Computing Cost-Effective Census Maps From Cell Phone Traces

Vanessa Frias-Martinez, Victor Soto, Jesus Virseda, and Enrique Frias-Martinez

Telefonica Research, Madrid, Spain
vanessa,vsoto,jvjerez,efm@tid.es

Abstract. Census maps contain important socio-economic information regarding the population of a country. Computing these maps is critical given that policy makers often times make important decisions based upon such information. However, the compilation of census maps requires extensive resources and becomes highly expensive, especially for emerging economies with limited budgets. On the other hand, the ubiquitous presence of cell phones, both in developed and emerging economies, is generating large amounts of digital footprints. These footprints can reveal human behavioral traits related to specific socio-economic characteristics. In this paper we propose a new tool, *CenCell*, to approximate census information from behavioral patterns collected through cell phone call records. The tool provides affordable census information by accurately classifying socio-economic levels from cell phone call records with classification rates of up to 70%.

1 Introduction

Census maps gather large amounts of information regarding the socio-economic status of households at a national scale. These maps contain information that characterizes various social and economic aspects like the educational level of the citizens or the access to electricity. Such information is aggregated and reported at various granularity levels, from a national scale, to states, all the way down to urban geographic areas of a few square kilometers. The accuracy of these maps is critical given that many policy decisions made by governments and international organizations are based upon variables measured through census maps. National Statistical Institutes compute such maps every five to ten years, and typically require a large number of enumerators that carry out interviews gathering information pertaining the main socio-economic characteristics of each household. All these prerequisites make the computation of census maps highly expensive, especially for budget-constraint emerging economies. To reduce costs, countries like Mexico or Guatemala have made cuts both in the number of interview questions and in the number of citizens interviewed, which unfortunately impacts the quality of the final census information [1, 2].

On the other hand, the ubiquitous presence of cell phones in emerging economies is generating large datasets of digital footprints. Data mining techniques applied to such datasets can be used to reveal cell phone usage patterns specific to socio-economic levels. Previous research has already shown that cell phone-based behavioral patterns

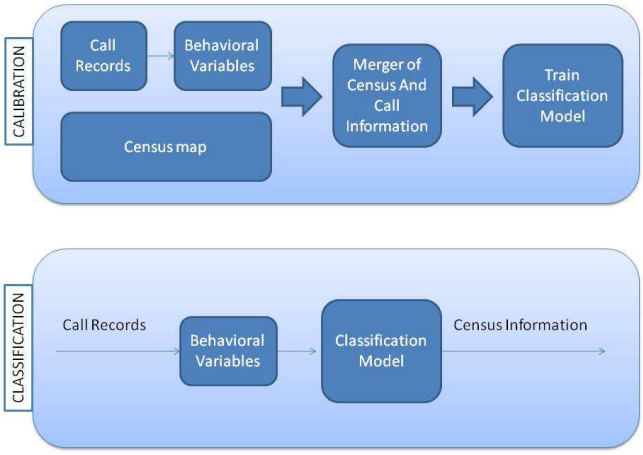


Fig. 1. *CenCell* Architecture.

might be correlated to specific socio-economic characteristics [3, 4]. For example, Eagle *et al.* showed correlations between the size of a cell phone social network and the socio-economic level of a person, and Frias *et al.* observed strong relationships between mobility and socio-economic indices [5].

In this paper, we propose a new tool for governments and policy makers that allows to compute affordable census maps by decreasing the number of geographical areas that need to be interviewed by the enumerators. The tool, called *CenCell*, is designed to allow institutions to approximate the census information of areas not covered by the enumerators using anonymized cell phone call records gathered by telecommunication companies. At its core, *CenCell* consists of a classification algorithm that determines the socio-economic level of a region based on the aggregated cell phone behavioral patterns of its citizens. Thus, *CenCell* significantly decreases the workload of the enumerators that carry out the interviews and as such, allows to reduce the budget allocated for the computation of census maps.

2 CenCell: General Architecture

Figure 1 shows the general architecture of the tool. It consists of two main components: (1) the *calibration* phase, which needs to be executed only once to set up the system for a region; and (2) the *classification* phase, which is executed every time census information is required for a specific geographical area in the region that was not covered by the enumerators through household surveys. The *calibration* phase needs two datasets, one containing anonymized cell phone call records for the region under study and another one containing the regional socio-economic levels computed by the local National Statistical Institute through household surveys. This phase first computes a set of cell

phone usage behavioral patterns from call records. Next, it combines both datasets to obtain a map that associates to each cellular tower in the region under study a set of cell phone behavioral variables and a socio-economic level. This map is used to train a classification model that will output the socio-economic levels of the areas not covered by the enumerators based on the cell phone behavioral patterns of its citizens (*classification* phase). It has to be noted that *CenCell* uses anonymized aggregated patterns of behavior so no individual information is used to build the classification models.

During the *calibration* phase *CenCell* tests the classification accuracy of a battery of supervised and unsupervised algorithms and selects the one with the best results, which will be used in the *classification* phase. Previous work explored the use of supervised techniques (SVMs and Random Forests) to forecast socio-economic levels from cell phone records [4]. However, given that *CenCell* computes the socio-economic levels (SELs) for each cellular tower based on a weighted average of overlay maps, the final values might be smoothed or blurred depending on the information distribution in the original maps. In an attempt to overcome this problem, *CenCell* also explores unsupervised techniques to identify groups of cell phone behaviors without prior knowledge of their socio-economic values. Sections 4 and 5 cover details about the different techniques explored by *CenCell* and its results. Additionally *CenCell's calibration* phase, computes classification models for different socio-economic level granularities. Although the SEL is a continuous variable, it is often times expressed as a discrete value through a letter (A, B, C, etc.). The granularity of the SELs *i.e.,* the number of SEL classes in which the continuous values are divided into, varies a lot across studies. Some researchers differentiate three socio-economic levels [6] while others prefer to use a larger range of values in their analyses [7]. To account for this need, *CenCell* outputs the best predictive technique for each granularity value in the *calibration* phase. As such, *CenCell* provides a *knob* that allows researchers to select a specific granularity depending on the classification error they are willing to accept. In Section 5, we delve more into results and implications of this approach. Finally, it is important to highlight that in order to build accurate classification models the tool needs that: (i) the area selected for the *calibration* phase is representative of the different socio-economic levels of the country and (ii) both call records and socio-economic variables correspond to a similar period in time.

On the other hand, the *classification* phase uses the models generated by the *calibration* phase to compute the socio-economic level of the geographical areas in the region that were not covered by the enumerators. This phase, which only requires access to anonymized aggregated calling records, can be executed as many times as needed and allows policy makers to compute affordable census maps without the need to held household surveys across all the region under study but rather in a few areas. Next, Sections 3 and 4 describe the *calibration* phase in detail.

3 Datasets and Information Merging

3.1 Call Detail Records

Cell phone networks are built using a set of base transceiver station (BTS) towers 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 using Voronoi tesselation. 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. From all the information contained in a CDR, *CenCell* 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 took place. Using call detail records, *CenCell* computes 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, 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 of each subscriber. These variables include the 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. 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. In total, *CenCell* computes 69 consumption variables, 192 social network variables and 18 mobility variables.

3.2 Census Data

CenCell uses the census maps collected by National Statistical Institutes (NSI) to gather socio-economic information about the population under study. NSIs carry out individual and household surveys at a national level every five to ten years. These surveys employ a large staff of enumerators that are responsible for interviewing every household head within their assigned geographical area. The enumerators have been especially trained to be able to gather all the required information in a proper manner. Although in some cities in emerging economies 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, NSIs only make public average values over specific geographical areas. These areas might represent states, cities, neighborhoods or *geographical units (GUs)*, the smallest geographical division, which divides cities into small areas of up to a few square kilometers (approximately blocks). The census variables gathered by NSIs are usually divided into three groups: *education variables*, *demographic variables* and *goods' ownership variables*. With this information NSIs compute the socio-economic level (SEL) of a region as a weighted average of all the census variables. As mentioned, SEL values are typically represented as a set of discrete values typified through letters.

3.3 Merging Call Records with Census Data

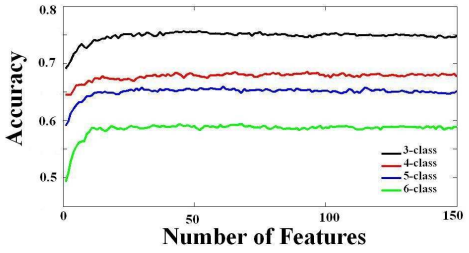
The objective of the *calibration* phase is to build a socio-economic classification model from cell phone records. To do so, we need to compute a training set that associates to each BTS cell tower: (1) aggregated cell phone behavioral variables for the citizens that

live within the cell tower coverage area and (2) the corresponding socio-economic level (SEL) for that same area. However, given that call records are gathered per BTS area and that the census information is reported per geographical unit (GU), the tool uses a three step protocol to merge the two sources of information [8, 4]: (Step 1) Associate the residential location of each subscriber to a BTS area; (Step 2) Compute the overlapping between Voronoi diagrams and Geographic Units (GUs). This mapping allows us to merge census and BTS maps so as to associate socio-economic levels to each BTS coverage area; and finally, (Step 3) For each BTS, compute the aggregated consumption, social and mobility variables of all the subscribers whose residential location is within that area. These three steps produce a map that associates to each BTS a socio-economic level as well as a set of variables (features) characterizing the average cell phone usage for that area. Next, we explain the ML techniques used in the *calibration* phase.

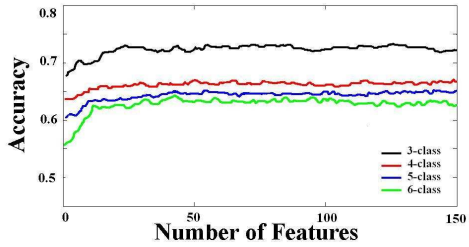
4 Classification Model Generation

The generation of the Classification Model represents the last step in the *calibration* phase. It builds a model that will allow the *classification* phase to approximate the socio-economic level of geographical areas that have not been covered by the evaluators to save budget. Specifically, it takes as input the $(SEL, features)$ dataset for a specific region and identifies the best classification technique for each socio-economic granularity. *CenCell* considers four different SEL granularities: three, four, five and six ranges (classes). For six SEL classes A covers range $[100 - 83)$, B $[83 - 66.4)$, C $[66.4 - 49.8)$, D $[49.8 - 33.2)$, E $[33.2 - 16.6)$ and F $[16.6 - 0]$. Smaller granularities follow a similar distribution in ranges. In terms of the features (cell phone variables), having a large number might boost classification, but it can also generate a lot of noise (*the curse of high dimensional datasets*). For that reason, *CenCell* first identifies the significance of each feature and applies classification techniques on the features ordered by their relevance. Specifically, it evaluates two different feature selection techniques: maxrel and mRMR (as difference mRMR-MID or quotient mRMR-MIQ) [9, 10].

Once the features have been ordered according to their significance, *CenCell* tests supervised SVMs as well as unsupervised EM Clustering. We selected SVMs since they have been successfully used in similar classification problems [11, 12]. On the other hand, unsupervised EM clustering was selected among all clustering unsupervised techniques, since populations have been previously shown to follow Gaussian distributions in terms of socio-economic variables [13]. *CenCell* evaluates each combination of technique and granularity and outputs the one that has the best predictive power for each SEL granularity. As a result, the tool provides policy makers with the possibility of selecting a granularity and classification quality according to their own interests. This step first partitions the BTS dataset with the ordered features and SELs for training and testing, containing 2/3 and 1/3 respectively. Using each supervised and unsupervised technique, *CenCell* computes the classification rate for each SEL granularity, from three classes (A, B and C) to six (A, B, C, D, E and F), and for each subset of ordered features in $n = \{1, \ldots, 279\}$ produced by Maxrel, mRMR-MIQ and mRMR-MID. *CenCell* implements the SVM using a Gaussian RBF kernel and identifying its two parameters (C and γ) through a 5-fold cross-validation over the training set



(a) SVM



(b) EM Clustering

Fig. 2. Accuracy versus number of features and SEL granularity for (a) supervised SVM and (b) unsupervised EM clustering. The feature selection technique used was mRMR-MIQ.

for each combination of technique, granularity, and subset of features. As for EM clustering, *CenCell* computes a mixture Gaussian models for each socio-economic level, granularity and subset of features until the log-likelihood values are maximized. During testing, each final cluster is labeled with the dominating socio-economic level.

5 Experimental Evaluation

The objective of this section is to evaluate the classification power of *CenCell* to determine the socio-economic level of regions that are not covered by household surveys to save budget. Our CDR dataset contains 6 months of cell phone calls, SMSs and MMSs from over $500,000$ pre-paid and contract subscribers from a large city in an emerging economy in Latin America. The city has a total of 920 BTS cellular towers and the subscribers represent a 20% of the total population. On the other hand, the census information was acquired from the local NSI and contained a total of 1200 GUs with their SEL expressed as a continuous value between 0 and 100. The city was selected because it covered all ranges of SELs. Our final dataset consists of 920 pairs $(SEL, features)$ that we divide into training (552 pairs) and testing sets (368 pairs).

The classification accuracies for each pair of technique and socio-economic granularity explored by *CenCell* are presented in Figure 2. We observe that SVMs achieve classification rates of up to 76% when differentiating three SEL classes and the top 38 ordered features selected by mRMR-MIQ. We also notice that as we increase the granularity of the SEL, the classification accuracy decreases reaching a value of 57% for

six SEL classes and the top 19 features. On the contrary, EM clustering achieves worse classification rates than SVM for granularities three, four and five. However, EM clustering yields better results when six socio-economic levels are differentiated. It shows accuracies of 63% compared to the 57% reached by SVMs, with 6 clusters and the top 20 features. We hypothesize that as the socio-economic granularity increases, the map-overlay technique in (Step 2) generates more blurred SEL levels thus making unsupervised techniques a more adequate approach. At the end of the *calibration* phase, *CenCell* would select SVMs as classification tools for granularities three, four and five; and would select EM clustering when six socio-economic levels are differentiated. To understand better the nature of the classification errors, we analyzed the confusion matrices for the best approaches selected by *CenCell*. These confusion matrices revealed that when incorrectly predicted, the output SEL class *tends to be* adjacent to the correct one. Such errors reflect an implicit order between the SELs which limits the impact of the errors on the computation of the maps.

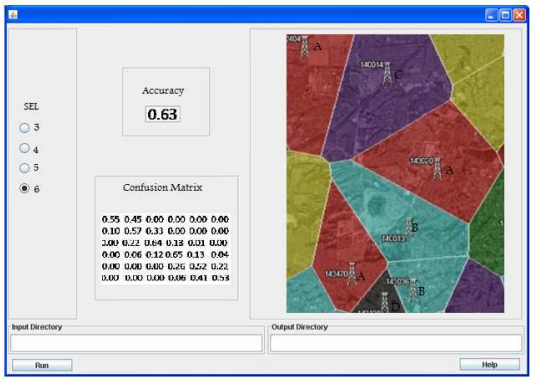


Fig. 3. *CenCell* graphical interface showing a zoomable map.

6 CenCell For Policy Makers

Socio-economic levels are crucial variables to measure the evolution of a society. However, obtaining socio-economic maps is a costly process, specially for emerging economies. *CenCell* allows policy makers to reduce the number of geographical areas where enumerators need to carry out interviews by approximating the socio-economic level of these areas with features extracted from calling records. The experimental evaluation of *CenCell* has shown that using socio-economic information from 66% of all the 920 geographical areas within a city, we achieve SEL classification rates of up to 76%. This means that policy makers could eliminate household surveys for the other 34% of areas in the city and compute fairly accurate approximations of their socio-economic values while saving budget. Given that the NSI of the city under study determines an average

census expense per survey area of approximately 904USD, *CenCell* would have saved around $332,672$USD to the local NSI.

Figure 3 shows the graphical interface for *CenCell*. The tool offers policy makers the possibility of selecting a specific granularity for their analyses (from three to six). Upon selection, *CenCell* determines the SEL of the regions that did not participate in the household surveys and builds a *zoomable* socio-economic map of the region with both census and approximate SEL values. Additionally, it shows the accuracy of the map together with the confusion matrix. As explained in the experimental section, the larger the granularity, the lower the accuracy of the maps computed by *CenCell* will be. However, such errors tend to be adjacent to the real values, which highly limits the error impact in the analysis. By exploring different SEL granularities, their associated maps and accuracies, policy makers can decide which combination is more suited for their analyses. As such, *CenCell*constitutes a powerful tool for socio-economic analyses that policy makers can use as a black box *i.e.*, without any need to understand the algorithms supporting the computation of the maps. Future work will focus on predicting socio-economic levels over time. For that purpose, we will work with time-series of household surveys and evaluate whether the evolution of census information over time can be modelled and predicted using calling records. Additionally, we will also work on providing more formal economic models to understand the budget savings that *CenCell* can bring about to emerging economies.

References

1. http://www.oem.com.mx/oem/notas/n1646250.htm.
2. http://www.cicred.org/Eng/Publications/pdf/c-c20.pdf.
3. N. Eagle. Behavioral inference across cultures: Using telephones as a cultural lens. *IEEE Intelligent Systems*, 23:4:62–64, 2008.
4. V. Soto, V.Frias-Martinez, J. Virseda, and E. Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *User Modelling, Adapt. and Pers., UMAP*, 2011.
5. V. Frias-Martinez and J. Virseda. On the relationship between socio-economic factors and cell phone usage. *ICTD, Atlanta, USA*, 2012.
6. J. Wyatt and K. Mattern. Low-ses students and college outcomes: The role of ap fee reductions. *College Board: AP Data and Records*, 2011.
7. E. Worrall, S. Basu, and K. Hanson. The relationship between socio-economic status and malaria: a review of the literature. *Health Economics For Developing Countries*, 2003.
8. V. Frias-Martinez, J. Virseda, A.Rubio, and E. Frias. Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. *ICTD*, 2010.
9. Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2005.
10. Chris H. Q. Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinformatics and Comp. Biol.*, 3(2):185–206, 2005.
11. Robert Burbidge and Bernard Buxton. An introduction to support vector machines for data mining. Technical report, UCL, 2001.
12. Enrique Frias-martinez and et al. Survey of data mining approaches to user modeling for adaptive hypermedia. *IEEE Transactions on Systems, Man and Cybernetics*, 36(6), 2006.
13. M. Yadollahi. Factors affecting family economic status. *European Journal of Scientific Research*, 37:94–109, 2009.