The Relevance of Cell Phone Traces for Predicting Socioeconomic Time Series

ABSTRACT

National Statistical Institutes typically hire large numbers of enumerators to carry out periodic surveys regarding the socioeconomic status of a society. Such approach suffers from two drawbacks:(i) the survey process is expensive, especially for emerging countries that struggle with their budgets and (ii) the socioeconomic indicators are computed *ex-post i.e.,* after socioeconomic changes have already happened. We propose the use of human behavioral patterns computed from calling records to predict future values of socioeconomic indicators. Our objective is to help institutions be able to forecast socioeconomic changes before they happen while reducing the number of surveys they need to compute. For that purpose, we explore a battery of different predictive approaches for time series and show that multivariate time-series models yield R-square values of up to 0.65 for certain socioeconomic indicators.

1. INTRODUCTION

The development of a society is typically measured through socioeconomic indicators. Variables such as the levels of employment, the gross domestic product (GDP) or the consumers' price index (CPI) provide insightful information regarding the socio-economic status of households at a national scale. Accurately computing such information is critical given that many policy decisions made by governments and international organizations are based upon such socioeconomic variables. For that purpose, National Statistical Institutes (NSIs) typically hire large numbers of enumerators that carry out periodic interviews to gather information pertaining the main socioeconomic indicators of a society. Such approach has two important disadvantages: (i) the survey process is expensive, especially for emerging countries that struggle to balance their budgets and (ii) the socioeconomic indicators are computed *ex-post i.e.,* after changes in the socioeconomic indicators have happened. In this paper, we explore an alternative model for the computation of socioeconomic indicators in budget-limited regions that allows to save on budget by reducing the frequency of the periodic surveys and that provides information *ex-ante i.e.,* forecasts socioeconomic indicators before they actually happen. As such, institutions working for social good will be able to use affordable forecasts to react before specific events like an increase in unemployment actually happen.

The ubiquitous presence of social media and cell phones is generating large datasets of web searches, tweets or call logs that reveal human behavioral footprints. Data mining techniques applied to such datasets have been used to extract usage patterns correlated to specific social, economic or health indicators. Ginsberg *et al.* presented *Google Flu Trends* and *Google Correlate* which use Google daily web search logs statistics related to various socioeconomic indicators to forecast its future values [9, 7]. For example, the authors showed that the Google web searches related to refinance index or mortgage rates accurately predict the time series of such indicators computed by different banking associations through surveys [9]. Similarly, the authors showed that the time series modeling daily web searches of words related to *influenza* can predict the weekly CDC reports on ILI (influenza-like illnesses) in the US [7]. Moving to Twitter logs, Ruiz *et al.* showed that Twitter activity is strongly correlated to time series from the financial domain and can act as predictors of different economic indicators [10]. In fact, modeling volumetric and social network features from Twitter logs, the authors showed that time series-based predictive models enhanced with Twitter activity improve the predictability of economic indicators.

As opposed to the use of Twitter logs or Google searches –that might be associated to a population with certain socioeconomic and literacy levels– cell phone calling logs have the advantage of representing large percentages of the population given the high penetration rates of cell phones in emerging regions. Previous research using cell phone calling logs has already shown that cell phone-based behavioral patterns are correlated to specific socio-economic character-

istics [3, 11]. For example, Eagle *et al.* showed correlations between the size of a cell phone social network and the socioeconomic level of a person, and Frias *et al.* observed strong relationships between mobility and socio-economic indices [6]. Additionally, Soto *et al.* showed that the socioeconomic level(SEL) of a region at a given moment in time, can be predicted from cell phone activity during the same time period with an 80% of accuracy using training sets containing both SELs and calling activity [11, 4]. Although this work threw some light onto the predictability of SELs using calling data, it focused on *predicting the present i.e.,* SEL values at a specific moment in time rather than forecasting future values.

To overcome these limitations, we focus our research on evaluating whether the time series of socioeconomic indicators computed by National Statistical Institutes (NSIs) using surveys can be forecasted using behavioral information extracted from calling records. For that purpose, we build time series modeling the calling behavior of a population over a period of time and evaluate its predictive power over the socioeconomic time series computed by the NSIs. By using calling records which are already collected by telecommunications companies for billing purposes, we seek to provide forecasting tools that will allow institutions to save on their survey expenses and to react before specific social changes happen. The main contributions of our paper are:

- To analyze the relationship between socioeconomic and calling records time series. The former obtained from NSIs through surveys and the latter computed from calling records over a period of time.
- To evaluate the predictive power of calling data time series to forecast socioeconomic indicators time series.

Our contributions extend previous work by predicting future time series values instead of simply *predicting the present*. The rest of the paper is organized as follows. Section 2 explains the related research and Section 3 describes the datasets used in the paper. Next, in Section 4 we present the methodological approach and in Section 5 we describe our results. We finish with conclusions and future work in Section 6.

2. RELATED WORK

In this section, we describe research findings that use human-generated datasets like web searches, tweets or call logs over a period of time to predict social, economic and health indicators computed through interviews by different institutions.

2.1 Web Search Logs

Ginsberg *et al.* developed *Google Flu Trends* which uses Google web search logs related to influenza to forecast the weekly CDC reports on ILI (influenza-like illnesses) in the US [7]. The authors divided the web searches into nine US regions and computed, for each region, time series for the number of weekly queries related to ILI. Using a linear model on the weekly ILI web searches from 2004 to 2007 the authors obtained a fit for the CDC data with a mean correlation across regions of $r = 0.9$. Additionally, to evaluate the forecasting power of Google's web searches, the authors proposed a regression on the regional web searches with untested data from 2008 and showed a mean correlation across regions of 0.97 with the CDC time series. Similar results were reported for the prediction of dengue incidence in Singapore and Bangkok computed by the ministries of health from google dengue-related searches [1].

These tools show that health-related web searches can predict the future number of ILI cases computed by the CDC with high accuracy. Based on such results, Google developed *Google Correlate* which extends *Google Flu Trends* by analyzing the predictability of different social and economic indicators [9] such as refinance index or mortgage rate from related web searches. The authors compare the predictability of different socioeconomic time series using both Auto-Regression (AR) and Vector Autor-Regression (VAR) models. AR models predict socieocomic time series exclusively from previous values and VAR models extend AR models by including web search information. Their results show that VAR models including search information increase the quality of the forecast reaching larger R-square values while reducing the MAE (mean absolute error).

2.2 Twitter Logs

Ruiz *et al.* studied the relationship between Twitter activity and time series from the financial domain [10]. The authors used a 6-month dataset of Twitter activity and extracted both activity and graph features. Activity features refer to volumetric measures of Twitter activity talking about companies and the stock market including number of tweets or number of hashtags; whereas graph features modeled the properties of the Twitter graph that is formed when users tweet or re-tweet about stock companies including number of nodes, edges, number of connected components or degree. These features, modeled over time, generate time series that can be compared against stock market data series to understand the relationship between both. The authors explored how the use of specific Twitter features could enhance the trading strategy of a trader at the stock market. For that purpose, they evaluated four trading strategies: (i) Random, random selection of the stocks; (ii) Fixed, select stock using a particular financial indicator of the company; (iii) Auto-Regression (AR): predict the stock with the largest benefits exclusively using previous stock data and (iv) Twitter-Augmented Regression (VAR): predict the stock with the largest benefits using both previous stock data and activity and graph Twitter features. The authors showed that the Twitter-augmented strategy with the feature *number of nodes* is the one that yields the highest benefits thus highlighting the importance of using external user-generated information to enhance financial models.

A similar approach was used by Zhang *et al.* to show the existence of correlations between the sentiment in specific Twitter posts and stock market indicators and to analyze the predictive power of microblogging logs with respect to specific economic indicators [12]. Using one year of re-tweets (*RT @*) originating from the US and containing both feeling- and economic-related words – such as *hope* or *dollar*– the authors built two time series: the number of re-tweets and the evolution of economic indicators NASDAQ, DJIA or S&P. The authors found statistically significant correlations between tweet statistics and changes in oil price or the DJIA. Additionally, using correlation and Granger's causality analysis the authors posit that Twitter posts might be able to forecast changes in economic indices one day in advance.

2.3 Cell Phone Records

There exists a large body of work analyzing the relationship between socioeconomic indicators and cell phone calling records [2, 5, 6]. Blumenstock *et al.* studied the impact that factors like gender or socio-economic status have on cell phone use in Rwanda [2]. The authors combined two datasets, one containing call detail records from a telco company in Rwanda and the other one containing socioeconomic 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. Similarly, Frias *et al.* showed that there exist differences between specific socioeconomic factors and how cell phones are used by citizens in an emerging economy in Latin America [6]. The authors combined cell phone calling records from an emerging region with socioeconomic information collected by the National Statistical Institute of the country through personal interviews and questionnaires. The results showed statistically significant differences between socioeconomic levels and the number of calls people make; between the education level and the reciprocity of the calls or between gender and the average distances travelled by citizens, among others.

Moving beyond statistical relationships, Soto *et al.* extended the previous research by proposing the use of Support Vector Machines (SVMs) and Random Forests to predict the socioeconomic level of a region based on cell phone usage patterns computed from call logs [11]. The authors use both call logs and socioeconomic indicators from 2010 and divide them into training and testing sets, reporting forecast accuracy rates of over 80%. However, it is important to highlight that this approach can only predict the present *i.e.,* predict the socioeconomic level of a region at a moment in time, based on the socioeconomic levels and call logs from other regions at that same moment in time. Based on the successful predictive approaches showed using Google and Twitter time series, our paper extends Soto's approach by moving from *predicting values in the present* to the prediction of future values in socioeconomic time series using calling logs. We posit that as opposed to Google or Twitter, the penetration rates of cell phones in emerging economies is larger across socioeconomic levels, and thus behavioral models built from these might be able to represent larger segments of the population.

3. DATASETS

In this section, we describe the two datasets involved in the analysis of the predictive power that call logs have with respect to socioeconomic indicators time series.

3.1 Socioeconomic Indicators

We gather socioeconomic time series from the local National Statistical Institute (NSI) of an emerging economy in Latin America. The local NSI periodically carries out surveys at each state of the country so as to obtain monthly values for a wide range of socioeconomic indicators. For analytical purposes, we focus our research on six socioeconomic indicator time series that are computed, monthly, at a state level including (i) Total assets, measuring both tangible and financial assets of the state (ii) Total number of employed citizens (iii) Total number of workers employed by private industries and organizations, (iv) Total number of civil servant employed by public institutions, (v) Total number of subcontracted workers and (vi) Total number of subcontracted civil servants. We retrieve the previous time series for one of the largest states in the emerging region under study and for a time period of 17 months, which is the range for which we also have cell phone calling logs available.

3.2 Cell Phone 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. 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 17 months (from February 2009 to June 2010) of daily cell phone calls from pre-paid and contract subscribers in a Latin American country. From these call detail records, we compute two groups of variables so as to model cell phone usage: consumption and mobility variables. The *consumption variables* characterize the general cell phone use statistics for a specific region and period of time. Specifically, we measure the number of input or output calls and its duration. The *mobility variables* characterize spatio-temporal mobility patterns with the granularity of the area of coverage of a BTS. Specifically, we measure the average number of BTSs used; the average *talk distance* or distance traveled by customers while talking on the phone; the average *route distance* or distance traveled by customers between phone calls; the average *total distance traveled* by customers during a period of time; the *radius of gyration* or distance between the used BTSs weighted by the number of calls made from each tower and the *diameter* or distance between all used BTSs. The radius of gyration can be considered a measure of the area where a person typically works and lives, whereas the diameter approximately represents the geographical area where a person spends most of her time (work and leisure).

Given that the selected socioeconomic time series from the NSI are computed at a state level with monthly frequency, both consumption and mobility variables for the population under study are obtained with the same geographical and temporal granularity *i.e.,* each variable represents an average monthly value for the customers that live within the same state. Further details related to the computation of the time series are explained in Section 5.

4. FORECASTING INDICATORS

Our aim is to understand whether calling variables can be used to predict future socioeconomic indicators and as such

help institutions in emerging regions save on their budgets as well as providing *ex-ante* information to react before socioeconomic changes actually happen. For that purpose, we analyze whether the consumption and mobility time series extracted from call logs can help to forecast the socioeconomic time series computed by the NSI. In this section we explain the methodology for our experimental analysis.

4.1 Building the Time Series

First, we need to compute a time series for each calling variable presented in the Datasets section. Using the 17 months of calling records, we compute the monthly time series for each consumption and mobility variable in the state S under study. For each calling variable x its time series $x_S = \{x_0, x_1, ...x_t\}$ is a chronological sequence of monthly measures where each x_i represents the monthly average of the calling variable X for state S during month i, where $i = 1...17$. Additionally, given that the type of contract a subscriber has with the telecommunications company might impact the behavior of the user, we compute two different time series per variable, one for the subscribers that have the pre-paid option $x_{S,prepaid}$ and one for the subscribers that have a contract with the company $x_{S,contract}$.

On the other hand, the NSI time series $n_S = \{n_0, n_1, ...n_t\}$ – computed by the NSI through surveys – are chronological sequences of monthly measures where each n_i represents the monthly average of a socioeconomic indicator for state S during month i, where $i = 1...17$.

4.2 Stationary Time Series

With both the socioeconomic and calling variables time series in hand, we start the time series analysis. Typically, time series analysis consists of two steps: (i) build a model $x_{S,t} = f(x_{(S,t-1)}, .., x_{(S,t-q)})$ that accurately represents the time series using past values and (ii) use the model to predict future values. Before running any type of analysis, it is critical to guarantee that the time series is stationary. A time series is *stationary* if its joint probability distribution does not change when it is shifted in time or space. As a result, the mean and covariance do not vary over time. If its autocovariance depends only on the lag, the time series is said to be *weakly stationary.*

Often times, socioeconomic time series are not (weakly) stationary processes. In order to test whether our socioeconomic and calling variables time series are stationary, we run two statistical tests: the *kpss* test, also known as the Kwiatkowski, Phillips, Schmidt and Shin test and the ADF or Augmented Dickey-Fuller test. Both tests look for trend stationarity in the time series by testing the null hypothesis of a unit root in a univariate time series. Additionally, we support our analysis by plotting the autocorrelation (ACF) and partial autocorrelation functions (PACF) of the time series, where a decay after a few peaks at the first lags should be observed to guarantee stationarity.

If a distribution does not pass the stationarity tests, we use two techniques to attempt to reach it: differencing and logging. Differencing consists on building a new distribution z such that each of its values is computed as the nth-order difference between an element and its n past elements *i.e.,* $z = \{x_1 - x_0, x_2 - x_1, ..., x_p - x_{p-1}\}$ for first-order difference $(n = 1)$ or $z = \{x_{t-1} - x_{t-2} - ... - x_0, ..., x_p - x_{p-1} - ... - x_{p-t}\}$ for $t-th$ order difference $(n = t)$. Whenever necessary, and in order to improve stationarity, the differenced values are represented as a percentage of change (incremental or decremental) with respect to the previous value dividing each difference by its last original value *i.e.,* $x_i = \frac{x_i - x_{i-1}}{x_{i-1}}$ being x_i and x_{i-1} consecutive elements of time series z. On the other hand, logging consists on computing the logarithm on each value of the distribution. In fact, log-transforming the data sometimes helps to stabilize the variance.

4.3 Leaders and Trailers

Once we have the socioeconomic and calling time series in a stationary state, we can compute their cross-correlation coefficients (CCF). The CCF reveals the correlations that might exist between two time series, and the temporal lags at which such correlations happen. As such, it is a necessary first analytical step and a good indicator of the potential predictive power that a *leader* time series might have over a *trailer* time series. The CCF(x,y) is defined as:

$$CCF(x, y) = \frac{E[(X_t - \mu_x)(Y_{t+\tau} - \mu_y)]}{\rho_x * \rho_y} \quad (1)$$

which represents the normalized cross-covariance of time series x and y and gives an understanding of the correlations between the two at positive and negative temporal lags (...,-2,-1,0,+1,+2,...) [10]. Strong correlations at negative lags imply that the time series y might be able to forecast time series x and vice versa, thus discerning between the leader and the trailer time series. Whenever no correlations are found, and in an attempt to get a better understanding of the relationship between the two series, we carry out an additional pre-whitening of the time series involved in the CCF computation. The main purpose is to clean the signal provided by the time series and detect correlations that might be hidden behind noise. The pre-whitening technique finds a fitting for time series x and determines its residuals, after which, these are substracted from time series y and the CCF is examined between the residuals for x and the pre-whitened y time series [8].

4.4 Predicting Time Series

The CCF allows to identify which calling time series are trailers of specific NSI time series. In this section, we present two techniques to build predictive models that can forecast future socioeconomic indicators using calling logs.

4.4.1 Multivariate Regression Models

Our first approach is a multivariate regression model where the NSI time series values are predicted exclusively using calling variables time series. Specifically, an NSI time series y is modeled as $y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon_i$ where x_j represents a calling variable time series. It is important to highlight that this multivariate model ignores the temporal order of the samples in the time series, and attempts to find a fitting solely based on tuples of values $(y_t, \{x_{(1,t)}, ..., x_{(p,t)}\})$ available at different times t. Next section explores a predictive model where the temporal order also plays a role.

4.4.2 Time Series Models

Time series analysis attempts to forecast future values of a series from its previous ones *e.g.,* the Dow Jones Industrial Average (DJIA) at time t can be predicted from its i previous values in the past as $DJIA_t = \alpha_1 DJIA_{(t-1)} + ... + \alpha_i DJIA_{(t-i)}$. Traditional time series analysis proposes

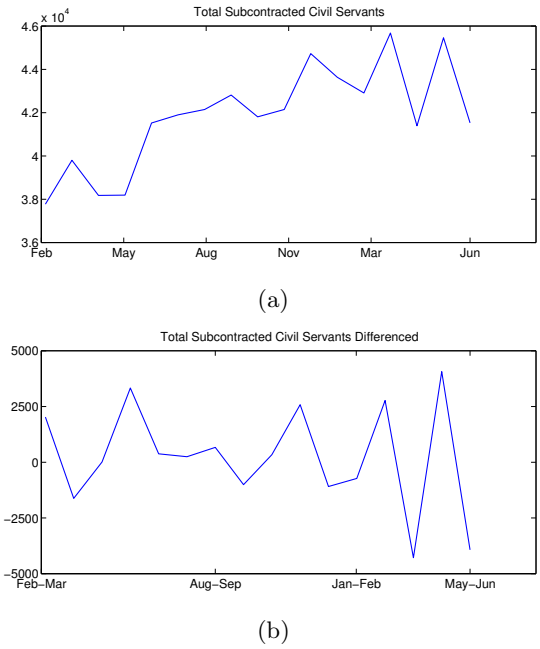


Figure 1: NSI series Total Subcontracted Civil Servants before (17 points) and after being stationarized with 1st-order differencing (16 points).

three approaches to fit and forecast a series: Auto-regressive model (AR), Moving average (MA) and Auto-regressive Moving average (ARMA). The AR(p) models a univariate time series where the value at time t is predicted from its p previous temporal consecutive values. The MA(q) models a time series where the forecasted value at time t depends on q previous unobserved white-noise values. Finally, the ARMA(p,q) is a combination of both AR and MA models, where the forecasted values are computed from p previous values of the time series and q previous value of a white noise distribution.

These models are *univariate i.e.,* solely based on previous values of the time series under study. However, it might be the case that other variables can potentially provide additional information to enhance the model that explains the evolution of a specific series. For instance, to predict the price of a stock we can use its previous values, but it might help to understand how the $S\&P$ time series evolves over time too. For that reason, the AR, MA and VARMA models have their corresponding *multivariate* variations namely VAR, VMA and VARMA. Taking as a baseline a simple AR(p) model $x_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t$ where x_t is the value of time series x at time t, we can compute its enhanced multivariate version VAR(p) as $y_t = c + \tau_1 y_{t-1} + ... + \tau_p y_{t-p} + \epsilon_t$ where y represents a vector of n time series y_i and τ_i are (nxn) coefficient matrices containing the model parameters for each individual time series. As such, the VAR(p) allows us to express each time series y_i as dependent not only on its own past values but also on the values that other variables have in the past. For example, a VAR(2) with $i = 1$ would be modeled as $y_{1,t} = c_1 + \tau^1_{11} y_{1,t-1} + \tau^1_{12} y_{2,t-1} + \tau^2_{11} y_{1,t-2} + \tau^2_{12} y_{2,t-2} + \epsilon_{1t}$ for time series y_1, whose values are forecasted based on its own past values at lags minus one and minus two lags as well as on the values of time series y_2 for the same temporal lags.

4.4.3 Forecast Evaluation

To evaluate these models, we divide the $17-month$ dataset into an $13-month$ training set and a $4-month$ testing set and report their R-square values, which measure the quality of the fit and the quality of the predictions that can be achieved using such model respectively. Additionally, we evaluate the predictive power for different horizons or steps ahead in time. A *one-step ahead forecast* ($h = 1$) simply predicts the next value of the series at time $t + 1$ using the previous real value at time t (extracted from the series computed by the NSI though surveys). On the other hand, an *n-step ahead forecast* ($h = n$) predicts the future value at time $t + n$ solely based on the $t + n - 1$ previous predictions and on the real value at time t.

From a policy perspective, a horizon of one would allow institutions to react upon specific events, however would not allow to save on budget given that the real value of the NSI series is always necessary to predict future values. On the other hand, a horizon of n means that institutions can not only react to changes in socioeconomic indicators before they happen, but also save on the budget allocated to compute these indicators since real survey values would only be necessary every n periods of time. This would imply saving the budget allocated to compute $n - 1$ surveys.

In the case of multivariate regressions, we explore all possible combinations of different calling variable time series and evaluate their predictive power. Given that multivariate regressions do not incorporate temporal information in the models, we simulate the $h = 1$ horizon by re-training the model after each prediction with the real NSI value and calling variables for that specific sample; whereas horizon $h = n$ is simulated re-training the model with the predicted NSI values and their calling variables for the last $n - 1$ samples and the real NSI value for the first sample. As for the multivariate time-series models, we use the univariate models for the NSI variables as a baseline and evaluate whether its multivariate counterparts –enhanced with consumption or mobility time series– improve the quality of the models and their predictive power. Unlike the multivariate regressions, there is no need to re-train given that the model itself already incorporates temporal information.

5. EXPERIMENTAL RESULTS

In this section, we describe the results obtained after running all the analyses described in the previous section. We used the econometrics toolbox in Matlab for all the tests and fittings.

5.1 Stationary Time Series

As mentioned earlier, in order to carry out time series analysis we first need to make sure that these are stationary. For each CDR and NSI series, we carry out the stationarity tests and whenever necessary we use differentiation or logging of the time series values. As an example, Figures 1(a) and 1(b) show the time series for the NSI variable *Total subcontracted civil servants* before and after making it stationary. Originally, we observe that the time series is non-stationary with an incremental pattern and varying mean and variances over the 17 points. This hypothesis is also confirmed by the the fact that none of the statistical tests

CDR / NSI	Assets	Employment	Workers	C. Servants	Sub. Workers	Sub. C.Servants
Input Calls	(6,0.5030)	(5,0.4850)	(5 , 0.5136)			
Output Calls		(-1,-0.6579)	(-3 , -0.4811)		(2, -0.5553)	(-1,-0.4321)
Duration InCalls	(6,0.5213)	(5,0.5062)	(5, 0.5109)			
Duration OutCalls		(-1,-0.6279)	(-3 , -0.4545)		(-2,-0.5752)	
N. BTS			(-3,0.687)	(1,-0.675)		
Talk Distance		(-3,0.676)				
Route Distance		(-1,0.685)	(-1,-0.654)			
Total Distance		(-1,0.891)	(1,-0.819)		(-2,0.708)	
Radius Gyration	(8,-0.643)	(9,0.621)	(-5,-0.731)	(-2,0.642)	(9,0.770)	
Diameter	(8,-0.651)	(9,0.698)	(-5,-0.739)	(-2,0.659)	(9,0.760)	(-1,0.642)

Table 1: CCF between NSI and CDR variable time series for subscribers in the state under study. Statistically significant correlations ($p < 0.01$) and lags at which these happen.

for stationarity reject the null hypothesis for the unit root. For that reason, we applied a differencing approach where each value in the time series is defined as the difference between itself and the next value. Figure 1(b) shows the end results. We observe that the time series has become more stationary, which is confirmed by the statistical tests. All of the calling and NSI time series required either a 1st-order differenciation or a percent change so as to convert them to stationary time series. Thus, the final size of each series is 16 points (12 for training and 4 for testing).

5.2 Cross-Correlations

In this section, we analyze the cross-correlation coefficients (CCF) between the CDR and the NSI variables. We seek relationships that happen at negative correlations which represent CDR variables that might be able to predict changes in the NSI variables before these actually happen. Table 1 shows the cross-correlation coefficients (CCF) between the CDR and the NSI variables at different lags in time. For each pair of variables, we report the lag at which the correlation is statistically significant ($p < 0.01$) and its correlation index. Results are only reported for contract customers because most of the CCF trends are very similar to the prepaid ones. However, it is important to clarify that in general the correlation values for pre-paid customers are smaller.

The consumption variables show that the number of ouput calls and their duration are correlated to the total employed, total number of workers, subcontracted workers and subcontracted civil servants at negative lags. Specifically, we observe negative CCFs meaning that the larger the number of workers for that state, the smaller the number of output calls and durations customers seem to show. This might be related to cell phones being used as job-seeking tools *i.e.,* cell phones show large number of output calls whenever the employment rates are low. On the other hand, input measures show correlations at positive lags which reveal that NSI variables (leaders) might be predictive of CDR variables (trailers).

In terms of mobility variables, traveled distances (talk, route and total) show positive correlations at negative lags for employment rates, meaning that whenever there is a increase in the distance travelled, such change might be predictive of an increase in the employment rates. Alternatively, number of BTSs, radius of gyration and diameter show correlations at negative lags for the total number of workers and subcontracted civil servants. These three variables show an interesting trend: larger values might be predictive of decreases in the total number of workers (negative CCF) and increases in the number of civil servants (positive CCF). Given that both measures give an approximation of the area where a person spends most of her time, we could interpret that workers that have a job tend to move in larger areas than when they are unemployed whereas civil servants decrease their area of mobility when they are employed.

To sum up, both consumption and mobility time series show potential predictive power over several NSI series including employment, total workers, total civil servants, subcontracted workers and subcontracted civil servants. These results show that behavioral variables computed from calling records can potentially be used as predictors of NSI variables computed through traditional surveys. Next, we evaluate various mathematical models to measure the predictive power of the calling variables time series.

5.3 Forecasting

5.3.1 Multivariate Regressions

In this section, we explore the predictive power of the calling time series over the NSI indicators using multivariate regression analysis, which only uses consumption and mobility series stripped of their temporal information to build the predictive models.

Tables 2 and 3 show the R-square values for the multivariate regressions on each NSI variable. The R-square training and testing values represent the percentage of the set that is explained with the fitted model and the quality of the predictions respectively. We only discuss results for one-step ahead predictions (h=1) since larger horizons did not yield significant results. Whenever the testing R-square is not reported, it is because it was a negative value implying that the fitted model is as bad as a horizontal line and has null predictive power. Additionally, we report results for regressions using sets of up to four different consumption variables (Table 2) and four different mobility variables (Table 3). In the cases when not all variables are used (subsets of two or three variables), we report the best fits. Finally, the tables are based on calling variables computed for customers with a contract. Results for pre-paid users were very similar, although these showed slightly smaller values. As such, it appears that contract customers share stronger behavioral patterns that approximate better the NSI variables.

Table 2 shows that the R-square training values for models built using one or more consumption time series are quite low, meaning that is is relatively hard to build a model that

R-square	CDR Series	Assets	Employment	Workers	C. Servants	Sub. Workers	Sub C.Servants
Train	2	0.20	0.18	0.05	0.18	0.18	0.12
	3	0.26	0.21	0.06	0.19	0.21	0.12
	4	0.27	0.22	0.08	0.21	0.21	0.12
Test(h=1)	2	-	≈ 0	-	≈ 0	≈ 0	-
	3	-	-	-	-	≈ 0	-
	4	-	-	-	-	-	-

Table 2: Training and Testing R-squares for Multivariate Regression Models with one or more Consumption Variables Time Series for customers with a contract in the state under study.

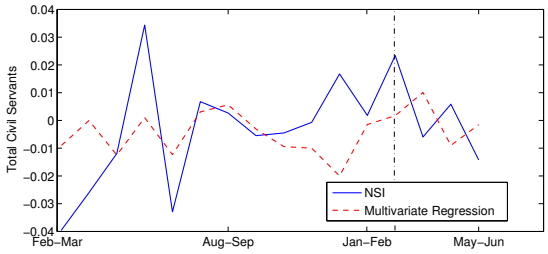
R-square	CDR Series	Assets	Employment	Workers	C. Servants	Sub. Workers	Sub C.Servants
Train	2	0.30	0.28	0.31	0.40	0.20	0.25
	3	0.59	0.38	0.40	0.52	0.33	0.30
	4	0.68	0.46	0.45	0.67	0.54	0.36
Test(h=1)	2	-	-	-	-	≈ 0	≈ 0
	3	-	-	-	-	≈ 0	≈ 0
	4	-	-	-	-	-	-

Table 3: Training and Testing R-squares for Multivariate Regression Models with one or more Mobility Variables Time Series for customers with a contract in the state under study.

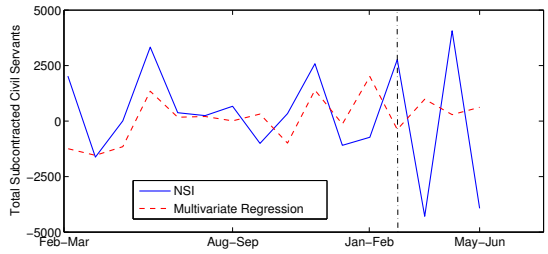
accurately describes the NSI time series. Although adding more than one time series seems to improve the R-square values during the training, these are still low. The highest values were reached for the total assets and total employed using four different calling time series with peak values around 0.27 or smaller. As a result, the predictive power of the multivariate regression model is almost non-existent. In fact, the R-square values at horizon one were all either negative or very close to zero which reveals that consumption calling behaviors are bad to forecast values at any horizon.

On the other hand, we observe that the mobility variables show higher training R-square values than the consumption variables. Specifically, total assets, total civil servants and total subcontracted workers reach the highest peaks with values of up to 0.68. Total workers and total number of subcontracted civil servants are the variables that are worst explained by the fitted models with R-square values of 0.46 and 0.36, respectively. As expected, reducing the number of regressors, also reduces the quality of the model across all NSI variables. Whenever less than four variables were considered, the combinations that yielded best results always included diameter and radius of gyration which appear to be the strongest variables to compute the best models. Although the training models appear to find decent fits for the NSI data, the testing results were also very poor with either negative or close to zero R-square testing values at any horizon.

To illustrate our findings, Figure 2 shows the NSI time series for the total civil servants and total subcontracted civil servants together with their forecast computed with a multivariate regression. Although the regressions do not take into account the time label, we selected to represent the pairs of value and fitted value along the time axis so as to be able to easily analyze the quality of the regressors' predictive power. Also, the values for the y axis represent the percent change (a) or differenced (b) time series, and not its original values which were re-computed to guarantee stationarity. Figure 2(a) shows the differenced total number of subcontracted workers computed by the NSI (solid line), the fitted values (dashed line) computed using a four-CDR vari-



(a)



(b)

Figure 2: Multivariate Regressions for (Percent Change) Total Civil Servants and (Differenced) Subcontracted Civil Servants using (a) consumption and (b) mobility series with $h = 1$. Predicted values from Feb-Mar onwards, after vertical line.

	Assets	Employment	Workers	C. Servants	Sub. Workers	Sub C.Servants
R-square Train	0.19	0.45	0.75	0.68	0.32	0.40
R-square Test (h=1)	-	-	-	-	**0.52**	-
R-square Test (h=2)	-	**0.23**	-	-	0.07	**0.30**

Table 4: ARMA fittings for NSI variables. The table shows the training and testing R-square values for horizons one-step and two-steps ahead.

able multivariate regression model with the training data (until Feb-Mar) and the predicted values (dashed line) using the model over the testing data (after vertical line). We observe that the fitting of the model is quite bad as it does not capture either the real volumes or the trends of the original data. Therefore, the model fails to predict real values and even incremental or decremental trends, modelling exactly the opposite trends. Similarly, Figure 2(b) shows the differenced total number of subcontracted employees computed by the NSI and its fitted and predicted values based on a combination of four mobility variables including talk distance and radius of gyration. In this case, although the training appears to be much better, the testing also fail to identify either real volumes or trends.

In general, both tables and figures show that multivariate regression models built with behavioral variables computed from calling records fail at forecasting socioeconomic indicators. These results might be related to the fact that our proposed regression models do not take into account the temporal information of the time series during the training of the model. We hypothesize that such models fail to identify specific temporal associations between calling behavioral patterns and socioeconomic indicators. Additionally, our regression models do not take into account previous information from the NSI indicators, which might also help to enhance the models. In an attempt to overcome these issues, next section explores the use of multivariate time-series models that take advantage of the temporal labels as well as of the information contained in the NSI time series.

5.3.2 Time Series Analysis

In this section, we evaluate the use of multivariate time-series analysis to forecast NSI socioeconomic series. For that purpose, we first analyze the predictive power of different AR, MA and ARMA models which only use past values of their own NSI time series to predict future ones. Taking these models as a baseline, we will compare them against their multivariate counterparts which we build using both NSI and calling variables time series.

In general, we found that the best univariate time-series model for the NSI indicators was an ARMA model, except for the case of the total number of civil servants where an AR model showed the best fitting. Both ARMA and AR models have p values of up to 3 and the q value was always one, which reveals that, in general, future NSI values can be forecasted looking at values from one to three months before. Table 4 shows the training and testing R-square values for the best univariate time-series models at horizons of one-step and two-steps ahead since larger values did not give significant results. We can see that total employment, total number of workers and total number of civil servants had the best training R-square values with 0.45, 0.75 and 0.68. In terms of predictive power, univariate time-series models do a good job for certain NSI time series, and certainly improve the results obtained with the multivariate regression models. At horizon one, the model for the subcontracted workers time series shows an R-square value of 0.5 whereas it looses almost all predictive power at horizon two. As such, the NSI series itself, could be used by institutions to forecast socioeconomic changes in the number of subcontracted workers one month in advance. On the other hand, the series total employed and number of subcontracted civil servants have significant predictive values at horizon two with R-squares of 0.23 and 0.30, respectively. Although these values are a little bit low, they could certainly be used to predict trends rather than real values. In any case, we use these results as a baseline to evaluate whether time series computed from calling records might improve the quality of the univariate time-series models.

Tables 5 and 6 show the training and testing R-square values for the multivariate time-series models built with one or more consumption and mobility time series respectively. Due to the reduced number of training points in our dataset, the models are limited to combinations of up to three different calling variables. We force the model to have at least one non-zero coefficient for the consumption or mobility variables so as to compare it against its univariate baseline. Testing R-square values are only reported at horizons one and two since larger values did not reveal significant results except for a few exceptions that are presented later on. For clarity purposes, we only present results for contract customers and do not specify the model details for each combination, but in general, the most common model was a $VAR(1)$, $VAR(2)$ and $VAR(3)$.

Table 5 shows that the best training R-square values are achieved when NSI models use only one consumption variable time series. In terms of variables, total employed, total civil servants and total subcontracted workers show the highest R-square training values of 0.82, 0.72 and 0.74, respectively. Regarding predictive power, we observe that number of employees, total subcontracted workers and total subcontracted civil servants show very good R-square testing values for one-step ahead predictions with values 0.65, 0.66 and 0.51 respectively across different subsets of calling series, which always included the number of output calls. Figure 3(a) shows the original (solid line), the trained model (dashed line until Feb-Mar) and the predicted values (after vertical dashed line, from Feb-Mar onwards) at $h = 1$ for the differenced total employed time series using three CDR variables. We observe that both trends and volumes are modeled quite well meaning that institutions could use these models to predict changes in the total number of employees with one month in advance.

On the other hand, two-step ahead predictions show smaller although still significant values for the same NSI series: total employed, total subcontracted workers and subcontracted civil servants with values of 0.31, 0.37 and 0.35, respectively. These models only used a unique calling time series which

R-square	CDR Series	Assets	Employment	Workers	C. Servants	Sub. Workers	Sub C.Servants
Train	1	0.6421	0.8262	0.5522	0.7222	0.7448	0.6276
	2	0.6140	0.8173	0.5471	0.7220	0.7438	0.6155
	3	0.6055	0.8169	0.4889	0.6806	0.7019	0.4781
Test(h=1)	1	-	**0.65**	-	-	0.53	**0.51**
	2	-	0.16	-	-	0.40	0.43
	3	-	0.10	-	-	**0.66**	0.38
Test(h=2)	1	-	**0.31**	-	-	**0.37**	**0.35**
	2	-	-	-	-	-	-
	3		-	-	-	-	-

Table 5: VARMA models computed with an NSI variable and one or more consumption variables for contract subscribers in the state under study. We evaluate one-step (h=1) and two-step (h=2) ahead predictions.

R-square	CDR Series	Assets	Employment	Workers	C. Servants	Sub. Workers	Sub C.Servants
Train	1	0.68	0.82	0.88	0.96	0.81	0.80
	2	0.68	0.82	0.87	0.96	0.79	0.79
	3	0.66	0.80	0.86	0.92	0.79	0.78
Test(h=1)	1	-	-	-	-	0.33	**0.38**
	2	-	-	-	-	**0.37**	0.36
	3		-	-	-	-	-
Test(h=2)	1	-	**0.18**	-	-	0.07	**0.04**
	2	-	-	-	**0.13**	**0.08**	-
	3						

Table 6: VARMA models computed with an NSI variable and one or more mobility variables for contract subscribers in the state under study. We evaluate one-step (h=1) and two-step (h=2) ahead predictions.

was either the number of output calls or its duration. Figure 3(b) shows the original, the trained model and the predicted values at $h = 2$ for the differenced total subcontracted workers. As shown, our two-steps ahead predictions could be used as a way to forecast trends (increase or decrease in a value) rather than absolute volumes which are sometimes under-estimated. Hence, institutions in emerging regions could use our models to forecast trends two months ahead which would allow them to carry out surveys only once every two months, thus saving budget. The latest collected survey data would then be used to update the multivariate time-series model so as to maintain the quality of its predictions. We acknowledge the limitations of our model at horizon two if real value predictions are desired, however we believe that being able to predict trends also has a lot of potential from a policy perspective. Finally, it is important to highlight that multivariate time-series models show higher predictive power than the univariate models at horizons one and two. We hypothesize that this is probably due to the fact that calling records provide complementary behavioral information to the NSI series that enhances the forecast power of the models.

Similarly, Table 6 shows that the best training and testing R-square values for the multivariate time-series models using mobility data. We observe that total employed, total workers and total civil servants show the highest training R-square values of 0.82, 0.88 and 0.96, respectively. In terms of predictive power, the total number of subcontracted workers and subcontracted civil servants have good R-square testing values for one-step ahead predictions – 0.38 and 0.37 respectively– meaning that institutions in emerging regions could predict changes in the number of subcontracted personnel before these happen and act accordingly. As an example, Figure 3(c) shows the original, the trained model and the predicted values (after vertical dashed line) at $h = 1$ for the differenced total subcontracted civil servants series based on two mobility series including radius of gyration (percent change). We observe that our model predicts quite well the general trend, although the real values are a bit under-estimated.

Two-step ahead predictions show smaller testing R-square values but still significant for total employed and total civil servants (0.18 and 0.13, respectively). Figure 3(d) shows the original, the fitted model and its predicted values (h=2) for the percent change total civil servants time series. Interestingly, mobility variables show less predictive power over certain NSI series than their univariate counterpart (recall that we force at least one mobility variable to have a non-zero coefficient). For example, R-square values for number of subcontracted workers (at $h = 1$) decreases from 0.52 to 0.37. Thus, it is fair to say that for some NSI variables, mobility variables do not appear to provide behavioral information that better complements the univariate models.

To sum up, multivariate time-series models can improve, in some cases, the predictive power of the traditional univariate series models by simply adding one or more calling variable time series to the model. In general, our results show good R-square values for one-step ahead predictions and acceptable values for two-step ahead predictions for certain NSI time series. Although some of the multivariate time-series models discussed might not be able to predict approximate real values, they can certainly forecast changes in the trends of the NSI series. Specifically, we achieve significant results for the number of employed people, the total number of subcontracted workers and the total number of subcontracted civil servants. Our models failed to provide

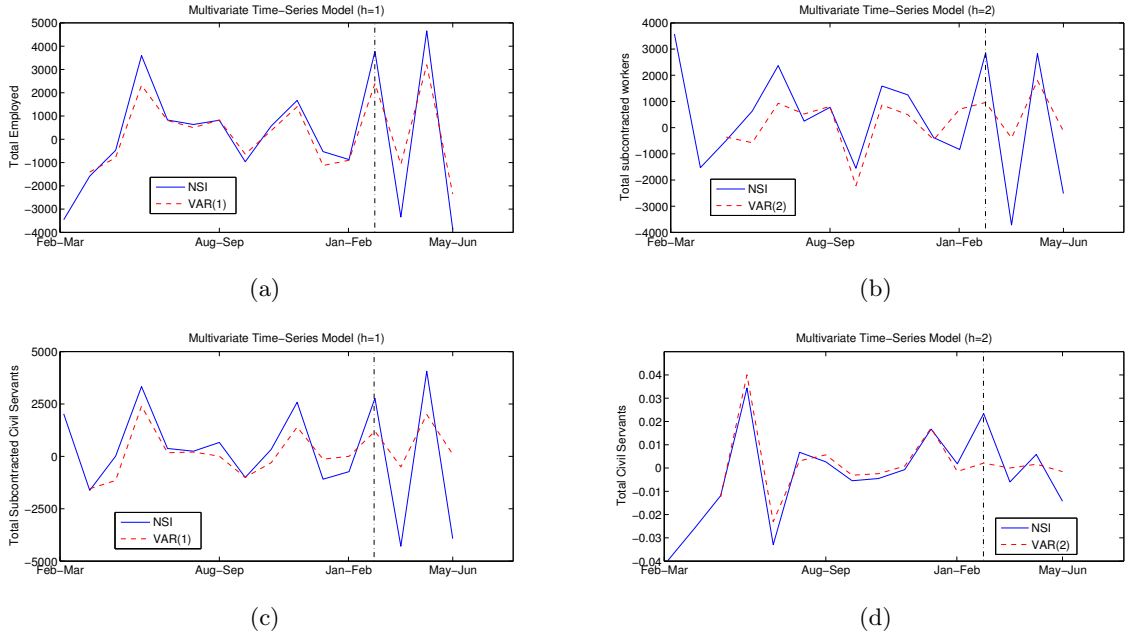


Figure 3: Multivariate Time-Series Models: (a) Total Employed (Diff., h=1) and (b) Total subcontracted Workers (Diff., h=2) with consumption series; (c) Total subcontracted civil servants (Diff., h=1) and (d) Total Civil Servants (% Change, h=2) with mobility series. Forecasts from Feb-Mar onwards (vertical line).

good predictive models for the other NSI series, which might be due to the limited size of our training and testing datasets (16 points in total). We will focus our future work on daily time series that introduce larger volumes of data to fit the models. We believe that these are preliminary results that might encourage local Statistical Institutes to consider calling logs provided by telecommunications companies as an affordable approach to forecast future values or trends for the socioeconomic indicators of their interest.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have analyzed whether behavioral variables extracted from calling records can be used to predict socioeconomic time series. Our objective is to provide institutions in emerging economies with a forecasting tool to reduce the number of surveys they have to run to gather such indicators. Additionally, we expect that such tool will also offer the capabilities to react to socioeconomic changes before they actually happen. For that purpose, we have evaluated different multivariate regression and multivariate time-series forecasting models using consumption and mobility time series extracted from calling records from an emerging economy in Latin America spanning a period of 17 months. Our results show that using multivariate time-series computed with different sets of consumption and mobility variables time series yield good predictive results, whereas multivariate regressions fail to provide good forecasts. Specifically, our models provide good R-square values for certain NSI variables at horizon one. As for horizon two, R-square values are lower, however, the models are useful to forecast trends instead of real values. In the future, we plan to analyze non-linear approaches such as NN or SVMs applied to time series.