (social groups). It is impor-



```
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 1. Procedure to compute statistical tests for all cell phone use and census variables.

tant 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 use and census variables.

To carry out the ANOVA tests, we first divide the range of each of the census variables into four quartiles ($q_1, \ldots, q_4$) 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 subsets, e.g., q_1 = (min,min+(max−min)/4). Each q_i represents a social group in our population associated with low, medium-low, medium-high, or high values for a specific census variable. The one exception to this is the SEL census variable, for which we use the ranges defined by the NSIs, which differentiate several letter-based groups. NSIs typically consider between four or six SEL values (six values define the 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 weekly cell phone usage variables across the social groups for a specific census variable.

Figure 1 shows the procedure 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, calling variables, census variables), we compute for each cell phone use variable *pvar,* and for each census variable, *cvar,* a list of 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 use variable values of *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 geographic 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, 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)* that contains the *cvar* census variable values for the same BTSs, as shown in Figure 1. Next, we report statistical results for each group of variables characterizing cell phone use

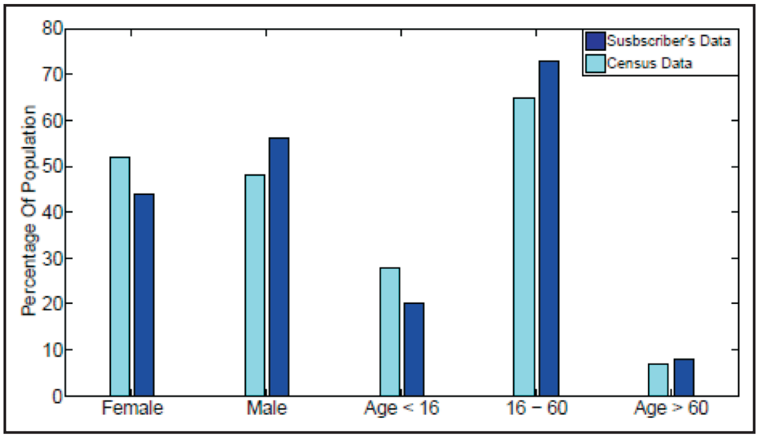


Figure 2. Histogram representing the percentage of subscribers divided by gender and age groups in our sample (dark blue) and for the whole emerging economy (light blue). We observe that both follow similar distributions.

(consumption, social, and mobility), as well as the census variables. Given the large number of pairs {*consumption variable, census variable*} that have to be evaluated, for purposes of clarity, we only report variables that gave significant statistical results. For the ANOVA tests, we report results with p <0 .05, and for Pearson's correlations, we will discuss moderate (0.4 < lrl < 0.7) and strong correlations (lrl > 0.7). Additionally, we only report voice-based mobility information, since we do not have geolocation information for SMS traffic.

3. Methodology Evaluation

To evaluate the proposed methodology, we use a CDR dataset containing five months of cell phone calls and SMSs from over 10 million prepaid and contract subscribers across 12 large and mid-sized cities in a Latin American country. We also gather the demographic and socioeconomic information shown in Table 1 as provided by the local NSI for the entire country. For the specific country under study, the NSI differentiates six SEL levels. Also, the GUs are highly granular, representing just a few city blocks each. With these datasets in hand, we compute the calling variables and merge these with the census data to obtain the pairs (*calling variable*$_i$, *census variable*$_j$) for each BTS in the cities under study.

3.1 Validity

The statistical analysis in the methodology determines whether there exist significant differences or correlations between cell phone use and census variables across different geographic regions. 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 prepaid and contract subscribers, that does not guarantee it represents the whole population. To be able to extend our findings beyond the subscribers in our sample to the entire population in large and middle-sized cities in the emerging economy under study, we need to guarantee that the sample represents 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 prepaid subscribers provide personal information regarding gender, age, and residential location, we can compare the subscribers in our study with the overall population. Figure 2 shows the distribution of age, and gender per cell phone subscriber in our sample, compared with the distribution of the same variables across the whole population obtained from the NSI's census dataset. Figure 3 shows the same comparison for the SEL variable. We can observe that the distributions are similar, with a few exceptions. In the case of gender and age groups, we observe that our subscriber distribution is slightly skewed toward male cell phone users. A similar bias is observed in the age groups. This is probably because, sometimes, the cell phone number is associated to the buyer of the cell phone (often the husband or 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 (see Figure 3), 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 with the lowest SEL, we might also lose some of the granu-

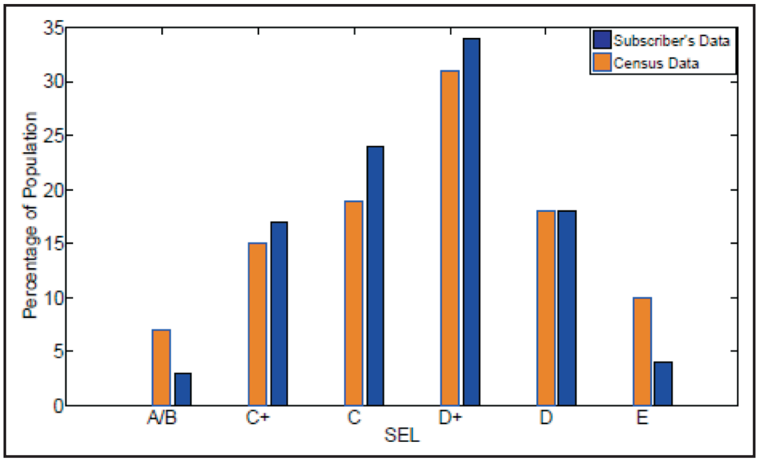


Figure 3. Histogram representing the percentage of subscribers in our sample from each socioeconomic level (dark blue) and the percentage of citizens in the country under study per socioeconomic level (orange).

larity for SEL E. Although not shown here, individual census variables in our sample showed similar distributions to the country's population. These similarities allow us to extend the findings in this section to all the large and middle-sized cities in the country, and not simply to the subscribers in our sample. However, it is important to clarify that these results are not necessarily applicable to rural environments. In any case, this is not a limitation arising from our methodology, but rather, from the data we use in the evaluation. If the calling records were to contain information regarding all citizens from urban and rural environments, and if we were able to prove that our sample followed a similar distribution to the entire population, then our methodology could output findings applicable to the population at large, in both cities and rural areas.

Next, we describe our qualitative and quantitative findings regarding the relationship among cell phone usage consumption, social and mobility variables, and socioeconomic parameters in large and middle-sized urban environments in an emerging region in Latin America. Although some of the results might seem obvious, it is important to clarify that our main contribution is to provide a large-scale methodology to reveal novel relationships, or to confirm or reject cell phone behaviors that have been well-researched at smaller scales. Section 5 presents a discussion of the results and a comparison against related work.

3.2 Consumption Variables

The *consumption variables* found to have statistical significance with respect to the census variables 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 related expenses (Expenses). In terms of SMS behavior, the total number of SMSs sent or received (Total), the output SMSs (Output), and the SMS expenses (Expenses) were all significant.

Table 2 shows the statistical results for the ANOVA tests and the Pearson's correlation. The numerical values in each cell represent ANOVA p-values, the asterisk (*) represents a moderate correlation, the double asterisk (**) represents a strong correlation, and (+) and (−) signs refer to a positive or a negative correlation between each two distributions, respectively. Recall that, for the ANOVA tests, we report results with $p < 0.05$, and for Pearson's correlations, we discuss moderate ($0.4 < |r| < 0.7$) and strong correlations ($|r| > 0.7$). Table 3, on the other hand, shows the magnitude of the changes provoked by the correlations. Specifically, for each pair of correlated variables, we show the percentage of change in the calling variable between the values for the lowest and highest quartiles of the census variable. For example, a value of 0.3 for a positive correlation between SEL levels and number of calls means that geographic regions with the highest SELs have an average of 30% more calls than regions with the lowest SELs.

Our main findings show that there exists a statistically significant difference between the socioeconomic level of a region and the average number of calls made by the subscribers within the same region. The average number of output calls also differs significantly depending on the socioeconomic level, and with even greater significance ($p < 0.01$). We also see that, as expected, the average expenses on calls are significantly different across socioeconomic levels. However, no significant difference is observed for call duration. In terms of SMSs, we observe statistically significant differences between the socioeconomic level and the average total num-

Table 2. Census Variables and Consumption Variables.

	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

Note: Numbers represent ANOVA p-values; (+) and (−) signs refer to a positive or negative correlation between each two distributions; (*) represents a moderate correlation (0.4 < |r| < 0.7); and (**) represents a strong correlation (|r| > 0.7).

ber 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 geographic area who have finished secondary school and the number of calls, the number of input calls, and the expenses, but no significant differences are found in SMS use.

The correlation results show moderate positive correlations between the SEL and both the number of calls and the expenses; i.e., regions with higher socioeconomic levels are associated with both a larger number of calls made by their citizens and larger amounts of money spent on calls. As we can observe in Table 3, regions with the highest SEL have, on average, 21% more calls and 19% more expenses than geographic regions with the lowest SELs. We also observe a positive correlation between the percentage of citizens who have completed secondary school and the number of calls and expenses, meaning that geographic regions with a larger number of citizens who possess a secondary diploma also have a larger number of calls (up to 12% more), as well as a maximum of 13% more SMSs consumed by the citizens who live in these areas.

The demographic variables also show interesting results. The percentage of middle-aged population (16–60) in a region seems to determine a significant difference in the number of calls made, as well as their associated expenses. We also found statistically significant differences in the average number of SMSs 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, and the average number of output SMSs. In fact, we observe a moderate negative correlation between the percentage of male population and the number of output calls; i.e., regions with larger percentages of male population are associated with lower numbers of output calls than less male-populated areas, with decreases of up to 11% fewer calls. The symmetric correlation is found for females, whereby the higher the female population in a region, the larger the number of output calls that are made, with a maximum difference of 10% more calls.

Finally, the goods variables that revealed statistically significant differences were the presence of a house with a cement floor, the presence of a PC in the household, and the presence of all goods in a household. We observe that the three variables show significant differences in the number of calls, the expenses associated with the calls, the average number of SMSs sent and received, and the associated expenses. In terms of correlations, geographic regions that have a larger presence of PCs at home appear to be moderately positively correlated to the number of cell phone calls and SMSs. In detail, regions with the largest number of PCs in households have a maximum of 14% more calls and 17% more SMSs than geographic regions with the lowest number of PCs at home. To summarize, we observe

Table 3. Magnitude of the Changes Provoked by the Correlations Between Census Variables and Consumption Variables.

	Calls				**SMS**		
Census Variable	**Total**	**Output**	**Duration**	**Expenses**	**Total**	**Output**	**Expenses**
SEL	0.21			0.19			
S.School	0.12			0.13			
Middle-Age							
Male		0.11					
Female		0.10					
Cement Floor							
PC	0.14				0.17		
All							

Note: Specifically, for each pair of correlated variables, we show the percentage of change in the calling variable between the values for the lowest and highest quartiles of the census variable.

that geographic regions with higher socioeconomic levels, including education and access to goods, tend to be correlated to higher consumption levels of calls, SMSs, and expenses, with increased activity of up to 21%. On the other hand, gender appears to have an impact on social aspects of the communication (duration of the calls), but not on general consumption levels.

3.3 Social Variables

The *social network variables* that revealed statistical significance with respect to socioeconomic information are the reciprocity of the communications and the physical distance between contacts.

Table 4 shows the results for the ANOVA and Pearson's correlation tests, and Table 5 shows the magnitude of the changes provoked by the correlations as the percentage of change between the lowest and highest quartiles. We observe that there exists a significant difference between the socioeconomic level of a geographic region and the number of reciprocal calls or SMSs sent and received. In fact, the statistical significance is observed when—on average—there exist one or more reciprocal calls. In the case of SMSs, the reciprocity needs to be two or higher to determine a statistical significance. We also observe that the physical distance is statistically different across diverse socioeconomic levels for both calls and SMSs.

In terms of correlations, we observe that the socioeconomic level of a geographic region 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. As Table 5 shows, higher SELs can have 12% more reciprocal R(5) calls than the lowest SELs. In terms of average physical distance between people who live within a geographic region and their contacts, we see a positive correlation with the SEL in the case of cell phone calls, but not for SMSs. In fact, regions with higher SELs might have physical distances from their contacts that are 16% higher than the lowest SEL regions. The percentage of population in a region that has finished secondary school also seems to determine statistically significant relationships with the number of reciprocal calls and the physical distance from social network contacts.

In fact, reciprocity of two calls/SMSs 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 that has attained a secondary diploma. The percentage of the population that has finished secondary school is found to be positively correlated to the number of reciprocal calls between contacts who call each other at least five times per week on average; i.e., the larger the percentage of population in a certain geographic region with a higher level of education, the more reciprocal calls we observe, with a maximum difference of 10% more calls. Similarly, we also observe that the larger the percentage of population that has completed secondary studies, the larger the physical distance between contacts; i.e., regions with higher percent-

Table 4. 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 +*

Numbers represent ANOVA p-values; (+) and (−) signs refer to a positive or a negative correlation between each two distributions; one asterisk (*) represents a moderate correlation ($0.4 < |r| < 0.7$); and two asterisks (**) represent a strong correlation ($|r| > 0.7$).

Table 5. Magnitude of the Changes Provoked by the Correlations Between 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	0.12	0.10		0.16	0.15			
S.School	0.10			0.15				
Middle-Age	0.05			0.06				
Male	0.07	0.06		0.08				
Female	0.09	0.07		0.08				
PC	0.13			0.12	0.15			0.12
All	0.14			0.13	0.10			0.13

ages of more educated citizens tend to have social networks that are more geographically dispersed (extending over areas up to 15% larger) than regions with populations with less education.

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 on the number of reciprocal calls, as well as on the physical distance. We observe that the middle-aged group (ages 16–60) shows negative correlations between the number of reciprocal calls and the age; i.e., the older the population in a geographic region, the fewer reciprocal calls appear to be made, with a maximum decrease in these calls of 5%. On the other hand, we can report that the older the population of a geographic region, the farther away voice contacts tend to be (up to 6% more dispersed for regions with the greatest elderly population). We also see that the higher the percentage of male population in a geographic area, the fewer reciprocal calls between subscribers (up to 7% less) and the greater the physical distance between voice contacts (up to 8% more dispersed). Inversely, regions with higher percentages of female population share both a higher number of reciprocal calls and smaller physical distances between the contacts, with maximum differences between top and bottom values of 9%, 7%, and 8%, respectively.

Finally, the *goods variables* that refer to the percentage of population with a PC in the household, as well as to the percentage of population with all amenities at home (electricity, water, PC, fridge, TV, and washing machine) reveal statistically significant differences in terms of both higher reciprocal calls and higher reciprocal SMSs. Analogously, the physical distance is also associated with significant differ-

Table 6. 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

Numbers represent ANOVA p-values; (+) and (−) signs refer to a positive or a negative correlation between each two distributions; (*) represents a moderate correlation ($0.4 < |r| < 0.7$); and (**) represents a strong correlation ($|r| > 0.7$).

Table 7. Magnitude of the Changes Provoked by the Correlations Between Census Variables and Mobility Variables.

	Calls			
Census Variable	**N.BTS**	**Dist.Travelled**	**Radius**	**Diameter**
SEL	0.38	0.25	0.36	0.21
P.School	0.23	0.21	0.22	0.20
Middle-Age	0.10	0.11	0.15	0.12
Male	0.15	0.16	0.14	0.15
Female	0.16	0.18	0.10	0.12
PC				
All				

ences. 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 a population. In fact, the higher the variables are for presence of a PC or of all amenities in a region, the higher the number of reciprocal calls made by the subscribers, and the greater the physical distances between these subscribers and their voice/SMS contacts are. Specifically, we observe that regions with the highest percentages of access to PCs can have 13% more reciprocal calls, and up to 15% more reciprocal SMSs. In general, it is thus fair to affirm that statistically significant differences between social variables and census data are not as extreme as the consumption variables, with changes in magnitude of up to 15% on average.

3.4 Mobility Variables

Table 6 shows that all the *mobility variables* we model show statistically significant differences with the census variables: 1) number of different BTSs visited by a person, 2) distance travelled by a person, 3) radius of gyration, and 4) diameter. On the other hand, Table 7 shows the magnitude of the changes provoked by these correlations.

We observe a strong positive correlation between the number of BTSs and the SEL, such that the higher the SEL of a region, the larger the number of BTSs used by citizens in that region. The magnitude of change between the lowest and the highest SELs is large, with the highest SEL regions showing up to 38% more BTSs visited than the poorest regions. A similar strong correlation is found for the radius of gyration, whereas distance travelled and diameter show moderate correlations. In general, we observe magnitude changes of 21–38% between the highest and lowest SELs, which are much higher than those obtained for the consumption or social variables. The percentage of citizens who have finished

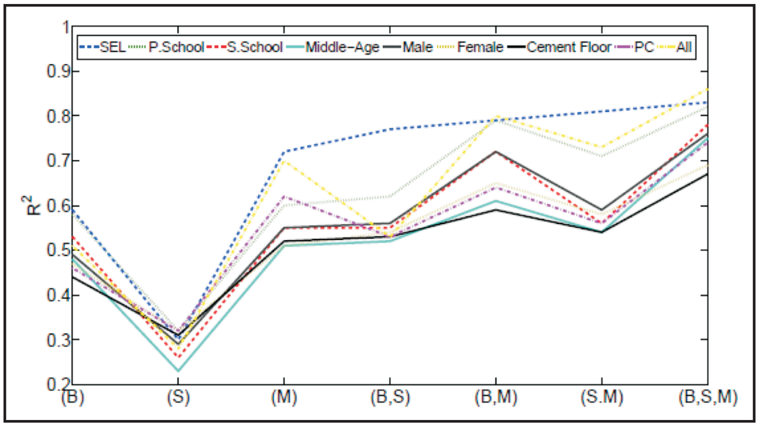


Figure 4. Multivariate regression results. Prediction of census variables from cell phone use consumption (B), social (S) and mobility variables (M). R^2 values represent the goodness of the fit.

primary studies also shows statistically significant differences across all mobility variables. We observe that geographic areas with larger numbers of educated people seem to be correlated to larger distances travelled (up to 21% larger distances), as well as to larger distances between home and work (radius values can increase up to 22%).

In terms of age, we observe similar results for the statistical differences and correlations. In fact, we see that geographic regions with younger populations have larger distances travelled (up to 11% more), as well as larger radii and diameters (up to 15% and 12% larger, respectively), meaning that the population in these regions might have longer commutes. 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, and the larger the distances travelled and radii and diameters observed. On the other hand, the higher the percentage of female population within a geographic region, the smaller the average number of BTSs visited, and the smaller the diameters, radii, and distances travelled. Gender also shows important magnitude changes across mobility variables, with differences between regions with very high or very low percentages of one gender changing their values up to 18%, as shown in Table 7. 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. In general, we observe that mobility variables are the ones that suffer the largest changes in magnitude for the highest and lowest values of regional census variables.

4. Predictive Models

In this section, we explore whether cell phone usage variables can be used to build predictive models for the census variables. Such models could provide *affordable tools to approximate* the variables of the census maps, which are expensive to compute, especially 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 consumption $\overline{B}$, social, $\overline{S}$, or mobility $\overline{M}$ cell phone use variables computed from the call records:

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

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, $\overline{B}$ refers to total number of calls, total number of output calls, duration, and expenses; $\overline{S}$ refers to reciprocal calls and physical distance, and $\overline{M}$ refers 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 4 shows the R^2 values for the prediction models built using the cell phone consumption variables $\overline{B}$, the cell phone social variables $\overline{S}$, and the mobility variables $\overline{M}$. We show results for the prediction of the census variables using only one group of variables (either $\overline{B}$, $\overline{S}$ or $\overline{M}$), or using any combination of the three groups.

Our main findings show that the combination of all three variable groups—$\overline{B}$, $\overline{S}$ and $\overline{M}$ build a highly predictive model, especially for the SEL (with $R^2 = 0.83$), the percentage of population with primary school education ($R^2 = 0.82$), and the percentage of population with *all* basic goods at home (with $R^2 = 0.86$). We also observe that the group of consumption and mobility variables ($\overline{B}$, $\overline{M}$) outperform the predictive accuracy of the social and mobility variables ($\overline{S}$, $\overline{M}$). In fact, adjusted R^2 values are considerably larger for ($\overline{B}$, $\overline{M}$) than for ($\overline{S}$, $\overline{M}$). However, the group of consumption and social variables ($\overline{B}$, $\overline{S}$) seems to be a worse predictor than the other combined pairs. Finally, in terms of individual groups of variables, the mobility variables $\overline{M}$ are, by far, the best predictive group across all census variables. As explained earlier, because our datasets only contain information regarding urban customers, these predictive results are only applicable to urban census maps. However, the same approach can be used over datasets containing both rural and urban customers to report predictive accuracies applicable to the population at large.

5. Discussion and Related Work

To further understand the potential of the novel methodology we have proposed, this section compares our evaluation results of cell phone use in an emerging Latin American region with related research for other regions. Our focus is to understand similarities and differences, as well as novel results. However, it is important to clarify that, although there exist many studies on cell phone use in emerging regions, the nature of the questions varies a lot from analysis to analysis, often making it difficult to draw direct comparisons. The discussion in this section provides mostly preliminary intuitions that compare our work with previous research, and as such, are solely based on theoretical comparisons between the works.

Our analyses regarding **consumption variables** show that regions with higher socioeconomic levels, including higher percentages of educated people or generalized access to PCs, are correlated to larger consumption variables in terms of number of calls, SMSs, and expenses. As such, our methodology provides a formal method to back a widespread assumption: *Geographic regions with higher socioeconomic levels have larger cell phone consumption levels.* Qualitative studies have found similar results: Crandall, Otieno, Mutuku, and Colaco's (2012) research in Kenya revealed that young people with no formal education use their cell phones less, with higher differences in the use of SMSs than calls. On the other hand, quantitative studies have also reported similar findings. Eagle (2008) analyzed calling records from the UK and found that regions with higher communication diversity were correlated with lower deprivation indexes, which are the UK measure of SELs. Another important highlight regarding consumption variables is that *geographic regions with higher percentages of female population share larger numbers of output cell phone calls.* This result is opposed to findings by Blumenstock and Eagle (2010), where interviews with individuals in Rwanda revealed that men reported a higher number of cell phone calls than women. Interestingly, the authors also used calling data from Rwanda to automatically compute gender differences in number of calls and found no statistically significant results. Our results might be revealing of cultural and gender differences across emerging regions on different continents.

In terms of **social networks,** our analyses highlight important positive correlations among socioeconomic levels, education, and gender, and the reciprocity of the calls or the physical distance between contacts. These findings reveal that *geographic regions with lower socioeconomic levels build weaker social network structures, which tend to be physically closer.* In a recent study, Castells, Fernández-Ardèvol, and Galperin (2011) reported the existence of asymmetric communication structures in low-income areas where cell phones are mostly used as a tool to receive calls. Our results back these findings and confirm that calling reciprocity is, in fact, a behavior that is much more present in high-income regions. Similarly, the qualitative analysis carried out by Donner (2007) in Kigali with 277 microentrepreneurs showed that users with higher educational levels were also more likely to add new contacts to their social networks. Our evaluation shows that regions with greater percentages of citizens who have finished primary or secondary school are correlated to larger numbers of highly reciprocal calls, which might be associated with users who feel comfortable using the phone as a way to both communicate and add contacts.

On the other hand, physical distance between cell phone contacts is a novel measure that, to our knowledge, has not been studied through qualitative studies in the past, probably due to its complex measurement. However, it is an important variable that can provide relevant information regarding family and business structures in emerging regions. Barrantes and Fernandez-Ardèvol (2012) reveal that, although the use of cell phones by farmers at fairs in the area of Puno, Peru, is common, the cell phone still remains most important in its capacity as a tool to communicate with relatives and friends. That regions with lower socioeconomic levels appear to have smaller physical distances between contacts might be due to the fact that relatives and families in these regions live in more geographically constrained areas.

Finally, our analysis on **mobility variables** clearly reveals important differences in mobility patterns across socioeconomic levels and gender: *Geographic areas with higher socioeconomic levels and larger male populations share larger mobility patterns.* Hanson and Johnston (1985) and Wachs (1987) showed that there exist important gender differences in the travel patterns of men and women in Sweden and the United States. Both found that women tend to make shorter work trips than their male counterparts. Additionally, Maden (1981) revealed that, in general, working women earn lower salaries and try to avoid longer commutes. But most important, women generally look for shorter commutes to be able to balance work and family. Our results, although coming from an emerging region, are consistent with these findings. As for socioeconomic levels, there also exist a considerable number of studies confirming our findings. For example, Poppola and Faborode (2011) collected interview data from more than 10,000 households in Ibadan, Nigeria, and revealed that higher socioeconomic levels were positively correlated to larger travel distances.

Overall, we believe that our main contribution is to provide a methodology that combines large-scale calling data with surveys from the NSIs to allow researchers in the social sciences to extract inferences at large scale (see section 2). We have evaluated the methodology with a cell phone dataset from an emerging region in Latin America, and we have shown that our results, although obvious in many cases, confirm results previously found through qualitative studies. These studies, based on interviews and surveys, have a smaller scale by nature. Thus, our methodology provides a back-up confirmation for social scientists looking to challenge their results with larger populations than they can capture with surveys. On the other hand, our methodology also uses novel variables (like the physical distance) that can reveal findings that might bring about new research questions of interest to social scientists.

Another important contribution of our work is the empirical evaluation of our methodology, as shown in section 3. We believe that many of the results presented here can be helpful to policy makers who are used to managing country-scale statistics. In fact, our results can shed light onto traveling patterns, social network structures, and cell phone use patterns of citizens in urban environments. Finally, researchers working on cell phone-based services can also benefit from the findings coming out of our evaluation for the development of personalized and meaningful *services for all.*

6. Conclusions and Future Work

We have presented a novel methodology to understand, at large-scale, the relationship among socioeconomic variables and the way people use cell phones. Our novel approach combines large datasets of cell phone records with countrywide census data gathered by NSIs to reveal large-scale findings without the need to carry out personal interviews or questionnaires. We have applied this methodology to the calling records of urban citizens of an emerging Latin American region. The main findings reveal moderate and strong correlations among the socioeconomic level of a person and the cell phone-related expenses incurred, the reciprocity of her or his communications, the physical distance between the caller and her or his contacts, or the geographic areas where the person typically moves about. Additionally, we have provided a model that allows for predicting a variety of socioeconomic variables exclusively from cell phone records. Such a model can be used as a cost-effective approximation of census information, which is expensive to compute, especially for emerging economies. Future work will focus on computing similar analyses for other emerging economies to understand cross-cultural differences in the use of cell phones, and on studying whether country-based differences exist across urban and rural areas. ■

References

Barrantes, R., & Fernandez-Ardèvol, M. (2012). Mobile phone use among market traders at fairs in Peru. *Information Technologies & International Development, 8*(3), 35–52.

Blumenstock, J. E., & Eagle, N. (2010, December 13–16). *Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda.* Fourth International Conference on Information and Communication Technologies and Development (ICTD), London, UK.

Castells, M., Fernández-Ardèvol, M., & Galperin, H. (2011). Síntesis de resultados y conclusions [Summary of results and conclusions]. In M. Fernández-Ardèvol, H. Galperin, & M. Castells (Eds.), *Mobile communications and socioeconomic development in Latin America* (pp. 319–350). Barcelona, Spain: Ariel.

Crandall, A., Otieno, A., Mutuku, L., & Colaco, J. (2012, November). *Mobile phone usage at the Kenyan base of the pyramid.* iHub Research. Retrieved from http://blogs.worldbank.org/ic4d/files/ic4d/mobile_phone_usage_kenyan_base_pyramid.pdf

Donner, J. (2007). The use of mobile phones by micro entrepreneurs in Kigali, Rwanda: Changes to social and business network. *Information Technologies & International Development, 3*(2), 3–19.

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

Frias-Martinez, V., Virseda, J., Rubio, A., & Frias-Martinez, E. (2010, December 13–16). *Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data.* Fourth International Conference on Information & Communication Technologies and Development (ICTD), London, UK.

Gonzalez, M., Hidalgo, C., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature, 453,* 479–482.

Hanson, S., & Johnston, I. (1985). Gender differences in work-trip length: Explanations and implication. *Urban Geography, 3,* 193–219.

Hughes, N., & Lonie, S. (2007). M-PESA: Mobile money for the "unbanked": Turning cellphones into 24-hour tellers in Kenya. *Innovations: Technology, Governance, Globalization, 2*(1–2), 63–81.

Lane, J. M., Carpenter, L. C., Whitted, T., & Blinn, J. F. (1980). Scan line methods for displaying parametrically defined surfaces. *Communications of the ACM, 23*(1), 23–34.

Madden, J. (1981). Why women work closer to home. *Urban Studies, 18,* 181–194.

Poppola, K., & Faborode, T. (2011, June 15–18). *Effect of socio-economic status on household all purpose travel patterns of men and women in Ibadan, Oyo State, Nigeria.* Fourth European Conference on African Studies (ECAS), Uppsala, Sweden.

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011). Prediction of socioeconomic levels using cellphone records. International Conference on User Modeling, Adaptation and Personalization, UMAP'11, July 11–15. Industrial Track, Girona, Spain.

Verclas, K. (2010). *Scaling mobile services for development: What will it take?* Retrieved from http://mobileactive.org/scaling-mobile-services-development-what-will-it-take

Voronoi, G. (1907). Nouvelles applications des paramètres continus à la théorie des formes quadratiques [New applications of continuous parameters to the theory of the quadratic form]. *Journal fur die Reine und Angewandte Mathematik, 133,* 97–178.

Wachs, M. (1987). Men, women, and wheels: The historical basis of sex differences in travel patterns. *Transportation Research Record, 1135,* 10–16.